Proceedings of the

International Congress of Mathematicians

Seoul 2014

SEOUL ICM 2014
INTERNATIONAL CONGRESS OF MATHEMATICIANS

Proceedings of the **International Congress of Mathematicians**

Seoul 2014

VOLUME II
Invited Lectures

Editors
Sun Young Jang
Young Rock Kim
Dae-Woong Lee
Ikkwon Yie

KM KYUNG MOON SA

Editors

Sun Young Jang, University of Ulsan
Young Rock Kim, Hankuk University of Foreign Studies
Dae-Woong Lee, Chonbuk National University
Ikkwon Yie, Inha University

Technical Editors

Young Rock Kim, The Korean TEX Society
Hyun Woo Kwon, The Korean TEX Society

Proceedings of the International Congress of Mathematicians
August 13–21, 2014, Seoul, Korea

Published by

KYUNG MOON SA Co. Ltd.
174, Wausan-ro Mapo-gu Seoul, Korea
Tel: +82-2-332-2004 Fax: +82-2-336-5193
E-mail: kyungmoon@kyungmoon.com
Homepage: www.kyungmoon.com

ISBN 978-89-6105-805-6
ISBN 978-89-6105-803-2 (set)

Printed in Korea

Contents

3. Number Theory

4. Algebraic and Complex Geometry

5. Geometry

6. Topology

1. Logic and Foundations

Model theory of difference fields and applications to algebraic dynamics

Zoé Chatzidakis

Abstract. This short paper describes some applications of model theory to problems in algebraic dynamics.

Mathematics Subject Classification (2010). Primary 03C60; Secondary 12H10, 14GXX.

Keywords. Model theory, difference fields, canonical base property.

1. Introduction

A few years ago, Hrushovski noticed that the model theory of difference fields could give a new proof of a result of M. Baker on algebraic dynamics. Baker's result deals with endomorphisms of $\mathbb{P}^1$ defined over a function field K, and shows that under certain conditions, the endomorphism of $\mathbb{P}^1$ is isomorphic (over some algebraic extension of K) to one defined over the constant field k of K. He answered thus a question of Szpiro and Tucker. Nothing was known for varieties of higher dimension. We started working on this together, were able to answer a question of Baker (1.7 in [1]) in case of function fields of characteristic 0, and got a descent result in some special cases: there is a bijective rational map from our original algebraic dynamics (V, ϕ) to one defined over the smaller field. Because our tools are difference fields, the maps we obtain are in general only birational isomorphisms and not isomorphisms when the dimension of the underlying variety is > 1. These results appeared in [4] and [5].

It turns out that another model-theoretic tool, the Canonical Base Property, a property enjoyed by existentially closed difference fields, allows one to obtain a fairly strong result in a more general context. Explaining what is now known is the object of section 4 of this paper.

Section 2 recalls some of the now classical results of the model theory of difference fields, as well as some more recent ones (e.g., 2.12). In section 3, we explain briefly the connection between our algebraic dynamics (V, ϕ) (where ϕ is rational dominant, not necessarily a morphism) and difference fields. In section 4, we introduce the Canonical Base Property, some of its history, give some of its consequences, and explain briefly the strategy to show that existentially closed fields of arbitrary characteristic enjoy it. Section 5 puts everything together.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. Difference fields and their model theory

2.1. Basic definitions. A difference ring is a ring R with a distinguished endomorphism σ. A *difference* field is a difference ring which is a field (note that the endomorphism will necessarily be injective). A difference ring becomes naturally a structure of the language $\mathcal{L} = \{+, -, \cdot, \sigma, 0, 1\}$, where $+, -, \cdot$ are interpreted as the usual binary operations, 0 and 1 are the usual constants, and σ is interpreted by the endomorphism. The difference ring is *inversive* if the endomorphism is onto. Every difference ring R has a unique up to R-isorphism *inversive closure*, or *inversive hull*, i.e., an inversive difference ring containing it, and which R-embeds into every inversive difference field containing R.

The *difference polynomial ring in the variables* $Y = (Y_1, \ldots, Y_n)$ *over* R, denoted $R[Y]_\sigma$, is the polynomial ring $R[\sigma^j(Y_i) \mid 1 \leq i \leq n, j \geq 0]$, endowed with the natural extension of σ defined by sending $\sigma^j(Y_i)$ to $\sigma^{j+1}(Y_i)$ for each i and j.

If K is a field, then zero-sets of elements of $K[Y_1, \ldots, Y_n]_\sigma$ generate the closed sets of a topology on K^n, and this topology is Noetherian. It is very similar to the Zariski topology. I will call the closed sets of this topology *σ-closed*.

All these results and more can be found in Richard Cohn's book [7].

2.2. The model theory of existentially closed difference fields. A difference field K is *existentially closed* if every finite system of difference equations with coefficients in K which has a solution in a difference field containing K, has a solution in K. Note that an existentially closed difference field is necessarily inversive and algebraically closed. Every difference field embeds into an existentially closed one, and the existentially closed difference fields form an elementary class, with theory usually called ACFA. These fields were first investigated in the 90's by Macintyre, Van den Dries and Wood, see [12]. An indepth study, concentrating on geometric stability properties of these fields was then started by Hrushovski and myself, later joined by Peterzil [3, 6]. I will now recall some of the classical results.

The theory ACFA expresses the following properties of the $\mathcal{L}$-structure K:

– K is algebraically closed, $\sigma \in \mathrm{Aut}(K)$;

– If U, V are irreducible (algebraic) varieties, with $U \subset V \times V^\sigma$, and such that U projects dominantly onto V and V^σ, then there is a such that $(a, \sigma(a)) \in U$. [Here V^σ denotes the variety obtained by applying σ to the defining equations of V.]

2.3. Notation. $\mathbb{N}$ denotes the set of non-negative integers. We will work in a large sufficiently saturated existentially closed difference field $\mathcal{U}$. If E is a field, then E^{alg} denotes the (field-theoretic) algebraic closure of E. If E is a difference subfield of $\mathcal{U}$, and a a tuple in $\mathcal{U}$, then $E(a)_\sigma$ denotes the difference field generated by a over E, i.e. $E(a)_\sigma = E(\sigma^i(a) \mid i \in \mathbb{N})$, and $E(a)_{\sigma^{\pm 1}}$ its inversive hull $E(a)_{\sigma^{\pm 1}} = E(\sigma^i(a) \mid i \in \mathbb{Z})$.

2.4. Some properties of ACFA and of its models. Most of the results here appear in [13] or in [3]. ACFA does not eliminate quantifiers, the problem coming from the fact that an automorphism of a field E needs not extend uniquely to the algebraic closure E^{alg} of E. However, this is the only obstacle, and one obtains that if E is an algebraically closed difference field, then ACFA $\cup$ qfDiag(E) is complete (Here qfDiag(E) denotes the quantifier-free diagramme of E in the language $\mathcal{L}(E)$ obtained by adjoining constant symbols for the elements of E). This last result has several important consequences:

(1) Completions of ACFA are obtained by describing the action of the automorphism on

the algebraic closure of the prime field. This implies that ACFA is decidable.

(2) If E is a difference subfield of a model $\mathcal{U}$ of ACFA, and a, b are tuples in $\mathcal{U}$, then $tp(a/E) = tp(b/E)$ if and only if there is an E-isomorphism $E(a)_\sigma^{alg} \to E(b)_\sigma^{alg}$ which sends a to b.

(3) If $A \subset \mathcal{U}$, then the model-theoretic algebraic closure $\mathrm{acl}(A)$ of A is the smallest inversive algebraically difference field containing A. The definable closure of A, $\mathrm{dcl}(A)$, is usually much larger than the inversive difference field generated by A: it is the subfield of $\mathrm{acl}(A)$ fixed by the elements of $\mathrm{Aut}(\mathrm{acl}(A)/A)$ which commute with σ.

(4) Let $S \subset \mathcal{U}^n$ be definable. Then there is a set $W \subset \mathcal{U}^{n+m}$ defined by difference equations such that the projection π on the first n coordinates defines a finite-to-one map from W onto S.

One can also show that any completion of the theory ACFA is supersimple (of SU-rank ω), and that it eliminates imaginaries. An important definable subset of $\mathcal{U}$, is the *fixed field*

$$\mathrm{Fix}(\sigma) := \{a \in \mathcal{U} \mid \sigma(a) = a\}.$$

It is a pseudo-finite field, and its induced structure is that of a pure field. It is also stably embedded, and therefore, if $S \subset \mathrm{Fix}(\sigma)^n$ is definable in $\mathcal{U}$ with parameters from $\mathcal{U}$, then it is of the form $S' \cap \mathrm{Fix}(\sigma)^n$, where S' is definable in the langauge of rings with parameters from $\mathrm{Fix}(\sigma)$.

In positive characteristic p, there are other definable automorphisms, which are built up using the definable Frobenius automorphism $\mathrm{Frob} : x \mapsto x^p$ and its powers Frob_q. More precisely, if $\tau = \sigma^n \mathrm{Frob}^m$, where $n \geq 1$, $m \in \mathbb{Z}$, then $\mathrm{Fix}(\tau)$ is a pseudo-finite field, stably embedded; the induced structure is that of a pure field if $n = 1$, but involves the automorphism σ if $n > 1$. We will also call $\mathrm{Fix}(\tau)$ a fixed field. One has the following result:
(1.12 in [3]) *Let τ be as above, (K, σ) a model of ACFA, and consider its reduct the difference field (K, τ). Then $(K, \tau) \models$ ACFA.*

2.5. Independence and SU-rank. As the theory is supersimple, every type is ranked by the rank SU, a rank based on forking (or non-independence). In what follows, A, B, C are subsets of $\mathcal{U}$, a is a tuple of elements of $\mathcal{U}$, and E is a difference subfield of $\mathcal{U}$.

Independence of A and B over C, denoted $A \downarrow_C B$, is characterized by the linear disjointness of the fields $\mathrm{acl}(CA)$ and $\mathrm{acl}(CB)$ over $\mathrm{acl}(C)$. A set D definable over E has finite SU-rank iff every tuple $a \in D$ has finite SU-rank over E, and then

$$\mathrm{SU}(D) = \sup\{\mathrm{SU}(a/E) \mid a \in D\}.$$

One shows easily the following:

- $\mathrm{SU}(a/E) = 0$ if and only if $a \in \mathrm{acl}(E)$.
- $\mathrm{SU}(a/E) \leq 1$ if and only if for every $B \supset E$, either a and B are independent over E, or $a \in \mathrm{acl}(B)$.
- If $\mathrm{tr.deg}(E(a)_\sigma/E) < \infty$, and F is a difference field containing E, then $a \downarrow_E F$ if and only if $\mathrm{tr.deg}(E(a)_\sigma/E) = \mathrm{tr.deg}(F(a)_\sigma/F)$.
- If $\mathrm{tr.deg}(E(a)_\sigma/E) < \infty$, then $\mathrm{SU}(a/E) \leq \mathrm{tr.deg}(E(a)_\sigma/E)$.

- $\mathrm{SU}(a/E) < \omega$ if and only if $\mathrm{tr.deg}(E(a)_\sigma/E) < \infty$.

If $\mathrm{SU}(a/E) < \omega$, then $tp(a/E)$ can be *analysed* in terms of types of SU-rank 1, and so types of SU-rank 1 determine the properties of $tp(a/E)$. This will be explained below in the paragraph on semi-minimal analyses. First, a few definitions:

Definition 2.6. Let T be a supersimple theory which eliminates imaginaries, $\mathbb{U}$ a sufficiently saturated model of T, and $S \subset \mathbb{U}^n$, $P \subset \mathbb{U}^m$ subsets which are invariant under $\mathrm{Aut}(\mathbb{U}/A)$ for some small subset A of $\mathbb{U}$. E.g. S is A-definable, or is a union of realisations of types over A.

(1) S is *one-based* if whenever $a_1,\ldots,a_\ell \in S$ and $B \supset A, C = \mathrm{acl}(Aa_1\ldots,a_\ell) \cap \mathrm{acl}(AB)$, then $(a_1,\ldots,a_\ell)$ and B are independent over C.

(2) A partial type is *one-based* if the set of its realisations is one-based.

(3) S is *internal to* P, resp. *almost-internal to* P, if for some finite set B, we have $S \subset \mathrm{dcl}(ABP)$, resp. $S \subset \mathrm{acl}(ABP)$.

(4) (difference field context) S is *qf-internal to* P if for some finite set B, if $a \in S$, then there is some tuple b of elements of P such that a is in the inversive difference field generated by ABb.

(5) If p, q are types, we say that p is internal, almost-internal, qf-internal, to q, if the set of realisations of p is internal, almost-internal, qf-internal, to the set of realisations of q.

The following is one of the major results in the model theory of difference fields, and is often called the *dichotomy theorem*:

Theorem 2.7 ([3, 6]). *Let q be a type of SU-rank 1 in a model $\mathcal{U}$ of ACFA. Then either q is one-based, or it is almost internal*[1] *to the generic type of $\mathrm{Fix}(\tau)$, where $\tau = \sigma$ if the characteristic is 0, and in positive characteristic, τ is of the form $\sigma^n\mathrm{Frob}^m$ for some $n \geq 1$, $m \in \mathbb{Z}$ relatively prime to n. Moreover, if the characteristic is 0 and q is one-based, then q is stable stably embedded.*

So, Theorem 2.7 tells us that if a type of SU-rank 1 is not one-based, then it is almost internal to $\mathrm{Fix}(\tau)$ for some definable τ. The property of being one-based is very strong, since it gives a criterion for independence. It also forbids the existence of two distinct group laws, such as in fields. Hrushovski and Pillay ([11]) showed that stable one-based groups of finite rank are particularly nice, and their result generalises partially to our context, as follows:

Theorem 2.8. *Let G be an algebraic group definable in a model $\mathcal{U}$ of ACFA, et let B be a quantifier-free definable subgroup of $G(\mathcal{U})$ which is one-based, and defined over some $E = \mathrm{acl}(E)$. Let X be a quantifier-free definable subset of B^n. Then X is a Boolean combination of cosets of E-definable subgroups of B^n.*

In particular, if Y is a subvariety of G^n, then $Y \cap B^n$ is a finite union of translates of quantifier-free definable subgroups of B^n.

If $\mathcal{U}$ has characteristic 0, the result extends to arbitrary definable group G and definable subsets X of B^n: they are Boolean combination of translates of definable subgroups of B^n, and these subgroups are defined over E.

[1] The original formulation is: non-orthogonal to

The following gives a useful characterization of types of SU-rank 1 which are almost-internal to $\mathrm{Fix}(\sigma)$:

Theorem 2.9. *Let $\mathcal{U}$ be a model of ACFA, $E = \mathrm{acl}(E)$ a difference subfield of $\mathcal{U}$ and b a tuple in $\mathcal{U}$, with SU$(b/E) = 1$. Then $tp(b/E)$ is almost-internal to the generic type of $\mathrm{Fix}(\sigma)$ if and only if*

$$\mathrm{tr.deg}(E(b)_\sigma/E) = 1 \text{ and } \{[E(b, \sigma^\ell(b)) : E(b)] \mid \ell \in \mathbb{Z}\} \text{ is bounded.}$$

2.10. Some consequences of the dichotomy. The fact that definable sets which are orthogonal to the fixed fields are one-based, is at the core of several applications to number theory, by Hrushovski ([9]) and by Scanlon ([17–19]). I will explain how its use gives a new proof of the conjecture of Manin-Mumford. Recall first the

Conjecture of Manin-Mumford. *Let A be an abelian variety defined over a number field k, and let $X \subset A$ be a subvariety. Then the Zariski closure of $X(k^{alg}) \cap \mathrm{Tor}(A)(k^{alg})$ is a finite union of translates of abelian subvarieties of A by torsion points.*

This conjecture, as well as several strengthenings (A a commutative algebraic group, k an arbitrary field, with similar conclusions) have been proved using different methods. The one by Hrushovski deals with an arbitrary commutative algebraic group G defined over a number field. One important point is that the torsion subgroup lives in the semi-abelian quotient of the group, and he shows that the number of components of the Zariski closure of $\mathrm{Tor}(G) \cap X$ is bounded by the number of components of the Zariski closure of $\mathrm{Tor}(H) \cap \pi(X)$, where H is the quotient of G by its maximal vector subgroup, and $\pi : G \to H$ is the natural map. Results of Mumford, together with a characterization by Hrushovski of one-based subgroups of abelian varieties or of $\mathbb{G}_m$, allow him to show that there is some $\sigma \in \mathrm{Aut}(\mathbb{Q})$ such that the torsion subgroup of G is contained in a quantifier-free definable subgroup B of G, which defines a one-based group in any existentially closed difference field containing $(\bar{\mathbb{Q}}, \sigma)$. This, together with 2.8 and a simple argument, give the result. Bounds on the complexity of the difference equations defining B give bounds on the number of cosets involved in the description.

The applications by Scanlon have a similar flavour.

2.11. The classical semi-minimal analysis. A standard result on supersimple theories states that if $tp(a/E)$ has finite SU-rank, then there are SU-rank 1 types $p_1, \ldots, p_n$, and tuples $a_1, \ldots, a_n$ such that $\mathrm{acl}(Ea) = \mathrm{acl}(Ea_1 \ldots, a_n)$, and for each i, $tp(a_i/Ea_{i-1})$ is almost-internal to p_i. Such a sequence $a_1, \ldots, a_n$ is called a *semi-minimal analysis of* $tp(a/E)$.

It may happen that one can choose the a_i's such that each $tp(a_i/E)$ is almost-internal to p_i; in that case, notice that $tp(a/E)$ is almost internal to the set S of realisations of the p_i's. This is a strong condition on $tp(a/E)$, and we will say in this case that $tp(a/E)$ is *almost-internal* [to types of rank 1].

One can refine the semi-minimal analysis a little and impose that the a_i's are in $\mathrm{dcl}(Ea)$, and that the types $tp(a_i/\mathrm{acl}(Ea_{i-1}))$ are internal to p_i, for all i. But, as mentioned above, in the case of difference fields, the definable closure is too large to hope obtain precise results on definable sets. After some work, and precise analysis of what internality to a fixed field

means, one obtains the following result:

Proposition 2.12 ([5, 2.13]). *Let E be an inversive difference field, a a tuple in $\mathcal{U}$ such that $\sigma(a) \in E(a)^{alg}$.*

(1) *Then there are $a_1, \ldots, a_n = a \in E(a)_\sigma$, such that, setting $A_i = E(a_{i-1})_\sigma$ for each i (with $a_0 = \emptyset$), $tp(a_i/A_i)$ satisfies one of the following:*

   (i) *$tp(a_i/A_i)$ is algebraic;*
   (ii) *$tp(a_i/A_i)$ is one-based;*
   (iii) *$tp(a_i/A_i)$ is qf-internal to $\mathrm{Fix}(\tau)$ for some $\tau = \mathrm{Frob}^m\sigma^n$.*

(2) *Furthermore, let $\ell \geq 1$ be an integer, $(\mathcal{U}', \sigma')$ a model of ACFA, and $f : (E(a)_\sigma, \sigma^\ell) \to (\mathcal{U}', \sigma')$ an embedding of difference fields. Then, if $a_1, \ldots, a_n$ are as in (1), we have similar results holding in $\mathcal{U}'$: $tp^{\mathcal{U}'}(f(a_i)/f(A_i))$ is algebraic in case* (i)*, one-based in case* (ii)*, and qf-internal to $\mathrm{Fix}(\tau^\ell)$ in case* (iii)*.*

The content of this proposition is very strong. Note that in particular it implies that whether the tuple a is "one-based over E" depends only on its quantifier-free type over E, not on the particular embedding of $E(a)_\sigma$ into a model of ACFA. This result decomposes the extension $E(a)_\sigma/E$ into a tower of field extensions, each one of a certain kind.

3. Difference fields and algebraic dynamics

Definition 3.1. An *algebraic dynamics* defined over a field K is given by a pair (V, ϕ) consisting of a (quasi-projective) variety defined over K, together with a rational dominant map $\phi : V \to V$.

Remarks 3.2. In the literature, ϕ is often assumed in addition to be a morphism. Moreover, one also often imposes that the morphism be *polarized*, i.e., that there is an ample vector bundle $\mathfrak{L}$ on V and an integer $q > 1$ such that $\phi^*\mathfrak{L} \simeq \mathfrak{L}^{\otimes q}$. These hypotheses have strong consequences which we will discuss later.
If L is a field extension of K, an algebraic dynamics (V, ϕ) gives naturally rise to one defined over L, by viewing V as defined over L. We will constantly use this remark, and always consider them as algebraic dynamics over a large ambient algebraically closed field $\mathcal{U}$ (while they may be defined over smaller subfields).
If V is not absolutely irreducible, it may become reducible when viewed over L, and for this reason **we will always assume that our varieties are absolutely irreducible**.

Definition 3.3. If (V, ϕ) and (W, ψ) are algebraic dynamics, a *morphism* $(V, \phi) \to (W, \psi)$ is a rational map $f : V \to W$ such that $f \circ \phi = \psi \circ f$. It is *dominant* if $f : V \to W$ is dominant.

(3.4) Let (V, ϕ) be as above, and consider the function field $K(V)$ of V. The map ϕ then yields an endomorphism ϕ^* of $K(V)$, which leaves K fixed, and is defined by $f \mapsto f \circ \phi$, for $\phi \in K(V)$ (We view the elements of $K(V)$ as partial functions on $V(K)$ taking their values in K).
The *degree* of the morphism ϕ is $\deg(\phi) = [K(V) : \phi^*K(V)]$

Another equivalent way of translating algebraic dynamics into the difference field context, is the following: let a be a generic of V over K, and define an endomorphism σ of $K(a)$ by letting σ be the identity on K, and setting $\sigma(a) = \phi(a)$. If $f : (V, \phi) \to (W, \psi)$ is a dominant morphism, then $b = f(a)$ will be a generic of W, and we will have $\sigma(b) = \psi(b)$. Thus dominant morphisms of algebraic dynamics correspond to inclusions of difference fields.

3.5. Applying the semi-minimal analysis. Applying 2.12, there are tuples $a_1, \ldots, a_n = a \in K(a)$, such that for each i, $\sigma(a_i) \in K(a_i) \subset K(a_{i+1})$, and $tp(a_i/K(a_{i-1}))$ is either algebraic, or qf-internal to $\mathrm{Fix}(\tau)$, or one-based.

These tuples a_i give rise to a fibration of (V, ϕ), namely, if V_i is the algebraic locus of a_i over K, ϕ_i the rational endomorphism of V_i such that $\sigma(a_i) = \phi_i(a_i)$ and $g_i : V_i \to V_{i-1}$ the rational map induced by the inclusion $K(a_{i-1}) \subset K(a_i)$, we obtain

$$(V, \phi) \xrightarrow{g_n} (V_{n-1}, \phi_{n-1}) \xrightarrow{g_{n-1}} \cdots \xrightarrow{g_2} (V_1, \phi_1).$$

Note that the fibers of these maps are not themselves algebraic dynamics: indeed, the map σ transports the fiber $f_n^{-1}(a_{n-1})$ to $f_n^{-1}(\sigma(a_{n-1})) = f_n^{-1}(\phi_{n-1}(a_{n-1}))$.

3.6. Internality to the fixed field $\mathbf{Fix}(\boldsymbol{\sigma})$. Assume that $tp(a_i/K(a_{i-1}))$ is internal to $\mathrm{Fix}(\sigma)$, and that $K(a_i)$ intersects the separable closure $K(a_{i-1})^s$ of $K(a_{i-1})$ in $K(a_{i-1})$. Then, over some L containing $K(a_{i-1})$ and linearly disjoint from $K(a_i)$ over $K(a_{i-1})$, there is a tuple b such that $L(a_i) = L(b)$ and $\sigma(b) = b$. This implies that $L(a_i) = L(\sigma(a_i))$. If $i = 1$, then we get that ϕ_1 is a birational map, i.e., has degree 1. If $i \geq 2$, we obtain that ϕ_i induces a birational map between $g_i^{-1}(a_{i-1})$ and $g_i^{-1}(\sigma(a_{i-1}))$, and we have $\deg(\phi_i) = \deg(\phi_{i-1})$.

3.7. Algebraic extensions. Note that if a_j is algebraic over $K(a_{j-1})$, then also $\deg(\phi_j) = \deg(\phi_{j-1})$.

4. The Canonical base property

This property was originally a property of compact complex manifolds, which was isolated (independently) by Campana and Fujiki. Work of Moosa and Pillay provided a translation of this property in model-theoretic terms ([13] and [15]); Pillay and Ziegler ([16]) showed that various enriched fields enjoy it. This property will be later called the Canonical Base Property, CBP for short, by Moosa and Pillay who investigate it further in [14], and ask several questions.

Definition 4.1. Let T be a theory which eliminates imaginaries, $\mathbb{U}$ a saturated model of T, $A \subset \mathbb{U}$ and a a tuple in $\mathbb{U}$, $p(x) = tp(a/A)$.

(1) If T is stable and p is stationary, then p is definable, that it, for every formula $\varphi(x, y)$, there is a formula $d_\varphi(y)$ (with parameters in A) such that for every tuple b in A (of the correct arity), $\mathbb{U} \models d_\varphi(b)$ if and only if $\varphi(x, b) \in p$. Furthermore, these definitions define a (consistent and complete) type over $\mathbb{U}$. The *canonical base* of p is the smallest definably closed subset of $\mathcal{U}$ over which one can find parameters for all the formulas $d_\varphi(y)$ (in other words, contains the code of all sets defined by the $d_\varphi(y)$). It is denoted by $\mathrm{Cb}(p)$ or $\mathrm{Cb}(a/A)$, and is contained in A.

(2) If T is unstable, but simple, then the definition of canonical base is more involved, see e.g. Wagner's book [20], as it is defined in terms of extension base. It is easier to define the algebraic closure of the canonical base, denoted $\overline{\mathrm{Cb}}(p)$ or $\overline{\mathrm{Cb}}(a/A)$: it is the smallest algebraically closed subset B of A such that a and A are independent over B. If T is supersimple, then $\overline{\mathrm{Cb}}(p)$ will be contained in the algebraic closure of finitely many realisations of p, and so will have finite SU-rank if p has. Note that this definition also makes sense for infinite tuples, and we will often use it for the infinite tuple enumerating the algebraic closure of a finite tuple.

Example 4.2. Consider the theory ACF of algebraically closed fields, say of characteristic 0 for simplicity, and let $\mathbb{U}$ be a large algebraically closed field, $A \subset \mathbb{U}$ a subfield, and a a tuple in $\mathbb{U}$. Assume that $A(a)$ is a regular extension of A, and consider the algebraic locus V of a over A. Then $\mathrm{Cb}(a/A)$ is simply the field of definition of V.

Example 4.3. Let a be a tuple in $\mathcal{U}$, E a difference subfield of $\mathcal{U}$. If X is a tuple of indeterminates of the same size as a, then one can consider the ideal I of $E[X]_\sigma$ of difference polynomials which vanish at a. As in classical geometry, this ideal has a smallest (difference) field of definition, i.e., there is a unique smallest difference subfield E_0 of E such that I is generated by its intersection with $E_0[X]_\sigma$. Then $\overline{\mathrm{Cb}}(a/E) = \mathrm{acl}(E_0)$.

Definition 4.4. Let T be a supersimple theory which eliminates imaginaries. We say that T has the *Canonical Base Property*, or ***CBP***, if whenever A and B are algebraically closed sets such that $\mathrm{SU}(A/A \cap B) < \omega$ and $B = \overline{\mathrm{Cb}}(A/B)$, then $tp(B/A)$ is almost-internal (to types of SU-rank 1).

4.5. Comments.

(1) Let $C = A \cap B$, and a, b finite tuples such that $A = \mathrm{acl}(Ca)$, $B = \mathrm{acl}(Cb)$. Then $\mathrm{SU}(A/C) = \mathrm{SU}(a/C)$. The notion of almost-internality is by definition preserved under passage to the algebraic closure, so there are a set $D = \mathrm{acl}(D)$ containing A and independent from B over A, and tuples $b_1, \ldots, b_n$ with $\mathrm{SU}(b_i/D) = 1$, such that $\mathrm{acl}(DB) = \mathrm{acl}(Db_1 \ldots b_n)$.

(2) The definition in the stable case deals with finite tuples a and b, assumes that $\mathrm{Cb}(a/b) = b$, and deduces that $tp(b/a)$ is internal to types of rank 1.

(3) If $tp(A/C)$ is one-based, then ... by definition of one-basedness, we know that A and B are independent over their intersection, and therefore $B = C$. To say it in another fashion: if $tp(a/E)$ is one-based, and B contains E, then $\overline{\mathrm{Cb}}(a/B) \subset \mathrm{acl}(Ea)$.

(4) Hrushovski, Palacin and Pillay give in [10] an example of an ω-stable theory of finite rank which does not have the CBP. This example is built up from the theory ACF of algebraically closed fields.

Theorem 4.6 (Pillay-Ziegler [16]).

(1) *The theory of differentially closed fields of characteristic* 0 *has the CBP* (*version for stable theories*).

(2) *The elementary theory of an existentially closed difference field of characteristic* 0 *has the CBP.*

Pillay and Ziegler have some additional partial results concerning types of rank 1 in separably closed fields, but not the full and hoped for result. Their proof uses jet spaces, and generalises only partially to positive characteristic, because of possible inseparability problems. In order to show that the result holds for existentially closed fields of arbitrary characteristic, one needs to show a decomposition result:

Theorem 4.7 (1.16 in [2]). *Let T be a supersimple theory, $\mathbb{U}$ a large model of T, A, B and $C = A \cap B$ algebraically closed subsets of $\mathbb{U}$ such that $\mathrm{SU}(A/C) < \omega$ and $B = \overline{\mathrm{Cb}}(A/B)$. Then there are $a_1, \ldots, a_n \in A$, types $p_1, \ldots, p_n$ of SU-rank 1 (maybe over some larger base set D which is independent from AB over C), such that $\mathrm{acl}(Ca_1 \ldots, a_n) = \mathrm{acl}(CA)$; and each $tp(a_i/C)$ has a semi-minimal analysis in which all components are almost-internal to the set of realisations of the $\mathrm{Aut}(\mathbb{U}/C)$-conjugates of p_i.*
Furthermore, each of the types p_i is non-one-based.

From this, one shows easily that it suffices to show the CBP for types whose semi-minimal analysis only involves one fixed non-one-based type of rank 1. In the particular case of existentially closed difference fields of positive characteristic p, we must therefore look at types analysable in terms of $\mathrm{Fix}(\tau)$, for the various possible τ. When $\tau = \sigma$, one shows the following:

Lemma 4.8. *Let a be a finite tuple in $\mathcal{U}$, of finite SU-rank over $E = \mathrm{acl}(E)$, and assume that the semi-minimal analysis of $tp(a/E)$ only involves $\mathrm{Fix}(\sigma)$-almost-internal types. Then there is a tuple $b \in E(a)_{\sigma^{\pm 1}}$ such that $E(a)_{\sigma^{\pm 1}}$ is separably algebraic over $E(b)$.*

Inspection of the proof of Pillay-Ziegler then shows that there is no problem when $\tau = \sigma$: their proof goes through verbatim. Working in the reduct $(\mathcal{U}, \tau)$ then allows to obtain the results for all types analysable in $\mathrm{Fix}(\tau)$. Using the dichotomy Theorem 2.7, this finishes the proof of

Theorem 4.9 (3.5 in [2]). *Existentially closed difference fields of any characteristic have the CBP.*

The CBP has several interesting consquences, which I will now list. Relative versions of these results exist.

Theorem 4.10 (References are to [2]). *Let T be a supersimple theory with the CBP, $\mathbb{U}$ a saturated model, and $A, B, C = A \cap B$ algebraically closed subsets of $\mathbb{U}$, with $\mathrm{SU}(A/C)$ finite.*

(1) (2.1) *If $B = \overline{\mathrm{Cb}}(A/B)$, then $tp(B/C)$ is almost-internal.*

(2) (2.2) *More generally, if $tp(B/A)$ is almost-internal, then so is $tp(B/C)$.*

(3) (2.4) *There is some $D = \mathrm{acl}(D)$ with $C \subseteq D \subseteq A$ such that whenever $E = \mathrm{acl}(E)$ is such that $tp(A/E)$ is almost-internal, then $E \subseteq D$.*

(4) (2.5) *If $B = \overline{\mathrm{Cb}}(A/B)$ and D is such that $tp(A/D)$ is almost-internal, then so is $tp(AB/D)$.*

(5) (2.10) *Let a_1, a_2, b_1, b_2 be tuples of finite SU-rank, $\mathcal{S}$ a set of types of SU-rank 1 and assume that*

- *$tp(b_2)$ is almost-internal to types in $\mathcal{S}$,*
- *$\mathrm{acl}(b_1) \cap \mathrm{acl}(b_2) = \mathrm{acl}(\emptyset)$,*

– $a_1 \downarrow_{b_1} b_2$ *and* $a_2 \downarrow_{b_2} b_1$,

– $a_2 \in \mathrm{acl}(a_1 b_1 b_2)$.

Then there is $e \subset \mathrm{dcl}(a_2 b_2)$ *such that* $tp(a_2/e)$ *is almost-internal to types in* $\mathcal{S}$ *and* $e \downarrow b_2$. *In particular, if* $tp(a_2/b_2)$ *is hereditarily orthogonal to all types in* $\mathcal{S}$, *then* $a_2 \in \mathrm{acl}(eb_2)$.

4.11. Comments. Here is an easy consequence of item (1): assume that $tp(A/C)$ is not almost-internal, has finite SU-rank, and that $A \cap B = C$. Then A and B are independent over C.
Item (4) answers a question of Moosa and Pillay ([14]).
Item (5) is a *descent result*, and is (together with 2.12) the main ingredient of the applications to algebraic dynamics by Hrushovski and myself. After some work, and use of Proposition 2.12, one refines the descent result 4.10(5) to obtain the following:

Theorem 4.12 (4.11 in [2]). *Let* K_1, K_2 *be fields intersecting in* k, *for* $i = 1, 2$, *and with algebraic closures intersecting in* k^{alg}, *let* V_i *be an absolutely irreducible variety and* $\phi_i : V_i \to V_i$ *a dominant rational map defined over* K_i. *Assume that* K_2 *is a regular extension of* k, *and that there is an integer* $r \geq 1$ *and a dominant rational map* $f : V_1 \to V_2$ *such that* $f \circ \phi_1 = \phi_2^{(r)} \circ f$. *Then there is a variety* V_0 *and a dominant rational map* $\phi_0 : V_0 \to V_0$, *all defined over* k, *a dominant map* $g : V_2 \to V_0$ *such that* $g \circ \phi_2 = \phi_0 \circ g$, *and* $\deg(\phi_0) = \deg(\phi_2)$.

5. Applications of the CBP to algebraic dynamics

The original result of Matthew Baker. Let k be a field, C a curve over k, and $K = k(C)$. Let $\phi : \mathbb{P}^1 \to \mathbb{P}^1$ be defined over K, and of degree $d \geq 2$. One can define a logarithmic height function on the points of $\mathbb{P}^1(K)$, called the Weil height, and which I will denote by h. For details, please see [1]. If $K = k(t)$, then the Weil height of a point $P \in \mathbb{P}(K)$ is simply the minimal degree of polynomials needed to represent the point P. One then defines the canonical height $\hat{h}(P)$ as:

$$\hat{h}(P) = \lim_{n \to \infty} h(\phi^{(n)}(P))/d^n.$$

[Here $\phi^{(n)}$ denotes the iteration n times of the map ϕ.] One verifies that $\hat{h}(\phi(P)) = d\hat{h}(P)$; moreover, there is a constant $C > 0$, such that for any point P, one has $|\hat{h}(P) - h(P)| < C$. Clearly, any *preperiodic point* P (i.e., such that for some integers $m > n$ one has $\phi^{(m)}(P) = \phi^{(n)}(P)$) must have $\hat{h}(P) = 0$. Baker's theorem shows that these are the only ones, unless, over some finite extension of K one has $(\mathbb{P}^1, \phi) \simeq (\mathbb{P}^1, \psi)$ for some ψ defined over k:

Theorem 5.1 ([1]). *Let* $k \subset K$ *and* ϕ *be as above. Assume that for no finite algebraic extension of* K', *there is an* $M \in \mathrm{PGL}_2(K')$ *such that* $M^{-1}\phi M$ *is defined over* k. *Then a point* $P \in \mathbb{P}^1(K)$ *satisfies* $\hat{h}(P) = 0$ *if and only if it is preperiodic.*

He shows moreover that there is a positive ϵ which bounds below the canonical height of non-preperiodic points of $\mathbb{P}^1(K)$.

5.2. The analogue for higher dimensional varieties. The setting: Let V be a quasi-projective variety defined over K, $\phi : V \to V$ a dominant rational map of degree $d \geq 2$. Once fixed an embedding of V into projective space, the Weil heights of points of $V(K)$ exist as before. (But to obtain the canonical height, additional conditions are necessary.) We assume that for some N, the points $P \in V(K)$ such that all $\phi^{(n)}(P)$, $n \geq 0$, have height $\leq N$, form a Zariski dense subset of V.

The hope: (V, ϕ) is isomorphic to some (W, ψ) defined over k.

5.3. The observation which makes things work. The following observation, due to Szpiro, is what allows model theory to play a role, since it gives a certain configuration which one can exploit.

Given some integer N, the points of $V(K)$ which have Weil height $\leq N$, form what we will call a *limited set*, i.e., there is some algebraic set U defined over k, a constructible map $\pi : U \to V$ (defined over K), such that $\pi(U(k))$ contains all points of $V(K)$ of Weil height $\leq N$, and π is injective on $U(k)$ (see e.g. section 3 of [4]). Consider the following sets:

$$V_0 = \pi(U(k));\ V_n = \bigcap_{0 \leq j \leq n} \phi^{-(j)}(V_0).$$

So, a point P will be in V_n if and only if each of $P, \phi(P), \ldots, \phi^{(n)}(P)$ has Weil height $\leq N$.

The map ϕ induces a (partially defined) constructible map ϕ^* on U. Namely, if $Q \in U(k)$, and $\phi\pi(Q) \in V_0$, then $\phi^*(Q)$ is defined by $\pi\phi^*(Q) = \phi\pi(Q)$. Assume that for the number N above, the sets V_n are Zariski dense in V. We now look at U_n, the Zariski closure of $\pi^{-1}(V_n) \cap U(k)$. These sets form a decreasing chain of Zariski closed infinite subsets of U, which must therefore stabilise at some integer n. Let $\tilde{U} \subset U_n$ be the union of all irreducible components W of U_n such that $\pi(W(k))$ is Zariski dense in V. Then, the constructible ϕ^* induces a permutation of the irreducible components of $\tilde{U}$ of maximal dimension, and for some $r \geq 1$, the constructible map $(\phi^*)^{(r)}$ yields a rational dominant endomap ψ of some irreducible component W of $\tilde{U}$ of maximal dimension. Note that $\pi(W(k))$ is still Zariski dense in V, but that π sends (W, ψ) to $(V, \phi^{(r)})$. It turns out that this is sufficient to obtain some results, using Theorem 4.12.

Theorem 5.4 ([5, 3.2], [2, 4.12]). *With assumption as in 5.2, let $\mathcal{U}$ be a model of ACFA containing K, and a a generic point of V over K satisfying $\sigma(a) = \phi(a)$.*

(1) *Assume that the semi-minimal analysis of $tp(a/K)$ does not involve $\mathrm{Fix}(\sigma)$. Then there is a bijective morphism $g : (V, \phi) \to (V_0, \phi_0)$ for some (V_0, ϕ_0) defined over k. In characteristic 0, this g is a birational isomorphism.*

(2) *In the general case, there is a dominant rational map $(V, \phi) \to (V, \phi_0)$ where (V, ϕ_0) is defined over k, and $\deg(\phi) = \deg(\phi_0)$.*

Sketch of Proof. I will use (the proof of) 4.11 in [2], and follow its notation. By the above discussion 5.3, we know that there is some algebraic dynamics (V_1, ϕ_1) defined over k, and which dominates $(V, \phi^{(r)})$ for some $r \geq 1$. Let $\mathcal{U}$ be a model of ACFA containing K, let a_2 be a generic of V satisfying $\sigma(a_2) = \phi(a_2)$. Applying 4.11 of [2] (with $K_1 = k$, $K_2 = K$ and $(V_2, \phi_2) = (V, \phi)$), there is $a_3 \in K(a_2)$ such that $\sigma(a_3) \in k(a_3)$. If V_0 is the algebraic locus of a_3 over k, and $\phi_0 \in k(V_0)$ is such that $\phi_0(a_3) = \sigma(a_3)$, then $\deg(\phi) = \deg(\phi_0)$, and there is a rational dominant map $(V, \phi) \to (V_0, \phi_0)$. This gives (2).

The proof of 4.11 in [2] shows that $tp(a_2/K(a_3)_{\sigma^{\pm 1}})$ is almost-internal to $\mathrm{Fix}(\sigma)$. Hence, in case (1), it must be algebraic. Thus $K(a_2)$ is a finite algebraic extension of $K(a_3)$. Let $\alpha \in K(a_2)$ be defined by $K(\alpha) = K(a_2) \cap K(a_3)^s$, so that $K(a_2)/K(\alpha)$ is purely inseparable.

Now, recall from the proof of 4.11 that there is some generic a_1 of V_1 over K, such that $a_2 \in K(a_1)$. Then $k(a_1)$ and $K(a_3)$ are linearly disjoint over $k(a_3)$, and because $K(\alpha)/K(a_3)$ is separable and $K(a_2) \subset K(a_1)$, it follows that $K(\alpha) = K(\beta)$ for some $\beta \in k(a_1)$. Then $\beta \in k(a_3)^s$. As $\sigma(a_2) \in K(a_2)$, we have $\sigma(\alpha) \in K(\alpha)$, hence $\sigma(\beta) \in k(\beta)$. Let $\tilde{V}$ be the algebraic locus of β over k, and $\tilde{\phi} \in k(\tilde{V})$ such that $\sigma(\beta) = \tilde{\phi}(\beta)$, g the rational map $V \to \tilde{V}$ such that $g(a_2) = \beta$. Then g is generically bijective, and sends (V, ϕ) to $(\tilde{V}, \tilde{\phi})$. In characteristic 0, we may take $\alpha = a_2$, and g is then birational. This finishes the proof of (1). □

5.5. Comments. The fact that we work with function fields only tells us about the generic behaviour of the algebraic dynamics, and does not allow us to show full isomorphisms, only birational isomorphisms.

Remark 5.6. If in addition to the hypotheses of 5.2, one assumes that the map ϕ is a polarised morphism with associated constant $q > 1$, then the conclusion of 5.4(1) holds, so that we get the full result. This follows from an observation made without proof in [4]. The proof I sketch below is due to Hrushovski.

Proof. First, note that the hypotheses imply, by a result of Fakhruddin [8], that we may assume that $V \subset \mathbb{P}^N$ for some N, and that the morphism ϕ on V is the restriction to V of a morphism $\psi : \mathbb{P}^N \to \mathbb{P}^N$. Suppose that the conclusion of 5.4(1) does not hold, and let $\mathcal{U}$ be a model of ACFA containing K.

Let $g : (V, \phi) \to (V_0, \phi_0)$ be given by 5.4, with $\deg(\phi) = \deg(\phi_0)$, let $a = a_2 \in \mathcal{U}$ be a generic of V satisfying $\sigma(a_2) = \phi(a_2)$ and let $a_3 = g(a_2)$ (a generic of V_0 satisfying $\sigma(a_3) = \phi_0(a_3)$). Equality of the degrees of ϕ and ϕ_0 implies that the restriction of ϕ to $S = g^{-1}(a_3)$ is an isomorphism. The variety $S' = \phi(S)$ equals S^σ, and therefore $\deg(S') = \deg(S)$. We will show the following:

If S is a subvariety of V, and $\deg(S) = \deg(\phi(S))$ (as subvarieties of $\mathbb{P}^N$), then the degree of the map ϕ restricted to S is $q^{\dim(S)}$.

Let $r = \dim(S)$, and let $L_1, \ldots, L_r$ be generic hyperplanes. Then $\deg(S') = S' \cdot L_1 \cdot \cdots \cdot L_r$, and also equals $|S' \cap L_1 \cap \cdots \cap L_r|$, the number of points of $S' \cap L_1 \cap \cdots \cap L_r$ counted with multiplicities. Pulling back by ϕ, we get

$$\begin{aligned}\deg(S) &= \deg(S')\deg(\phi|S) = |\phi^{-1}(S') \cap \phi^{-1}(L_1) \cap \cdots \cap \phi^{-1}(L_r)| \\ &= S \cdot qL_1 \cdots qL_r = q^r \deg(S)\end{aligned}$$

(here we use $\phi^* L_i = qL_i$). As $\deg(S) = \deg(S')$, the restriction of ϕ to S has degree q^r.

As $\phi|S$ is birational and therefore of degree 1, we must have $r = 0$. This implies that S is finite, i.e., that a_2 is algebraic over $K(a_3)$, and we conclude as in 5.4(1). □

Acknowledements. The author thanks MSRI for its support during the spring 2014.

References

[1] Matthew Baker, *A finiteness theorem for canonical heights attached to rational maps over function fields*, J. Reine Angew. Math. **626** (2009), 205–233.

[2] Z. Chatzidakis, *A note on canonical bases and one-based types in supersimple theories*, Confluentes Mathematici Vol. 4, No. 3 (2012) 1250004 (34 pages). DOI: 10.1142/S1793744212500041.

[3] Z. Chatzidakis and E. Hrushovski, *Model theory of difference fields*, Trans. Amer. Math. Soc. **351** (1999), 2997–3071.

[4] ———, *Difference fields and descent in algebraic dynamics, I*, Journal of the IMJ, **7** (2008) No 4, 653–686.

[5] ———, *Difference fields and descent in algebraic dynamics, II*, Journal of the IMJ, **7** (2008) No 4, 687–704.

[6] Z. Chatzidakis, E. Hrushovski, and Y. Peterzil, *Model theory of difference fields, II: Periodic ideals and the trichotomy in all characteristics*, Proceedings of the London Math. Society (3) **85** (2002), 257–311.

[7] R.M. Cohn, *Difference algebra*, Tracts in Mathematics **17**, Interscience Pub. 1965.

[8] Najmuddin Fakhruddin, *Questions on self maps of algebraic varieties*, J. Ramanujan Math. Soc. **18** (2003), no. 2, 109–122.

[9] E. Hrushovski, *The Manin-Mumford conjecture and the model theory of difference fields*, Ann. Pure Appl. Logic **112** (2001), no. 1, 43–115.

[10] Ehud Hrushovski, Daniel Palacin, and Anand Pillay, *On the canonical base property*, preprint 2012. arXiv:1205.5981.

[11] E. Hrushovski and A. Pillay, *Weakly normal groups*, in: Logic Colloquium **85**, North Holland 1987, 233–244.

[12] A. Macintyre, *Generic automorphisms of fields*, APAL **88** Nr 2-3 (1997), 165–180.

[13] Rahim Moosa, *On saturation and the model theory of compact Kähler manifolds*, J. Reine Angew. Math. **586** (2005), 1–20.

[14] R. Moosa and A. Pillay, *On canonical bases and internality criteria*, vol. **52** no. 3 (2008), 901–917.

[15] Anand Pillay, *Model-theoretic consequences of a theorem of Campana and Fujiki*, Fundamenta Mathematicae, **174**, (2002), 187–192.

[16] A. Pillay and M. Ziegler, *Jet spaces of varieties over differential and difference fields*, Selecta Math. (N.S.) **9** (2003), no. 4, 579–599.

[17] T. Scanlon, *p-adic distance from torsion points of semi-abelian varieties*, J. für Reine u. Ang. Mat. **499** (1998), 225–236.

[18] ———, *The conjecture of Tate and Voloch on p-adic proximity to torsion*, International Mathematics Research Notices, 1999, no. 17, 909–914.

[19] ———, *Diophantine geometry of the torsion of a Drinfeld module*, J. Number Theory **97** (2002), no. 1, 10–25.

[20] F. Wagner, *Simple theories*, Kluwer Academic Pub., Dordrecht 2000.

UFR de Mathématiques, UMR 7586,Université Paris 7, Case 7012, 75205 Paris cedex 13, France
E-mail: zoe@math.univ-paris-diderot.fr

Logic and operator algebras

Ilijas Farah

Abstract. The most recent wave of applications of logic to operator algebras is a young and rapidly developing field. This is a snapshot of the current state of the art.

Mathematics Subject Classification (2010). 03C20, 03C98, 03E15, 03E75, 46L05, 46L10.

Keywords. Classification of C*-algebras, tracial von Neumann algebras, logic of metric structures, Borel reducibility, ultraproducts.

1. Introduction

The connection between logic and operator algebras in the past century was sparse albeit fruitful. Dramatic progress has brought set theory and operator algebras closer together over the last decade. A number of long-standing problems in the theory of C*-algebras were solved by using set-theoretic methods, and solutions to some of them were even shown to be independent from ZFC. There is much to be said about these developments (as witnessed in three almost disjoint recent survey papers [30, 45, 96]), but that is not what this paper is about. New applications of logic to operator algebras are being found at such a pace that any survey is bound to become obsolete within a couple of years. Instead of presenting an encyclopaedic survey, I shall proceed to describe the current developments (many of them from the unpublished joint work, [33, 34]) and outline some possible directions of research. The choice of the material reflects my interests and no attempt at completeness has been made. Several results proved by operator algebraists without using logic that have logical content are also included.

'Logic' in the title refers to model theory and (mostly descriptive) set theory, with a dash of recursion theory in a crucial place.

2. Operator algebras

Let $\mathcal{B}(H)$ denote the Banach algebra of bounded linear operators on a complex Hilbert space H equipped with the operation * of taking the adjoint. A *C*-algebra* is a Banach algebra with involution which is *-isomorphic to a subalgebra of $\mathcal{B}(H)$ for some H. Notably, all algebraic isomorphisms between C*-algebras are isometries. All C*-algebras considered here will be unital, unless otherwise specified. A *von Neumann algebra* is a unital subalgebra of $\mathcal{B}(H)$ which is closed in the weak operator topology. An algebra isomorphic to a von

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Neumann algebra is called a *W*-algebra*. Standard terminology from operator theory is imported into operator algebras, and in particular positivity of self-adjoint operators plays an important role.

I only have something to say about those von Neumann algebras that have a trace. A *normalized trace* (on a von Neumann algebra or a unital C*-algebra) is a unital positive functional such that $\tau(ab) = \tau(ba)$ for all a and b. We shall only consider unital algebras and normalized traces. A trace on a von Neumann algebra is automatically continuous in the weak operator topology. A tracial infinite-dimensional von Neumann algebra with a trivial center is a *II_1 factor*. The terminology comes from von Neumann's type classification, in which the unique I_n factor is $M_n(\mathbb{C})$; we shall not consider other types of factors.

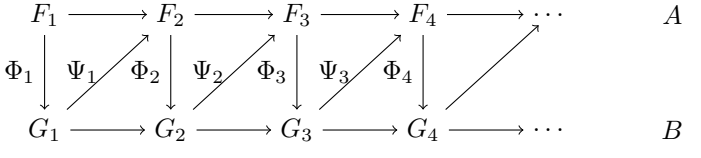
If τ is a trace on an operator algebra A then the ℓ_2-norm $\|a\|_2 = \tau(a^*a)^{1/2}$ turns A into a pre-Hilbert space. The algebra A is represented on this space by the left multiplication; this is the *GNS representation* corresponding to τ. If A is a C*-algebra, then the weak closure of the image of A is a tracial von Neumann algebra. If A is simple and infinite-dimensional, this algebra is a II_1 factor. A GNS representation can be associated to an arbitrary positive unital functional (*state*).

The category of abelian C*-algebras is equivalent to the category of locally compact Hausdorff spaces and the category of abelian von Neumann algebras with a distinguished trace is equivalent to the category of measure algebras. Because of this, these two subjects are considered to be noncommutative (or quantized) topology and measure theory, respectively.

There is only one (obvious, spatial) way to define the tensor product of von Neumann algebras. A C*-algebra A is *nuclear* if for every C*-algebra B there is a unique C*-norm on the algebraic tensor product of A and B. The importance of this notion is evident from a variety of its equivalent characterizations (see [11]), one of them being Banach-algebraic amenability. Although by a result of Junge and Pisier (see [11]) there is finite subset $F \Subset \mathcal{B}(H)$ such that no nuclear C*-algebra includes F, these algebras are ubiquitous in a number of applications.

For more on C*-algebras and von Neumann algebras see [8, 11, 55].

2.1. Intertwining. A *metric structure* is a complete metric space (A, d) equipped with functions $f\colon A^n \to A$ and predicates $p\colon A^n \to \mathbb{R}$, all of which are assumed to be uniformly continuous on d-bounded sets. Consider two separable complete metric structures A and B. Assume we have partial isometric homomorphisms $\Phi_n\colon F_n \to G_n$, $\Psi_n\colon G_n \to F_{n+1}$ for $n \in \mathbb{N}$ such that $F_n \subseteq F_{n+1} \subseteq A$ and $G_n \subseteq G_{n+1} \subseteq B$ for $n \in \mathbb{N}$ and $\bigcup_n F_n$ and $\bigcup_n G_n$ are dense in A and B respectively. Furthermore assume that in the following diagram



the n-th triangle commutes up to 2^{-n}. Then $\Phi\colon \bigcup_n F_n \to B$ defined by $\Phi(a) = \lim_n \Phi_n(a)$ and $\Psi\colon \bigcup_n G_n \to A$ defined by $\Psi(b) = \lim_n \Psi_n(b)$ are well-defined isometric homomorphisms. Their continuous extensions to A and B are respectively an isomorphism from A onto B and vice versa.

Variations of this method for constructing isomorphisms between C*-algebras comprise *Elliott's intertwining argument*. In *Elliott's program* for classification of separable, nuclear, unital and simple C*-algebras maps Φ_n and Ψ_n are obtained by lifting morphism between the K-theoretic invariants (so-called *Elliott invariants*) of A and B. The first result along these lines was the Elliott–Bratteli classification of separable AF algebras (i.e., direct limits of finite-dimensional C*-algebras) by the ordered K_0. Remarkably, for A and B belonging to a rather large class of nuclear C*-algebras this method shows that any morphism between Elliott invariants lifts to a morphism between the algebras. Elliott conjectured that the separable, nuclear, unital and simple algebras are classified by K-theoretic invariant known as the Elliott invariant. This bold conjecture was partially confirmed in many instances. See [77] for more on the early history of this fascinating subject.

Examples of separable, nuclear, unital and simple C*-algebras that limit the extent of Elliott's classification program were given in [78] and [90]. Algebras defined in [90] have a remarkable additional property. Not only do the nonisomorphic algebras A and B have the same Elliott invariant, but in addition they cannot be distinguished by any homotopy-invariant continuous functor. We shall return to these examples in §4.3. The revised Elliott program is still one of the core subjects in the study of C*-algebras (see [25]).

2.2. Strongly self-absorbing (s.s.a.) algebras. An infinite-dimensional C*-algebra is *UHF* (uniformly hyperfinite) if it is an infinite tensor product of full matrix algebras $M_n(\mathbb{C})$. If A is UHF, then every two unital copies of $M_n(\mathbb{C})$ in it are unitarily conjugate and therefore every endomormphism of A is a point-norm limit of inner automorphisms. The *generalized natural number* of A has as its 'divisors' all n such that $M_n(\mathbb{C})$ embeds unitally into A. Glimm proved that this is a complete isomorphism invariant for the separable UHF algebras.

If A is UHF then it has a unique trace τ. The tracial von Neumann algebra corresponding to the τ-GNS representation of A (§2) is the *hyperfinite II_1 factor, R*, and it does not depend on the choice of A. It is the only injective II_1 factor and it has played a key role in the classification of injective factors [19].

Two *-homomorphisms Φ and Ψ from A into B are *approximately unitarily equivalent* if there is a net of inner automorphisms α_λ, for $\lambda \in \Lambda$, of B such that

$$\lim_\lambda \alpha_\lambda \circ \Phi(a) = \Psi(a)$$

for all $a \in A$ (convergence is taken in the operator norm for C*-algebras and in the ℓ_2-norm for tracial von Neumann algebras). If $A \otimes B \cong A$ we say that A is *B-absorbing* and if $A \otimes A \cong A$ then we say that A is *self-absorbing*. Here and in what follows, we will often be providing two definitions at once, one for von Neumann algebras and another for C*-algebras. The difference comes in the interpretation of $\otimes$, either as the von Neumann (spatial) tensor product $\bar{\otimes}$ or as the C*-algebra minimal (spatial) tensor product $\otimes$. *McDuff* factors are the R-absorbing II_1 factors. A separable C*-algebra D is *strongly self-absorbing* (s.s.a.) [92] if there is an isomorphism $\Phi\colon D \to D \otimes D$ and map $\mathrm{id} \otimes 1_D : D \to D \otimes D$ is approximately unitarily equivalent to Φ. The definition of strongly self-absorbing is modified to II_1 factors following the convention stated above, by replacing $\|\cdot\|$ with $\|\cdot\|_2$ and $\otimes$ with $\bar{\otimes}$.

The hyperfinite factor R is the only s.s.a. tracial von Neumann algebra with separable predual (Stefaan Vaes pointed out that this was essentially proved in [19, Theorem 5.1(3)]). A UHF algebra A is s.s.a. if and only if it is self-absorbing. However, the latter notion is in general much stronger. For any unital C*-algebra A the infinite tensor product $\bigotimes_{\mathbb{N}} A$ is

self-absorbing but not necessarily s.s.a. Every s.s.a. C*-algebra D is simple, nuclear and unital [23].

Three s.s.a. algebras are particularly important. The *Jiang–Su* algebra $\mathcal{Z}$ is an infinite-dimensional C*-algebra which is indistinguishable from $\mathbb{C}$ by its Elliott invariant. Conjecturally, $\mathcal{Z}$-absorbing infinite-dimensional separable, nuclear, unital and simple algebras are classifiable by their Elliott invariant. The Cuntz algebra $\mathcal{O}_2$ is the universal algebra generated by two partial isometries with complementary ranges. The Cuntz algebra $\mathcal{O}_\infty$ is the universal unital C*-algebra generated by partial isometries v_n, for $n \in \mathbb{N}$, with orthogonal ranges. The first step in the Kirchberg–Phillips classification of purely infinite separable, nuclear, unital and simple algebras was Kirchberg's result that every such algebra is $\mathcal{O}_\infty$-absorbing and that $\mathcal{O}_2$ is A-absorbing for every separable, nuclear, unital and simple algebra (see [77]).

3. Abstract classification

A *Polish space* is a separable, completely metrizable topological space. A subset of a Polish space is *analytic* if it is a continuous image of some Polish space. Essentially all classical classification problems in mathematics (outside of subjects with a strong set-theoretic flavour) can be modelled by an analytic equivalence relation on a Polish space. Moreover, the space of classifying invariants is also of this form, and computation of the invariant is usually given by a Borel measurable map. This is indeed the case with C*-algebras and the Elliott invariant [43].

If E and F are equivalence relations on Polish spaces, E is *Borel-reducible* to F, $E \leq_B F$, if there exists a Borel-measurable $f\colon X \to Y$ such that $x \mathrel{\mathrm{E}} y$ if and only if $f(x) \mathrel{\mathrm{E}} f(y)$. One can interpret this as stating that the classification problem for E is not more difficult than the classification problem for F. Following Mackey, an equivalence relation E Borel-reducible to the equality relation on some Polish space is said to be *smooth*. By the *Glimm–Effros* dichotomy the class of non-smooth Borel-equivalence relations has an initial object [52], denoted E_0. It is the tail equality relation on $\{0,1\}^{\mathbb{N}}$. While the Glimm–Effros dichotomy was proved by using sophisticated tools from effective descriptive set theory, the combinatorial core of the proof can be traced back to work of Glimm and Effros on representations of locally compact groups and separable C*-algebras. See [47, 57] for more on (invariant) descriptive set theory.

When is an equivalence relation classifiable? Many non-smooth equivalence relations are considered to be satisfactorily classified. An example from the operator algebras is the Elliott–Bratteli classification of separable AF algebras by countable abelian ordered groups. A rather generous notion is being 'classifiable by countable structures.' Hjorth's theory of turbulence [54] provides a powerful tool for proving that an orbit equivalence relation is not classifiable by countable structures.

Sasyk and Törnquist have proved that every class of injective factors that was not already satisfactorily classified is not classifiable by countable structures [80, 81]. By combining results of [24, 44, 48, 71, 79], one proves that the following isomorphism relations are Borel-equireducible.

(a) Isomorphism relation of separable C*-algebras.

(b) Isomorphism relation of Elliott–classifiable separable, nuclear, unital and simple algebras.

(c) Isometry relation of separable Banach spaces.

(d) Affine homeomorphism relation of metrizable Choquet simplices.

(e) Isometry relation of Polish spaces.

Each of these equivalence relations (as well as the isometry of a class of separable metric structures of any given signature) is Borel-reducible to an orbit equivalence relation of a Polish group action [24].

Being Borel-reducible to an orbit equivalence relation is, arguably, the most generous definition of being concretely classifiable. Conjecturally, E_1, the tail-equivalence relation on $[0,1]^{\mathbb{N}}$, is an initial object among Borel equivalence relations not Borel-reducible to an orbit equivalence relation [58]. Notably, the isomorphism of separable Banach spaces is the $\leq_B$-terminal object among analytic equivalence relations [46].

The answer to the question 'When is an equivalence relation classifiable' is frequently of somewhat sociological nature. It is notable that the isomorphism relation of abelian unital C*-algebras (generally considered intractable) is Borel-reducible to the isomorphism relation of Elliott-classifiable AI algebras (for which there is a satisfactory classification relation). Also, as pointed out by David Fremlin, most analysts find that normal operators are satisfactorily classified up to conjugacy by the spectral theorem, although they are not classifiable by countable structures.

Nevertheless, the theory of Borel-reducibility is a great example of a situation in which logic provides concrete obstructions to sweeping conjectures. For example, the classification of countable abelian torsion free groups of rank $n+1$ is strictly more complicated than the classification of countable abelian torsion free groups of rank n for every n [89]. (Notably, the proof of this result uses *Popa superrigidity* of II_1 factors, [75].) This theory was recently successfully applied to (non)classification of automorphisms of group actions on operator algebras ([59], automorphisms of C*-algebras [60, 64] and subfactors [9].

A partial Borel-reducibility diagram of classification problems in operator algebras is given in Figure 3.1. For an explanation of terminology see [30, §9]. I am indebted to Marcin Sabok for pointing out that the isomorphism of countable structures of any signature is Borel-reducible to the isomorphism relation of separable AF algebras [15].

Borel-reduction of equivalence relations as defined above does not take into the account the functorial nature of the classification of C*-algebras. Some preliminary results on Borel functorial classification were obtained by Lupini.

4. Model-theoretic methods

Until recently there was not much interaction between model theory and operator algebras (although model theory was fruitfully applied to the geometry of Banach spaces, (see [53]). Recent emergence of the logic of metric structures [5], originally introduced only for bounded metric structures, created new opportunities for such interactions. It was modified to allow operator algebras in [36].

4.1. Logic of metric structures. Model theory can roughly be described as the study of axiomatizable classes of structures and sets definable in them. Axiomatizable properties can be expressed in syntactic terms, but they are also characterized by preservation under ultraproduts and ultraroots (see §6). A category $\mathcal{C}$ is *axiomatizable* if there exists a first-

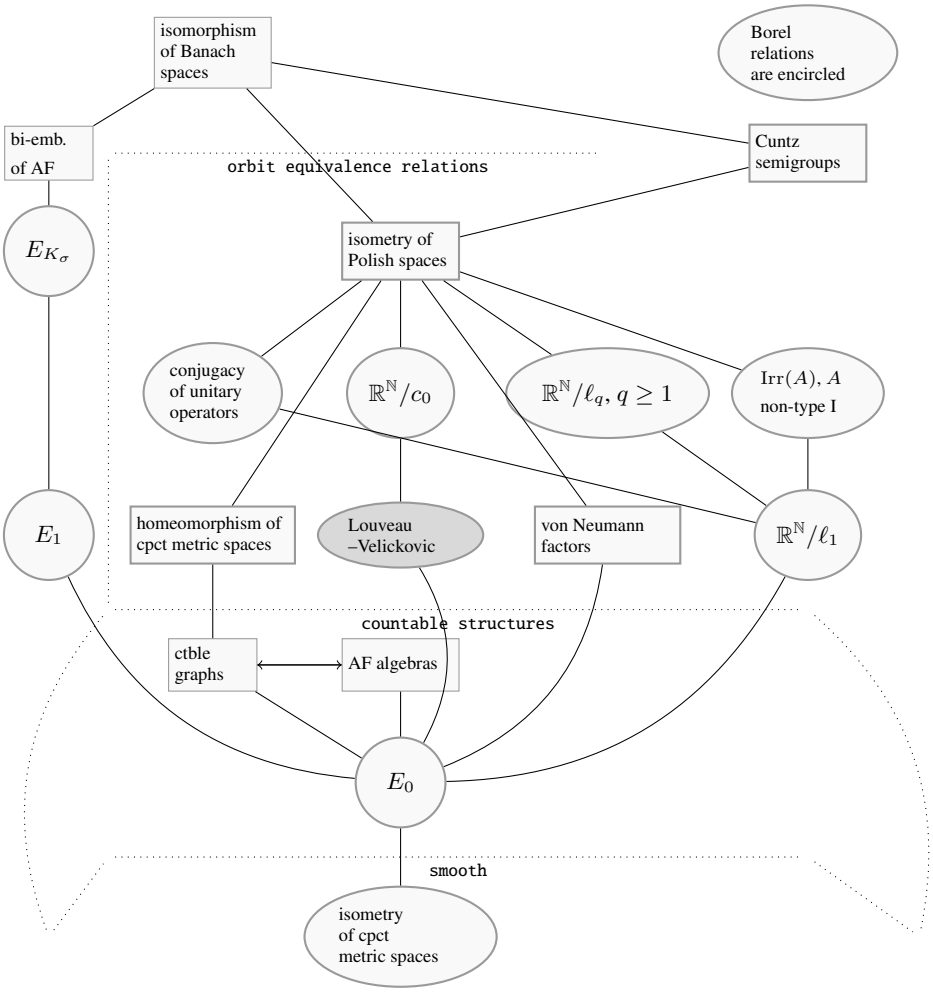


Figure 3.1.

order theory $\mathbf{T}$ such that the category $\mathfrak{M}(\mathcal{C})$ of all models of $\mathbf{T}$ is equivalent to the original category.

Classical model theory deals with discrete structures, and its variant suitable for metric structures as defined in §2.1 was introduced in [5]. In this logic interpretations of formulas are real-valued, propositional connectives are real-valued functions, and quantifiers are $\sup_x$ and $\inf_x$. Each function and predicate symbol is equipped with a modulus of uniform continuity. This modulus is a part of the language. If the diameter of the metric structures is fixed, then every formula has its own modulus of uniform continuity, respected in all relevant metric structures. Formulas form a real vector space equipped with a seminorm, $\|\phi(\bar{x})\| = |\sup \phi(\bar{a})^A|$ where the supremum is taken over all metric structures A of the given language and all tuples $\bar{a}$ in A of the appropriate type. Formulas are usually required to have range in $[0, \infty)$ (or $[0, 1]$ in the bounded case) but allowing negative values results in equivalent logic; see also [4]. The *theory* of a model is the kernel of the functional $\phi \mapsto \phi^A$,

where ϕ ranges over all sentences (i.e., formulas with no free variables) of the language. This kernel uniquely defines the functional, which can alternatively be identified with the theory. The weak*-topology on this space is also known as the *logic topology*. If the language is countable then the space of formulas is separable and the spaces of theories and types (see §4.4) are equipped with compact metric topology.

Two metric structures are *elementarily equivalent* if their theories coincide. A formula is *existential* if it is of the form $\inf_{\bar{x}} \phi(\bar{x})$ for some quantifier-free formula $\phi(\bar{x})$. The *existential theory* of A is $\mathrm{Th}_\exists(A) = \{\psi \in \mathrm{Th}(A) : \psi \text{ is existential}\}$.

There are several equivalent ways to adapt the logic of metric structures to operator algebras and to unbounded metric structures in general [3, 36]. Axiomatizability is defined via equivalence of categories as above, but model $M(A)$ associated with A has more (albeit artificial) structure. It is equipped with *domains of quantification*, bounded subsets of A on which all functions and predicates are uniformly continuous (with a fixed modulus of uniform continuity) and over which quantification is allowed. It is the existence of category $\mathfrak{M}(\mathcal{C})$, and not its particular choice, that matters.

In the simplest version of $M(A)$ quantification is allowed only over the (operator norm) n-balls of the algebra. The notion of *sorts* over which one can quantify corresponds to those functors from the model category into metric spaces with uniformly continuous functions that commute with ultraproducts (see [36, 2]). For example, $M(A)$ can be taken to consist of all matrix algebras $M_n(A)$ for $n \in \mathbb{N}$, as well as completely positive, contractive maps between them and finite-dimensional algebras. This is important because nuclearity is equivalently characterized as the CPAP, the *completely positive approximation property* (see [11] and [12]).

C*-algebras are axiomatized as Banach algebras with an involution that satisfy the *C*-equality*, $\|aa^*\| = \|a\|^2$, by the Gelfand–Naimark and Segal (GNS mentioned earlier) theorem. Abelian C*-algebras are obviously axiomatized by $\sup_{x,y} \|xy - yx\|$ and non-abelian C*-algebras are slightly less obviously axiomatized by $\inf_{\|x\|\le 1} |1 - \|x\|| + \|x^2\|$ (a C*-algebra is nonabelian if and only if it contains a nilpotent element).

The proof that the tracial von Neumann algebras with a distinguished trace are also axiomatizable ([36], first proved in [6]) goes deeper and uses Kaplansky's Density Theorem. Again, quantification is allowed over the (operator norm) unit ball and the metric is the ℓ_2 metric $\|a\|_2 = \tau(a^*a)^{1/2}$. The operator norm is not continuous with respect to the ℓ_2 metric and it therefore cannot be added to II_1 factors as a predicate.

There are elementarily equivalent but nonisomorphic separable unital AF algebras. This is proved by using descriptive set theory. The association $A \mapsto \mathrm{Th}(A)$ is Borel, and hence the relation of elementary equivalence is smooth (§3). The category of AF algebras is equivalent to the category of their ordered K_0 groups. By the Borel version of this result and the fact that the isomorphism of dimension groups is not smooth the conclusion follows.

The following proposition is taken from [34].

Proposition 4.1.

(1) *For every separable, nuclear, unital and simple C*-algebra there exists an elementarily equivalent, separable, non-nuclear, C*-algebra.*

(2) *The reduced group C*-algebra of the free group with infinitely many generators $C^*_r(F_\infty)$ is not elementarily equivalent to a nuclear C*-algebra.*

Instead of providing a genuine obstruction, this proposition precipitated some of the most interesting progress in the field.

Here is a simple but amusing observation. The *Kadison–Kastler* distance between subalgebras of $\mathcal{B}(H)$ is the Hausdorff (norm) distance between their unit balls. For every sentence ϕ the map $A \mapsto \phi^A$ is continuous with respect to this metric. Therefore the negation of an axiomatizable property is stable under small perturbations of an algebra (see [18] and references thereof for more on perturbations of C*-algebras).

4.2. Elementary submodels. If A is a submodel of B, it is said to be an *elementary submodel* if $\phi^A = \phi^B \restriction A^n$ for every n and every n-ary formula ϕ. The Downwards Löwenheim–Skolem theorem implies that every model has a separable elementary submodel. Its C*-algebraic variant is known as 'Blackadar's method' and is used to provide separable examples from known nonseparable examples (see [8, II.8.5] and [72]).

Proposition 4.2. *Assume A is a C*-algebra and B is its elementary submodel. Then B is a C*-algebra with the following properties.*

(1) *Every trace of B extends to a trace of A.*

(2) *Every ideal of B is of the form $I \cap B$ for some ideal I of A.*

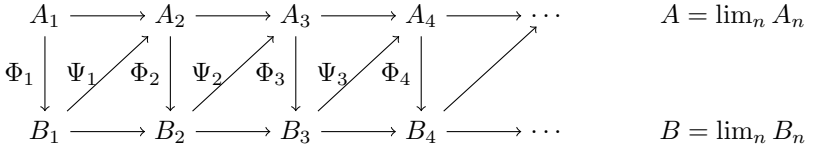
(3) *Every character of B extends to a character of A.*

(4) *If A is nuclear so is B.*

In particular, B is monotracial and/or simple if and only if A has these properties. It should be noted that neither of these properties is axiomatizable, because neither of them is preserved under taking ultrapowers (see [76] for the nonaxiomatizability of having a unique trace and §6 for the ultrapowers).

A drastic example of a property that does not persist to elementary submodels is given in Theorem 8.1.

4.3. Intertwining again. We return to Elliott's intertwining argument (§2.1):



If the maps Φ_n are expected to converge to an isomorphism, it is necessary that they approximate elementary maps. For a formula $\phi(\bar{x})$ and a tuple $\bar{a}$ in the domain of Φ_n one must have $\phi(\bar{a})^A = \lim_n \phi(\Phi_n(\bar{a}))^B$. Even more elementarily, the algebras A and B ought to be elementarily equivalent (no pun intended). Every known counterexample to Elliott's program involves separable, nuclear, unital and simple algebras with the same Elliott invariant, but different theories. For example, the radius of comparison was used in [91] to distinguish between continuum many nonisomorphis separable, nuclear, unital and simple algebras with the same Elliott invariant, and it can be read off from the theory of an algebra [34].

This motivates an outrageous conjecture, that the following question has a positive answer.

Question 4.3. *Assume that separable, nuclear, unital and simple algebras A and B have the same Elliott invariant and are elementarily equivalent. Are A and B necessarily isomorphic?*

Since being $\mathcal{Z}$-stable is axiomatizable (see §6.1), the revised Elliott conjecture that all $\mathcal{Z}$-stable separable, nuclear, unital and simple algebras are classified by their Elliott invariant is a special case of a positive answer to Question 4.3. All known nuclear C*-algebras belong to the so-called *bootstrap class*, obtained by closing the class of type I algebras under operations known to preserve nuclearity (see [8]). An (expected) negative answer to Question 4.3 would require new examples of separable, nuclear, unital and simple algebras. Can model-theoretic methods provide such examples?

4.4. Omitting types. Let $\mathbb{F}_n$ be the set of formulas whose free variables are included in $\{x_1, \dots, x_n\}$. An *n-type* is a subset $\mathbf{t}$ of $\mathbb{F}_n$ such that for every finite $\mathbf{t}_0 \Subset \mathbf{t}$ and for every $\varepsilon > 0$ there are a C*-algebra A and n-tuple $\bar{a}$ in the unit ball of A such that $|\phi(\bar{a})| < \varepsilon$ for all $\phi \in \mathbf{t}_0$. By applying functional calculus one sees that this definition is equivalent to the apparently more general standard definition ([5]) in which types consist of arbitrary closed conditions. An n-type $\mathbf{t}$ is *realized* by n-tuple $\bar{a}$ in the unit ball of A if $\phi(\bar{a})^A = 0$ for all ϕ in $\mathbf{t}$. It is *omitted* in A if it is not realized by any n-tuple in A_1. Łoś's theorem implies that every type is realized in an ultraproduct; we shall return to this in §6.0.1 but presently we are concerned with omitting types.

The omitting types theorem of classical ('discrete') model theory [67] provides a simple condition for omitting a type in a model of a given theory. A predicate p is *definable* if for every $\varepsilon > 0$ there exists a formula $\phi(\bar{x})$ which up to ε approximates the value of p. A type is definable if the distance function to its realization in a saturated model (§6.0.1, §7.1) is definable. By the omitting types theorem of [5] a type is omissible if and only if it is not definable, with the additional stipulation that it be *complete* (i.e., maximal under the inclusion). While a definable type is never omissible even if it is incomplete, Ben Yaacov has isolated types that are neither definable nor omissible. His example was simplified by T. Bice.

Theorem 4.4 ([39]).

(1) *There is a theory $\mathbf{T}$ in a separable language such that the set of types omissible in some model of $\mathbf{T}$ is a complete Σ^1_2 set.*

(2) *There are a complete theory $\mathbf{T}$ in a separable language and a countable set $\mathbf{P}$ of types such that for every finite $\mathbf{P}_0 \Subset \mathbf{P}$ there exists a model M of $\mathbf{T}$ that omits all types in $\mathbf{P}_0$, but no model of $\mathbf{T}$ omits all types in $\mathbf{P}$.*

Therefore the question of whether a type is omissible in a model of a given metric theory is by (1) far from being Borel or even analytic and therefore intractable, and by (2) separately omissible types over a complete theory are not necessarily jointly omissible. Both results stand in stark contrast to the situation in classical model theory.

The idea that the omitting types theorem can be used in the study of C*-algebras emerged independently in [14] and [83]. A sequence $\mathbf{t}_n$, for $n \in \mathbb{N}$, of m-types is *uniform* if there are formulas $\phi_j(\bar{x})$ for $j \in \mathbb{N}$ with the same modulus of uniform continuity such that $\mathbf{t}_n = \{\phi_j(\bar{x}) \geq 2^{-n} : j \in \mathbb{N}\}$ for every n. In this situation, the interpretation of the infinitary formula $\phi(\bar{x}) = \inf_j \phi_j(\bar{x})$ is uniformly continuous in every model (with a fixed modulus of uniform continuity) and moreover $\sup_{\bar{x}} \phi(\bar{x})^A = 0$ if and only if A omits all $\mathbf{t}_n$.

Nuclearity, simplicity, as well as many other important non-axiomatizable properties of C*-algebras (including nuclear dimension or decomposition rank $\leq n$; see [99]) are characterized by omitting a uniform sequence of types. The classical theory of omitting types

applies to such types unchanged: a uniform sequence of types is omissible in a model of a complete theory $\mathbf{T}$ if and only if none of the types is isolated [39]. As an extra, this characterization shows that one can find a separable elementary submodel of a nonnuclear algebra that is itself nonnuclear by assuring that it includes a tuple that realizes the relevant type

4.5. Strongly self-absorbing algebras II. These algebras have remarkable model-theoretic properties. Every s.s.a. algebra D is a *prime model* of its theory (it elementarily embeds into every other model of its theory) and every unital morphism of D into another model of its theory is elementary (§4.2).

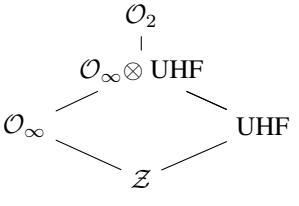
Proposition 4.5. *If D and E are s.s.a. algebras then the following are equivalent.*

(1) *E is D-*absorbing*: $E \otimes D \cong E$.*

(2) *D is isomorphic to a subalgebra of E.*

(3) $\mathrm{Th}_\exists(D) \subseteq \mathrm{Th}_\exists(E)$.

The implications from (1) to (2) and from (2) to (3) are always true, but both converses fail in general. S.s.a. algebras are as rare as they are important and the following diagram represents all known s.s.a. algebras, given in the order defined by either clause of Proposition 4.5.



Finding an s.s.a. algebra other than the ones in the diagram would refute the revised Elliott program.

5. Tracial von Neumann algebras

Many of the pathologies that plague (or enrich, depending on the point of view) the theory of C*-algebras are not present in von Neumann algebras.

By a result of McDuff, the relative commutant of a II_1 factor in its ultrapower is trivial, nontrivial and abelian, or the factor tensorially absorbs R (see Proposition 6.1). Each of these three classes is nonempty, and there is presently no other known method for distinguishing theories of II_1 factors (see [37]).

The hyperfinite II_1 factor R is a canonical object and every embedding of R into a model of its theory is elementary (§4.5). However, there are embeddings between models of the theory of R that are not elementary (i.e., the theory of R is not *model-complete*), and in particular this theory does not allow the elimination of quantifiers [31, 51]. This may be an indication that we do not have the right language for the theory of II_1 factors. The obstruction for the elimination of quantifiers extracted in [31] from [56] is removed by adding a predicate for the unitary conjugacy relation. As this is a definable relation, adding such predicate

affects only syntactical structure of the language. It is not clear whether adding finitely, or even countably, many such predicates could make the theory of R model-complete. This may suggest that the theory of R is as complicated as the first-order arithmetic or ZFC.

Given a II_1 factor M and a projection p in M, are M and its corner pMp elementarily equivalent? By the Keisler–Shelah theorem, this is equivalent to asking whether these algebras have isomorphic ultrapowers. A positive answer would imply that all free group factors $L(F_n)$, for $n \geq 2$, are elementarily equivalent, giving a 'poor man's' solution to the well-known problem whether the free group factors are isomorphic (see [22]). On the other hand, a negative answer would provide a continuum of distinct theories of II_1 factors that are corners of $L(F_2)$. A deeper analysis of the model theory of II_1 factors is likely to involve Voiculescu's free probability.

In recent years theories of C*-algebras and von Neumann algebras are increasingly considered as inseparable. Some of the most exciting progress on understanding tracial C*-algebras was initiated in [68]. We shall return to this in §6.1, but see also [12].

6. Massive algebras I: Ultraproducts

We now consider algebras that are rarely nuclear and never separable, but are nevertheless indispensable tools in the study of separable nuclear algebras.

Ultraproducts emerged independently in logic and in functional analysis (more precisely, in the theory of II_1 factors) in the 1950's (see the introduction to [88]). If (A_n, d_n), for $n \in \mathbb{N}$, are bounded metric structures of the same signature and $\mathcal{U}$ is an ultrafilter on $\mathbb{N}$, then the *ultraproduct* $\prod_{\mathcal{U}} A_n$ is defined as follows. On the product structure $\prod_n A_n$ consider the quasi-metric

$$d_{\mathcal{U}}((a_n),(b_n)) = \lim_{n\to\mathcal{U}} d_n(a_n, b_n).$$

Since every function symbol f has a fixed modulus of uniform continuity, it defines a uniformly continuous function on the quotient metric structure $\prod_n A_n/\sim_{d_{\mathcal{U}}}$. This structure is the *ultraproduct* of A_n, for $n \in \mathbb{N}$, associated to the ultrafilter $\mathcal{U}$. It is denoted by $\prod_{\mathcal{U}} A_n$.

In the not necessarily bounded case one replaces $\prod_n A_n$ with $\{(a_n) \in \prod_n A_n : a_n$ belong to the same domain of quantification$\}$. With our conventions, in the operator algebra case this is the ℓ_∞-product usually denoted $\prod_n A_n$. The nontrivial fact that an ultrapower of tracial von Neumann algebras is a tracial von Neumann algebra is an immediate consequence of the axiomatizability.

The usefulness of ultraproducts draws its strength largely from two basic principles. The first one is *Łoś's theorem*, stating that for any formula $\phi(\bar{x})$ we have

$$\phi(\bar{a})^{\prod_{\mathcal{U}} A_n} = \lim_{n\to\mathcal{U}} \phi(\bar{a}_n)^{A_n}.$$

This in particular implies that the diagonal embedding of A into its ultrapower is elementary (§4.2), and therefore the theory is preserved by taking ultrapowers. The second principle will be discussed in §6.0.1.

This may be a good place to note two results in abstract model theory that carry over to the metric case [5]. A category $\mathbb{K}$ with an appropriately defined ultraproduct construction is closed under the elementary equivalence if and only if it is closed under isomorphisms, ultraproducts, and ultraroots (i.e., $A^{\mathcal{U}} \in \mathbb{K}$ implies $A \in \mathbb{K}$). By the Keisler–Shelah theorem, two models are elementarily equivalent if and only if they have isomorphic ultrapowers. Both results require considering ultrafilters on arbitrarily large sets (see [86]).

The fact that it is easier to prove that an ultraproduct of C*-algebras is a C*-algebras than that an ultraproduct of tracial von Neumann algebras is a tracial von Neumann algebra is reflected in the fact that it is easier to prove that the C*-algebras are axiomatizable than that the tracial von Neumann algebras are axiomatizable.

All ultrafilters considered here concentrate on $\mathbb{N}$ and are nonprincipal. It is not possible to construct such an ultrafilter in ZF alone, as a (rather weak) form of the Axiom of Choice is required for its construction. However, results about separable C*-algebras and separably acting II_1 factors proved using ultrafilters can be proved without appealing to the Axiom of Choice, by standard absoluteness arguments.

An ultrapower of an infinite-dimensional, simple, unital C*-algebra is by Łoś's theorem unital. It is, however, nonseparable, not nuclear, and it is simple only under exceptional circumstances. This shows that none of these three properties is axiomatizable (cf. Proposition 4.1). Nevertheless, separable, nuclear, unital and simple C*-algebras can be constructed by using the Henkin construction and omitting types theorem ([39], see §4.4).

6.0.1. Countable saturation. We define the second important property of massive algebras. If a type (see §4.4) is allowed to contain formulas with parameters from an algebra A we say that it is a type *over* A.

An algebra A is *countably saturated* if every countable type $\mathbf{t}(\bar{x})$ over A is realized in A if and only if it is consistent. (These algebras are sometimes said to be $\aleph_1$*-saturated*. The latter terminology is more conveniently extended to higher cardinalities.) Every ultrapower associated to a nonprincipal ultrafilter on $\mathbb{N}$ is countably saturated. A weakening of countable saturation suffices for many purposes (see §7), and we shall return to full saturation in §7.1.

6.1. Relative commutants. In the theory of operator algebras even more important than the ultrapower itself is the *relative commutant* of the algebra inside the ultrapower,

$$A' \cap A^{\mathcal{U}} = \{b \in A^{\mathcal{U}} : ab = ba \text{ for all } a \in A\}.$$

The current prominence of ultrapowers as a tool for studying separable algebras can be traced back to McDuff ([70]) and the following proposition (generalized to s.s.a. algebras in [92]).

Proposition 6.1. *If D is strongly self-absorbing and A is separable, then A is D-absorbing if and only if D embeds into $A' \cap A^{\mathcal{U}}$.*

The nontrivial, converse, implication uses the following (a lemma in model theory that I learned from Wilhelm Winter) proved using the intertwining argument.

Lemma 6.2. *If $A \subseteq B$ are separable metric structures and $B^{\mathcal{U}}$ has a sequence of isometric automorphisms α_n such that $\lim_n \alpha_n(a) = a$ for all $a \in A$ and $\lim_n \operatorname{dist}(\alpha_n(b), A) = 0$ for all $b \in B$, then A and B are isometrically isomorphic.*

Noting that all nonprincipal ultrafilters on $\mathbb{N}$ 'look the same' and in particular that the choice of $\mathcal{U}$ in Proposition 6.1 is irrelevant as long as it is a nonprincipal ultrafilter on $\mathbb{N}$, one may ask the following.

Question 6.3. *If M is a separable metric structure, does the isomorphism type of $M^{\mathcal{U}}$ (and $M' \cap M^{\mathcal{U}}$, if M is a Banach algebra) depend on $\mathcal{U}$ at all?*

If M is a Hilbert space or a measure algebra, then a simple argument (using Maharam's theorem in the latter case) gives a negative answer. Also, Continuum Hypothesis (CH) im-

plies negative answer to both questions for an arbitrary separable M (see §7.1). Therefore, the question is whether CH can be removed from this proof.

Question 6.3 for relative commutants was asked by McDuff [70] and Kirchberg ([61]) in the case of McDuff factors and C*-algebras, respectively. In [49] it was proved that, under some additional assumptions on M, CH is equivalent to the positive answer to either of these questions! This was achieved by using only results from classical ('discrete') model theory. By using the logic of metric structures and Shelah's non-structure theory, the full result was proved in [35] and [41].

Theorem 6.4. *Assume CH fails. If M is a separable C*-algebra or a McDuff factor with a separable predual, then M has $2^{2^{\aleph_0}}$ nonisomorphic ultrapowers and $2^{2^{\aleph_0}}$ nonisomorphic relative commutants associated to nonprincipal ultrafilters on $\mathbb{N}$.*

Let's zoom out a bit. A complete first-order theory $\mathbf{T}$ has the *order property* if there exist $n \geq 1$ and a $2n$-ary formula $\phi(\bar{x}, \bar{y})$ such that for every m there is a model $\mathfrak{M}$ of $\mathbf{T}$ which has a 'ϕ-chain' of length at least m. A *ϕ-chain* is a sequence $\bar{x}_i, \bar{y}_i$, for $i \leq m$, such that

$$\phi(\bar{x}_i, \bar{y}_j) = 0 \text{ if } i \leq j \text{ and } \phi(\bar{x}_i, \bar{y}_j) = 1 \text{ if } i > j.$$

This is the metric version of one of the important non-structural properties of theories in Shelah's stability theory ([85] and [35]). The theory of any infinite-dimensional C*-algebra and of any II_1 factor has the order property. This is proved by continuous functional calculus and by utilizing noncommutativity, respectively. However, the theories of abelian tracial von Neumann algebras do not have the order property, essentially by applying Maharam's theorem on measure algebras.

Theorem 6.5. *Suppose that A is a separable structure in a separable language.*

(1) *If the theory of A does not have the order property then all of its ultrapowers associated to nonprincipal ultrafilters on $\mathbb{N}$ are isomorphic.*

(2) *If the theory of A has the order property then the following are equivalent:*

   (a) *A has fewer than $2^{2^{\aleph_0}}$ nonisomorphic ultrapowers associated with nonprincipal ultrafilters on $\mathbb{N}$.*

   (b) *all ultrapowers of A associated to nonprincipal ultrafilters on $\mathbb{N}$ are isomorphic.*

   (c) *the Continuum Hypothesis holds.*

6.2. Model theory of the relative commutant. The notion of a relative commutant does not seem to have a useful generalization in the abstract model theory and its model-theoretic properties are still poorly understood.

While the structure of relative commutants of II_1 factors in their ultrapowers provides the only known method for distinguishing their theories, every infinite-dimensional separable C*-algebra has a nontrivial relative commutant in its ultrapower ([61], also [35]). The relative commutant of the Calkin algebra (§7) in its ultrapower is trivial [61] and the relative commutant of $\mathcal{B}(H)$ may or may not be trivial, depending on the choice of the ultrafilter [40].

It is not difficult to see that the existential theory of $A' \cap A^{\mathcal{U}}$ depends only on the theory of A. However, a result of [61] implies that there is a separable C*-algebra A elementarily

equivalent to $\mathcal{O}_2$ such that $A' \cap A^{\mathcal{U}}$ and $\mathcal{O}_2 \cap \mathcal{O}_2^{\mathcal{U}}$ have different $\forall\exists$-theories. (An $\forall\exists$-*sentence* is one of the form $\sup_{\bar{x}} \inf_{\bar{y}} \phi(\bar{x}, \bar{y})$ where ϕ is quantifier-free.) In the following all ultrafilters are nonprincipal ultrafilters on $\mathbb{N}$.

Proposition 6.6. *Assume A is a separable C*-algebra.*

(1) *For all $\mathcal{U}$ and $\mathcal{V}$, the algebras $A' \cap A^{\mathcal{U}}$ and $A' \cap A^{\mathcal{V}}$ are elementarily equivalent.*

(2) *For every separable $C \subseteq A' \cap A^{\mathcal{U}}$ we have $\mathrm{Th}_\exists(A' \cap C' \cap A^{\mathcal{U}}) = \mathrm{Th}_\exists(A' \cap A^{\mathcal{U}})$.*

(3) *If D is a separable unital subalgebra of $A' \cap A^{\mathcal{U}}$ then there are $\aleph_1$ commuting copies of D inside $A' \cap A^{\mathcal{U}}$.*

An entertaining proof of (1) can be given by using basic set theory. Collapse $2^{\aleph_0}$ to $\aleph_1$ without adding reals. Then $\mathcal{U}$ and $\mathcal{V}$ are still ultrafilters on $\mathbb{N}$ and one can use saturation to find an isomorphism between the ultrapowers that sends A to itself. The theories of two algebras are unchanged, and therefore by absoluteness the result follows. Clause (3) is an immediate consequence of (2) and it is a minor strengthening of a result in [61].

When A is not $\mathcal{Z}$-stable, the relative commutant of A can have characters even if it is simple ([62]). In the case when algebra A is nuclear and $\mathcal{Z}$-stable, $A' \cap A^{\mathcal{U}}$ inherits some properties from A. For example, each of the traces on $A' \cap A^{\mathcal{U}}$ extends to a trace on $A^{\mathcal{U}}$ by [68] (cf. Proposition 4.2). The relative commutants of s.s.a. algebras are well-understood; the following was proved in [33].

Proposition 6.7. *If D is a s.s.a. algebra and $\mathcal{U}$ is a nonprincipal ultrafilter on $\mathbb{N}$, then $D' \cap D^{\mathcal{U}}$ is an elementary submodel of $D^{\mathcal{U}}$. Moreover, CH implies that these two algebras are isomorphic.*

6.3. Expansions and traces. If a metric structure A is expanded by adding a new predicate τ, its ultrapower $A^{\mathcal{U}}$ expands to the ultrapower of the expanded structure $(A, \tau)^{\mathcal{U}}$ which still satisfies Łoś's theorem and is countably saturated.

If A is a unital tracial C*-algebra then its traces form a weak*-compact convex subset $T(A)$ of the dual unit ball. For $\tau \in T(A)$ denote the tracial von Neumann algebra associated with the τ-GNS representation (§2) by N_τ. If A is simple and infinite-dimensional and τ is an extremal trace then N_τ is a factor, and if A is in addition nuclear and separable then N_τ is isomorphic to the hyperfinite factor R. This is because A is nuclear if and only if its weak closure in every representation is an injective von Neumann algebra, and R is the only injective II_1 factor with a separable predual. The following was proved in [68] and improved to the present form in [62].

Proposition 6.8. *If A is separable and $\tau \in T(A)$, then the quotient map from $A' \cap A^{\mathcal{U}}$ to $N_\tau' \cap (N_\tau)^{\mathcal{U}}$ is surjective.*

If $b \in A^{\mathcal{U}}$ is such that its image is in the commutant of N_τ', then by countable saturation one finds a positive element c of norm 1 such that $\tau(c) = 0$ and $c(a_n b - b a_n) = (a_n b - b a_n)c = 0$ for all a_n in a fixed countable dense subset of A. The fact that the type of such c is consistent follows from the fact that the image of b is in N_τ'. Then $(1-c)b(1-c)$ is in $A' \cap A^{\mathcal{U}}$ and it has the same image under the quotient map as b.

Proposition 6.8 precipitated remarkable progress on understanding tracial C*-algebras, the most recent results of which are [69] and [82].

7. Massive algebras II: Coronas

Another class of massive C*-algebras (with no analogue in von Neumann algebras) has special relevance to the study of separable algebras. If A is a non-unital C*-algebra, the *multiplier algebra* of A, $M(A)$, is the noncommutative analogue of the Čech–Stone compactification of a locally compact Hausdroff space. It is the surjectively universal unital algebra containing A as an essential ideal. The *corona* (or *outer multiplier*) algebra of A is the quotient $M(A)/A$. Some examples of coronas are the Calkin algebra $\mathcal{Q}(H)$ (the corona of the algebra of compact operators) and the *asymptotic sequence algebra* $\ell_\infty(A)/c_0(A)$ for a unital A. The latter algebra, as well as the associated *central sequence algebra* $A' \cap \ell_\infty(A)/c_0(A)$ are sometimes used in classification of C*-algebras instead of the metamathematically heavier ultrapowers and the corresponding relative commutants. While Łoś's theorem miserably fails for the asymptotic sequence algebra, all coronas and corresponding relative commutants share some properties of countably saturated algebras. The simplest of these properties is being SAW*: for any two orthogonal separable subalgebras A and B of a corona there exists a positive element c such that $ca = a$ for all $a \in A$ and $cb = 0$ for all $b \in B$.

7.0.1. Quantifier-free saturation. An algebra C is *quantifier-free saturated* if every countable type over C consisting only of quantifier-free formulas is consistent if and only if it is realized in C. An algebra C is *countably degree-1 saturated* if every countable type over C consisting only of formulas of the form $\|p\|$, where p is a *-polynomial of degree 1, is consistent if and only if it is realized in C. A dummy variable argument shows that the degree-2 saturation is equivalent to quantifier-free saturation. By refining an argument introduced by Higson, the following was proved in [32].

Theorem 7.1. *If A is a corona of a separable non-unital C*-algebra, or a relative commutant of a separable subalgebra of such corona, then A is countably degree-1 saturated.*

A very interesting class of countable degree-1 saturated C*-algebras was isolated in [94].

7.0.2. A sampler of properties of countable degree-1 saturated algebras. Assume C is countably degree-1 saturated (the results below also apply to tracial von Neumann algebras, and in this case (1), (3) and (5) do not even require countable degree-1 saturation).

(1) C has SAW* as well as every other known countable separation property [32].

(2) A separable algebra A is isomorphic to a unital subalgebra of C if and only if $\mathrm{Th}_\exists(A) \subseteq \mathrm{Th}_\exists(C)$.

(3) A representation of a group Γ in A is a homomorphism $\pi\colon \Gamma \to (\mathrm{GL}(A), \cdot)$. It is *unitarizable* if there is an invertible $h \in A$ such that $h^{-1}\pi(g)h$ is a unitary for all $g \in \Gamma$. Conjecturally unitarizability of all uniformly bounded representations of a group Γ on $\mathcal{B}(H)$ is equivalent to the amenability of Γ (see [74]). If Γ is a countable amenable group, then every uniformly bounded representation π of Γ in C is unitarizable [17].

(4) C is not isomorphic to the tensor product of two infinite-dimensional algebras ([26] for the ultraproducts of II_1 factors and [50] for the general result). Therefore an ultrapower or a corona is never isomorphic to a nontrivial tensor product and the separability assumption is needed in Proposition 6.1.

(5) ('Discontinuous functional calculus.') If a is a normal operator, then by the *continuous functional calculus* for every continuous complex-valued function g on the spectrum,

$\mathrm{sp}(a)$, of a the naturally defined $g(a)$ belongs to the abelian algebra generated by a.

Proposition 7.2. *Assume C is countably degree-1 saturated and $B \subseteq \{a\}' \cap C$ is separable, $U \subseteq \mathrm{sp}(a)$ is open, and $g\colon U \to \mathbb{C}$ is a bounded continuous function. Then there exists $c \in C \cap C^*(B,a)'$ such that for every $f \in C_0(\mathrm{sp}(a))$ we have*

$$cf(a) = (gf)(a).$$

If moreover g is real-valued then c can be chosen to be self-adjoint.

The 'Second Splitting Lemma' ([10, Lemma 7.3]) is a special case of the above when C is the Calkin algebra, $a = h_0$ is self-adjoint, and the range of g is $\{0,1\}$.

7.0.3. Failure of saturation. While the asymptotic sequence algebras, as well as some abelian coronas, are fully countably saturated [42], this is not true for sufficiently noncommutative coronas. By a K-theoretic argument N. C. Phillips constructed two unital embeddings of the CAR algebra into the Calkin algebra $\mathcal{Q}(H)$ that are approximately unitarily equivalent, but not conjugate by a unitary ([32, §4]). This gives a countable quantifier-free type over $\mathcal{Q}(H)$ that is consistent but not realized. Even coronas of separable abelian C*-algebras provide a range of different saturation properties (see [42]).

7.1. Automorphisms. A metric model A is *saturated* if every type over A whose cardinality is smaller than the *density character* $\chi(A)$ of A (i.e., the smallest cardinality of a dense subset) which is consistent is realized in A. The Continuum Hypothesis (CH) implies that all countably saturated models of cardinality $2^{\aleph_0}$ are saturated. A transfinite back-and-forth argument shows that any two elementarily equivalent saturated models of the same density character are isomorphic and that a saturated model A has $2^{\chi(A)}$ automorphisms. By a counting argument, most of these automorphisms are outer and moreover nontrivial when 'trivial automorphism' is defined in any reasonable way; see [20] for a (lengthy) discussion. This explains the effectiveness of CH as a tool for resolving problems of a certain form. A deeper explanation is given in Woodin's celebrated Σ^2_1-absoluteness theorem (see [100]).

By the above, CH implies that an ultrapower $A^{\mathcal{U}}$ of a separable, infinite-dimensional algebra has automorphisms that do not lift to automorphisms of $\ell_\infty(A)$. Much deeper is a complementary series of results of Shelah, to the effect that if ZFC is consistent then so is the assertion that any isomorphism between ultraproducts of models with the strong independence property lifts to an isomorphism of the products of these models [87]. No continuous version of this result is known. One difficulty in taming ultrapowers is that the ultrafilter is not a definable object; in particular Shelah's results apply only to a carefully constructed ultrafilter in a specific model of ZFC.

Motivated by work on extension theory and a very concrete question about the unilateral shift, in [10] it was asked whether the Calkin algebra has outer automorphisms. Since the Calkin algebra is not countably saturated (§7.0.3) it took some time before such an automorphism was constructed using CH [73]. This is one of the most complicated known CH constructions, involving an intricate use of EE-theory to extend isomorphisms of direct limits of separable subalgebras. A simpler proof was given in [28, §1], and the method was further refined in [20]. Instead of following the usual back-and-forth construction in which isomorphisms between separable subalgebras are recursively extended, one uses CH to embed the first derived limit of an inverse system of abelian groups into the outer automorphism group.

Forcing axioms imply that the Calkin algebra has only inner automorphisms [28]. Conjecturally, for every non-unital separable C*-algebra the assertion that its corona has only (appropriately defined) 'trivial' automorphisms is independent of ZFC (see [20]). Even the abelian case of this conjecture is wide open [42].

The 'very concrete question' of Brown–Douglas–Fillmore alluded to two paragraphs ago is still wide open: Is there an automorphism of $\mathcal{Q}(H)$ that sends the image of the unilateral shift $\dot{s}$ to its adjoint? Fredholm index obstruction shows that such an automorphism cannot be inner. Since the nonexistence of outer automorphisms of $\mathcal{Q}(H)$ is relatively consistent with ZFC, so is a negative answer to the BDF question. Every known automorphism α of $\mathcal{Q}(H)$ in every model of ZFC has the property that its restriction to any separable subalgebra is implemented by a unitary. Both $\dot{s}$ and $\dot{s}^*$ are unitaries with full spectrum and no nontrivial roots. It is, however, not even known whether $\dot{s}$ and $\dot{s}^*$ have the same (parameter-free) type in $\mathcal{Q}(H)$; a positive answer would provide a strong motivation for the question of whether $\mathcal{Q}(H)$ is countably homogeneous.

7.2. Gaps. A *gap* in a semilattice $\mathcal{B}$ is a pair $\mathcal{A}$, $\mathcal{B}$ such that $a \wedge b = 0$ for all $a \in \mathcal{A}$ and all $b \in \mathcal{B}$ but there is no c such that $c \wedge a = a$ and $c \wedge b = 0$ for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$. There are no countable gaps in a countably saturated Boolean algebra such as $\mathcal{P}(\mathbb{N})/\operatorname{Fin}$, the quotient of $\mathcal{P}(\mathbb{N})$ over the ideal Fin of finite sets. In 1908 Hausdorff constructed a gap in $\mathcal{P}(\mathbb{N})/\operatorname{Fin}$ with both of its sides of cardinality $\aleph_1$. Later Luzin constructed a family of $\aleph_1$ orthogonal elements in $\mathcal{P}(\mathbb{N})/\operatorname{Fin}$ such that any two of its disjoint uncountable subsets form a gap. It should be emphasized that both results were proved without using CH or any other additional set-theoretic axioms.

Hausdorff's and Luzin's results show that $\mathcal{P}(\mathbb{N})/\operatorname{Fin}$ is not more than countably saturated. In particular, if the Continuum Hypothesis fails then the obvious back-and-forth method for constructing automorphisms of $\mathcal{P}(\mathbb{N})/\operatorname{Fin}$ runs into difficulties after the first $\aleph_1$ stages. In one form or another, gaps were used as an obstruction to the existence of morphisms in several consistency results in analysis, notably as obstructions to extending a partial isomorphism ([84, §V], [21, 28]).

Two subalgebras A and B of an ambient algebra C form a *gap* if $ab = 0$ for all $a \in A$ and $b \in B$, but there is no positive element c such that $ca = a$ and $cb = 0$ for all $a \in A$ and all $b \in B$. The gap structure of $\mathcal{P}(\mathbb{N})/\operatorname{Fin}$ can be imported into the Calkin algebra, but the gap structure of the latter is also much richer [101].

However, the failure of higher saturation in coronas is also manifested in a genuinely noncommutative fashion. A countable family of commuting operators in a corona of a separable algebra can be lifted to a family of commuting operators if and only if this is true for each one of its finite subsets.

Proposition 7.3. *In $M_2(\ell_\infty/c_0)$ there exists a family of $\aleph_1$ orthogonal projections such that none of its uncountable subsets can be lifted to a commuting family of projections in $M_2(\ell_\infty)$.*

This was stated in [45] for the Calkin algebra in place of (barely noncommutative) $M_2(\ell_\infty/c_0)$, but the proof given there clearly gives the stronger result. The combinatorial essence for the proof of Proposition 7.3 echoes Luzin's original idea. One recursively constructs projections p_γ in $M_2(\ell_\infty)$ so that $p_\gamma p_{\gamma'}$ is compact but $\|[p_\gamma, p_{\gamma'}]\| > 1/4$ for all $\gamma \neq \gamma'$. Then the image this family in the corona is as required, as a counting argument shows that no uncountable subfamily can be simultaneously diagonalized.

Recall that every uniformly bounded representation of a countable amenable group in a countably degree-1 saturated algebra is unitarizable (Proposition 7.2). This is false for uncountable groups. This was proved in [17] and improved to the present form in [93] using Luzin's gap.

Proposition 7.4. *There is a uniformly bounded representation π of $\bigoplus_{\aleph_1}\mathbb{Z}/2\mathbb{Z}$ on $M_2(\ell_\infty/c_0)$ such that the restriction of π to a subgroup is unitarizable if and only if the subgroup is countable.*

The construction of Kadison–Kastler-near, but not isomorphic, nonseparable algebras in [16] involves what at the hindsight can be considered as a gap. It is not known whether there is a separable example (see [18] for several partial positive results).

8. Nonseparable algebras

Not surprisingly, the theory of nonseparable algebras hides surprises and problems not present in the separable case; see [95].

8.1. Nonseparable UHF algebras. Uniformly hyperfinite (UHF) algebras are defined as tensor products of full matrix algebras (§2.2). However, there are two other natural ways to define uniformly hyperfinite: as (i) an inductive limit of a net of full matrix algebras, or (ii) as an algebra in which every finite subset can be arbitrarily well approximated by a full matrix subalgebra. These three notions, given in the order of decreasing strength, coincide in the separable unital case. Dixmier asked whether separability is needed for this conclusion. The answer is that in every uncountable density character, UHF and (i) differ, but that one needs an algebra of density character $\aleph_2$ in order to distinguish between (i) and (ii) [38]. An extension of methods of [38] resulted in a nuclear, simple C*-algebra that has irreducible representations on both separable and nonseparable Hilbert space [27]. This is in contrast with the transitivity of the space of irreducible representations of a separable simple C*-algebra [63].

8.2. Representation theory. Representation theory of separable algebras has deeply affected development of the classical descriptive set theory, as evident from the terminology of both subjects (terms 'smooth' and 'analytic' have the same, albeit nonstandard in other areas of mathematics, meaning). Extension of the work of Glimm and Effros on representation theory combined with methods from logic initiated the abstract classification theory (§3). The representation theory of nonseparable algebras was largely abandoned because some of the central problems proved to be intractable (see the introduction to [1]). One of these stumbling blocks, *Naimark's problem*, was partially solved in [1] (see also [96]). By using a strengthening of CH (Jensen's $\diamondsuit_{\aleph_1}$ principle) and a deep result on representation theory of separable C*-algebras (an extension of [63] mentioned above), Akemann and Weaver constructed a C*-algebra that has a unique (up to spatial equivalence) irreducible representation on a Hilbert space, but is not isomorphic to the algebra of compact operators on any Hilbert space. An extension of [1] shows that $\diamondsuit_{\aleph_1}$ implies the existence of a simple C*-algebra with exactly m inequivalent irreducible representations. By a classical result of Glimm (closely related to the Glimm–Effros dichotomy), a simple separable C*-algebra with two inequivalent representations has $2^{\aleph_0}$ inequivalent representations. It is not known

whether a counterexample to Naimark's problem can be found in ZFC alone or by using an axiom other than $\diamondsuit_{\aleph_1}$ (such as $\diamondsuit_\kappa$ for $\kappa > \aleph_1$). The fact that every forcing notion that adds a new real number destroys all ground-model examples is a bit of an annoying teaser.

Cyclic representations of C*-algebras are, via the GNS construction (§2), in a natural bijective correspondence with their states (i.e., positive unital functionals). Pure (i.e., extremal) states are noncommutative versions of ultrafilters. The space of nonprincipal ultrafilters on $\mathbb{N}$, (along with the associated quotient structure $\mathcal{P}(\mathbb{N})/\operatorname{Fin}$) is arguably the most important set-theoretically malleable object known to man. The study of pure states on $\mathcal{B}(H)$ (i.e., 'quantized ultrafilters') has already produced some surprising results ([2, 7]; also see [66]).

8.3. Amenable operator algebras. A prominent open problem in the theory of operator algebras is whether every algebra of operators on a Hilbert space which is amenable is isomorphic to a C*-algebra. By using Proposition 7.4, one obtains the following [17, 93]).

Theorem 8.1. *There exists a nonseparable amenable subalgebra of $M_2(\ell_\infty)$ which is not isomorphic to a C*-algebra. None of its nonseparable amenable subalgebras is isomorphic to a C*-algebras, yet it is an inductive limit of separable subalgebras (even elementary submodels) each of which is isomorphic to a C*-algebra. Moreover, for every $\varepsilon > 0$ such an algebra can be found in an ε-Kadison–Kastler neighbourhood of a C*-algebra.*

The question whether there exists a separable counterexample remains open; see [65].

9. Concluding remarks

The most recent wave of applications of logic to operator algebras started by work of Nik Weaver and his coauthors, in which several long-standing problems were solved by using additional set-theoretic axioms (see [96]). Although we now know that the answers to some of those problems (such as the existence of outer automorphisms of the Calkin algebra) are independent from ZFC, statements of many prominent open problems in operator algebras are absolute between models of ZFC and therefore unlikely to be independent (see the appendix to [29] for a discussion).

Nevertheless, operator algebras do mix very well with logic. Jon Barwise said "As logicians, we do our subject a disservice by convincing others that the logic is first-order and then convincing them that almost none of the concepts of modern mathematics can really be captured in first-order logic." Remarkably, some of the deepest results on the structure of C*-algebras have equivalent formulation in the language of (metric) first-order logic (this applies e.g., to [97] and [98]).

In many of the developments presented here methods from logic were blended with highly nontrivial operator-algebraic methods. Good examples are the proof that the theory of R does not allow elimination of quantifiers [51] the key component of which comes from [13], the already mentioned use of [56], and blending of $\diamondsuit_{\aleph_1}$ with the transitivity of pure state space of separable simple algebras [63] in [1].

Finally, some results in pure logic were motivated by work on operator algebras. Examples are Theorem 6.5, which is new even for discrete structures, and negative and positive results on omitting types (§4.4).

Acknowledements. I am indebted to Bradd Hart, Isaac Goldbring, Aaron Tikuisis and Yemon Choi for a number of remarks on the first draft of the present paper that considerably improved it in many ways. This work was partially supported by NSERC.

References

[1] C. Akemann and N. Weaver, *Consistency of a counterexample to Naimark's problem*, Proc. Natl. Acad. Sci. USA **101** (2004), no. 20, 7522–7525.

[2] ———, *$\mathcal{B}(H)$ has a pure state that is not multiplicative on any masa*, Proc. Natl. Acad. Sci. USA **105** (2008), no. 14, 5313–5314.

[3] I. Ben Yaacov, *Continuous first order logic for unbounded metric structures*, Journal of Mathematical Logic **8** (2008), 197–223.

[4] ———, *On a C*-algebra formalism for continuous first order logic*, preprint, available at http://math.univ-lyon1.fr/homes-www/begnac/., 2008.

[5] I. Ben Yaacov, A. Berenstein, C.W. Henson, and A. Usvyatsov, *Model theory for metric structures*, Model Theory with Applications to Algebra and Analysis, Vol. II (Z. Chatzidakis et al., eds.), London Math. Soc. Lecture Notes Series, no. 350, Cambridge University Press, 2008, pp. 315–427.

[6] I. Ben Yaacov, C.W. Henson, M. Junge, and Y. Raynaud, *Preliminary report - vNA and NCP*, preprint, 2008.

[7] T. Bice, *Filters in C^*-algebras*, Canad. J. Math. **65** (2013), no. 3, 485–509.

[8] B. Blackadar, *Operator algebras*, Encyclopaedia of Mathematical Sciences, vol. 122, Springer-Verlag, Berlin, 2006, Theory of C^*-algebras and von Neumann algebras, Operator Algebras and Non-commutative Geometry, III.

[9] A. Brothier and S. Vaes, *Families of hyperfinite subfactors with the same standard invariant and prescribed fundamental group*, arXiv preprint arXiv:1309.5354 (2013).

[10] L.G. Brown, R.G. Douglas, and P.A. Fillmore, *Unitary equivalence modulo the compact operators and extensions of C^*-algebras*, Proceedings of a Conference on Operator Theory, Lecture Notes in Math., vol. 345, Springer, 1973, pp. 58–128.

[11] N. Brown and N. Ozawa, *C^*-algebras and finite-dimensional approximations*, Graduate Studies in Mathematics, vol. 88, American Mathematical Society, Providence, RI, 2008.

[12] N.P. Brown, *The symbiosis of C*- and W*-algebras*, Contemporary Mathematics (2011), 121–155.

[13] ———, *Topological dynamical systems associated to II_1 factors*, Adv. Math. **227** (2011), no. 4, 1665–1699, With an appendix by Narutaka Ozawa.

[14] K. Carlson, E. Cheung, I. Farah, A. Gerhardt-Bourke, B. Hart, L. Mezuman, N. Sequeira, and A. Sherman, *Omitting types and AF algebras*, Arch. Math. Logic **53** (2014), 157–169.

[15] R. Camerlo and S. Gao, *The completeness of the isomorphism relation for countable Boolean algebras*, Trans. Amer. Math. Soc. **353** (2001), no. 2, 491–518.

[16] M.-D. Choi and E. Christensen, *Completely order isomorphic and close C*-algebras need not be *-isomorphic*, Bulletin of the London Mathematical Society **15** (1983), no. 6, 604–610.

[17] Y. Choi, I. Farah, and N. Ozawa, *A nonseparable amenable operator algebra which is not isomorphic to a C^*-algebra*, Forum Math. Sigma **2** (2014), 12 pages.

[18] E. Christensen, A.M. Sinclair, R.R. Smith, S.A. White, and W. Winter, *Perturbations of nuclear C^*-algebras*, Acta Math. **208** (2012), no. 1, 93–150.

[19] A. Connes, *Classification of injective factors. Cases II_1, II_∞, III_λ, $\lambda \neq 1$*, Ann. of Math. (2) **104** (1976), 73–115.

[20] S. Coskey and I. Farah, *Automorphisms of corona algebras and group cohomology*, Trans. Amer. Math. Soc. (to appear).

[21] H.G. Dales and W.H. Woodin, *An introduction to independence for analysts*, London Mathematical Society Lecture Note Series, vol. 115, Cambridge University Press, 1987.

[22] K. Dykema, *Interpolated free group factors*, Pacific J. Math **163** (1994), no. 1, 123–135.

[23] E. G. Effros and J. Rosenberg, *C*-algebras with approximately inner flip*, Pacific J. Math **77** (1978), no. 2, 417–443.

[24] G.A. Elliott, I. Farah, V. Paulsen, C. Rosendal, A.S. Toms, and A. Törnquist, *The isomorphism relation of separable C*-algebras*, Math. Res. Letters (to appear), preprint, arXiv:1301.7108.

[25] G.A. Elliott and A.S. Toms, *Regularity properties in the classification program for separable amenable C^*-algebras*, Bull. Amer. Math. Soc. **45** (2008), no. 2, 229–245.

[26] J. Fang, L. Ge, and W. Li, *Central sequence algebras of von Neumann algebras*, Taiwanese Journal of Mathematics **10** (2006), no. 1, pp–187.

[27] I. Farah, *Graphs and CCR algebras*, Indiana Univ. Math. Journal **59** (2010), 1041–1056.

[28] ———, *All automorphisms of the Calkin algebra are inner*, Annals of Mathematics **173** (2011), 619–661.

[29] ———, *Absoluteness, truth, and quotients*, Infinity and Truth (C.T. Chong et al., eds.), Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, vol. 25, World Scientific, 2013, pp. 1–24.

[30] I. Farah, *Selected applications of logic to classification problem of C*-algebras*, IMS 2012 Singapore Graduate Summer School Lecture Notes (C.T. Chong et al., eds.), to appear.

[31] I. Farah, I. Goldbring, B. Hart, and D. Sherman, *Existentially closed II_1 factors*, arXiv preprint arXiv:1310.5138, 2013.

[32] I. Farah and B. Hart, *Countable saturation of corona algebras*, C.R. Math. Rep. Acad. Sci. Canada **35** (2013), 35–56.

[33] I. Farah, B. Hart, L. Robert, and A. Tikuisis, *work in progress*, (2014).

[34] I. Farah, B. Hart, L. Robert, A. Tikuisis, A. Toms, and W. Winter, *Logic of metric structures and nuclear C*-algebras*, Oberwolfach Reports **Report No. 43/2013**, to appear.

[35] I. Farah, B. Hart, and D. Sherman, *Model theory of operator algebras I: Stability*, Bull. London Math. Soc. **45** (2013), 825–838.

[36] ———, *Model theory of operator algebras II: Model theory*, Israel J. Math. (to appear), arXiv:1004.0741.

[37] ———, *Model theory of operator algebras III: Elementary equivalence and II_1 factors*, Bull. London Math. Soc. (to appear).

[38] I. Farah and T. Katsura, *Nonseparable UHF algebras I: Dixmier's problem*, Adv. Math. **225** (2010), no. 3, 1399–1430.

[39] I. Farah and M. Magidor, *Omitting types in continuous logic is hard*, in preparation, 2014.

[40] I. Farah, N.C. Phillips, and J. Steprāns, *The commutant of $L(H)$ in its ultrapower may or may not be trivial*, Math. Annalen **347** (2010), 839–857.

[41] I. Farah and S. Shelah, *A dichotomy for the number of ultrapowers*, Journal of Mathematical Logic **10** (2010), 45–81.

[42] ———, *Rigidity of continuous quotients*, (2014), arXiv:1401.6689.

[43] I. Farah, A.S. Toms, and A. Törnquist, *The descriptive set theory of C*-algebra invariants*, Int. Math. Res. Notices **22** (2013), 5196–5226, Appendix with Caleb Eckhardt.

[44] ———, *Turbulence, orbit equivalence, and the classification of nuclear C*-algebras*, J. Reine Angew. Math. **688** (2014), 101–146.

[45] I. Farah and E. Wofsey, *Set theory and operator algebras*, Appalachian set theory 2006-2010 (J. Cummings and E. Schimmerling, eds.), Cambridge University Press, 2013, pp. 63–120.

[46] V. Ferenczi, A. Louveau, and C. Rosendal, *The complexity of classifying separable Banach spaces up to isomorphism*, J. Lond. Math. Soc. (2) **79** (2009), no. 2, 323–345.

[47] S. Gao, *Invariant descriptive set theory*, Pure and Applied Mathematics (Boca Raton), vol. 293, CRC Press, Boca Raton, FL, 2009.

[48] S. Gao and A.S. Kechris, *On the classification of Polish metric spaces up to isometry*, Mem. Amer. Math. Soc. **161** (2003), no. 766, viii+78.

[49] L. Ge and D. Hadwin, *Ultraproducts of C^*-algebras*, Recent advances in operator theory and related topics (Szeged, 1999), Oper. Theory Adv. Appl., vol. 127, Birkhäuser, Basel, 2001, pp. 305–326.

[50] S. Ghasemi, *SAW* algebras are essentially non-factorizable*, Glasgow Math. Journal (2012), preprint, arXiv:1209.3459.

[51] I. Goldbring, B. Hart, and T. Sinclair, *The theory of tracial von Neumann algebras does not have a model companion*, J. Symbolic Logic **78** (2013), no. 3, 1000–1004.

[52] L.A. Harrington, A.S. Kechris, and A. Louveau, *A Glimm–Effros dichotomy for Borel equivalence relations*, Journal of the American Mathematical Society **4** (1990), 903–927.

[53] C.W. Henson and J. Iovino, *Ultraproducts in analysis*, Analysis and Logic (Catherine Finet and Christian Michaux, eds.), London Mathematical Society Lecture Notes Series, no. 262.

[54] G. Hjorth, *Classification and orbit equivalence relations*, Mathematical Surveys and Monographs, vol. 75, American Mathematical Society, 2000.

[55] V.F.R. Jones, *Von Neumann algebras*, 2010, lecture notes, http://math.berkeley.edu/∼vfr/.

[56] K. Jung, *Amenability, tubularity, and embeddings into R^ω*, Mathematische Annalen **338** (2007), no. 1, 241–248.

[57] A.S. Kechris, *Classical descriptive set theory*, Graduate texts in mathematics, vol. 156, Springer, 1995.

[58] A.S. Kechris and A. Louveau, *The structure of hypersmooth Borel equivalence relations*, Journal of the American Mathematical Society **10** (1997), 215–242.

[59] D. Kerr, H. Li, and M. Pichot, *Turbulence, representations, and trace-preserving actions*, Proc. Lond. Math. Soc. (3) **100** (2010), no. 2, 459–484.

[60] D. Kerr, M. Lupini, and N.C. Phillips, *Borel complexity and automorphisms of C*-algebras*, preprint, 2014.

[61] E. Kirchberg, *Central sequences in C^*-algebras and strongly purely infinite algebras*, Operator Algebras: The Abel Symposium 2004, Abel Symp., vol. 1, Springer, Berlin, 2006, pp. 175–231.

[62] E. Kirchberg and M. Rørdam, *Central sequence C*-algebras and tensorial absorption of the Jiang–Su algebra*, Journal für die reine und angewandte Mathematik (Crelle's Journal) (2012).

[63] A. Kishimoto, N. Ozawa, and S. Sakai, *Homogeneity of the pure state space of a separable C^*-algebra*, Canad. Math. Bull. **46** (2003), no. 3, 365–372.

[64] M. Lupini, *Unitary equivalence of automorphisms of separable C*-algebras*, arXiv preprint arXiv:1304.3502 (2013).

[65] L. W. Marcoux and A. I. Popov, *Abelian, amenable operator algebras are similar to C*-algebras*, arXiv:1311.2982 (2013).

[66] A. Marcus, D.A. Spielman, and N. Srivastava, *Interlacing families II: Mixed characteristic polynomials and the Kadison–Singer problem*, arXiv:1306.3969 (2013).

[67] D. Marker, *Model theory*, Graduate Texts in Mathematics, vol. 217, Springer-Verlag, New York, 2002.

[68] H. Matui and Y. Sato, *Strict comparison and $\mathcal{Z}$-absorption of nuclear C*-algebras*, Acta mathematica **209** (2012), no. 1, 179–196.

[69] ———, *Decomposition rank of UHF-absorbing C*-algebras*, arXiv:1303.4371, 2013.

[70] D. McDuff, *Central sequences and the hyperfinite factor*, Proc. London Math. Soc. **21** (1970), 443–461.

[71] J. Melleray, *Computing the complexity of the relation of isometry between separable Banach spaces*, Math. Log. Q. **53** (2007), no. 2, 128–131.

[72] N. Ozawa, *Dixmier approximation and symmetric amenability for C*-algebras*, J. Math. Sci. Univ. Tokyo **20** (2013), 349–374.

[73] N.C. Phillips and N. Weaver, *The Calkin algebra has outer automorphisms*, Duke Math. Journal **139** (2007), 185–202.

[74] G. Pisier, *Similarity problems and completely bounded maps*, Lecture Notes in Mathematics, vol. 1618, Springer, 2001.

[75] S. Popa, *Strong rigidity of II_1 factors arising from malleable actions of w-rigid groups. I*, Invent. Math. **165** (2006), no. 2, 369–408.

[76] L. Robert, *Nuclear dimension and sums of commutators*, arXiv:1309.0498 (2013).

[77] M. Rørdam, *Classification of nuclear C^*-algebras*, Encyclopaedia of Math. Sciences, vol. 126, Springer-Verlag, Berlin, 2002.

[78] ———, *A simple C^*-algebra with a finite and an infinite projection*, Acta Math. **191** (2003), 109–142.

[79] M. Sabok, *Completeness of the isomorphism problem for separable C*-algebras*, arXiv preprint arXiv:1306.1049 (2013).

[80] R. Sasyk and A. Törnquist, *Borel reducibility and classification of von Neumann algebras*, Bulletin of Symbolic Logic **15** (2009), no. 2, 169–183.

[81] ———, *Turbulence and Araki–Woods factors*, Journal of Functional Analysis **259** (2010), no. 9, 2238–2252.

[82] Y. Sato, S. White, and W. Winter, *Nuclear dimension and $\mathcal{Z}$-stability*, arXiv preprint arXiv:1403.0747 (2014).

[83] P. Scowcroft, *Some model-theoretic correspondences between dimension groups and AF algebras*, Ann. Pure Appl. Logic **162** (2011), no. 9, 755–785.

[84] S. Shelah, *Proper forcing*, Lecture Notes in Mathematics 940, Springer, 1982.

[85] ———, *Classification theory and the number of nonisomorphic models*, second ed., Studies in Logic and the Foundations of Mathematics, vol. 92, North-Holland Publishing Co., Amsterdam, 1990.

[86] ———, *Vive la différence I: Nonisomorphism of ultrapowers of countable models*, Set theory of the continuum, Springer, 1992, pp. 357–405.

[87] ———, *Vive la différence III*, Israel J. of Math. **166** (2008), no. 1, 61–96.

[88] D. Sherman, *Notes on automorphisms of ultrapowers of II_1 factors*, Studia Math. **195** (2009), 201–217.

[89] S. Thomas, *On the complexity of the classification problem for torsion-free abelian groups of finite rank*, Bull. Symbolic Logic **7** (2001), no. 3, 329–344.

[90] A.S. Toms, *On the classification problem for nuclear C^*-algebras*, Ann. of Math. (2) **167** (2008), no. 3, 1029–1044.

[91] ———, *Comparison theory and smooth minimal C^*-dynamics*, Comm. Math. Phys. **289** (2009), no. 2, 401–433.

[92] A.S. Toms and W. Winter, *Strongly self-absorbing C^*-algebras*, Trans. Amer. Math. Soc. **359** (2007), no. 8, 3999–4029.

[93] A. Vignati, *An algebra whose subalgebras are characterized by density*, arXiv:1402.1112, 2014.

[94] Dan-Virgil Voiculescu, *Countable degree-1 saturation of certain C*-algebras which are coronas of Banach algebras*, arXiv:1310.4862, 2013.

[95] N. Weaver, *A prime C^*-algebra that is not primitive*, J. Funct. Anal. **203** (2003).

[96] ———, *Set theory and C^*-algebras*, Bull. Symb. Logic **13** (2007), 1–20.

[97] W. Winter, *Decomposition rank and $\mathcal{Z}$-stability*, Invent. Math. **179** (2010), 229–301.

[98] ———, *Nuclear dimension and $\mathcal{Z}$-stability of pure C*-algebras*, Invent. Math. **187** (2012), 259–342.

[99] W. Winter and J. Zacharias, *The nuclear dimension of C*-algebras*, Advances in Mathematics **224** (2010), no. 2, 461–498.

[100] W.H. Woodin, *Beyond Σ^2_1 absoluteness*, Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002) (Beijing), Higher Ed. Press, 2002, pp. 515–524.

[101] B. Zamora-Aviles, *Gaps in the poset of projections in the Calkin algebra*, Israel J. Math (to appear).

Department of Mathematics and Statistics, N520 Ross, 4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada

E-mail: ifarah@yorku.ca

Amalgamation functors and homology groups in model theory

John Goodrick, Byunghan Kim, and Alexei Kolesnikov

Abstract. We introduce the concept of an *amenable class of functors* and define homology groups for such classes. Amenable classes of functors arise naturally in model theory from considering types of independent systems of elements. Basic lemmas for computing these homology groups are established, and we discuss connections with type amalgamation properties.

Mathematics Subject Classification (2010). Primary 03C45; Secondary 55N35.

Keywords. Amalgamation functors, homology groups, model theory, Hurewicz correspondence, groupoids.

This paper abstracts the model theoretic results from [6] to a more general category-theoretic context. Namely, we introduce the concept of an *amenable class of functors*. It is a class of functors from a family of finite sets which is closed under subsets into a fixed image category satisfying a short list of axioms. We show that most of the general results proved in [6] hold in the broader *amenable* context. In addition we give some fundamental lemmas and examples which supplement the results of [6].

In Section 1, we introduce the notion of an amenable class of functors into a fixed category and we define the homology groups $H_n(\mathcal{A}, B)$ for an amenable class $\mathcal{A}$ and an object B in the image category. Model theory provides the best examples of amenable classes of functors, as described in [6].

In Section 2, given a rosy structure, we introduce the notion of the *type* homology groups, in contrast to the *set* homology groups defined in [6]. We show that the two homology groups are isomorphic.

Section 3 supplies some basic sufficient conditions for triviality of the homology groups and gives some additional examples of homology groups from model theory.

In Section 4 we outline some ongoing investigations related to our homology theory.

1. Simplicial homology in a category

In this section, we generalize the homology groups for rosy theories defined in [6] to a more general category-theoretic setting. Then we aim to provide a general framework for homology group computations. This section uses model theory only as a source of examples.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

1.1. Basic definitions and facts. Throughout this section, $\mathcal{C}$ denotes a category. If s is a set, then we consider the power set $\mathcal{P}(s)$ of s to be a category with a single inclusion map $\iota_{u,v} : u \to v$ between any pair of subsets u and v with $u \subseteq v$. A subset $X \subseteq \mathcal{P}(s)$ is called *downward-closed* if whenever $u \subseteq v \in X$, then $u \in X$. In this case we consider X to be a full subcategory of $\mathcal{P}(s)$. An example of a downward-closed collection that we will use often below is $\mathcal{P}^-(s) := \mathcal{P}(s) \setminus \{s\}$. We use ω for the set of natural numbers.

We are interested in a family of functors $f : X \to \mathcal{C}$ for downward-closed subsets $X \subseteq \mathcal{P}(s)$ for various finite subset sets s of the set of natural numbers. For $u \subseteq v \in X$, we shall write $f^u_v := f(\iota_{u,v}) \in \mathrm{Mor}_{\mathcal{C}}(f(u), f(v))$. Before specifying the desirable closure properties of the collection $\mathcal{A}$ of such functors, we need some auxiliary definitions.

Definition 1.1.

(1) Let X be a downward closed subset of $\mathcal{P}(s)$ and let $t \in X$. The symbol $X|_t$ denotes the set $\{u \in \mathcal{P}(s \setminus t) \mid t \cup u \in X\} \subseteq X$.

(2) For s, t, and X as above, let $f : X \to \mathcal{C}$ be a functor. Then *the localization of f at t* is the functor $f|_t : X|_t \to \mathcal{C}$ such that

$$f|_t(u) = f(t \cup u)$$

and whenever $u \subseteq v \in X|_t$,

$$(f|_t)^u_v = f^{u \cup t}_{v \cup t}.$$

(3) Let $X \subset \mathcal{P}(s)$ and $Y \subset \mathcal{P}(t)$ be downward closed subsets, where s and t are finite sets of natural numbers. Let $f : X \to \mathcal{C}$ and $g : Y \to \mathcal{C}$ be functors.

We say that f and g are *isomorphic* if there is an order-preserving bijection $\sigma : s \to t$ with $Y = \{\sigma(u) : u \in X\}$ and a family of isomorphisms $\langle h_u : f(u) \to g(\sigma(u)) : u \in X\rangle$ in $\mathcal{C}$ such that for any $u \subseteq v \in X$, the following diagram commutes:

$$\begin{array}{ccc} f(u) & \xrightarrow{h_u} & g(\sigma(u)) \\ \downarrow{\scriptstyle f^u_v} & & \downarrow{\scriptstyle g^{\sigma(u)}_{\sigma(v)}} \\ f(v) & \xrightarrow{h_v} & g(\sigma(v)) \end{array}$$

In the definition if σ is an arbitrary bijection, then f and g are said to be *weakly isomorphic*.

Remark 1.2. If X is a downward closed subset of $\mathcal{P}(s)$ and $t \in X$, then $X|_t$ is a downward closed subset of $\mathcal{P}(s \setminus t)$. Moreover $X|_t$ does not depend on the choice of s.

Definition 1.3. Let $\mathcal{A}$ be a non-empty set of functors $f : X \to \mathcal{C}$ such that $X \subseteq \mathcal{P}(\omega)$ is finite and downward closed and $\mathcal{C}$ is a fixed category (called the *image category of* $\mathcal{A}$). We say that $\mathcal{A}$ is *amenable* if it satisfies all of the following properties:

(1) (Invariance under weak isomorphisms) If $f : X \to \mathcal{C}$ is in $\mathcal{A}$ and $g : Y \to \mathcal{C}$ is weakly isomorphic to f, then $g \in \mathcal{A}$.

(2) (Closure under restrictions and unions) If $X \subseteq \mathcal{P}(s)$ is downward-closed and $f : X \to \mathcal{C}$ is a functor, then $f \in \mathcal{A}$ if and only if for every $u \in X$, we have that $f \restriction \mathcal{P}(u) \in \mathcal{A}$.

(3) (Closure under localizations) Suppose that $f : X \to \mathcal{C}$ is in $\mathcal{A}$ for some $X \subseteq \mathcal{P}(s)$ and $t \in X$. Then $f|_t : X|_t \to \mathcal{C}$ is also in $\mathcal{A}$.

(4) (Extensions of localizations are localizations of extensions) Suppose that $f : X \to \mathcal{C}$ is in $\mathcal{A}$ and $t \in X \subseteq \mathcal{P}(s)$ is such that $X|_t = X \cap \mathcal{P}(s \setminus t)$. Suppose that the localization $f|_t : X \cap \mathcal{P}(s \setminus t) \to \mathcal{C}$ has an extension $g : Z \to \mathcal{C}$ in $\mathcal{A}$ for some $Z \subseteq \mathcal{P}(s \setminus t)$. Then there is a map $g_0 : Z_0 \to \mathcal{C}$ in $\mathcal{A}$ such that $Z_0 = \{u \cup v : u \in Z, v \subseteq t\}$, $f \subseteq g_0$, and $g_0|_t = g$.

Remark 1.4. Model theory supplies the best examples of the amenable collection of functors. For example, as in [6] we could take $\mathcal{C}$ to be all boundedly (or algebraically) closed subsets of the monster model of a first-order theory, and let $\mathcal{A}$ be all functors which are "independence-preserving" (in Hrushovski's terminology [11]) and such that every face $f(u)$ is the bounded (or algebraic) closure of its vertices; then $\mathcal{A}$ is amenable. We may further restrict $\mathcal{A}$ by requiring, for instance, that all the "vertices" $f(\{i\})$ of functors $f \in \mathcal{A}$ realize the same type, or by placing further restrictions on "edges" $f(\{i, j\})$ and other higher-dimensional faces. We can also take $\mathcal{C}$ to be a category of *types* of the closed subsets of the model. These examples will be explained more precisely in Section 2.

Definition 1.5. Suppose that $f : X \to \mathcal{C}$ is a functor from a downward-closed collection X of sets and $B \in \mathrm{Ob}(\mathcal{C})$. If $f(\emptyset) = B$ then we say that f *is over* B. Let $\mathcal{A}_B$ denote the set of all functors $f \in \mathcal{A}$ that are over B.

Remark 1.6. It is easy to see that condition (2) in Definition 1.3 is equivalent to the conjunction of the following two conditions:

(1) (Closure under restrictions) If $f : X \to \mathcal{C}$ is in $\mathcal{A}$ and $Y \subseteq X$ with Y downward-closed, then $f \restriction Y$ is also in $\mathcal{A}$.

(2) (Closure under unions) Suppose that $f : X \to \mathcal{C}$ and $g : Y \to \mathcal{C}$ are both in $\mathcal{A}$ and that $f \restriction X \cap Y = g \restriction X \cap Y$. Then the union $f \cup g : X \cup Y \to \mathcal{C}$ is also in $\mathcal{A}$.

For instance, if these two conditions are true and $f : X \to \mathcal{C}$ is a functor from a downward-closed set X such that $f \restriction \mathcal{P}(u) \in \mathcal{A}$ for every $u \in X$, then if $u_1, \ldots, u_n$ are maximal sets in X, we can use closure under unions $(n-1)$ times to see that $f \in \mathcal{A}$ (since it is the union of the functors $f \restriction \mathcal{P}(u_i)$).

For the remainder of this section, we fix a category $\mathcal{C}$ and a non-empty amenable collection $\mathcal{A}$ of functors mapping into $\mathcal{C}$. As we mentioned in the above remark, every functor in $\mathcal{A}$ can be described as the union of functors whose domains are $\mathcal{P}(s)$ for various finite sets s. Such functors will play a central role in this paper.

Definition 1.7. Let $n \geq 0$ be a natural number. A (*regular*) n*-simplex in* $\mathcal{C}$ is a functor $f : \mathcal{P}(s) \to \mathcal{C}$ for some set $s \subseteq \omega$ with $|s| = n + 1$. The set s is called the *support of* f, or $\mathrm{supp}(f)$.

Let $S_n(\mathcal{A}; B)$ denote the collection of all regular n-simplices in $\mathcal{A}_B$. Then let $S(\mathcal{A}; B) := \bigcup_n S_n(\mathcal{A}; B)$ and $S(\mathcal{A}) := \bigcup_{B \in \mathrm{Ob}(\mathcal{C})} S(\mathcal{A}; B)$.

Let $C_n(\mathcal{A}; B)$ denote the free abelian group generated by $S_n(\mathcal{A}; B)$; its elements are called n*-chains in* $\mathcal{A}_B$, or n*-chains over* B. Similarly, we define $C(\mathcal{A}; B) := \bigcup_n C_n(\mathcal{A}; B)$ and $C(\mathcal{A}) := \bigcup_{B \in \mathrm{Ob}(\mathcal{C})} C(\mathcal{A}; B)$. The *support of a chain* c is the union of the supports of all the simplices that appear in c with a non-zero coefficient.

The adjective "regular" in the definition above is to emphasize that none of our simplices are "degenerate:" their domains must be *strictly* linearly ordered. It is more usual to allow for degenerate simplices, but for our purposes, this extra generality does not seem to be useful. Since all of our simplices will be regular, we will omit the word "regular" in what follows.

Now the rest of the development of the homology theory in this section will be exactly the same as the particular case of model theory described in the first section of [6]. The proofs are exactly the same and the reader will notice that the list of axioms for amenable family of functors singles out basic technical properties which enable the arguments in section 1 of [6] work. For the sake of completeness, we list here the all the definitions, lemmas, and theorems we will need for later sections but without giving proofs.

We begin with the notion of boundary operators used to define homology groups in our context.

Definition 1.8. If $n \geq 1$ and $0 \leq i \leq n$, then the *ith boundary operator* $\partial_n^i : C_n(\mathcal{A}; B) \to C_{n-1}(\mathcal{A}; B)$ is defined so that if f is an n-simplex with support $s = \{s_0 < \cdots < s_n\}$, then

$$\partial_n^i(f) = f \restriction \mathcal{P}(s \setminus \{s_i\})$$

and extended linearly to a group map on all of $C_n(\mathcal{A}; B)$.

If $n \geq 1$ and $0 \leq i \leq n$, then the *boundary map* $\partial_n : C_n(\mathcal{A}; B) \to C_{n-1}(\mathcal{A}; B)$ is defined by the rule

$$\partial_n(c) = \sum_{0 \leq i \leq n} (-1)^i \partial_n^i(c).$$

We write ∂^i and ∂ for ∂_n^i and ∂_n, respectively, if n is clear from context.

Definition 1.9. The kernel of ∂_n is denoted $Z_n(\mathcal{A}; B)$, and its elements are called $(n$-*)cycles*. The image of ∂_{n+1} in $C_n(\mathcal{A}; B)$ is denoted $B_n(\mathcal{A}; B)$. The elements of $B_n(\mathcal{A}; B)$ are called $(n$-*)boundaries*.

It can be shown (by the usual combinatorial argument) that $B_n(\mathcal{A}; B) \subseteq Z_n(\mathcal{A}; B)$, or more briefly, "$\partial_n \circ \partial_{n+1} = 0$." Therefore we can define simplicial homology groups relative to $\mathcal{A}$:

Definition 1.10. The *nth* (*simplicial*) *homology group of* $\mathcal{A}$ *over* B is

$$H_n(\mathcal{A}; B) = Z_n(\mathcal{A}; B)/B_n(\mathcal{A}; B).$$

There are two natural candidates for the definition of the boundary of a 0-simplex. One possibility is to define $\partial_0(f) = 0$ for all $f \in S_0(\mathcal{A}; B)$. Another possibility is to extend the definition of an n-simplex to $n = -1$; namely a (-1)-simplex f is an object $f(\emptyset)$ in $\mathcal{C}$. Then the definition of a boundary operator extends naturally to the operator $\partial_0 : f \in S_0(\mathcal{A}; B) \mapsto B$.

As we show in Lemma 3.1, computing the group H_0 in a specific context using the first definition gives $H_0 \cong \mathbb{Z}$ while using the second definition we get $H_0 = 0$. Thus, the difference between the approaches is parallel to that between the homology and reduced homology groups in algebraic topology [1].

Next we define the amalgamation properties. We use the convention that n denotes the set $\{0, 1, \ldots, n-1\}$.

Definition 1.11. Let $\mathcal{A}$ be an amenable family of functors into a category $\mathcal{C}$ and let $n \geq 1$.

(1) $\mathcal{A}$ has *n-amalgamation* if for any functor $f : \mathcal{P}^-(n) \to \mathcal{C}$, $f \in \mathcal{A}$, there is an $(n-1)$-simplex $g \supseteq f$ such that $g \in \mathcal{A}$.

(2) $\mathcal{A}$ has *n-complete amalgamation* or *n-CA* if $\mathcal{A}$ has k-amalgamation for every k with $1 \leq k \leq n$.

(3) $\mathcal{A}$ has *strong 2-amalgamation* if whenever $f : \mathcal{P}(s) \to \mathcal{C}$, $g : \mathcal{P}(t) \to \mathcal{C}$ are simplices in $\mathcal{A}$ and $f \restriction \mathcal{P}(s \cap t) = g \restriction \mathcal{P}(s \cap t)$, then $f \cup g$ can be extended to a simplex $h : \mathcal{P}(s \cup t) \to \mathcal{C}$ in $\mathcal{A}$.

(4) $\mathcal{A}$ has *n-uniqueness* if for any functor $f : \mathcal{P}^-(n) \to \mathcal{C}$ in $\mathcal{A}$ and any two $(n-1)$-simplices g_1 and g_2 in $\mathcal{A}$ extending f, there is a natural isomorphism $F : g_1 \to g_2$ such that $F \restriction \mathrm{dom}(f)$ is the identity.

Remark 1.12.

(1) There is a mismatch that n-amalgamation refers to the existence of $(n-1)$-simplex extending its boundary. But this numbering is coherent with historical developments of amalgamation theory in model theory and homology theory in algebraic topology.

(2) The definition of n-amalgamation can be naturally extended to $n = 0$: $\mathcal{A}$ has 0-amalgamation if it contains a functor $f : \{\emptyset\} \to \mathcal{C}$. This holds in any amenable family of functors.

Definition 1.13. We say that an amenable family of functors $\mathcal{A}$ is *non-trivial* if $\mathcal{A}$ has 1-amalgamation, and satisfies the strong 2-amalgamation property.

The following remark is immediate from the definitions.

Remark 1.14. Any non-trivial amenable collection of functors $\mathcal{A}$ contains an n-simplex for each $n \geq 1$.

Everywhere below, we only deal with non-trivial amenable families of functors.

1.2. Computing homology groups. As in [6] we introduce special kinds of n-chains which are useful for computing homology groups.

Definition 1.15. If $n \geq 1$, an *n-shell* is an n-chain c of the form

$$\pm \sum_{0 \leq i \leq n+1} (-1)^i f_i,$$

where $f_0, \ldots, f_{n+1}$ are n-simplices such that whenever $0 \leq i < j \leq n+1$, we have $\partial^i f_j = \partial^{j-1} f_i$.

Definition 1.16. If $n \geq 1$, and *n-fan* is an n-chain of the form

$$\pm \sum_{i \in \{0,..,\widehat{k},\ldots,n+1\}} (-1)^i f_i$$

for some $k \leq n+1$, where the f_i are n-simplices such that whenever $0 \leq i < j \leq n$, we have $\partial^i f_j = \partial^{j-1} f_i$. In other words, an n-fan is an n-shell missing one term.

If c is an n-fan, then ∂c is an $(n-1)$-shell; and $\mathcal{A}$ has n-amalgamation if and only if every $(n-2)$-shell in $\mathcal{A}$ is the boundary of an $(n-1)$-simplex in $\mathcal{A}$. And $\mathcal{A}$ has n-uniqueness if and only if every $(n-2)$-shell in $\mathcal{A}$ is the boundary of at most one $(n-1)$-simplex in $\mathcal{A}$ up to isomorphism.

As mentioned earlier, we now state without proofs a series of lemmas and theorems analogous to those in [6], Section 1. In particular, we state two "prism lemmas" (1.25 and 1.27) and a result that every element of a homology group is the equivalence class of a shell (Theorem 1.28).

Lemma 1.17. *If $n \geq 2$ and $\mathcal{A}$ has n-CA, then every $(n-1)$-cycle is a sum of $(n-1)$-shells. Namely, for each $c \in Z_{n-1}(\mathcal{A}; B)$, $c = \sum_i \alpha_i f_i$, there is a finite collection of $(n-1)$-shells $c_i \in Z_{n-1}(\mathcal{A}; B)$ such that $c = \sum_i (-1)^n \alpha_i c_i$.*

Moreover, if S is the support of the chain c and m is any element not in S, then we can choose $\sum_i \alpha_i c_i$ so that its support is $S \cup \{m\}$.

Corollary 1.18. *Assume $\mathcal{A}$ has n-CA for some $n \geq 2$. Then $H_{n-1}(\mathcal{A}; B)$ is generated by*

$$\{[c] \,:\, c \text{ is an } (n-1)\text{-shell over } B\}.$$

In particular, if any $(n-1)$-shell over B is a boundary, then so is any $(n-1)$-cycle.

Corollary 1.19. *If $\mathcal{A}$ has n-CA for some $n \geq 3$, then $H_{n-2}(\mathcal{A}; B) = 0$.*

Corollary 1.18 will be strengthened to Theorem 1.28.

Definition 1.20. If $n \geq 1$, an *n-pocket* is an n-cycle of the form $f - g$, where f and g are n-simplices with support S (where S is an $(n+1)$-element set).

Lemma 1.21. *Suppose that $f, g \in S_n(\mathcal{A})$ are* isomorphic *functors such that $\partial_n f = \partial_n g$. Then the n-pocket $f - g$ is a boundary.*

Lemma 1.22. *Suppose that $n \geq 1$ and $\mathcal{A}$ has $(n+1)$-amalgamation. Then for any n-fan*

$$g = \pm \sum_{i \in \{0,\ldots,\widehat{k},\ldots,n+1\}} (-1)^i f_i$$

there is some n-simplex f_k and some $(n+1)$-simplex f such that $g + (-1)^k f_k = \partial f$.

The next lemma says that n-pockets are equal to n-shells, "up to a boundary."

Lemma 1.23. *Assume that $\mathcal{A}$ has the $(n+1)$-amalgamation property for some $n \geq 1$. For any $B \in \mathcal{C}$, any n-shell in $\mathcal{A}_B$ with support $n+2$ is equivalent, up to a boundary in $B_n(\mathcal{A}; B)$, to an n-pocket in $\mathcal{A}_B$ with support $n+1$. Conversely, any n-pocket with support $n+1$ is equivalent, up to a boundary, to an n-shell with support $n+2$.*

From Corollary 1.18 and Lemma 1.23 we derive the following:

Corollary 1.24. *If $\mathcal{A}$ has 3-amalgamation, then $H_2(\mathcal{A}; B)$ is generated by equivalence classes of 2-pockets.*

Lemma 1.25 (Prism lemma). *Let $n \geq 1$. Suppose that $\mathcal{A}$ has $(n+1)$-amalgamation. Let $f - f'$ be an n-pocket in $\mathcal{A}_B$ with support s, where $|s| = n+1$. Let t be an $(n+1)$-element set disjoint from s. Then given n-simplex g in $\mathcal{A}_B$ with the domain $\mathcal{P}(t)$, there is an n-simplex g' such that $g - g'$ forms an n-pocket in $\mathcal{A}_B$ and is equivalent, modulo $B_n(\mathcal{A}; B)$, to the pocket $f - f'$. We may choose g' first and then find g to have the same conclusion.*

Corollary 1.26. *Let $n \geq 1$. Suppose $\mathcal{A}$ has $(n+1)$-CA. The group $H_n(\mathcal{A}; B)$ is generated by equivalence classes n-shells with support $n+2$.*

We have a shell version of the prism lemma as well:

Lemma 1.27 (Prism lemma, shell version). *Let $\mathcal{A}$ satisfy $(n+1)$-amalgamation for some $n \geq 1$. Suppose that an n-shell $f := \sum_{0 \leq i \leq n+1} (-1)^i f_i$ and an n-fan*

$$g^- := \sum_{i \in \{0,\ldots,\hat{k},\ldots,n+1\}} (-1)^i g_i$$

are given, where f_i, g_i are n-simplices over B, $\text{supp}(f) = s$ with $|s| = n+2$, and $\text{supp}(g^-) = t = \{t_0, ..., t_{n+1}\}$, where $t_0 < ... < t_{n+1}$ and $s \cap t = \emptyset$. Then there is an n-simplex g_k over B with support $\partial_k t := t \setminus \{t_k\}$ such that $g := g^- + (-1)^k g_k$ is an n-shell over B and $f - g \in B_n(\mathcal{A}; B)$.

The next theorem gives an even simpler standard form for elements of $H_n(\mathcal{A}; B)$.

Theorem 1.28. *If $\mathcal{A}$ has $(n+1)$-CA for some $n \geq 1$, then*

$$H_n(\mathcal{A}; B) = \{[c] : c \text{ is an } n\text{-shell over } B \text{ with support } n+2\}.$$

Now using Theorem 1.28 and Lemma 1.23, we obtain the following:

Corollary 1.29. *If $(n+1)$-CA (for some $n \geq 1$) holds in $\mathcal{A}$, then*

$$H_n(\mathcal{A}; B) = \{[c] : c \text{ is an } n\text{-pocket in } \mathcal{A} \text{ over } B \text{ with support } n+1\}.$$

2. Type versus set homology groups in model theory

In this section, we define some amenable classes of functors that arise in model theory. Namely given either a complete rosy theory T or a complete type $p \in S(A)$ in a rosy theory, we will define both the "type homology groups" $H_n^t(T)$ (or $H_n^t(p)$) and the "set homology groups" $H_n^{set}(T)$ (or $H_n^{set}(p)$). As noted, $H_n^{set}(p)$ and the classes of p-set-functors were already introduced in [6] and the properties of those were the motivation for Definition 1.3. As we show below, these definitions will lead to isomorphic homology groups (Proposition 2.12).

We make the same assumptions on our underlying theory T as in [6]: **in what follows, we assume that $T = T^{\text{eq}}$ is a complete rosy theory** (e.g. stable, simple, or o-minimal) and we work in its fixed large saturated model $\mathfrak{C} = \mathfrak{C}^{\text{eq}}$. The reason for this is so that we have a nice independence notion [3]. Throughout, "$\downarrow$", "independence" or "non-forking" will mean thorn-independence. So if T is simple then we assume it has elimination of hyperimaginaries in order for non-forking independence to be equal to thorn-independence [3]. But the assumptions are for convenience not for full generality. For example if T is simple, then one may assume the independence is usual non-forking in $\mathfrak{C}^{\text{heq}}$ while replacing acl by bdd and so on. Moreover there are non-rosy examples having suitable independence notions that fit in our amenable category context (see [10] and [13]).

We refer the reader to [12, 20] and to [3, 18] for general background on simple and rosy theories, respectively.

2.1. Type homology. We will work with $*$-types – that is, types with possibly infinite sets of variables – and to avoid some technical issues, we will place an absolute bound on the cardinality of the variable sets of the types we consider. Fix some infinite cardinal $\kappa_0 \geq |T|$. We will assume that every $*$-type has no more than κ_0 free variables. We also fix a set $\mathcal{V}$ of variables such that $|\mathcal{V}| > \kappa_0$ and assume that all variables in $*$-types come from the set $\mathcal{V}$ (which is a "master set of variables.") We work in a monster model $\mathfrak{C} = \mathfrak{C}^{\text{eq}}$ which is saturated in some cardinality greater than $2^{|\mathcal{V}|}$. We let $\bar{\kappa} = |\mathfrak{C}|$. As we will see in the next section, the precise values of κ_0 and $|\mathcal{V}|$ will not affect the homology groups.

Given a set A, strictly speaking we should write "**a** complete $*$-type of A" instead of "**the** complete $*$-type of A" – there are different such types corresponding to different choices for associating each element of A with a variable from $\mathcal{V}$, and this distinction is crucial for our purposes.

If X is any subset of the variable set $\mathcal{V}$, $\sigma : X \to \mathcal{V}$ is any injective function, and $p(\overline{x})$ is any $*$-type such that $\overline{x}$ is contained in X, then we let

$$\sigma_* p = \{\varphi(\sigma(\overline{x})) : \varphi(\overline{x}) \in p\}.$$

Definition 2.1. If A is a small subset of the monster model, then $\mathcal{T}_A$ is the category such that

(1) The objects of $\mathcal{T}_A$ are are all the complete $*$-types in T over A, including (for convenience) a single distinguished type $p_\emptyset$ with no free variables;

(2) $\text{Mor}_{\mathcal{T}_A}(p(\overline{x}), q(\overline{y}))$ is the set of all injective maps $\sigma : \overline{x} \to \overline{y}$ such that $\sigma_*(p) \subseteq q$.

Note that this definition gives a notion of two types $p(\overline{x})$ and $q(\overline{y})$ being "isomorphic:" namely, that q can be obtained from p by relabeling variables.

Definition 2.2. If $A = \text{acl}(A)$ is a small subset of the monster model, a *closed independent type-functor based on* A is a functor $f : X \to \mathcal{T}_A$ such that:

(1) X is a downward-closed subset of $\mathcal{P}(s)$ for some finite $s \subseteq \omega$.

(2) Suppose $w \in X$ and $u, v \subseteq w$. Recall our notational convention $f^u_w := f(\iota_{u,w})$. Let us write $\overline{x}_w$ to be the variable set of $f(w)$. Then whenever $\overline{a}$ realizes the type $f(w)$ and $\overline{a}_u$, $\overline{a}_v$, and $\overline{a}_{u\cap v}$ denote subtuples corresponding to the variable sets $f^u_w(\overline{x}_u)$, $f^v_w(\overline{x}_v)$, and $f^{u\cap v}_w(\overline{x}_{u\cap v})$, then

$$\overline{a}_u \underset{A \cup \overline{a}_{u\cap v}}{\downarrow} \overline{a}_v.$$

(3) For all non-empty $u \in X$ and any $\overline{a}$ realizing $f(u)$, we have (using the notation above) $\overline{a} = \text{acl}\left(A \cup \bigcup_{i\in u} \overline{a}_{\{i\}}\right)$.

(The adjective "closed" in the definition refers to the fact that, by (3), all the types $f(u)$ are $*$-types of algebraically closed tuples.)

Let $\mathcal{A}^t(T; A)$ denote all closed independent type-functors based on A.

Remark 2.3. It follows from the definition above and the basic properties of nonforking that if f is a closed independent type-functor based on A and $u \in \text{dom}(f)$ is a non-empty set of size k, then any realization $\overline{a}$ of $f(u)$ is the algebraic closure of an AB-independent set $\{\overline{a}_1, \ldots, \overline{a}_k\}$, where B is the subtuple of $\overline{a}$ corresponding to the variables $f^\emptyset_u(\overline{x}_\emptyset)$ and each $\overline{a}_i$ is the subtuple corresponding to the variables in $f^{\{i\}}_u(\overline{x}_{\{i\}})$.

Definition 2.4. If $A = \mathrm{acl}(A)$ is a small subset of the monster model and $p \in S(A)$, then a *closed independent type-functor in* p is a closed independent type-functor $f : X \to \mathcal{T}_A$ based on A such that if $X \subseteq \mathcal{P}(s)$ and $i \in s$, then $f(\{i\})$ is the complete $*$-type of $\mathrm{acl}(AC \cup \{b\})$ over A, where C is some realization of $f(\emptyset)$ and b is some realization of a nonforking extension of p to AC.

Let $\mathcal{A}^t(p)$ denote all closed independent type-functors in p.

Now using the basic independence properties of rosy theories, it is not hard to verify amenability of the above families of functors. In particular one may consult the proof of [6, 1.19]

Proposition 2.5. *The sets $\mathcal{A}^t(T; A)$ and $\mathcal{A}^t(p)$ are non-trivial amenable families of functors.*

Definition 2.6. If A is a small algebraically closed subset of $\mathfrak{C}$, then we write $S_n\mathcal{T}_A$ as an abbreviation for $S_n(\mathcal{A}^t(T; A); p_\emptyset)$ (the collection of closed n-simplices in $\mathcal{A}^t(T; A)$ over the distinguished type $p_\emptyset$), $B_n\mathcal{T}_A$ and $Z_n\mathcal{T}_A$ for the boundary and cycle groups, and $H^t_n(T; A)$ for the homology group $H_n(\mathcal{A}^t(T; A); p_\emptyset)$.

Similarly if $p \in S(A)$, then we use the abbreviations $S_n\mathcal{T}(p)$ for $S_n(\mathcal{A}^t(p); p_\emptyset)$; and $B_n\mathcal{T}(p)$, $Z_n\mathcal{T}(p)$, and $H^t_n(p)$.

2.2. Set homology.

Definition 2.7. Let A be a small subset of $\mathfrak{C}$. By $\mathcal{C}_A$ we denote the category of all subsets *containing* A of $\mathfrak{C}$ of size no more that κ_0, where morphisms are partial elementary maps over A (that is, fixing A pointwise).

For a functor $f : X \to \mathcal{C}_A$ and $u \subseteq v \in X$, we write $f^u_v(u) := f^u_v(f(u)) \subseteq f(v)$.

Definition 2.8. A *closed independent set-functor based on* $A = \mathrm{acl}(A)$ is a functor $f : X \to \mathcal{C}_A$ such that:

(1) X is a downward-closed subset of $\mathcal{P}(s)$ for some finite $s \subseteq \omega$.

(2) For all non-empty $u \in X$, we have that $f(u) = \mathrm{acl}(A \cup \bigcup_{i\in u} f^{\{i\}}_u(\{i\}))$ and the set $\{f^{\{i\}}_u(\{i\}) : i \in u\}$ is independent over $f^\emptyset_u(\emptyset)$.

Let $\mathcal{A}^{set}(T; A)$ denote all closed independent set-functors based on A.

Now we recall the following in [6].

Definition 2.9. If $A = \mathrm{acl}(A)$ is a small subset of the monster model and $p \in S(A)$, then a *closed independent set-functor in* p is a closed independent set-functor $f : X \to \mathcal{C}_A$ based on A such that if $X \subseteq \mathcal{P}(s)$ and $i \in s$, then $f(\{i\})$ is a set of the form $\mathrm{acl}(C \cup \{b\})$ where $C = f^\emptyset_{\{i\}}(\emptyset) \supseteq A$ and b realizes some non-forking extension of p to C.

Let $\mathcal{A}^{set}(p)$ denote all closed independent set-functors in p.

Just as in the previous subsection, we have:

Proposition 2.10. *The sets $\mathcal{A}^{set}(T; A)$ and $\mathcal{A}^{set}(p)$ are non-trivial amenable families of functors.*

Definition 2.11. If A is a small subset of $\mathfrak{C}$, then we write $S_n\mathcal{C}_A$ to denote $S_n(\mathcal{A}^{set}(T;A);A)$ (the collection of closed n-simplices in $\mathcal{A}^{set}(T;A)$ *over* A), and similarly we write $B_n\mathcal{C}_A$ and $Z_n\mathcal{C}_A$ for the boundary and cycle groups over A, and use the notation $H_n^{set}(T;A)$ for the homology group $H_n(\mathcal{A}^{set}(T;A);A)$.

If $A = \mathrm{acl}(A)$ and $p \in S(A)$, then we use similar abbreviations

$$S_n\mathcal{C}(p) := S_n(\mathcal{A}^{set}(p);A), B_n\mathcal{C}(p), Z_n\mathcal{C}(p), \text{ and } H_n^{set}(p).$$

Proposition 2.12.

(1) *For any n and any set A, $H_n^t(T;A) \cong H_n^{set}(T;A)$.*

(2) *For any n and any complete type $p \in S(A)$, $H_n^t(p) \cong H_n^{set}(p)$.*

Proof. The idea is that we can build a correspondence $F : S\mathcal{C}_A \to S\mathcal{T}_A$ which maps each set-simplex f to its "complete $*$-type" $F(f)$. Note that this will involve some non-canonical choices: namely, which variables to use for $F(f)$, and in what order to enumerate the various sets in f (since our variable set $\mathcal{V}$ is indexed and thus implicitly ordered). We will write out a proof of part (1) of the proposition, and part (2) can be proved similarly by relativizing to p.

Let $S_{\leq n}\mathcal{C}_A$ and $S_{\leq n}\mathcal{T}_A$ denote, respectively, $\bigcup_{i\leq n} S_i\mathcal{C}_A$ and $\bigcup_{i\leq n} S_i\mathcal{T}_A$. We will build a sequence of maps $F_n : S_{\leq n}\mathcal{C}_A \to S_{\leq n}\mathcal{T}_A$ whose union will be F. Given such an F_n, let $\overline{F}_n : C_{\leq n}\mathcal{C}_A \to C_{\leq n}\mathcal{T}_A$ be its natural extension to the class of all set-k-chains over A for $k \leq n$.

Claim 2.13. *There are maps $F_n : S_{\leq n}\mathcal{C}_A \to S_{\leq n}\mathcal{T}_A$ such that:*

(1) *F_{n+1} is an extension of F_n;*

(2) *If $f \in S_{\leq n}\mathcal{C}_A$ and $\mathrm{dom}(f) = \mathcal{P}(s)$, then $\mathrm{dom}(F_n(f)) = \mathcal{P}(s)$ and $[F_n(f)](s)$ is a complete $*$-type of $f(s)$ over A;*

(3) *For any $k \leq n$, any $f \in S_k\mathcal{C}_A$, and any $i \leq k$, $F_n(\partial^i f) = \partial^i [F_n(f)]$; and*

(4) *F_n is surjective, and in fact for every $g \in S_k\mathcal{T}_A$ (where $0 \leq k \leq n$), there are **more** than $2^{|\mathcal{V}|}$ simplices $f \in S_k\mathcal{C}_A$ such that $F_n(f) = g$.*

Proof. We prove the claim by induction on n. The case where $n = 0$ is simple: only conditions (2) and (4) are relevant, and note that we can insure (4) because the monster model $\mathfrak{C}$ is $(2^{|\mathcal{V}|})^+$-saturated and there are at most $2^{|\mathcal{V}|}$ elements of $S_0\mathcal{T}_A$. So suppose that $n > 0$ and we have $F_0, \ldots, F_n$ satisfying all these properties, and we want to build F_{n+1}. We build F_{n+1} as a union of a chain of partial maps from $S_{\leq n+1}\mathcal{C}_A$ to $S_{\leq n+1}\mathcal{T}_A$ extending F_n (that is, functions whose domains are subsets of $S_{\leq n+1}\mathcal{C}_A$).

Subclaim 2.14. *Suppose that $F : X \to S_{\leq n+1}\mathcal{T}_A$ is a function on a set $X \subseteq S_{\leq n+1}\mathcal{C}_A$ of size at most $(2^{|\mathcal{V}|})^+$ and that F satisfies (1) through (3). Then for any simplex $g \in S_{n+1}\mathcal{T}_A$, there is an extension F_0 of F satisfying (1) through (3) such that $|\mathrm{dom}(F_0)| \leq (2^{|\mathcal{V}|})^+$ and:*

$$(*) \text{ There are } (2^{|\mathcal{V}|})^+ \text{ distinct } f \in S_{n+1}\mathcal{C}_A \text{ such that } F'(f) = g.$$

Proof. Let $\partial g = g_0 - g_1 + \ldots + (-1)^n g_n$ (where $g_i = \partial^i g$), and let $\mathcal{P}(s)$ be the domain of g. By induction, each g_i is the image under F_n of $(2^{|\mathcal{V}|})^+$ different n-simplices in $\mathcal{C}_A$; let $\langle f_i^j : j < (2^{|\mathcal{V}|})^+ \rangle$ be a sequence of distinct simplices such that for every $j < (2^{|\mathcal{V}|})^+$, $F_n(f_i^j) =$

g_i. By saturation of the monster model, for each $j < (2^{|\mathcal{V}|})^+$ we can pick an $(n+1)$-simplex $f_j \in \mathcal{C}_A$ with domain $\mathcal{P}(s)$ such that $\partial f_j = f_0^j - f_1^j + \ldots + (-1)^n f_n^j$ and $\mathrm{tp}(f_j(s)) = g(s)$. Then the f_j are all distinct, so we can let $F_0 = F \cup \{(f_j, g) : j < (2^{|\mathcal{V}|})^+\}$. □

Now by the subclaim, we can use transfinite induction to build a **partial** map $F' : S_{\leq n+1}\mathcal{C}_A \to S_{\leq n+1}\mathcal{T}_A$ satisfying (1) through (4) (also using the fact that there only (at most) $2^{|\mathcal{V}|}$ different simplices in $S_{n+1}\mathcal{T}_A$ and the fact that the union of a chain of partial maps from $S_{\leq n+1}\mathcal{C}_A$ to $S_{\leq n+1}\mathcal{T}_A$ satisfying conditions (2) and (3) will still satisfy these conditions).

Finally, we can extend F' to a function on all of $S_{\leq n+1}\mathcal{C}_A$ by a second transfinite induction, extending F' to each $f : \mathcal{P}(s) \to \mathcal{C}_A$ in $\mathcal{C}_A$ one at a time; to ensure that properties (2) and (3) hold, we just have to pick $F_{n+1}(f)$ to be some $(n+1)$-simplex with the same domain $\mathcal{P}(s)$ whose n-faces are as specified by F_n and such that $[F_{n+1}(f)](s)$ is a complete $*$-type of $f(s)$ over A. □

Now let $F = \bigcup_{n<\omega} F_n$. By property (3) above, it follows that for any chain $c \in C\mathcal{C}_A$, we have $\overline{F}(\partial c) = \partial\left[\overline{F}(c)\right]$. Hence $\overline{F}$ maps $Z_n\mathcal{C}_A$ into $Z_n\mathcal{T}_A$ and $B_n\mathcal{C}_A$ into $B_n\mathcal{T}_A$, and so $\overline{F}$ induces group homomorphisms $\varphi_n : H_n^{set}(T; A) \to H_n^t(T; A)$. Verifying that φ_n is injective amounts to checking that whenever $\overline{F}(c) \in B_n\mathcal{T}_A$, the set-chain c is in $B_n\mathcal{C}_A$, but this is staightforward: if, say, $\overline{F}(c) = \partial c'$, then we can pick a set-simplex $\hat{c}$ "realizing" c' such that $\partial\hat{c} = c$. Finally, condition (4) implies that φ_n is surjective, so $H_n^{set}(T; A) \cong H_n^t(T; A)$. □

Remark 2.15. Since Proposition 2.12 is true for any choices of κ_0, $\mathcal{V}$, and the monster model $\mathfrak{C}$ as long as $|T| \leq \kappa_0 < |\mathcal{V}|$ and $2^{|\mathcal{V}|} \leq |\mathfrak{C}|$, it follows that our homology groups (with the restriction of the set A) do not depend on the choices of κ_0, $|\mathcal{V}|$, or the monster model.

Without specifying a base set A, one could also define $C_n(T)$ to be the direct sum $\bigoplus_{i<\bar{\kappa}} C_n\mathcal{C}_{A_i}$ where $\{A_i | i < \bar{\kappa}\}$ is the collection of all small subsets of $\mathfrak{C}$, and similarly $Z_n(T)$, $B_n(T)$, and $H_n(T) := Z_n(T)/B_n(T)$. Then the boundary operator ∂ sends n-chains to $(n-1)$-chains componentwise. Hence it follows $H_n(T) = \bigoplus_{i<\bar{\kappa}} H_n(T; A_i)$. This means the homology groups defined without specifying a base set depends on the choice of monster model, and so this approach would not give invariants for the theory T.

2.3. An alternate definition of the set homology groups. In our definition of the set homology groups $H_n^{set}(T; A)$ and $H_n^{set}(p)$ (where $p \in S(A)$), we have been assuming that the base set A is fixed pointwise by all of the elementary maps in a set-simplex – this is built into our definition of $\mathcal{C}_A$. It will turn out that we get an equivalent definition of the homology groups if we allow the base set to be "moved" by the images of the inclusion maps in a set-simplex, as we will show in this subsection.

Definition 2.16.

(1) A *set-n-simplex weakly over A* is a set-n-simplex $f : \mathcal{P}(s) \to \mathcal{C}(= \mathcal{C}_\emptyset)$ such that $f(\emptyset) = A$.

(2) If $p \in S(A)$, then a set-n-simplex $f : \mathcal{P}(s) \to \mathcal{C}$ is *weakly of type p* if $f(\emptyset) = A$, and for every $i \in s$,

$$f(\{i\}) = \mathrm{acl}\left(f^\emptyset_{\{i\}}(A) \cup \{a_i\}\right)$$

for some a_i such that $\mathrm{tp}(a_i/f^{\emptyset}_{\{i\}}(A))$ is a conjugate of p.

Let $S'_n\mathcal{C}_A$ be the collection of all set-n-simplices weakly over A. Note that the boundary operator ∂ takes an n-simplex weakly over A to a chain of $(n-1)$-simplices weakly over A, and so we can define "weak set homology groups over A," which we denote $H'_n(T;A)$. Similarly, we can define $H'_n(p)$, the "weak set homology groups of p," from chains of set-simplices that are weakly of type p.

Proposition 2.17.

(1) *For any n and any $A \in \mathcal{C}$, $H'_n(T;A) \cong H^{set}_n(T;A)$.*

(2) *For any n and any complete type $p \in S(A)$, $H'_n(p) \cong H^{set}_n(p)$.*

Proof. As usual, the two parts have identical proofs, and we only prove the second part.

We will identify $S'_0\mathcal{C}_A$ as a big single *complex* as follows. Due to our cardinality assumption, for each $n < \omega$, there are $\bar{\kappa}$-many 0-simplices in $S'_0\mathcal{C}_A$ having the common domain $\mathcal{P}(\{n\})$. Then we consider the following domain set $\mathcal{D}_0 = \{\emptyset\} \cup \{\{(n,i)\} |\ n < \omega, i < \bar{\kappa}\}$. Now as said we identify $S'_0\mathcal{C}_A$ as a *single* functor F'_0 from $\mathcal{D}_0$ to $\mathcal{C}$ such that $F'_0(\emptyset) = A$, and $F'_0(\{(n,i)\}) = (f')^n_i(\{n\})$ where $(f')^n_i \in S'_0\mathcal{C}_A$ is the corresponding 0-simplex with $((f')^n_i)^{\emptyset}_{\{n\}} = (F'_0)^{\emptyset}_{\{(n,i)\}}$. Similarly we consider $S_0\mathcal{C}_A$ as a functor F_0 from $\mathcal{D}_0$ to $\mathcal{C}_A$ such that $F_0(\emptyset) = A$, and $F_0(\{(n,i)\}) = f^n_i(\{n\}) \equiv (f')^n_i(\{n\})$ where $f^n_i \in S_0\mathcal{C}_A$ is the corresponding 0-simplex over A with $(f^n_i)^{\emptyset}_{\{n\}} = (F_0)^{\emptyset}_{\{(n,i)\}}$. Now F'_0 and F_0 are naturally isomorphic by η^0 with $\eta^0_{\emptyset}$ =the identity map of A, and suitable $\eta^0_{\{(n,i)\}}$ sending $(f')^n_i(\{n\})$ to $f^n_i(\{n\})$.

Now for $S'_1\mathcal{C}_A$, note that for each pair $(f')^{n_0}_{i_0}, (f')^{n_1}_{i_1}$ with $n_0 < n_1$, there are $\bar{\kappa}$-many 1-simplices f'_j in $S'_1\mathcal{C}_A$ having the common domain $\mathcal{P}(\{n_0, n_1\})$ with $\partial^0 f'_j = (f')^{n_1}_{i_1}$ and $\partial^1 f'_j = (f')^{n_0}_{i_0}$. Hence we now put the domain set

$$\mathcal{D}_1 = \mathcal{D}_0 \cup \{\{(n_0,i_0),(n_1,i_1),j\} |\ n_0 < n_1 < \omega; i_0, i_1, j < \bar{\kappa}\}.$$

Then we identify $S'_1\mathcal{C}_A$ as a functor F'_1 from $\mathcal{D}_1$ to $\mathcal{C}$ such that $F'_1 \restriction \mathcal{D}_0 = F'_{01}$, and $F'_1(\{(n_0,i_0),(n_1,i_1),j\})$ corresponds jth 1-simplex having $(f')^{n_0}_{i_0}, (f')^{n_1}_{i_1}$ as 0-faces. Similarly we try to identify $S'_1\mathcal{C}_A$ as a functor F_1 from $\mathcal{D}_1$ to $\mathcal{C}_A$, extending F_0. But to make F'_1 and F_1 isomorphic, we need extra care when defining F_1. For each $j < \bar{\kappa}$ and a set $a'_j = f'_j(\{n_0, n_1\})$ of corresponding 1-simplex f'_j, assign an embedding $\eta^1_j = \eta^1_{\{(n_0,i_0),(n_1,i_1),j\}}$ sending a'_j to a_j, extending the inverse of $(f'_j)^{\emptyset}_{\{n_0,n_1\}}$. Then we define

$$F_1(\{(n_0,i_0),(n_1,i_1),j\}) = a'_j \text{ and } (F_1)^{\{(n_k,i_k)\}}_{\{(n_0,i_0),(n_1,i_1),j\}} = \eta^1_j \circ (f'_j)^{\{n_k\}}_{\{n_0,n_1\}} \circ (\eta^0_{\{(n_k,i_k)\}})^{-1}.$$

Now then clearly η^1 with $\eta^1 \restriction \mathcal{D}_0 = \eta^0$ is an isomorphism between F'_1 and F_1.

By iterating this argument we can respectively identify $S'_n\mathcal{C}_A$ and $S_n\mathcal{C}_A$, as functors F'_n and F_n having the same domain $\mathcal{D}_n$ extending $\mathcal{D}_1$. Moreover we can also construct an isomorphism η^n, extending η^1, between F'_n and F_n. Note that each $x \in \mathcal{D}_n - \mathcal{D}_{n-1}$ corresponds an n-simplex $f' \in S'_n\mathcal{C}$, and η^n_x corresponds an n-simplex over A $f \in S_n\mathcal{C}$. This correspondence $f' \mapsto f$ induces a bijection from $C'_n\mathcal{C}_A$ to $C_n\mathcal{C}_A$, mapping $c' \mapsto c$, which indeed is an isomorphism of the two groups. Notice that by the construction, if an n-shell c' is the boundary of some $(n+1)$-simplex f', then c is the boundary of f. In general, it follows $(\partial d)' = \partial d'$ (*). Thus this correspondence also induces an isomorphism between $Z'_n(T;A)$

and $Z_n(T;A)$. Moreover it follows from (*) that the correspondence sends $B'_n(T;A)$ to $B_n(T;A)$: Let $c' = \partial d' \in B'_n(T;A)$. Then by (*), we have $c = \partial d \in B_n(T;A)$. Conversely for $c' \in Z'_n(T;A)$, assume $c = \partial e \in B_n(T;A)$. Now for e', again by (*), $\partial e' = c'$. Hence we have $c' \in B'_n(T;A)$. □

3. Basic facts and examples

From now on, we will usually drop the superscripts t and set from $H^t_n(p)$ and $H^{set}_n(p)$ defined in Section 2, since these groups are isomorphic, and use "$H_n(p)$" to refer to the isomorphism class of these two groups. In computing the groups below, we generally use $H^{set}_n(p)$ rather than $H^t_n(p)$.

3.1. Computing H_0. In this subsection, we observe that H_0 does not give any information, since it is always isomorphic to $\mathbb{Z}$, if $\partial_0(f) = 0$ for any 0-simplex f; or is trivial if $\partial_0(f)$ is defined to be $f(\emptyset)$:

Lemma 3.1.

(1) *If $\partial_0(f) = 0$, then for any complete type p over an algebraically closed set A, $H_0(p) \cong \mathbb{Z}$ and for any small subset A of $\mathfrak{C}$, $H_0(T;A) \cong \mathbb{Z}$.*

(2) *If $\partial_0(f) = f(\emptyset)$, then both groups in (1) are trivial.*

Proof. Both parts of the lemma can be proved by essentially the same argument, so we only write out the proof for the group $H_0(p)$ in (1).

For the proof we will define an augmentation map ϵ as in topology. Since we can add parameters to the language for A, we can assume that $A = \emptyset$.

Define $\epsilon : C_0\mathcal{C}(p) \to \mathbb{Z}$ by $\epsilon(c) = \sum_i n_i$ for a 0-chain $c = \sum_i n_i f_i$ of type p. Then ϵ is a homomorphism such that $\epsilon(b) = 0$ for any 0-boundary b (since $\epsilon(\partial f) = 0$ for any 1-simplex f). Thus ϵ induces a homomorphism $\epsilon_* : H_0(p) \to \mathbb{Z}$. Note that *any* 0-chain c is in $Z_0(p)$, so clearly ϵ_* is onto. We claim that ϵ_* is one-to-one, i.e. $\ker \epsilon_* = B_0(p)$. Given a 0-chain $c = \sum_{i\in I} n_i f_i$ such that $\epsilon_*(c) = \sum_{i\in I} n_i = 0$, we shall show c is a boundary. Pick some natural number m greater than every k_i where $\operatorname{dom} f_i = \mathcal{P}(\{k_i\})$. Let $a_i = \operatorname{acl}(a_i) = f_i(\{k_i\})$. Then choose a realizing p such that $a \downarrow \{a_i : i \in I\}$. Now let g_i be a closed 1-simplex of p such that $\operatorname{dom} g_i = \mathcal{P}(\{k_i, m\})$, $g_i(\{k_i\}) = a_i$, and $g_i(\{m\}) = a$. Then $\partial g_i = c_m - f_i$, where c_m is the 0-simplex such that $c_m(\emptyset) = \emptyset$ and $c_m(\{m\}) = a$. Then $c + \partial(\sum_i n_i g_i) = \sum_i n_i f_i + \sum_i n_i(c_m - f_i) = (\sum_i n_i)c_m = 0$. Hence c is a 0-boundary, and $H_0(p) \cong \mathbb{Z}$. □

3.2. Amalgamation properties. The amalgamation properties in Definition 1.11 can be specialized to the context of model theory, yielding the usual notion of n-amalgamation (as in [11]).

Definition 3.2.

(1) If A is a small subset of $\mathfrak{C}$, then T has the *n-amalgamation property over* (*based on*, resp.) A if for every $(n-2)$-shell c *over* (*based on*, resp.) A, there is an $(n-1)$-simplex f such that $c = \partial f$.

(2) A complete type p has the *n-amalgamation* if any closed functor $f : \mathcal{P}^-(n) \to \mathcal{C}_A$ in p can be extended to an $(n-1)$-simplex.

(3) Similarly, "n-uniqueness" *over* A, *based on* A, or *of the type* p can be defined and so can be the notion of "n-CA".

Remark 3.3.

(1) Amalgamation properties *based on* A is equivalent to amalgamation properties *over* all $B \supseteq A$, which implies n-amalgamation for any type p. A stable theory has 4-amalgamation over any model M, as noted in [2]. However, it need not have 4-amalgamation based on M. For suppose that T is a stable theory in which there is a definable groupoid $\mathcal{G}$ which has unboundedly many connected components, each of which is not almost retractable (see [4]). Then if $M \models T$, a is the name of a connected component of $\mathcal{G}$ which does not intersect M (noting that these are equivalence classes which live in T^{eq}), and $B = \mathrm{acl}(Ma)$, then T does not have 4-amalgamation over the set B.

(2) Similarly, if p has n-amalgamation, then so does any non-forking extension, but the converse need not hold even in a stable theory; see Remark 1.8 of [6].

(3) As is well known, if T is simple then T has 3-CA; and if T is stable, then T has 2-uniqueness by stationarity. A non-simple rosy theory cannot have 3-amalgamation [16] but it may have n-amalgamation for all $n \geq 4$ (e.g. the theory of dense linear ordering).

Now we can restate Corollary 1.28 as:

Fact 3.4. *Assume T has n-CA based on $A = \mathrm{acl}(A)$ for $n \geq 2$. Then*

$$H_{n-1}(T; A) = \{[c] |\ c \text{ is an } (n-1)\text{-shell over } A \text{ with support } n+1\}.$$

and

$$H_{n-1}(p) = \{[c] |\ c \text{ is an } (n-1)\text{-shell of } p \text{ with support } n+1\}.$$

So it follows:

Fact 3.5. *Suppose $n \geq 3$.*

(1) *If T has n-CA based on $A = \mathrm{acl}(A)$, then $H_{n-2}(T; A) = 0$.*

(2) *If $p \in S(A)$ (where $A = \mathrm{acl}(A)$) has n-CA, then $H_{n-2}(p) = 0$.*

However, the converse of the above fact is false in general: the theory of the random tetrahedron-free hypergraph does not have 4-amalgamation, but all of its homology groups are trivial ([6, 1.32]).

Fact 3.6. *If T is simple, then $H_1(T; A) = 0$ and $H_1(p) = 0$ for any strong type p in T.*

The fact above is extended to any rosy theory in [14].

3.3. More examples. Homology groups of some examples are already given in [6, Section 1.2]. There $H_2(p)$ of a strong type p in a stable theory is computed too. In this subsection,

we compute some homology groups for o-minimal examples.

Example 3.7. Let p be the unique 1-type over $\emptyset$ in the theory T_{dlo} of dense linear ordering (without end points). Due to weak elimination of imaginaries it is a strong type. We show that $H_n(p) = 0$ for every $n \geq 1$, even though it does not have 3-amalgamation. It is not hard to see that p has n-amalgamation for all $n \neq 3$. Now we claim that, just like in Claim 1.33 in [5], any n-cycle is a sum of n-shells. The proof will be similar, and we use the same notation. We want to construct the edges h_{ij}. The trick this time is to take a^* greater than all the points of the form $a' = g_{ij}(\{k\})$. Then given any edge $\{b, c\} = g_{ij}(\{k, \ell\})$, where either $b < c$ or $c < b$, pick $a > b, c$. Then since $\mathrm{tp}(a'a^*) = \mathrm{tp}(ba) = \mathrm{tp}(ca)$, the construction of h_{ij} on this level is compatible. For the rest of the construction, use n-amalgamation.

Due to the claim and $(n + 2)$-amalgamation, all of the groups $H_n(p)$ are 0 for $n \neq 1$. Furthermore, $H_1(p) = 0$ because any 1-shell is the boundary of a 2-fan (choose a point greater than all the vertices of all the terms in the 1-shell).

Example 3.8. In [14], it is shown that for any strong type p in $\mathfrak{C}^{\mathrm{eq}}$ of a rosy theory, if it is a Lascar type too then $H_1(p) = 0$. But the reason for the triviality of $H_1(p)$ can be arbitrarily complicated. Here we argue that if p is a complete 1-type over $A = \mathrm{acl}(A)$ in the home sort of an o-minimal theory then $H_1(p) = 0$ due to a rather simple reason. Now fix such a p in an o-minimal theory.

Lemma 3.9. *Assume p is non-algebraic. Then there is a type $q(x, y) \in S(A)$ such that:*

(1) *whenever $(a, b) \models q(x, y)$, then a and b are A-independent, and each realizes p; and*

(2) *for any pair (a, b) of A-independent realizations of p, there is a third realization c of p such that c is A-independent from ab and both (a, c) and (b, c) realize q.*

Proof. Recall that since T is o-minimal, any A-definable unary function $f(x)$ is either eventually increasing (that is, there is some point c such that if $c < x < y$ then $f(x) < f(y)$, eventually decreasing, or eventually constant. If f is eventually constant with eventual value d, then $d \in \mathrm{dcl}(A)$.

We say an A-definable function $f(x_1, \ldots, x_n)$ *bounded within* p if for any realizations $c_1, \ldots, c_n \models p$, there is d realizing p such that $d > f(c_1, \ldots, c_n)$. We call a pair of realizations (a, b) of p an *extreme pair* if whenever $f(x)$ is bounded within p, then $b > f(a)$.

First note that by the compactness theorem, for any a realizing p, there is a b realizing p such that (a, b) is an extreme pair. Also, if $b \in \mathrm{dcl}(aA) = \mathrm{acl}(aA)$, then there is an A-definable function $f : p(\mathfrak{C}) \to p(\mathfrak{C})$ such that $b = f(a)$, so since there is no maximal realization c of p (because such a realization c would be in $\mathrm{dcl}(A)$ and we are assuming that p is non-algebraic), it follows that (a, b) is **not** an extreme pair. So any extreme pair is algebraically independent over A and hence thorn-independent (see [18]).

Claim 3.10. *Any two extreme pairs have the same type over A.*

Proof. It is enough to check that if (a, b) and (a, c) are two extreme pairs, then $\mathrm{tp}(b/Aa) = \mathrm{tp}(c/Aa)$. By o-minimality, any Aa-definable set X is a finite union of intervals, and the endpoints $\{d_1, \ldots, d_n\}$ of these intervals lie in $\mathrm{dcl}(Aa)$. So $d_i = f(a)$ for some A-definable function f, and as we already observed $b, c \neq d_i$. Hence it suffices to see $b > d_i$ iff $c > d_i$. Now by the definition of an extreme pair,

$$\forall x \models p \ \exists y \models p \, [y > f(x)] \Rightarrow b > f(a) = d_i.$$

Also,

$$\exists x \models p \ \forall y \models p\,[y \leq f(x)] \Rightarrow \forall x \models p \ \forall y \models p\,[y \leq f(x)]$$

because any two realizations of p are conjugate under an automorphism in $\mathrm{Aut}(\mathfrak{C}/A)$ which permutes $p(\mathfrak{C})$, and so

$$\exists x \models p \ \forall y \models p\,[y \leq f(x)] \Rightarrow b \leq f(a) = d_i.$$

The same reasoning applies with c in place of b, so

$$\begin{aligned} b > d_i = f(a) &\Leftrightarrow \forall x \models p \ \exists y \models p\,[y > f(x)] \\ &\Leftrightarrow c > f(a) = d_i. \end{aligned}$$

□

Let $q(x, y) = \mathrm{tp}(a', b'/A)$ for some extreme pair. Condition (2) of the definition of weak 3-amalgamation can be ensured by picking $c \models p$ so that $c > g(a, b)$ for any A-definable function $g(y, z)$ bounded within p, which is possible by the compactness theorem. □

The two conditions in Lemma 3.9 clearly mean that p has weak 3-amalgamation defined in [14]. Because of this or direct observation it follows $H_1(p) = 0$.

4. Work in progress

Here we summarize some work in progress concerning our homology groups.

In [6] the following was conjectured: Let T be stable having $(n+1)$-CA (over any algebraically closed set), and $p \in S(A)$ with $A = \mathrm{acl}(A)$. Then for every $n \geq 1$,

$$H_n(p) \cong \Gamma_n(p) := \mathrm{Aut}(\widetilde{a_0 \ldots a_{n-1}}/\bigcup_{i=0}^{n-1} \overline{\{a_0 \ldots a_{n-1}\} \setminus \{a_i\}}),$$

where $\overline{a}$ denotes $\mathrm{acl}(aA)$; $\mathrm{Aut}(C/B)$ denotes the group of elementary permutations of the set C fixing B pointwise; $\{a_0, \ldots, a_n\}$ is A-independent, $a_i \models p$; and

$$\widetilde{a_0 \ldots a_{n-1}} := \overline{a_0 \ldots a_{n-1}} \cap \mathrm{dcl}(\bigcup_{i=0}^{n-1} \overline{\{a_0 \ldots a_n\} \setminus \{a_i\}}).$$

In [6], the conjecture is proved when $n = 1, 2$. We plan to publish a proof for all n in the forthcoming preprint [8]. We may call this the Hurewicz correspondence since the result connects the homology groups to something analogous to a homotopy group, as in algebraic topology. To accomplish this, we needed to generalize the notion of groupoids to higher dimensions, and the vertex groups of the higher groupoids should be isomorphic to the groups $\Gamma_n(p)$ defined above. We could not find suitable generalization in the literature fit in our needs, so in [7] we define *n-ary polygroupoids*. A 2-ary polygroupoid is just an ordinary groupoid. In an n-ary polgroupoid, the "morphisms" live in fibers above ordered n-tuples of objects, and there is a sort of n-ary composition rule on these morphisms. Composition is only possible under certain compatibility conditions, and there are axioms generalizing invertibility and associativity for ordinary groupoids. In [7], we show that in any stable

first-order theory that has k-uniqueness for all $k \leq n$ but fails $(n+1)$-uniqueness, there is an n-ary polygroupoid (definable in a mild extension of the language) which witnesses the failure of $(n+1)$-uniqueness.

In [14], as mentioned above it is proved that $H_1(p) = 0$ for any strong type p in a rosy theory as long as p is a Lascar type too; so any 1-shell in p is the boundary of a 2-chain. However, in contrast to the case of simple theories, we construct a series of types in rosy examples showing that there is no uniform bound for the minimal lengths of the 2-chains in the types having 1-shell boundaries. For this and its own research interests, in [14] and [17], all the possible 2-chains having the same 1-shell boundary are classified in a non-trivial amenable collection of functors. In this classification, the following results are obtained, among others: Any 2-chain with a 1-shell boundary is equivalent (preserving the boundary) to either an NR-type or an RN-type 2-chain with a support of size 3. Combinatorial and algebraic criteria determining the two types are given. A planar 2-chain is equivalent to a Lascar 2-chain.

In [4] and [5], from the failure of 3-uniqueness of a strong type p in a stable theory, a way of constructing canonical relatively definable groupiods is introduced. The profinite limit of vertex groups of the groupoids will be the automorphism group $\Gamma_2(p)$, and this seems to play a role in our setting analogous to that of a fundamental group; however, unlike $\pi_1(X)$ in topology, $\Gamma_2(p)$ is always abelian, since $\Gamma_2(p) \cong H_2(p)$. But in [15], a different canonical "fundamental" group for the type p is constructed which seems to give more information: this new group need not be abelian, and the group $\Gamma_2(p)$ is in the center of the new group.

In [6, 2.29], given an arbitrary profinite group G, only a brief sketch is given how to build a type p_G in a stable theory T_G such that $H_2(p_G) \cong G$. In [9], a more detailed proof is supplied.

Sustretov has recently found connections between 4-amalgamation and Galois cohomology in the preprint [19]. It would be very interesting to know if his work could be related to the computation of the homology groups discussed in this article.

Acknowledements. The 2nd author was supported by NRF of Korea grant 2013R1A1A2073 702, and Samsung Science Technology Foundation grant BA1301-03. The third author was partially supported by NSF grant DMS-0901315.

References

[1] Glen E. Bredon, *Topology and Geometry.* Springer-Verlag, NewYork, 1993.

[2] Tristram de Piro, Byunghan Kim, and Jessica Millar, *Constructing the hyperdefinable group from the group configuration*, Journal of Math. Logic, **6** (2006), 121–139.

[3] Clifton Ealy and Alf Onshuus, *Characterizing rosy theories*, Journal of Symbolic Logic, **72** (2007), 919–940.

[4] John Goodrick and Alexei Kolesnikov, *Groupoids, covers, and* 3*-uniqueness in stable theories*, Journal of Symbolic Logic, **75** (2010) 905–929.

[5] John Goodrick, Byunghan Kim, and Alexei Kolesnikov, *Amalgamation functors and boundary properties in simple theories*, Israel Journal of Mathematics, **193** (2013), 169–207.

[6] ———, *Homology groups of types in model theory and the computation of $H_2(p)$*, Journal of Symbolic Logic, **78** (2013), 1086–1114.

[7] ———, *Type-amalgamation properties and polygroupoids in stable theories*, Submitted.

[8] ———, *Homology groups of types in stable theories and the Hurewicz correspondence*, In preparation.

[9] ———, *Characterization of the second homology group of a stationary type in a stable theory*, To appear in *Proceedings of 13th Asian Logic Conference*.

[10] Gwyneth Harrison-Shermoen, *Independence Relations in Theories with the Tree Property.* Ph.D. thesis, University of California, Berkeley, 2013.

[11] Ehud Hrushovski, *Groupoids, imaginaries and internal covers*, Turkish Journal of Mathematics, **36** (2012), 173-198.

[12] Byunghan Kim, *Simplicity Theory*, Oxford University Press, 2014.

[13] Byunghan Kim and Hyeung-Joon Kim, *Notions around tree property 1*, Annals of Pure and Applied Logic, **162** (2011), 698–709.

[14] Byunghan Kim, SunYoung Kim, and Junguk Lee, *A classification of 2-chains having 1-shell boundaries in rosy theories*, Submitted.

[15] ———, *Non-commutative groupoids obtained from the failure of 3-uniqueness in stable theories*, In preparation.

[16] Byunghan Kim and Anand Pillay, *Simple theories*, Annals of Pure and Applied Logic, **88** (1997) 149–164.

[17] SunYoung Kim and Junguk Lee, *More on classification of 2-chains having 1-shell boundaries in rosy theories*, In preparation.

[18] Alf Onshuus, *Properties and consequences of thorn-independence*, Journal of Symbolic Logic, **71** (2006), 1–21.

[19] Dmitry Sustretov, *Elimination of generalised imaginaries and Galois cohomology*, Preprint; arXiv:1312.2273.

[20] Frank O. Wagner, *Simple theories.* Kluwer Academic Publishers, 2000.

Department of Mathematics, Universidad de los Andes, Bogotá, Colombia
E-mail: jr.goodrick427@uniandes.edu.co

Department of Mathematics, Yonsei University, Seoul, Korea
E-mail: bkim@yonsei.ac.kr

Department of Mathematics, Towson University, MD, USA
E-mail: akolesnikov@towson.edu

Definability in non-archimedean geometry

François Loeser

Abstract. We discuss several situations involving valued fields for which the model-theoretic notion of definability plays a central role. In particular, we consider applications to p-adic integration, diophantine geometry and topology of non-archimedean spaces.

Mathematics Subject Classification (2010). Primary 03C10, 03C65, 03C98,12J10, 14G22, 22E35; Secondary 03C64, 03C68, 11S80, 11F85, 14G20, 14T05, 20G25, 22E50.

Keywords. Valued fields, p-adic integration, non-archimedean geometry, motivic integration, Berkovich spaces.

1. Introduction

After the groundbreaking work of Ax-Kochen [2] and Eršov [25] in the sixties and of Denef [16] in the eighties, a wide array of applications of model theory of valued fields is now flourishing, ranging over topics as diverse as counting subgroups, the Langlands program and singularity theory. In all these applications the concept of definability in first order logic is central. In this survey, we shall focus on three such applications, each using the notion of definability in the context of valued fields in an essential way.

We start by presenting several transfer theorems for p-adic integrals. Such results allow to transfer statements over $\mathbb{Q}_p$ to statements over $\mathbb{F}_p(\!(t)\!)$ and vice versa. A first result, obtained in collaboration with R. Cluckers deals with identities between integrals with parameters. In work with R. Cluckers and T. Hales it was shown how it can used for the integrals occuring in the fundamental lemma. We shall also present more recent results obtained by R. Cluckers, J. Gordon and I. Halupczok on transfering local integrability or uniform boundedeness statements and some of their applications to p-adic harmonic analysis. In the next section, we shall explain how by working in a definable setting one can deduce global bounds from local bounds on differentials, despite the totally disconnected nature of non-archimedean valued fields and present some diophantine applications. This is recent joint work with R. Cluckers and G. Comte. The last section is about the topology of non-archimedean spaces. We shall present our work with E. Hrushovski on stable completion of algebraic varieties over a valued field, a model-theoretic analogue of the Berkovich analytification. A fundamental statement is that the stable completion of an algebraic variety is pro-definable. We shall explain how using this approach one can prove new tameness results for the topology of Berkovich spaces.

The present overview is far from being exhaustive, for instance it completely leaves out important work of Hrushovski and Kazhdan on motivic integration [29, 30], and some of its

Proceedings of the International Congress of Mathematicians, Seoul, 2014

recent applications [32, 35].

2. Definability and integration

2.1. Definable sets. A language $\mathcal{L}$ is a set consisting of symbols for constants (= 0-ary functions), n-ary functions and n-ary relations. Basic examples are the ring language $\{0, 1, +, -, \times, =\}$, the order language $\{<, =\}$, or the ordered abelian group language $\{<, 0, +, -, =\}$.

An $\mathcal{L}$-structure consists of a set M together with interpretations for symbols in $\mathcal{L}$. One requires that $=$ is interpreted by equality in M. A subset of M^n is said to be definable if it is of the form

$$\{(a_1, \cdots, a_n) \in M^n : \varphi(a_1, \cdots, a_n) \text{ holds }\}$$

with φ a first-order formula in $\mathcal{L}$ with n-free variables. When the formula φ involves parameters running over some $A \subset M$, one says the subset is A-definable. A map between A-definable sets is said to be A-definable if its graph is. In this way one defines the category Def_A of A-definable sets. All these notions extend naturally to many-sorted languages.

2.2. *p*-adic integrals. In his breakthrough paper [16] on the rationality of the Poincaré series associated to the p-adic points on a variety, Denef proved the following general rationality result for p-adic integrals:

Theorem 2.3. *Let X be a definable subset of $\mathbb{Q}_p^n$ and $g : X \to \mathbb{Q}_p$ be a bounded definable function. Then the integral*

$$\int_X |g|^s |dx|$$

is a rational function of p^{-s}.

Here definability refers to the ring language with parameters in $\mathbb{Q}_p$ (or, which amounts to the same here, any standard valued ring language, for instance $\mathcal{L}_{k,\Gamma}$ considered in 4.8). The proof relies on Macintyre's quantifier elimination theorem [36] for $\mathbb{Q}_p$.

For X a definable subset of $\mathbb{Q}_p^n$, denote by $\mathcal{C}_p(X)$ the $\mathbb{Q}$-algebra generated by functions of the form $|g|$ and $\mathrm{val}(g)$ with $g : X \to \mathbb{Q}_p$ definable. In the paper [17] in which he extended his rationality result to the setting of integrals with parameters, Denef proved the following result about stability under integration for functions in $\mathcal{C}_p$.

Theorem 2.4. *Let X be a definable subset of $\mathbb{Q}_p^n$. Let $\varphi \in C_p(X \times \mathbb{Q}_p^m)$. Assume for any $x \in X$, the function $\varphi_x : \lambda \mapsto \varphi(x, \lambda)$ is integrable. Then the function $x \mapsto \int_{\mathbb{Q}_p^m} \varphi_x |d\lambda|$ belongs to $\mathcal{C}_p(X)$.*

In [18], Denef proved a general cell decomposition theorem for $\mathbb{Q}_p$-definable sets, providing direct proofs of Theorems 2.3 and 2.4 and also of Macintyre's quantifier elimination theorem. The natural question of uniformity in p in Denef's Theorem 2.3 has been addressed by Pas in [39] and by Macintyre in [37]. In the paper [39] a three sorted language has been introduced, nowadays called the Denef-Pas language $\mathcal{L}_{\mathrm{DP}}$. In this language, there are three sorts of variables:

- variables running over the valued field for which the language is the ring language

- variables running over the residue field sort for which the language is the ring language
- variables running over the value group sort for which the language is the language of ordered groups.

and two additional symbols ac and val from the valued field sort to the residue field and value group sort, respectively. For $\mathbb{Q}_p$ or $k(\!(t)\!)$ the angular component map ac is interpreted as the first non zero coefficient in the p-adic, resp. t-adic, expansion and val as the valuation. In this setting, Pas proved a cell decomposition theorem which is uniform in p in [39]. In particular, this provides a new elementary proof of the following version of the classical result of Ax-Kochen-Eršov.

Theorem 2.5. *Let φ be a sentence (that is, a formula with no free variable) in the language $\mathcal{L}_{DP}$. For all but finitely prime numbers p, φ is satisfied in $\mathbb{F}_p(\!(t)\!)$ if and only if it is satisfied in $\mathbb{Q}_p$.*

2.6. Motivic integrals. In the series of papers [10] and [11] in collaboration with Raf Cluckers we have developed a general framework for motivic integration on definable sets in the Denef-Pas language. More precisely let k be a field of characteristic zero and set $K = k(\!(t)\!)$. We consider K as a structure for the Denef-Pas language. For any definable subset S of K^m (or more generally of $K^m \times k^n \times \mathbb{Z}^r$), we define in [10] an algebra $\mathcal{C}(S)$ of "constructible motivic functions" on S. For such functions one defines inductively the notion of being integrable and the value of the integral, using the cell decomposition theorem of Pas [39], and one proves an analogue of Theorem 2.4 in this context. Working in a relative setting is essential here. One of the main advantage of working in the definable setting over previous constructions as those in [19] or [20], is that there is no need anymore to consider completions of Grothendieck rings. Also, we are able to state and prove Fubini and change of variables theorems in full generality, and to deal with integrals with parameters. For more detailed, though accessible, presentations of this theory, we refer to the introduction of [10] and to the paper [12].

2.7. Transfer theorems for constructible motivic functions. Let $\mathbf{F}$ be a number field with ring of integers $\mathcal{O}$. Let $\mathcal{C}_{\mathcal{O}}$ denote the collection of triples (F, ι, ϖ), where F is non-archimedean local field, $\iota : \mathcal{O} \to F$ a ring homomorphism and ϖ a uniformizer in F. We denote by k_F the residue field of F and by q_F the cardinality of k_F. For $M > 0$, we denote by $\mathcal{C}_{\mathcal{O},M}$ the subcollection of triples (F, ι, ϖ) with F of residue characteristic $> M$.

Assume now $k = \mathbf{F}$, and fix a definable subset S of K^n. For some M large enough, for any (F, ι, ϖ) in $\mathcal{C}_{\mathcal{O},M}$ one may consider the specialization S_F of S in F^n obtained by specializing the formulas defining S using ι and sending t to ϖ. Similarly, for M large enough, a function φ in $\mathcal{C}(S)$ may be specialized to a function on S_F which we shall denote by φ_F.

In [11], we prove the following:

Theorem 2.8. *Let $\psi \in \mathcal{C}(S \times K^m)$ and $\psi' \in \mathcal{C}(S \times K^{m'})$. Then, there exists $M > 0$ such that, for every F_1 and F_2 in $\mathcal{C}_{\mathcal{O},M}$ such that $k_{F_1} \simeq k_{F_2}$,*

$$\int_{F_1^m} (\psi_{F_1})_x |d\lambda| = \int_{F_1^{m'}} (\psi'_{F_1})_x |d\lambda'|$$

for every $x \in S_{F_1}$ if and only if

$$\int_{F_2^m} (\psi_{F_2})_x |d\lambda| = \int_{F_2^{m'}} (\psi'_{F_2})_x |d\lambda'|$$

for every $x \in S_{F_2}$.

In particular, when $\mathbf{F} = \mathbb{Q}$, we get that, for almost all p, the identity

$$\int_{\mathbb{Q}_p^m} (\psi_{\mathbb{Q}_p})_x |d\lambda| = \int_{\mathbb{Q}_p^{m'}} (\psi'_{\mathbb{Q}_p})_x |d\lambda'|$$

holds for every $x \in S_{\mathbb{Q}_p}$ if and only if

$$\int_{\mathbb{F}_p((t))^m} (\psi_{\mathbb{F}_p((t))})_x |d\lambda| = \int_{\mathbb{F}_p((t))^{m'}} (\psi'_{\mathbb{F}_p((t))})_x |d\lambda'|$$

holds for every $x \in S_{\mathbb{F}_p((t))}$. Note that Theorem 2.5 can be viewed as a special case of Theorem 2.8 when $m = m' = 0$ and S is the definable subset of K^0 defined by the sentence φ.

In work with Cluckers and Hales [12] we have shown that Theorem 2.8 applies in particular to the integrals occurring in the fundamental lemma, both in the unweighted and weighted case. This is performed by representing all the data entering into the fundamental lemma within the general framework of identities of motivic integrals of constructible functions. This provides alternative proofs of results of Waldspurger in [46] and [47] and is of special interest in view of Ngô's proof of the fundamental lemma over local fields of positive characteristic [38]. One advantage of our approach is that it may be applied quite directly to other versions of the fundamental lemma, as in [50].

Another important property of motivic constructible functions is that they satisfy strong uniform boundedness statements, as proved by Cluckers, Gordon, Halupczok in the appendix B of [44]:

Theorem 2.9. *Let S be a definable set and let $\varphi \in \mathcal{C}(S \times \mathbb{Z}^n)$.*

(1) *There exist integers a and b, M, such that for every F in $\mathcal{C}_{\mathcal{O},M}$, if there exists a set-theoretical function $\alpha : \mathbb{Z}^n \to \mathbb{R}$ such that $|\varphi_F(s,\lambda)|_{\mathbb{R}} \leq \alpha(\lambda)$ on $S_F \times \mathbb{Z}^n$, then $|\varphi_F(s,\lambda)|_{\mathbb{R}} \leq q_F^{a+b||\lambda||}$ on $S_F \times \mathbb{Z}^n$, with $||\lambda|| = \sum_i |\lambda_i|$.*

(2) *Given integers a and b, there exists M, such that whether the bound*

$$|\varphi_F(s,\lambda)|_{\mathbb{R}} \leq q_F^{a+b||\lambda||}$$

holds or not on the whole of $S_F \times \mathbb{Z}^n$ depends only on k_F, for F in $\mathcal{C}_{\mathcal{O},M}$.

In the same paper they show this result may be applied to provide uniform bounds for orbitals integrals that are used in an essential way in the paper [44].

2.10. Transfer theorems for exponential constructible motivic functions. In [11], we extend the construction of algebras constructible motivic functions $\mathcal{C}(S)$, to take in account motivic versions of exponential functions, by constructing the algebra $\mathcal{C}^{exp}(S)$ of exponential constructible motivic functions on S for any definable set S. The formalism developed in [10] for $\mathcal{C}(S)$ carries over to $\mathcal{C}^{exp}(S)$.

Given an non-archimedean field F, one denotes by $\mathcal{D}_F$ the set of additive characters on F that are trivial on the maximal ideal and nontrivial on the valuation ring. Now, given φ in $\mathcal{C}^{exp}(S)$, for any F in $\mathcal{C}_{\mathcal{O},M}$ and any character θ in $\mathcal{D}_F$, one may specialize φ to a function $\varphi_{F,\theta}$ on S_F.

In this setting, Theorem 2.8 may be generalized as follows:

Theorem 2.11. *Let $\psi \in \mathcal{C}^{exp}(S \times K^m)$ and $\psi' \in \mathcal{C}^{exp}(S \times K^{m'})$. Then, there exists $M > 0$ such that, for every F_1 and F_2 in $\mathcal{C}_{\mathcal{O},M}$ such that $k_{F_1} \simeq k_{F_2}$,*

$$\int_{F_1^m} (\psi_{F_1,\theta})_x |d\lambda| = \int_{F_1^{m'}} (\psi'_{F_1,\theta})_x |d\lambda'|$$

for every $x \in S_{F_1}$ and any $\theta \in \mathcal{D}_{F_1}$ if and only if

$$\int_{F_1^m} (\psi_{F_1,\theta})_x |d\lambda| = \int_{F_1^{m'}} (\psi'_{F_1,\theta})_x |d\lambda'|$$

for every $x \in S_{F_1}$ and any $\theta \in \mathcal{D}_{F_1}$.

In the paper [13], Cluckers, Gordon, Halupczok prove the following remarkable transfer theorem for (local) integrability and boundedness:

Theorem 2.12. *Let S be a definable subset of K^m and let $\varphi \in \mathcal{C}^{exp}(S)$. There exists $M > 0$ such that, for fields F in $\mathcal{C}_{\mathcal{O},M}$, the validity of the statement that $\varphi_{F,\theta}$ is (locally) integrable, resp. (locally) bounded, for all $\theta \in \mathcal{D}_F$ depends only on the isomorphism class of k_F.*

Using Theorem 2.12, Cluckers, Gordon, Halupczok have been able in [14] to transfer Harish-Chandra's theorems on local integrability of characters of irreducible admissible representations of connected reductive p-adic groups from characteristic zero to (large) positive characteristic. An important ingredient in their approach is the definability of the Moy-Prasad filration subgroups, which they have proved in a number of important special cases.

3. Definability and non-archimedean diophantine geometry

3.1. Lipschitz functions. A C^1-function on an interval in $\mathbb{R}$ which has bounded derivative is automatically Lipschitz continuous. It is well known that such a result cannot hold for general C^1-functions over the p-adics since $\mathbb{Q}_p$ is total disconnectedness. However, under some definability conditions it is still possible to get results of this kind, as we shall explain now.

Let K be a field endowed with a discrete valuation for which it is complete. In this section, by definable we shall mean definable in the ring language L_K with parameters in K (in this case definable sets are also called semi-algebraic sets), or in the analytic language L_K^{an} which is obtained by adding to L_K a symbol for each restricted power series f in $K\{x_1, \ldots, x_m\}$, for $m \geq 1$. Such a symbol is interpreted as the function $K^m \to K$ which is zero outside $\mathcal{O}_K^m$ and given by $x \mapsto f(x)$ for $x \in \mathcal{O}_K^m$. In this case definable sets are also called subanalytic sets.

Let X be a subset of K^m. We say a function $f : X \to K$ is C-Lipschitz if for every x and y in X, $|f(x) - f(y)| \leq C|x - y|$. We say it is locally C-Lipschitz if for each point x_0 in X, the restriction of f to some neighborhood of x_0 is C-Lipschitz.

In the paper [7] with Cluckers and Comte we prove the following:

Theorem 3.2. *Let X be a definable subset of $\mathbb{Q}_p^m$ and let $f : X \to \mathbb{Q}_p$ be a definable map. Assume f is locally C-Lipschitz. Then there exists a finite partition of X into definable sets X_i and C' such that the restriction of f to each X_i is C'-Lipschitz.*

This statement is a p-adic analogue of a theorem of Kurdyka for real subanalytic sets [34]. In [9] Cluckers and Halupczok proved that it is in fact always possible to take $C' = C$.

3.3. A p-adic analogue of the Yomdin-Gromov lemma. A very efficient tool in diophantine geometry is the so-called determinant method which was developed by Bombieri and Pila in the influential paper [6] about the number of integral points of bounded height on affine algebraic and transcendental plane curves. Basically, the method consists in using a determinant of a suitable set of monomials evaluated at the integral points, in order to construct a family of auxiliary polynomials vanishing at all integral points on the curve within a small enough box. Building on the estimates in [6] for algebraic curves, Pila proved in [40] bounds on the number of integral (resp. rational) points of bounded height on affine (resp. projective) algebraic varieties of any dimension, improving on previous results by S. D. Cohen using the large sieve method [15].

In [41], Pila and Wilkie proved a general estimate for the number of rational points on the transcendental part of sets definable in an o-minimal structure; this has been used in a spectacular way by Pila to provide an unconditional proof of some cases of the André-Oort Conjecture [42]. Lying at the heart of Pila and Wilkie's approach is the possibility of having uniform - in terms of number of parametrizations and in terms of bounds on the partial derivatives - C^k-parametrizations. These parametrizations are provided by an o-minimal version of Gromov's algebraic parametrization Lemma [26], itself a refinement of a previous result of Yomdin [48],[49]. Such C^k-parametrizations enter the determinant method via Taylor approximation.

In the work [8] with Cluckers and Comte we provide a version of the Yomdin-Gromov lemma and the Pila-Wilkie theorem valid over $\mathbb{Q}_p$. At first sight one may have doubts such a statement could exist, since there seem there is no way for a global Taylor formula to make sense in this framework. However Theorem 3.2 which provides a version of first-order Taylor approximation, piecewise globally, in the definable p-adic setting is an encouraging sign. In [8], instead of generalizing this result to higher order, we show directly the existence of uniform C^k-parametrizations that do satisfy Taylor approximation, which is enough for our purpose.

Our p-adic analogue of the Yomdin-Gromov lemma is the following statement:

Theorem 3.4. *Let $n \geq 0$, $m \geq 0$ and $r \geq 0$ be integers and let $X \subset \mathbb{Z}_p^n$ be a subanalytic set of dimension m. Then there exists a finite collection of subanalytic functions $g_i : P_i \subset \mathbb{Z}_p^m \to X$ such that the union of the $g_i(P_i)$ equals X, the g_i have C^r norm bounded by 1, and the g_i may be approximated by Taylor polynomials of degree $r-1$ with remainder of order r, globally on P_i.*

For the precise definition of the C^r norm and of approximation by Taylor polynomials of certain degree and with certain error we refer to [8].

3.5. A p-adic analogue of the Pila-Wilkie theorem. For X a subset of $\mathbb{Q}_p^n$ and $T > 1$ a real number, write $X(\mathbb{Q}, T)$ for the set consisting of points $(x_1, \cdots, x_n)$ in $X \cap \mathbb{Q}^n$ such that one can write x_i as a_i/b_i where a_i and $b_i \neq 0$ are integers with $|a_i|_{\mathbb{R}} \leq T$ and $|b_i|_{\mathbb{R}} \leq T$.

For X a subset of $\mathbb{Q}_p^n$, write X^{alg} for the subset of X consisting of points x such that

there exists an algebraic curve $C \subset \mathbb{A}^n_{\mathbb{Q}_p}$ such that $C(\mathbb{Q}_p) \cap X$ is locally at x of dimension 1.

We prove in [8] the following p-adic analogue of the Pila-Wilkie theorem:

Theorem 3.6. *Let $X \subset \mathbb{Q}_p^n$ be a subanalytic set of dimension m with $m < n$. Let $\varepsilon > 0$ be given. Then there exist an integer $C = C(\varepsilon, X) > 0$ and a semialgebraic set $W = W(\varepsilon, X) \subset \mathbb{Q}_p^n$ such that $W \cap X$ lies inside X^{alg}, and such that for each T, one has*

$$\#(X \setminus W)(\mathbb{Q}, T) \leq CT^{\varepsilon}.$$

3.7. Results over $\mathbb{C}[\![t]\!]$. In the paper [8] we also obtain results when $K = \mathbb{C}(\!(t)\!)$. For instance a version of Theorem 3.2 still holds over $\mathbb{C}(\!(t)\!)$ (with $C' = C$), if one replaces "a finite partition of X" by "a partition parametrized by $\mathbb{C}^r$, for some r". For this to make sense one has to enlarge the language to have (higher) angular components maps à la Denef-Pas, see [8] for more details. Similarly, a version of Theorem 3.4 over $\mathbb{C}(\!(t)\!)$ is also proved in [8]. We end this section by stating a diophantine application of this result.

For each positive integer r one denotes by $\mathbb{C}[t]_{<r}$ the set of complex polynomials of degree $< r$. When A is a subset of $\mathbb{C}(\!(t)\!)^n$, one denotes by A_r the set $A \cap (\mathbb{C}[t]_{<r})^n$ and by $n_r(A)$ the dimension of the Zariski closure of A_r in $(\mathbb{C}[t]_{<r})^n \simeq \mathbb{C}^{nr}$.

Let X be an algebraic subvariety of $\mathbb{A}^n_{\mathbb{C}(\!(t)\!)}$ of dimension m. One can prove that for any $r > 0$, $n_r(X) \leq rm$. When X is linear this "trivial" estimate is the best possible. However, we prove in [8] that as soon as X has degree $d \geq 2$, the following non-trivial bound holds:

Theorem 3.8. *Let X be an irreducible subvariety of $\mathbb{A}^n_{\mathbb{C}(\!(t)\!)}$ of dimension m and degree $d \geq 2$. Then, for every positive integer r, one has*

$$n_r(X) \leq r(m-1) + \left\lceil \frac{r}{d} \right\rceil.$$

This result is a geometric analogue of a result of Pila in [40] on the number of integral (resp. rational) points of bounded height on affine (resp. projective) algebraic varieties of any dimension. Pila's proof proceeds by reducing to the case of curves which was considered by Bombieri and Pila in [6].

4. Definability and topology

In this section we present a model-theoretic approach to proving topological tameness properties in non-archimedean geometry which we developed in collaboration with Ehud Hrushovski [31].

4.1. o-minimality. It is by now quite well known that o-minimal geometry provides an efficient framework for the study of topology arising from an ordered structure, in particular in the context of ordered fields. Let us recall that an infinite structure M which is totally ordered by a binary relation $<$ is said to be o-minimal if every definable subset $X \subset M$, with parameters in M, is a finite union of intervals and points. Sets definable in a o-minimal structure have nice topological properties. For instance, for o-minimal expansions of the field $\mathbb{R}$ of real numbers, and $n \in \mathbb{N}$, definable subsets of $\mathbb{R}^n$ have a finite number of connected components which furthermore are definable, they are locally contractible and triangulable;

in particular they have the homotopy type of a finite simplicial complex. Classical examples of subsets of $\mathbb{R}^n$ definable in a o-minimal structure include semi-algebraic sets, subanalytic sets, or sets definable in the language of ordered rings with an exponential function. Another class of examples of o-minimal structures, playing an important role in our work, is provided by divisible ordered abelian groups Γ. In this last setting definable subsets of Γ^n essentially correspond to piecewise linear sets. An important feature of this model-theoretic framework for tameness is that it is particularly well adapted to proving uniformity statements for the topology of definable sets varying in definable families, for instance finiteness of homotopy types occuring in a given such family.

4.2. Valued fields. By a valued field we mean a field K, together with a surjective multiplicative map $\mathrm{val} : K^\times \to \Gamma$, with $\Gamma = (\Gamma, 0, +, <)$ an ordered abelian group such that $\mathrm{val}(x+y) \geq \min(\mathrm{val}(x), \mathrm{val}(y))$. We extend val to a map $\mathrm{val} : K \to \Gamma_\infty$, with Γ_∞ the disjoint union of Γ with a distinguished element ∞ which is larger than any element of Γ and absorbing for the addition. We shall denote by $\mathcal{O}_K$ the valuation ring of K and by $\mathcal{M}_K$ the maximal ideal of K.

4.3. Berkovich spaces. Let K be a valued field such that Γ is a subgroup of $(\mathbb{R}, +)$. Then $x \mapsto |x| = e^{-\mathrm{val}(x)}$ defines an absolute value $|\cdot| : K \to \mathbb{R}_{\geq 0}$. One says K is ultrametric if it is complete for this norm.

In [3], Berkovich introduced a general notion of analytic spaces over an ultrametric field K. In particular, for any algebraic variety V over K one may consider its Berkovich analytification V^{an}. In case V is affine with ring of regular functions $K[V]$, let us define V^{an} as a topological space. As a set V^{an} is the set of multiplicative seminorms on $K[V]$ extending the absolute value on K. There is a natural embedding $V^{an} \subset \mathbb{R}^{K[V]}$ and one endows V^{an} with the topology induced by the product topology on $\mathbb{R}^{K[V]}$. For an arbitrary algebraic variety V over K, one defines V^{an} by glueing. This construction is functorial: any morphism of algebraic variety $f : V \to W$ gives rise to a morphism $f^{an} : V^{an} \to W^{an}$. Note that $V(K)$ may be naturally identified with a subset of V^{an}. When V is affine, this is done by assigning to a point a in $V(K)$ the seminorm $f \mapsto |f(a)|$.

4.4. Some previously known topological properties of Berkovich spaces. Already in [3] Berkovich proved that general analytic spaces (including analytifications of algebraic varieties) have excellent general topological properties, in particular they are locally compact and locally path-connected.

More recently, in his paper [4], Berkovich proved that the general fibre of any polystable formal scheme admits a strong deformation retraction to a finite polyhedron, and using de Jong's results on alterations he deduced that any smooth analytic space is locally contractible.

On the other hand, Ducros proved in [21] that semi-algebraic subsets of V^{an}, i.e. subsets which are Zariski locally boolean combinations of subsets defined by inequalities $|f| \bowtie \lambda|g|$ with f, g in $K[V]$ and $\lambda \in \mathbb{R}_{\geq 0}$, where $\bowtie \in \{<, >, \leq, \geq\}$, have only a finite number of connected components, each of them semi-algebraic.

Another statement with an o-minimal flavour us the following. Let X be a compact analytic space and let f be an analytic function on X. For every $\varepsilon \geq 0$, let X_ε denote the set of points x in X such that $|f(x)| \geq \varepsilon$. According to Abbes and Saito under the assumption that f is invertible [1] and to Poineau in general [43], there is a finite partition of $\mathbb{R}_{\geq 0}$ into intervals such that on each of these intervals the natural map $\pi_0(X_{\varepsilon'}) \to \pi_0(X_\varepsilon)$

is a bijection whenever $\varepsilon \leq \varepsilon'$.

4.5. Statement of results. The results recalled in 4.4 provide rather strong evidence that there should exist general tameness results for the topology of non-archimedean spaces, quite analogous to the ones available in the o-minimal world. In the paper [31], we prove the following general statements on the topology of analytifications of algebraic varieties:

Theorem 4.6. *Let K be an ultrametric[1] field. Let V be a quasi-projective variety over K and let X be a semi-algebraic subset of V^{an}.*

(1) *There exists a strong homotopy retraction $h : [0,1] \times X \to X$ onto a closed subset of X which is homeomorphic to a compact finite polyhedral complex.*

(2) *The space X is locally contractible (one may drop the assumption V quasi-projective here).*

(3) *Let $f : V \to W$ be a morphism of algebraic varieties over K. Then the set of homotopy types of fibers of the map $f^{an}|_X : X \to W^{an}$ is finite.*

(4) *Let $f : V \to \mathbb{A}^1_K$ a morphism. For every $\varepsilon \geq 0$, let X_ε denote the set of points x in X such that $|f(x)| \geq 0$. Then there exists a finite partition of $\mathbb{R}_{\geq 0}$ into intervals such that the natural map $X_{\varepsilon'} \hookrightarrow X_\varepsilon$ is a homotopy equivalence whenever $\varepsilon \leq \varepsilon'$ belong to the same interval.*

4.7. Model-theoretic preliminaries. We shall deal with a complete theory T having quantifier elimination and work in a fixed universe $\mathbb{U}$, by which we mean a large very saturated and homogeneous model. All models M (and parameter sets A) we shall consider will be small substructures (resp. subsets) of $\mathbb{U}$.

If A is a small subset of $\mathbb{U}$, the definable closure $\mathrm{dcl}(A)$ is the set of all elements c in $\mathbb{U}$ such that there exists a formula $\varphi(x)$ with one free variable and parameters in A such that c is the only element of $\mathbb{U}$ such that $\varphi(c)$ holds. If X is a C-definable set and $C \subset A$, we write $X(A)$ for $X(\mathbb{U}) \cap \mathrm{dcl}(A)$.

A basic notion we shall use is that of a definable type. Let assume for simplicity of notation that there is only one sort. Let B be a set of parameters. Let $c = (c_1, \cdots, c_n)$ be a finite tuple of elements of $\mathbb{U}$. The set of all B-formulas satisfied by c in some model of T containing the c_i's is denoted by $\mathrm{tp}(c/B)$ and called the type of c over B. Such a set of formulas is called an n-type over B. In the special case where all c_i's already belong to B one says the type is realized (over B). Let $A \subset M$. We say an n-type p over M is A-definable if for every formula $\varphi(x_1, \cdots, x_n, y_1, \cdots, y_m)$ without parameters, there exists a formula $\varphi_p(y_1, \cdots, y_m)$ with parameters in A, such that for any $(b_1, \cdots, b_m)$ in M^m, $\varphi(x_1, \cdots, x_n, b_1, \cdots, b_m)$ belongs to p if and only if $\varphi_p(b_1, \cdots, b_m)$ holds in M. The mapping $\varphi \mapsto \varphi_p$ is called a defining scheme for p. If p is such an A-definable type over M, for any model M' containing M one can extend p to an A-definable type over M', by using the same defining scheme. Thus, we will not care about a specific M anymore when dealing A-definable types. Note that a realized type over A is always A-definable. These definitions extend naturally to many-sorted languages.

Let X be a C-definable set with $C \subset A$. We say that an A-definable type p is on X if the formula expressing that $x \in X$ belongs to the type p. We denote by $S_{X,def}(A)$ the set of A-definable types on X and set $S_{X,def} = \cup_A S_{X,def}(A)$. Any C-definable map

[1] In fact the completeness hypothesis on K plays no role here.

$f : X \to Y$ between C-definable sets induces a natural push-forward maps

$$f_* : S_{X,def}(A) \to S_{Y,def}(A) \text{ and } f_* : S_{X,def} \to S_{Y,def}.$$

4.8. The language. Classically, to study valued fields one considers a 3-sorted language $\mathcal{L}_{k,\Gamma}$ (or one of its variants) with sorts VF, Γ and $\mathbf{k}$ for the valued field, value group and residue field sorts, with respectively the ring, ordered abelian group and ring language, and additional symbols for the valuation val and the map $\mathrm{Res} : \mathrm{VF}^2 \to k$ sending (x, y) to the residue of xy^{-1} if $\mathrm{val}(x) \geq \mathrm{val}(y)$ and $y \neq 0$ and to 0 otherwise. We consider ACVF, the theory of algebraically closed fields with non trivial valuation such that val is surjective in this language. This theory become complete once the characteristic of the valued field and of its residue field are both fixed. It is a classical result of A. Robinson that ACVF admits quantifier elimination. Note that this result has already nice consequences in non-archimedean geometry. For instance in the paper of Ducros [23] it is used to give an alternate proof of the Bieri-Groves theorem [5].

We shall use an expansion $\mathcal{L}_{\mathcal{G}}$ of this language introduced by Haskell, Hrushovski and Macpherson in [27]. It has additional sorts S_n and T_n for $n \geq 1$, coding respectively n-dimensional lattices over the valuation ring, and elements in the reduction modulo the maximal ideal of such lattices. The main result of [27] is that ACVF has elimination of imaginaries in the language $\mathcal{L}_{\mathcal{G}}$ (which was not the case in the original language $\mathcal{L}_{k,\Gamma}$). A theory T is said to have elimination of imaginaries in a given language if all quotients of definable sets by definable equivalence relations are representable by definable sets. It is also proved in [27] that ACVF still has elimination of quantifiers in $\mathcal{L}_{\mathcal{G}}$.

One should note that expanding the language from $\mathcal{L}_{k,\Gamma}$ to $\mathcal{L}_{\mathcal{G}}$ does not create new definable sets in the sorts VF, Γ and $\mathbf{k}$. If V is an algebraic variety over a valued field, we may define definable subsets of V by requiring that their intersection with any affine open is a definable set.

Given a valued field F, a in F and α in $\mathrm{val}(F)$, resp. α in $\mathrm{val}(F^\times)$, one denotes by $B(a, \alpha)$ and $B^o(a, \alpha)$ respectively the closed and open ball of center a and valuative radius α. They are definable sets defined respectively by the formulas $\mathrm{val}(x - a) \geq \alpha$ and $\mathrm{val}(x - a) > \alpha$. If B is a ball defined over a model K of ACVF, the type expressing that $x \in B$ and $x \notin B'$ for every K-definable ball B' strictly contained in B is a K-definable type, called the generic type of B, and denoted by p_B.

Remark 4.9. Note that the set of all closed balls for K running over all models of ACVF (contained in $\mathbb{U}$) is definable in $\mathcal{L}_{\mathcal{G}}$ (without parameters). Indeed, it suffices to prove that the set of all closed balls of finite valuative radius is definable in $\mathcal{L}_{\mathcal{G}}$, and this follows from the following observation: given a, a' in K and b, b' in $K^\times$, the balls $B(a, \mathrm{val}(b))$ and $B(a', \mathrm{val}(b'))$ are equal if and only if the two-dimensional $\mathcal{O}_K$-lattices generated by $((b, 0), (a, b))$ and by $((b', 0), (a', b'))$ are equal. More precisely, there exists a definable set D in $\mathcal{L}_{\mathcal{G}}$ such that for any $A \subset \mathbb{U}$, $D(A)$ is in natural bijection with the set of A-definable closed balls.

4.10. Stably dominated types. In [28], Haskell, Hrushovski and Macpherson introduced within a general model-theoretic framework the notion of stably dominated types. Roughly speaking, a stably dominated type is a definable type which is "controlled by its stable part". In ACVF, stable domination is equivalent to being orthogonal to Γ in the following sense. Let X be a C-definable set and let $p \in S_{X,def}(A)$, for $C \subset A$. We shall say that p is orthog-

onal to Γ if for every model M of ACVF containing A, every tuple c such that $p = \mathrm{tp}(c/M)$, and every M-definable map $f : X \to \Gamma_\infty$, $f(c) \in \mathrm{val}(M)$. We denote by $\widehat{X}(A)$ the set of A-definable types on X that are orthogonal to Γ and by $\widehat{X}$ the union of all the sets $\widehat{X}(A)$, for $A \subset \mathbb{U}$. We call $\widehat{X}$ the stable completion of X.

Examples 4.11.

1. Realized types are stably dominated, i.e. for any definable set X there is a natural inclusion $\iota : X \to \widehat{X}$.
2. A type over Γ_∞^n is stably dominated if and only it is realized, i.e. $\iota : \Gamma_\infty^n \to \widehat{\Gamma_\infty^n}$ is a bijection.
3. The generic type of a ball is stably dominated if and only if the ball is closed.

It follows from Remark 4.9 and Example 4.11 (3) that, given a valued field F, there is a natural bijection ϑ between $\widehat{\mathbb{A}^1_F}$ and a definable set D, inducing, for any $A \subset \mathbb{U}$, a bijection between $\widehat{\mathbb{A}^1_F}(A)$ and $D(A)$. This is a special case of Theorem 4.14, but before going any further, we should introduce the notion of a pro-definable set. One defines the category ProDef_C of pro-definable sets over C as the category of pro-objects in the category of C-definable sets indexed by a small directed partially ordered set. Thus, if $X = (X_i)_{i\in I}$ and $Y = (Y_j)_{i\in J}$ are two such pro-objects

$$\mathrm{Hom}_{\mathrm{ProDef}_C}(X, Y) = \varprojlim_j \varinjlim_i \mathrm{Hom}_{\mathrm{Def}_C}(X_i, Y_j).$$

Elements of $\mathrm{Hom}_{\mathrm{ProDef}_C}(X, Y)$ will be called C-pro-definable morphisms between X and Y. By a result of Kamensky [33], the functor of "taking $\mathbb{U}$-points" induces an equivalence of categories between the category ProDef_C and the sub-category of the category of sets whose objects and morphisms are inverse limits of $\mathbb{U}$-points of definable sets indexed by a small directed partially ordered set. By pro-definable, we mean pro-definable over some C. We shall thus freely identify a pro-definable set $X = (X_i)_{i\in I}$ with the set $X(\mathbb{U}) = \varprojlim_i X_i(\mathbb{U})$. For any set B with $C \subset B \subset \mathbb{U}$, we set $X(B) = X(\mathbb{U}) \cap \mathrm{dcl}(B) = \varprojlim_i X_i(B)$.

Definition 4.12. Let X be a pro-definable set.

(1) X is called strict pro-definable if it can be written as a pro-definable set with surjective transition morphisms.

(2) X is called iso-definable if it is in pro-definable bijection with a definable set.

(3) $Y \subset X$ is called relatively definable if there exists $i \in I$ and a definable subset W of X_i such that $Y = \pi_i^{-1}(W)$, with π_i the canonical projection $X \to X_i$.

Theorem 4.13. *Let X be a B-definable set. Then $\widehat{X}$ may be canonically endowed with the structure of a strict B-pro-definable set. In particular, there exists a strict B-pro-definable set E such that for any $B \subset A$, there is a canonical identification $\widehat{X}(A) = D(A)$.*

For curves we have the following stronger statement:

Theorem 4.14. *Let C be an algebraic curve over a valued field K and let X be a definable subset of C. Then $\widehat{X}$ is iso-definable.*

For $C = \widehat{\mathbb{P}^1}$ the result follows from the description of $\widehat{\mathbb{A}^1}$ in terms of closed balls given above. The proof in the general case uses Riemann-Roch and Theorem 4.13.

Remark 4.15. The previous statement is optimal since one can show that, for X a definable subset of K^n, $\widehat{X}$ is iso-definable if and only the dimension of the Zariski closure of X is ≤ 1.

Lemma-Definition 4.16. Let $f : X \to Y$ be a map between B-definable sets. Then the map $f_* : S_{X,def} \to S_{Y,def}$ restricts to a B-pro-definable map $\widehat{f} : \widehat{X} \to \widehat{Y}$. In this way we have a functor from the category of B-definable sets to the category of B-pro-definable sets.

Let X be a definable subset. If Y is a definable subset of X, then $\widehat{Y}$ is a relatively definable subset of $\widehat{X}$. The set of realized types in $\widehat{X}$, which can be identified with $X(\mathbb{U})$ is iso-definable and relatively definable in $\widehat{X}$. Its points are called simple points of $\widehat{X}$.

4.17. $\widehat{V}$ as a topological space. We endow $\widehat{\mathbb{A}^n}$ with the coarsest topology such that for every polynomial $F \in \mathbb{U}[x_1, \cdots, x_n]$, the map $\widehat{\mathrm{val} \circ F} : \widehat{\mathbb{A}^n} \to \Gamma_\infty$ is continuous, where the topology on Γ_∞ is the order topology. For any definable subset X of $\mathbb{A}^n$, we endow $\widehat{X}$ with the induced topology. If V is an algebraic variety over a valued field K, we define the topology on $\widehat{V}$ by gluing: it is the unique topology inducing the previous topology on $\widehat{U}$ for U an affine open in V. If X is a definable subset of V, we endow the relatively definable subset $\widehat{X}$ with the induced topology.

We have the following basic properties:

Proposition 4.18. *Let V be an algebraic variety defined over a valued field K. Then:*

(1) *The topology on $\widehat{V}$ is pro-definable in the following sense: there exists a small set I, and for each $i \in I$, a K-definable family $U_i = (U_{i,b})_{b \in \mathbb{U}}$ of relatively definable subsets of $\widehat{V}$, such that the sets $U_{i,b}$, for $b \in \mathbb{U}$ and $i \in I$ generate the topology on $\widehat{V}$.*

(2) *The topology on $\widehat{V}$ is Hausdorff.*

(3) *The subset of simple points is dense in $\widehat{V}$.*

(4) *The induced topology on the set of simple points is the valuation topology.*

In general, we shall call pro-definable sets with a pro-definable topology, pro-definable spaces.

More generally, consider the map $\pi : V \times \mathbb{A}^m \to V \times \Gamma_\infty^m$ which is the identity on the V factor and val on the remaining ones. It induces a map $\widehat{\pi} : \widehat{V \times \mathbb{A}^m} \to \widehat{V \times \Gamma_\infty^m}$ and we endow $\widehat{V \times \Gamma_\infty^m}$ with the direct image topology, making it a pro-definable space. One shows that the canonical map $\widehat{V \times \Gamma_\infty^m} \to \widehat{V} \times \widehat{\Gamma_\infty^m} = \widehat{V} \times \Gamma_\infty^m$ is an homeomorphism.

4.19. Definable compactness. The usual notion of compactness is not well suited to the present setting as shown by the following example. Let K be a valued field with $\mathrm{val}(K^\times) = \mathbb{Q}$. Fix $\varepsilon \in \mathrm{val}(\mathbb{U}^\times)$ such that $0 < \varepsilon < \alpha$ for every positive α in $\mathbb{Q}$. Let C be set defined by the formula $0 \leq \mathrm{val}(x) \leq 1$. For $\alpha \in \mathbb{Q} \cap [0, 1]$ let U_α be defined by $\alpha - \varepsilon < \mathrm{val}(x) < \alpha + \varepsilon$. The family of open sets $\widehat{U_\alpha}$ is a cover of $\widehat{C}$ with no finite subcover.

To remidy this we shall introduce the notion of definable compactness for pro-definable spaces. Let us note that the definition we gave of a definable type still makes sense on pro-definable set.

Definition 4.20. Let X be a pro-definable space.

(1) Let p be a definable type on X. We say $a \in X$ is a limit of p if for every relatively definable neighborhood W of a, the formula expressing $x \in W$ belongs to p.

(2) We say X is definably compact if every definable type on X has a limit.

Note that if X is Hausdorff, limits are unique when they exist.

Let V be a closed subvariety of $\mathbb{A}^m$. A subset $X \subset V$ is said to be bounded in V if it is contained in a product of closed balls. For an arbitrary variety V, a definable subset $X \subset V$ is said to be bounded, if one may write $V = \cup_{i=1}^n V_i$ with V_i open and affine and $X = \cup_{i=1}^n X_i$, with X_i bounded in V_i. A subset of $V \times \Gamma_\infty^m$ will be said to be bounded if its preimage in $V \times \mathbb{A}^m$ is. Finally, a pro-definable subset $X \subset \widehat{V} \times \Gamma_\infty^m$ will be said to be bounded if there exists a bounded definable subset W of $V \times \Gamma_\infty^m$ such that $X \subset \widehat{W}$.

Theorem 4.21. *Let X be a pro-definable subset of $\widehat{V} \times \Gamma_\infty^m$. Then X is definably compact if and only if it is closed and bounded.*

Corollary 4.22. *A variety V over a valued field is complete if and only if $\widehat{V}$ is definably compact.*

4.23. Γ-internality. We shall now define an important class of subsets of $\widehat{V} \times \Gamma_\infty^m$ which "look like o-minimal sets".

Definition 4.24. A subset Z of $\widehat{V} \times \Gamma_\infty^m$ is said to be Γ-internal if it is iso-definable and there is a definable subset D of some Γ_∞^n and a surjective pro-definable map $D \to Z$.

The iso-definability condition is crucial here, and cannot be replaced by just requiring pro-definability. This definition is purely definable and does not say anything a priori about the topology of Z. The following embedding result shows that being Γ-internal imposes strong restrictions on the topology:

Theorem 4.25. *Let Z be a Γ-internal subset of $\widehat{V} \times \Gamma_\infty^m$. Then there exists an injective continuous definable map $f : Z \hookrightarrow \Gamma_\infty^n$ for some n. If Z is definably compact, such an f is an homeomorphism.*

If V and Z are defined over some set of parameters A, one cannot in general expect such an f to be defined, because it should be respect the Galois action. However the following holds:

Proposition 4.26. *Assume V and Z are defined over some set of parameters A in the* VF *and Γ sorts. Then there exists a finite A-definable set w and an injective continuous A-definable map $f : Z \hookrightarrow \Gamma_\infty^w$.*

4.27. Paths and definable connectedness. The mapping $[0, \infty] \to \widehat{\mathbb{P}^1}$ sending t to the generic type of the ball $B(0, t)$ may be seen as a path connecting 0 and the generic type $p_\mathcal{O}$ of the closed unit ball. Similarly the mapping $[0, \infty] \to \widehat{\mathbb{P}^1}$ sending t to the generic type of the ball $B(1, t)$ connects 1 and $p_\mathcal{O}$. By composing these paths one connects the point 0 and 1. However a technical issue occurs here. Since multiplication is not part of the structure Γ_∞, there is no way to identify the space obtained by gluing two copies of $[0, \infty]$ at 0 with an interval. We are thus led to consider generalized intervals, that is spaces obtained by

concatening a finite number of closed intevals in Γ_∞ either with the order from Γ_∞ or with the reverse order.

We denote by $I = [i_I, e_I]$ such a generalized interval. A path $\gamma : I \to \widehat{V} \times \Gamma_\infty^m$ is a continuous (pro)-definable map.

Let V be an algebraic variety over some valued field. We say a strict pro-definable subset Z of $\widehat{V}$ is definably connected if it contains no clopen strict pro-definable subsets other than $\emptyset$ and Z. We say that Z is definably path connected if for any two points a and b of Z there exists a definable path in Z connecting a and b. Clearly definable path connectedness implies definable connectedness. When V is quasi-projective and $Z = \widehat{X}$ with X a definable subset of V, the reverse implication will eventually follow from Theorem 4.32.

We have the following GAGA type theorem:

Theorem 4.28. *Let V be an algebraic variety over some valued field. Then $\widehat{V}$ is definably connected if and only if V is geometrically connected.*

4.29. Strong retractions for curves. Let $I = [i_I, e_I]$ be a generalized interval. A continuous pro-definable map $H : I \times \widehat{X} \to \widehat{Y}$ is called a definable homotopy between the maps $H_i = H_{|\{i_I\}\times\widehat{X}}$ and $H_e = H_{|\{e_I\}\times\widehat{X}}$, viewed as maps $\widehat{X} \to \widehat{Y}$. A definable homotopy $H : I \times \widehat{X} \to \widehat{X}$ is called a strong deformation retraction onto the set $\Sigma \subset \widehat{X}$ if $H_i = \mathrm{Id}_{\widehat{X}}$, $H(t, x) = x$ for every $t \in I$ and every $x \in \Sigma$ and $H_e(\widehat{X}) = \Sigma$.

There is a canonical strong deformation retraction of $\widehat{\mathbb{P}^1}$ onto the point $p_{\mathcal{O}}$ which is described as follows. Using the two standards affine charts, one may write each point of $\widehat{\mathbb{P}^1}$ as $p_{B(a,\alpha)}$ with $a \in \mathbb{P}^1(\mathbb{U})$ and $\alpha \geq 0$. The homotopy is given by taking $I = [\infty, 0]$ (thus $i_I = \infty$ and $e_I = 0$) and setting $\psi(t, p_{B(a,\alpha)}) = p_{B(a,\min(t,\alpha))}$.

More generally, given any finite subset D in $\mathbb{P}^1(\mathbb{U})$, let C_D be the image of $I \times (D \cup p_{\mathcal{O}})$ under ψ. The set C_D is a closed Γ-internal subset of $\widehat{\mathbb{P}^1}$. Set $\gamma(a) = \max\{t \in I; \psi(t, a) \in C_D\}$. Then $\psi_D : I \times \widehat{\mathbb{P}^1} \to \widehat{\mathbb{P}^1}$ sending (t, a) to $\psi(\max(\gamma(a), t), a)$ is a strong deformation retraction of $\widehat{\mathbb{P}^1}$ onto C_D.

Theorem 4.30. *Let C be an algebraic curve over a valued field K. There exists a strong deformation retraction, defined over K, $H : [0, \infty] \times \widehat{C} \to \widehat{C}$ onto a Γ-internal subset of $\widehat{C}$.*

Let us sketch the proof. A standard outward path on $\widehat{\mathbb{A}^1}$ at $x = p_{B(a,\alpha)}$ is given by $t \mapsto p_{B(a,t)}$ for $t \in (\beta, \alpha]$ for some $\beta < \alpha$. Now if $g : C \to \mathbb{A}^1$ is finite, with C a curve, by an outward path starting at $x \in \widehat{C}$, we mean a continuous definable lifting of a standard outward path starting at $g(x)$. One proves that for any $x \in \widehat{C}$ there exists at least one outward path starting at x and one says that x is branching if there is more than one outward path starting at x. A key lemma states that the number of such branching points is finite. For the proof of the theorem we may assume C is projective and consider $f : C \to \mathbb{P}^1$ finite and generically étale. One considers a finite set $D \subset \mathbb{P}^1$, defined over K, such that f is étale above the complement of D and C_D contains all the branching points, with respect to the restriction of g over both standards affine charts. One concludes the proof by showing that ψ_D lifts to the strong deformation retraction we are looking for.

4.31. The main theorem. We may now state the main result from [31]:

Theorem 4.32. *Let K be a valued field and $A = (K, G)$ with G a subset of Γ containing $\mathrm{val}(K)$. Let V a quasi-projective variety defined over K, X an A-definable subset of V. Assume given finitely many A-definable functions $\xi_i : X \to \Gamma_\infty$ and an action of a finite algebraic group over K on V leaving X globally invariant. Then there exists an A-definable strong deformation $H : I \times \widehat{X} \to \widehat{X}$ onto a Γ-internal subset Υ of $\widehat{X}$ such that:*

(1) *The set Υ embeds homeomorphically into Γ_∞^w for some finite A-definable set.*

(2) *H respects the functions ξ_i and is equivariant with respect to the group action.*

The structure of the proof goes as follows. One uses induction on the dimension of V. One start by reducing to the case where $X = V$ is projective equidimensional. One fixes an hypersurface $D_0 \subset V$ containing the singular locus of V and such that there exists an equivariant étale morphism $V \setminus D_0 \to \mathbb{A}^n$. Some further geometric considerations allow to reduce to the case when there is a morphism $u : V \to U = \mathbb{P}^{n-1}$, whose restriction to D_0 is finite, and a Zariski dense open subset U_0 of U such that, setting $V_0 = u^{-1}(U_0)$, $u|V_0$ factorizes as $q \circ f$ with $f : V_0 \to E_0 = U_0 \times \mathbb{P}^1$ a finite morphism and $q : E_0 \to U_0$ the projection.

Over U_0 the situation is that of a relative curve. Performing the curve construction in this relative setting provides a strong deformation retraction

$$H_{curves} : [0, \infty] \times \widehat{V_0 \cup D_0} \longrightarrow \widehat{V_0 \cup D_0}$$

fixing pointwise $\widehat{D_0}$ and with image a relatively Γ-internal set Υ_{curves}. By using the induction hypothesis (note that even if one starts with V without group action and no ξ_i's, they are needed at this stage of this induction), one constructs a definable homotopy $I \times \widehat{U} \to \widehat{U}$ whose restriction lifts to a strong deformation retraction

$$H_{base} : I \times \Upsilon_{curves} \longrightarrow \Upsilon_{curves}.$$

A third homotopy, which we call "inflation" is used to get out of of the complement of $\widehat{V_0 \cup D_0}$. On $\widehat{\mathbb{A}^n}$ one may consider the standard homotopy given by "increasing the polyradius". Using an appopriate stopping time function one gets another homotopy which we may lift, via the étale map $V \setminus D_0 \to \mathbb{A}^n$, to an homotopy

$$H_{inf} : [0, \infty] \times \widehat{V} \longrightarrow \widehat{V_0 \cup D_0}$$

fixing pointwise $\widehat{D_0}$.

After composing these three homotopies, one gets an homotopy $H' : I' \times \widehat{V} \to \widehat{V}$ that almost does the job, except that because of the use of inflation, we cannot insure that the points of the image of H' are all kept pointwise fixed by H' for all time values. To remedy this issue, we have to construct a fourth homotopy, H_Γ whose construction lies purely in the tropical Γ-internal world, so that the composition $H = H_\Gamma \circ H'$ finally satisfies the conclusion of the theorem.

4.33. Back to Berkovich spaces. A type $p = \mathrm{tp}(c/A)$ is said to be almost orthogonal to Γ is $\Gamma(Ac) = \Gamma(A)$.

Let F be a valued field with $\mathrm{val}(F^\times) \subset \mathbb{R}$. We consider the structure $\mathbb{F} = (F, \mathbb{R})$, where $\mathbb{R}$ belongs to the Γ-sort. Let V be a variety defined over F and X an $\mathbb{F}$-definable subset of V.

One defines $B_X(\mathbb{F})$ as the set of types over $\mathbb{F}$ lying on X and almost orthogonal to Γ. Similarly as for the Berkovich analytification and the stable completion, one endows $B_X(\mathbb{F})$ with a topology coming from the topology on $\mathbb{R}$. When F is complete, $B_V(\mathbb{F})$ and V^{an} are canonically homeomorphic.

By a result of Kaplansky, there exists a unique field F^{max}, up to $\mathbb{F}$-automorphism, which is a maximally complete algebraically closed non trivially valued field containing F, and has value group $\mathbb{R}$ and residue field the algebraic closure of the residue field of F.

The following proposition provides the link allowing to deduce the results about Berkovich spaces stated in Theorem 4.6 from Theorem 4.32 and its relative variants.

Proposition 4.34. *Let X be an $\mathbb{F}$-definable subset of some F-variety. Restriction of types induces a continuous, surjective and closed map $\pi : \widehat{X}(F^{max}) \to B_X(\mathbb{F})$.*

(1) *Let $f : \widehat{X} \to \widehat{Y}$ be a continuous $\mathbb{F}$-pro-definable map, with Y an $\mathbb{F}$-definable subset of some F-variety. Then there exists a unique continuous map $\tilde{f} : B_X(\mathbb{F}) \to B_Y(\mathbb{F})$ such that $\pi \circ f = \tilde{f} \circ \pi$.*

(2) *Let $H : I \times \widehat{X} \to \widehat{X}$ be a definable strong deformation retraction. Then $\tilde{H} : I(\mathbb{R}_\infty) \times B_X(\mathbb{F}) \to B_X(\mathbb{F})$ is a strong deformation retraction.*

(3) *$B_X(\mathbb{F})$ is compact if and only if $\widehat{X}$ is definably compact.*

Acknowledgements. The research of the author has been partially supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC Grant agreement no. 246903/NMNAG. The results presented in this survey are the outcome of long term collaborations with Raf Cluckers, Georges Comte and Ehud Hrushovski. The author would like to express them his heartfelt gratitude.

References

[1] A. Abbes and T. Saito, *Ramification of local fields with imperfect residue fields*, Amer. J. Math. **124** (2002), 879–920.

[2] J. Ax, S. Kochen, *Diophantine problems over local fields. I.*, Amer. J. Math. **87** (1965), 605–630.

[3] V.G. Berkovich, *Spectral theory and analytic geometry over non-Archimedean fields*, Mathematical Surveys and Monographs, 33. American Mathematical Society, Providence, RI, 1990.

[4] ———, *Smooth p-adic analytic spaces are locally contractible*, Invent. Math. **137** (1999), 1–84.

[5] R. Bieri and J. Groves, *The geometry of the set of characters induced by valuations* J. Reine Angew. Math. **347** (1984), 168–195.

[6] E. Bombieri, J. Pila, *The number of integral points on arcs and ovals*, Duke Math. J. **59** (1989), 337–357.

[7] R. Cluckers, G. Comte, and F. Loeser, *Lipschitz continuity properties for p-adic semi-algebraic and subanalytic functions*, Geom. Funct. Anal. **20** (2010), 68–87.

[8] ———, *Non-archimedean Yomdin-Gromov parametrizations and points of bounded height*, arXiv:1404.1952.

[9] R. Cluckers and I. Halupczok, *Approximations and Lipschitz continuity in p-adic semi-algebraic and subanalytic geometry*, Selecta Math. **18** (2012), 825–837.

[10] R. Cluckers and F. Loeser, *Constructible motivic functions and motivic integration*, Invent. Math., **173** (2008), 23–121.

[11] ———, *Constructible exponential functions, motivic Fourier transform and transfer principle*, Ann. Math. **171** (2010), 1011–1065.

[12] R. Cluckers, T. Hales, and F. Loeser, *Transfer Principle for the Fundamental Lemma*, in "On the Stabilization of the Trace Formula", edited by L. Clozel, M. Harris, J.-P. Labesse and B.-C. Ngô, International Press (2011), 309–347.

[13] R. Cluckers, J. Gordon, and I. Halupczok, *Integrability of oscillatory functions on local fields: transfer principles*, arXiv:1111.4405.

[14] ———, *Local integrability results in harmonic analysis on reductive groups in large positive characteristic*, arXiv:1111.7057.

[15] S. D. Cohen, *The distribution of Galois groups and Hilbert's irreducibility theorem*, Proc. London Math. Soc. **43** (1981), 227–250.

[16] J. Denef, *The rationality of the Poincaré series associated to the p-adic points on a variety*, Invent. Math., **77** (1984), 1–23.

[17] ———, *On the evaluation of certain p-adic integrals*, in Séminaire de théorie des nombres, Paris 1983–84, 25–47, Progr. Math., **59**, Birkhäuser Boston, Boston, MA, 1985.

[18] ———, *p-adic semi-algebraic sets and cell decomposition*, J. Reine Angew. Math., **369** (1986), 154–166.

[19] J. Denef, F. Loeser, *Germs of arcs on singular algebraic varieties and motivic integration*, Invent. Math. **135** (1999), 201–232.

[20] ———, *Definable sets, motives and p-adic integrals*, J. Amer. Math. Soc., **14** (2001), 429–469.

[21] A. Ducros, *Parties semi-algébriques d'une variété algébrique p-adique*, Manuscripta Math. **111** (2003), 513–528.

[22] ———, *Espaces analytiques p-adiques au sens de Berkovich*, Séminaire Bourbaki. Vol. 2005/2006. Astérisque **311** (2007), 137–176.

[23] ———, *Espaces de Berkovich, polytopes, squelettes et thÃl'orie des modèles*, Confluentes Math. **5** (2013), 57 pages.

[24] ———, *Les espaces de Berkovich sont modérés, d'après E. Hrushovski et F. Loeser*, arXiv:1210.4336.

[25] J. Eršov, *On the elementary theory of maximal normed fields*, Dokl. Akad. Nauk SSSR **165** (1965), 21–23.

[26] M. Gromov, *Entropy, homology and semialgebraic geometry*, Séminaire Bourbaki, vol. 1985/1986, Astérisque 145–146 (1987), 225–240.

[27] D. Haskell, E. Hrushovski, and D. Macpherson, *Definable sets in algebraically closed valued fields: elimination of imaginaries*, J. Reine Angew. Math. **597** (2006), 175–236.

[28] ———, *Stable domination and independence in algebraically closed valued fields*, Lecture Notes in Logic, 30. Association for Symbolic Logic, Chicago, IL; Cambridge University Press, Cambridge, 2008.

[29] E. Hrushovski and D. Kazhdan, *Integration in valued fields*, in Algebraic geometry and number theory, Progress in Mathematics 253, 261–405 (2006), Birkhäuser.

[30] ———, *The value ring of geometric motivic integration, and the Iwahori Hecke algebra of* SL_2. *With an appendix by Nir Avni*, Geom. Funct. Anal. **17** (2008), 1924–1967.

[31] E. Hrushovski and F. Loeser, *Non-archimedean tame topology and stably dominated types*, to appear in Annals of Mathematics Studies, arXiv:1009.0252.

[32] ———, *Monodromy and the Lefschetz fixed point formula*, arXiv:1111.1954.

[33] M. Kamensky, *Ind- and pro- definable sets*, Ann. Pure Appl. Logic **147** (2007),180–186.

[34] K. Kurdyka, *On a subanalytic stratification satisfying a Whitney property with exponent* 1. Real algebraic geometry (Rennes, 1991), 316–322, Lecture Notes in Math., 1524, Springer, Berlin, 1992.

[35] Q.T. Lê, *Proofs of the integral identity conjecture over algebraically closed fields*, arXiv:1206.5334.

[36] A. Macintyre, *On definable subsets of p-adic fields*, J. Symbolic Logic **41** (1976), 605–610.

[37] ———, *Rationality of* p*-adic Poincaré series: uniformity in* p, Ann. Pure Appl. Logic **49** (1990), 31–74.

[38] B. C. Ngô, *Le lemme fondamental pour les algèbres de Lie*, Publ. Math. Inst. Hautes Études Sci. **11** (2010), 1–169.

[39] J. Pas, *Uniform p-adic cell decomposition and local zeta functions*, J. Reine Angew. Math. **399** (1989), 137–172.

[40] J. Pila, *Density of integral and rational points on varieties*, Columbia University Number Theory Seminar (New York, 1992). Astérisque **228** (1995), 183–187.

[41] J. Pila and A. Wilkie, *The rational points of a definable set*, Duke Math. J. **133** (2006), 591–616.

[42] J. Pila, *o-minimality and the André-Oort conjecture for* $\mathbb{C}^n$, Annals of Math. **173** (201), 1779–1840.

[43] J. Poineau, *Un résultat de connexité pour les variétés analytiques p-adiques: privilège et noethérianité*, Compos. Math. **144** (2008), 107–133.

[44] S. Shin and N. Templier, *Sato-Tate theorem for families and low-lying zeros of automorphic L-functions* Appendix A by R. Kottwitz; Appendix B by R. Cluckers, J. Gordon and I. Halupczok, arXiv:1208.1945.

[45] L. van den Dries, *Tame topology and o-minimal structures*, Cambridge Univ. Press, New York, 1998.

[46] J.-L. Waldspurger, *Endoscopie et changement de caractéristique*, J. Inst. Math. Jussieu **5** (2006), 423–525.

[47] ———, *Endoscopie et changement de caractéristique: intégrales orbitales pondérées*, Ann. Inst. Fourier **59** (2009), 1753 – 1818.

[48] Y. Yomdin, *Volume growth and entropy*, Israel J. Math. **57** (1987), 285–300.

[49] ———, *$\mathcal{C}^k$-resolution of semialgebraic mappings. Addendum to: "Volume growth and entropy"*, Israel J. Math. **57** (1987), 301–317.

[50] Z. Yun, with appendix by J. Gordon, *The fundamental lemma of Jacquet-Rallis*, Duke Math. J., **156** (2011), 167–227.

Sorbonne Universités, UPMC Univ Paris 06, UMR 7586 CNRS, Institut Mathématique de Jussieu, F-75005, Paris, France

E-mail: Francois.Loeser@upmc.fr

Computability theoretic classifications for classes of structures

Antonio Montalbán

Abstract. In this paper, we survey recent work in the study of classes of structures from the viewpoint of computability theory. We consider different ways of classifying classes of structures in terms of their global properties, and see how those affect the structures inside the class. On one extreme, we have the classes that are Σ-small. These are the classes which realize only countably many $\exists$-types, and are characterized by having tame computability theoretic behavior. On the opposite end, we look at various notions of completeness for classes which imply that all possible behaviors occur among their structures. We introduce a new notion of completeness, that of being on top for effective-bi-interpretability, which is stronger and more structurally oriented than the previously proposed notions.

Mathematics Subject Classification (2010). Primary 03D45; Secondary 03C57.

Keywords. Sigma small classes, back-and-forth relations, rice relations, low property, bf-ordinal, effective-bi-interpretability.

1. Introduction

In this paper, we survey recent work on the study of classes of structures from the viewpoint of computability theory. By classes of structures we mean classes like the one of fields or of p-groups or of linear orderings. Our general objective is to consider global properties of the classes and derive properties about their individual structures.

Computable structure theory is an area inside computability theory and logic that is concerned with the computable aspects of mathematical objects and constructions. In particular, we are interested in the interplay between structure and complexity, or in other words, in understanding how the algebraic properties of a structure interact with its computational properties. For instance, we ask questions like the following: What kind of information can be encoded into an isomorphism type of a structure? How difficult is it to represent a certain structure? How difficult is it to recognize it?

When we consider classes of structures, there are two ends of the spectrum. On the one end are the classes which have some global property restricting the behavior of their structures. On the other end are the classes which are complete in the sense that they allow all possible behaviors to happen. Let us say a bit more about these two extremes.

Tame classes. In Section 2, we will review some concepts we will need later. One notion of simplicity is that of a class having a bound on the Scott rank of its structures. These classes are not necessarily that simple from a computational viewpoint, and much less from

Proceedings of the International Congress of Mathematicians, Seoul, 2014

the viewpoint of Borel equivalence relations. However, for natural classes this bound tends to be quite low, which makes them easier to analyze. The Scott rank of a structure is related to the number of Turing jumps necessary to fully understand it, and hence, the lower the Scott rank, the more manageable the structure.

A second notion of simplicity, one we believe is the most relevant to computability theory, is that of Σ-smallness, or actually effective Σ-smallness. This notion is studied in Section 3. The author started studying such classes in [46],[1] although the term "Σ-small" is new.

Definition 1.1. A class of structures $\mathbb{K}$ is Σ-*small* if it realizes countably many $\exists$-types, that is, if the set

$$\{\exists\text{-}tp_{\mathcal{A}}(\bar{a}) : \mathcal{A} \in \mathbb{K}, \bar{a} \in \mathcal{A}^{<\omega}\}$$

is countable, where

$$\exists\text{-}tp_{\mathcal{A}}(\bar{a}) = \{\varphi : \varphi(\bar{x}) \text{ is a first order existential formula with } \mathcal{A} \models \varphi(\bar{a})\}.$$

We remark that knowing the $\exists$-type of $\bar{a}$ is equivalent to knowing what finite substructures we can find in $\mathcal{A}$ extending $\bar{a}$. This is not entirely correct when $\mathcal{L}$ is infinite, where we need to consider sub-structures which only mention only a finite number of the symbols in $\mathcal{L}$. Another remark is that the types above are without parameters.

We start Section 3 by developing the effective version of this notion. The effectiveness assumption is not that strong, as it holds of all the examples we have analyzed. A large list of examples can be found in Subsection 3.1. We then study the role of Σ-small classes in many topics that have been widely studied in computable structure theory: Richter's extendibility condition, jumps of structures, the low property, the categoricity property and the Turing ordinal. As more evidence towards its naturalness, we will see in Theorems 3.14 and 3.15 how Σ-smallness induces a strong dichotomy on classes.

Complete classes. In Section 4, we review various ways of mapping structures from one class into another. For each of these reducibilities we have classes that are *on top* in the sense that all other classes can be reduced to it. We start with the well-known notion of Borel reducibility, and then move on to effective reducibility and Turing-computable reducibility, and to classes that are complete in the sense of Hirschfeldt, Khoussainov, Shore and Slinko.

In Section 5, we develop a new and stronger notion of reducibility based on the idea of *effective-bi-interpretability* between structures. We do not know that much about this new notion, but we do show that it preservers even more computational properties than all the previous reducibilities.

Properties on a cone. There are many properties in computability theory which tend to behave nicely when we have a nice natural class of structures, but that do not in general. One can often build strange and unnatural classes of structures where these nice behaviors do not occur. In this paper, we are interested in properties that hold of natural classes. Since we cannot quantify over "all natural classes," we often use the technical device of considering *properties on a cone*.

When we have a computability theoretic property P, we can often consider its relativization P^X for a given oracle $X \in 2^\omega$. We then consider the properties *relatively-P*, which

[1] In [46] we used the phrase "$\mathbb{K}$ has a computable 1-back-and-forth structure" for what we now say "$\mathbb{K}$ is effectively Σ-small."

means that P^X holds for all X, and P *on a cone*, which means that there is a $Z \in 2^\omega$ such that P^X holds for all $X \geq_T Z$. When we have a proof that a natural property P holds (or does not hold) when applied to a natural object, this proof almost always relativizes. Thus, we have a proof of relatively-P, and in particular, of P on a cone. So, if our objects are natural, we should not care whether we are using P, relatively-P, or P on a cone. However, due to the unnatural examples, many results can only be proved in general if we consider the properties *on a cone*. Such results usually call for a further analysis of its degree of effectiveness – we will not concentrate on this here.

Disclaimer. This paper does not pretend to be exhaustive. What it attempts is to convey the author's viewpoint, unifying many ideas that have been floating around for a while. The choice of topics and how much attention they receive is purely motivated by the author's taste, the author's own work, and the new ideas the author wants to develop.

Background and notation. We only consider countable structures throughout, so "structure" means "countable structure." We only consider relational languages, as we do not lose any generality for our purposes. The languages we consider are all computable: that is, if $\mathcal{L}$ consists of relations R_i for $i \in I$, where $I \subseteq \omega$ and R_i has arity $a(i)$, the function $a\colon I \to \omega$ is computable. (This only matters when $\mathcal{L}$ is infinite.)

A *presentation* of a structure $\mathcal{A}$, or a *copy* of $\mathcal{A}$, is just a structure $\mathcal{B}$ isomorphic to $\mathcal{A}$ whose domain is a subset of ω. This allows us to use everything we know about computable functions on ω to study $\mathcal{B}$. Given a presentation $\mathcal{A} = (A; R_i^{\mathcal{A}}, i \in I)$, with $A \subseteq \omega$, we let

$$D(\mathcal{A}) \;=\; A \oplus \bigoplus_{i\in I} R_i^{\mathcal{A}} \;\subseteq\; \omega \sqcup \bigsqcup_{i\in I} \omega^{a(i)}.$$

Via standard coding, we then think of $D(\mathcal{A})$ as a subset of ω, or equivalently a sequence in 2^ω. Note that $D(\mathcal{A})$ is essentially the atomic diagram of $\mathcal{A}$. When we say that the presentation $\mathcal{A}$ computes X, or is computable in Y, we mean that $D(\mathcal{A})$ computes X or is computable in Y. By a *class of $\mathcal{L}$-structures* we mean a set $\mathbb{K}$ of presentations of $\mathcal{L}$-structures which is closed under isomorphism. We often think of $\mathbb{K}$ and of $\{D(\mathcal{A}) : \mathcal{A} \in \mathbb{K}\} \subseteq 2^\omega$ as the same thing, and hence treat $\mathbb{K}$ as a class of reals.

We will often consider the infinity language $\mathcal{L}_{\omega_1,\omega}$, where countably infinite conjunctions and disjunctions are allowed, and its computable version, where these conjunctions and disjunctions must be computable. See [1, Sections 6 and 7]. We use $\Sigma^{\mathtt{in}}_\alpha$ to denote the infinitary Σ_α formulas and $\Sigma^{\mathtt{c}}_\alpha$ to denote the computably infinitary Σ_α formulas. For the *Scott rank* of a structure $\mathcal{A}$ we use the following definition: $SR(\mathcal{A})$ is the least α such that every automorphism orbit in $\mathcal{A}$ is $\Sigma^{\mathtt{in}}_\alpha$-definable without parameters. There are various definitions of Scott rank in the literature that give slightly different values (see [1, Section 6.7]). The reason we prefer ours is that it matches better with other complexity measures used in computability theory and descriptive set theory (see [43]).

2. Axiomatization and the isomorphism problem

The first measure of the complexity for a class is in terms of its complexity as a set of reals. This is directly connected with the complexity of the class in terms of its axiomatizations:

Theorem 2.1 (Lopez-Escobar [37]). *Let $\mathbb{K} \subseteq 2^\omega$ be a class of presentations of structures closed under isomorphisms. Then $\mathbb{K}$ is $\mathbf{\Sigma}^0_\alpha$ in the Borel hierarchy if and only if $\mathbb{K}$ is axiomatizable by an infinitary $\Sigma^{\mathtt{in}}_\alpha$ sentence.*

The lightface version of this theorem is also true: $\mathbb{K}$ is lightface Σ^0_α if and only if it is axiomatizable by a computably infinitary $\Sigma^{\mathtt{c}}_\alpha$ formula [61].

Not all nice classes of structures are $\mathcal{L}_{\omega_1,\omega}$-axiomatizable, or equivalently Borel, as for instance the class of ordinals, which is Π^1_1-complete. We, however, are mostly interested in $\mathcal{L}_{\omega_1,\omega}$-axiomatizable classes. Most of the natural classes we consider are actually $\Pi^{\mathtt{c}}_2$-axiomatizable, so we will sometimes make this assumption when we prove general results.

The second measure of complexity is the difficulty in telling apart different structures in $\mathbb{K}$. This is captured by the set

$$\{\langle D(\mathcal{A}), D(\mathcal{B})\rangle : \mathcal{A}, \mathcal{B} \in \mathbb{K}, \mathcal{A} \cong \mathcal{B}\} \quad \subseteq \quad (2^\omega)^2,$$

usually called the *isomorphism problem* for $\mathbb{K}$. For a Borel class of structures, this set is $\mathbf{\Sigma}^1_1$. For some classes, like linear orderings, this problem is Σ^1_1-complete. For other classes, this problem is quite simple, like $\mathbb{Q}$-vector spaces for which it is Π^0_3-complete. If we assume $\mathsf{ZFC}+\forall X\ (X^\sharp \text{ exists})$, then non-Borel isomorphism problem must be $\mathbf{\Sigma}^1_1$-complete. This follows from Wadge's theorem (see [52, Lemma 7D.3]), as every set that is not $\mathbf{\Pi}^1_1$, is $\mathbf{\Sigma}^1_1$-hard.

Theorem 2.2 ([6, Corollary 7.14]). *Let $\mathbb{K}$ be an $\mathcal{L}_{\omega_1,\omega}$-axiomatizable class. The following are equivalent:*

(1) *The isomorphism problem for $\mathbb{K}$ is Borel.*

(2) *$\mathbb{K}$ has bounded Scott rank.*

When we say that $\mathbb{K}$ has *bounded Scott rank* we mean $\{SR(\mathcal{A}) : \mathcal{A} \in \mathbb{K}\}$ has a supremum $\beta < \omega_1$. For example, the classes of $\mathbb{Q}$-vector spaces and of algebraically closed fields have bound 2 on their Scott ranks. The classes of equivalence structures and of torsion free abelian groups of finite rank have bound 3.

A simple remark is that if $\mathbb{K}$ has countably many structures (up to isomorphism, of course), it has bounded Scott rank. It follows from the model-theoretic Martin's conjecture that if $\mathbb{K}$ has a first order axiomatization and countably many structures, then $\omega+\omega$ is a bound for the Scott ranks of the structures in $\mathbb{K}$ (see [22] for the statement of Martin's conjecture). For more on how high this bound can be see [40, 57].

3. Σ-Small classes of structures

In this section, we see how the notion of Σ-small class connects with a lot of well-known concepts in computable structure theory.

Before looking at examples among familiar classes, let us introduce the effective version of this definition. If $\mathbb{K}$ is a natural class of structures and it is Σ-small, we have a natural countable collection of $\exists$-types. It is then reasonable to expect that one can list, compare and manipulate these types. An effectively Σ-small class is one where we can do this computably:

Definition 3.1. A Σ-small class $\mathbb{K}$ is *effectively* Σ*-small* if there is a computable list $\{p_i : i \in \omega\}$ of computable $\exists$-types listing all the $\exists$-types realized in $\mathbb{K}$ without repetitions, where the operations of erasing and permuting variables are computable, and deciding inclusion of $\exists$-types is also computable.[2]

3.1. Examples. All the classes of structures below are Σ-small. It is worth remarking that most of the examples mentioned below have been attractive to computability theorists for a long time because they enjoy nice computability properties other classes do not.

Vector spaces (over a fixed computable field F). If $\mathbb{K}$ has only countably many structures, it is clearly Σ-small. Proving that F-vector spaces are effectively Σ-small requires understanding their $\exists$-types, which is not hard to do.

Algebraically closed fields. Same as above.

Differentially closed fields of characteristic 0 (DCF_0). They are Σ-small because they are ω-stable. The class of models of an ω-stable theory is always Σ-small, as even using countably many parameters and full first-order types, there are still countably many types. It has not been verified whether DCF_0 is effectively Σ-small or not.

Abelian p-groups. That Abelian p-groups are effectively Σ-small follows from work of Khisamiev [32].

Equivalence structures. We refer the reader to [46, Section 4.2] for an analysis of the $\exists$-types on equivalence structures.

Trees (as partial orderings). By 'trees' we mean downward closed subsets of $\omega^{<\omega}$. That they are effectively Σ-small in the language of partial orderings follows from Richter's work [56]. Let us remark that a key tool in her proof is Kruskal's theorem [35] on the well-quasi-ordering of finite trees.

Trees of finite height (as graphs). The proof is like the case above using Kruskal's theorem for finite trees of a fixed height.

Linear orderings. All an $\exists$-type can say about a tuple $\bar{a} = \langle a_0, ..., a_{k-1} \rangle$ is the order among the elements of the tuple and, for each $n \in \omega$ and each $i, j < |\bar{a}|$, whether there are at least n elements between a_i and a_j. Thus, existential types are determined by the number of elements between the elements of the tuple, and hence there are countably many of them. One can also use this to prove they are effectively Σ-small.

Linear orderings with an added relation for adjacency. When we add the adjacency relation, the $\exists$-types get a bit more complicated, but they are still effective and countable (see [46, Section 4.1]).

Boolean algebras. All an $\exists$-type can say about a tuple is how many elements are below each Boolean combination of the elements of the tuple. These are, again, not that difficult to analyze.

[2]The exact list of properties that are required for a class to be *effectively* Σ-small is currently work in progress, and so far they are motivated from what we see in applications.

Boolean algebras with an added relation that identifies atoms. What makes Boolean algebras particularly interesting is that they remain Σ-small even if we add to them any $\Sigma^{\mathtt{in}}_{<\omega}$ relation. For instance, we can add all of the relations used by Knight and Sob [36] (atom, atomless, infinite, atomic, 1-atomic, atominf, $\sim$-inf, Int$(\omega+\eta)$, infatomicless, 1-atomless, and nomaxatomless) and they remain effectively Σ-small. An in-depth analysis of the $\Sigma^{\mathtt{in}}_n$-types of Boolean algebras was done by Harris and Montalbán in [29]. The fact that there are countably many of them uses key ideas from work of Flum and Ziegler [21].

Generalized Boolean algebras. These are distributive lattices with 0 and where every interval $[a,b]$ is a Boolean algebra. They are usually known in Russia as Ershov algebras. That they are effectively Σ-small follows from work of Khisamiev [32].

3.2. Richter's computable extendibility condition. In her Ph.D. thesis[3] [55], Linda Richter introduced the computable extendibility condition in order to show that there are structures that do not have Turing degree as defined by Jockusch. (A structure $\mathcal{A}$ has Turing degree $\mathbf{x}$ if $\mathbf{x}$ computes a copy of $\mathcal{A}$, and every copy of $\mathcal{A}$ computes $\mathbf{x}$.)

Definition 3.2 ([56, Section 3]). A structure $\mathcal{A}$ has the *computable extendibility condition* if each $\exists$-type realized in $\mathcal{A}$ is computable. A structure $\mathcal{A}$ has the *c.e. extendibility condition* if each $\exists$-type realized in $\mathcal{A}$ is c.e.

Richter's original definition was not in terms of types but in terms of finite structures extending a fixed tuple. As we mentioned right after Definition 1.1, these formulations are equivalent. The c.e. extendibility condition was not considered by Richter, but we include it here because it makes Theorem 3.3 below more rounded. In Russia, structures with the c.e. extendibility condition are said to be *locally constructivizable*.

Of course, if $\mathbb{K}$ is effectively Σ-small, then every structure $\mathcal{A}$ in $\mathbb{K}$ satisfies the computable extendibility condition. On the other hand, if every structure in $\mathbb{K}$ satisfies the c.e. extendibility condition, then $\mathbb{K}$ is Σ-small, because there are only countably many c.e. sets. Furthermore, the proofs in the literature that linear orderings, Boolean algebras and trees (as posets) in [55, 56] and p-groups and generalized Boolean algebras in [32] satisfy the computable extendibility condition are essentially proofs that these classes are effectively Σ-small.

The reason Richter introduced this notion is to prove the following theorem and its corollary below.

Theorem 3.3 (Essentially Richter). *Let $\mathcal{A}$ be any structure. The following are equivalent:*

(1) *$\mathcal{A}$ has the c.e. extendibility condition.*

(2) *Every set $X \subseteq \omega$ which is c.e. in every presentation of $\mathcal{A}$ is already c.e.*

Proof. That (1) implies (2) is essentially the same proof as [56, Theorem 3.1]. For the other direction, notice that every $\exists$-type realized in $\mathcal{A}$ is c.e. in every presentation of $\mathcal{A}$. □

Corollary 3.4. *If $\mathcal{A}$ has the c.e. extendibility condition and has Turing degree $\mathbf{x}$, then $\mathbf{x} = \mathbf{0}$.*

We can read Theorem 3.3 as saying that structures in an effectively Σ-small class cannot directly encode any non-trivial information. In a sense, Σ-smallness is not only a sufficient,

[3] directed by Carl Jockusch

but also a necessary condition for this to be the case. The following theorem shows that if $\mathbb{K}$ is not Σ-small, quite the opposite happens: every real is coded by some structure in $\mathbb{K}$ in a left-c.e. way. We recall that $X \in 2^\omega$ is *left-c.e. in* $A \in 2^\omega$ if the set $\{\sigma \in 2^{<\omega} : \sigma \leq_{lex} X\}$ is c.e. in A, or equivalently, if there is an A-computable approximation to X from the left.

Theorem 3.5 ([46], Theorem 3.1). *Let $\mathbb{K}$ be an $\mathcal{L}_{\omega_1,\omega}$-axiomatizable class of structures. The following are equivalent:*

(1) *$\mathbb{K}$ is not Σ-small.*

(2) *Relative to every oracle on a cone, the following holds: For every $Y \in 2^\omega$, there is a structure $\mathcal{A} \in \mathbb{K}$ such that Y is left-c.e. in every copy of $\mathcal{A}$.*

Part (2) of the theorem can be strengthened by changing "left-c.e." to just "c.e." in most cases. However, an example where left-c.e.ness is required is constructed in [45, Section 2.2].

3.3. Complete sets of r.i.c.e. relations. Another advantage of Σ-small classes is that they have nice structural jumps. We will see below how, when we have a Σ-small class $\mathbb{K}$, we usually have a nice and simple set of relations that give us all the structural information about the jump of the structures in $\mathbb{K}$. Understanding these complete sets of relations is usually very useful in applications.

Before talking about the jump, we need a notion of c.e.-ness among the relations on a structure. We will then look at complete relations among these and use them to define the jump of a structure. An equivalent notion of jump for a structure was originally defined by I. Soskov [4], although he used a very different format. The definition we use here is in the spirit of that introduced in [44] (see [47, Definition 1.2] for more historical remarks).

Definition 3.6. A relation $R \subseteq \mathcal{A}^{<\omega}$ is *relatively intrinsically computably enumerable (r.i.c.e.)* if, on every copy $(\mathcal{B}, R^{\mathcal{B}})$ of $(\mathcal{A}, R)$, we have that $R^{\mathcal{B}}$ (viewed as a subset of $\omega^{<\omega}$) is c.e. in $D(\mathcal{B})$.

Example 3.7. Over a $\mathbb{Q}$-vector space, the relation of linear dependence is r.i.c.e.; Over a ring, the relation that holds of $(r_0, ..., r_k)$ if the polynomial $r_0 + r_1 x + ... + r_k x^k$ has a root is r.i.c.e.

This definition gives a notion of c.e.-ness that we can use to define other standard concepts from computability theory on the subsets of $\mathcal{A}^{<\omega}$.

Definition 3.8. A relation $R \subseteq \mathcal{A}^{<\omega}$ is *relatively intrinsically (r.i.) computable* if it and its complement are both r.i.c.e. R is *r.i. computable in* $Q \subseteq \mathcal{A}^{<\omega}$ if R is r.i. computable in $(\mathcal{A}, Q)$. A partial function $f : \mathcal{A}^{<\omega} \rightharpoonup \mathcal{A}^{<\omega}$ is *partial r.i. computable* if its graph is a r.i.c.e. subset of $(\mathcal{A}^{<\omega})^2$.

Remark 3.9. The use of subsets of $\mathcal{A}^{<\omega}$ not only allows us to consider sequences of subsets of $\mathcal{A}^n$ for all n uniformly, but essentially all finite objects that can be built over $\mathcal{A}$. For instance, given $Q \subseteq (\mathcal{A}^{<\omega})^2$, let us define $R \subseteq \mathcal{A}^{<\omega}$ by $\bar{b} \in R$ iff $|\bar{b}| = \langle n, m\rangle$ for some $n, m \in \omega$ and $((b_0, ..., b_{n-1}), (b_n, ..., b_{n+m-1})) \in Q$. We then have that Q is r.i.c.e. (as in Definition 3.6 but for subsets of $(\mathcal{A}^{<\omega})^2$) if and only if R is r.i.c.e. In a similar way, we can code subsets of $(\mathcal{A}^{<\omega})^{<\omega}$ by subsets of $\mathcal{A}^{<\omega}$. Given $R, Q \subseteq \mathcal{A}^{<\omega}$, we define $R \oplus Q$ by $\bar{b} \in R \oplus Q$ if either $|\bar{b}| = 2n$ and $\bar{b} \upharpoonright n \in R$ or $|\bar{b}| = 2n + 1$ and $\bar{b} \upharpoonright n \in Q$. We can also

encode a set $X \subseteq \omega$ by a set $\vec{X} \subset \mathcal{A}^{<\omega}$ by letting $\bar{b} \in \vec{X}$ if and only if $|\bar{b}| \in X$. With a bit more work, we can encode any subset of $HF(\mathcal{A})$ (the hereditarily finite extension of $\mathcal{A}$) as a subset of $\mathcal{A}^{<\omega}$ (see [47, Section]).

R.i.c.e. relations can be characterized in a purely syntactic way, without referring to the different copies of the structure, using $\Sigma^{\mathtt{c}}_1$ formulas. Recall that a $\Sigma^{\mathtt{c}}$ formula is just a computable disjunction of $\exists$-formulas over a finite set of free variables.

Theorem 3.10 (Ash, Knight, Manasse, Slaman [2]; Chisholm [10]). *Let $\mathcal{A}$ be a structure, and $R \subseteq A^{<\omega}$ a relation on it. The following are equivalent:*

(1) *R is r.i.c.e.*

(2) *R is uniformly definable by $\Sigma^{\mathtt{c}}_1$ formulas with parameters from $\mathcal{A}$. That is, there is a tuple $\bar{a} \in \mathcal{A}^{<\omega}$ and a computable sequence of $\Sigma^{\mathtt{c}}_1$ formulas $\varphi_i(x_1, ..., x_{|\bar{a}|}, y_1, ..., y_i)$, for $i \in \omega$, such that*

$$(\forall \bar{b} \in \mathcal{A}^{<\omega})\ \ \bar{b} \in R \iff \mathcal{A} \models \varphi_{|\bar{b}|}(\bar{a}, \bar{b}).$$

Definition 3.11. A relation $R \subseteq \mathcal{A}^{<\omega}$ is *r.i.c.e. complete in $\mathcal{A}$* if it is r.i.c.e. and every other r.i.c.e. relation $Q \subseteq \mathcal{A}^{<\omega}$ is r.i. computable from R.

R.i.c.e. complete relations always exist. For instance, we can consider the analog of Kleene's predicate K: If we let $\varphi_{i,j}(y_1, ..., y_j)$ be the ith $\Sigma^{\mathtt{c}}_1$ formula of arity j, then the relation $\vec{K}^{\mathcal{A}} \subseteq \mathcal{A}^{<\omega} \times \omega$ defined by

$$(\bar{b}, i) \in \vec{K}^{\mathcal{A}} \iff \mathcal{A} \models \varphi_{i,|\bar{b}|}(\bar{b})$$

is r.i.c.e. complete.

Definition 3.12 ([47]). We define *the jump of $\mathcal{A}$* to be the structure $\mathcal{A}' = (\mathcal{A}, \vec{K}^{\mathcal{A}})$. Given a class $\mathbb{K}$, we let $\mathbb{K}' = \{\mathcal{A}' : \mathcal{A} \in \mathbb{K}\}$.

If $X \subseteq \omega$ is a c.e. set, then $\vec{X} \subseteq \mathcal{A}^{<\omega}$ (as in Remark 3.9) is clearly r.i.c.e. It follows that $\vec{0}'$ must be r.i. computable in every r.i.c.e. complete relation. On some structures, like $(\omega; 0, 1, +)$, $\vec{0}'$ is r.i.c.e. complete, but this is always not the case. If $\mathcal{A}$ is a linear ordering, then $co\text{-}Adj \oplus \vec{0}'$ is r.i.c.e. complete, where $co\text{-}Adj$ is the complement of the adjacency relation. For most linear orderings $co\text{-}Adj$ is not r.i. computable from $\vec{0}'$.

Definition 3.13. We say that R is *structurally r.i.c.e. complete* if $R \oplus \vec{0}'$ is r.i.c.e. complete. We then say that $(\mathcal{A}, R)$ is *a structural jump* of $\mathcal{A}$.

So, for a linear ordering L, $(\mathcal{L}, co\text{-}Adj)$ is a structural jump.

The author showed in [46] that if $\mathbb{K}$ is Σ-small, there is a countable sequence of $\Sigma^{\mathtt{in}}_1$ formulas which define a structurally complete r.i.c.e. relation in all the structures in $\mathbb{K}$ relative to every oracle on a cone.

Theorem 3.14 ([46]). *Let $\mathbb{K}$ be effectively Σ-small, and let $\{p_i : i \in \omega\}$ be a computable list of all the $\exists$-types realized in $\mathbb{K}$. Then, the $\Sigma^{\mathtt{c}}_1$ formulas*

$$\varphi_i \equiv \bigvee \{\psi : \psi \text{ is an } \exists\text{-formula, and } \psi \notin p_i\},$$

for $i \in \omega$ define a structurally r.i.c.e. complete relation on all structures in $\mathbb{K}$.

In most natural examples, we can find simpler structurally r.i.c.e. complete relations than the one given by Theorem 3.14. For instance, on $\mathbb{Q}$-vector spaces and algebraically closed fields, the relations of linear dependence and algebraic dependence are structurally r.i.c.e. complete, and on Boolean algebras, the not-atom relation is structurally r.i.c.e. complete.

It is also shown in [46] that this is not the case when $\mathbb{K}$ is not Σ-small: there is no sequence of formulas which works for all structures simultaneously.

Theorem 3.15. *If $\mathbb{K}$ is not Σ-small, there is no computable sequence of Σ^c_1-formulas defining a structurally r.i.c.e. complete relation simultaneously on all structures in $\mathbb{K}$.*

3.4. The low property. The low property has been studied for various classes in the last couple of decades. Only recently has it been looked at a general setting.

Definition 3.16. A class $\mathbb{K}$ has the *low property* if every low presentation $\mathcal{A} \in \mathbb{K}$ has a computable copy.

We recall that a set $X \subseteq \omega$ is *low* if X' is computable from $0'$. A presentation $\mathcal{A}$ is *low* if $D(\mathcal{A})$ is.

Jockusch and Soare [30] proved that the class of linear orderings does not have the low property, that is, that there is a low linear ordering without a computable copy. Downey and Jockusch [13] proved that the class of Boolean algebras has the low property. In that paper, they asked the following question, that is still open despite the efforts of various researches:

Question 1. Does every low_n Boolean algebras have a computable copy?

Some partial results are known. Thurber [59] proved that Boolean algebras have the low_2 property and Knight and Stob [36] the low_4 property. The low_5 property is still open. Harris and Montalbán showed that the difficulty at level 5 is not just that it needs one more jump, but a qualitatively new behavior: to show this behavioral difference is essential, they produced a low_5 Boolean algebra not $0^{(7)}$-isomorphic to any computable one – for $n = 1, 2, 3, 4$, it was known that every low_n Boolean algebra is $0^{(n+2)}$-isomorphic to a computable one.

Let us review some of the other examples. The class of equivalence structures does not have the low property. However, the class of equivalence structures with infinitely many infinite equivalence classes has the low property, as it follows from [7, Lemmas 2.2.(c) and 2.3]. Even if linear orderings do not have the low property, some sub-classes do. For instance, the class of all linear orderings where all elements have successor and predecessors does. More examples can be found in [3, 18, 19]. The class of linear orderings with only finitely many descending sequences (up to equivalence) was proved to have the low_n property for all n by Kach and Montalbán [33] (where two sequences are equivalent if they determine the same cut). It is open whether scattered linear orderings have the low property. The class of ordinals not only has the low property, but the low_α-property for all computable ordinals α. Classes that have the low_α-property for all $\alpha < \omega_1^{CK}$ are said to satisfy *hyperarithmetic-is-recursive*, that is, every hyperarithmetic structure in the class has a computable copy. We will get back to these classes in Theorem 3.26.

All the examples of classes with the low property we know are Σ-small. This happens for a reason:

Theorem 3.17. *Let $\mathbb{K}$ be a Π^{in}_2-class. If $\mathbb{K}$ has the low property on a cone, then $\mathbb{K}$ is Σ-small.*

Sketch of the proof. It follows from the author's construction in [46, Lemma 2.9 and Theorem 3.1] that if $\mathbb{K}$ is not Σ-small, there is an oracle relative to which the following happens:

For every $X \in 2^\omega$ and Y c.e. in X, there is a structure $\mathcal{A}$ in $\mathbb{K}$ computable from X and such that Y is left-c.e. in every copy of $\mathcal{A}$. All we need to do now is observe that there is a set that is c.e. over a low set but not left-c.e. For instance, Chaitin's Ω relativized to low 1-random real R is c.e. in R and, since it is 2-random, is not of c.e. degree. Thus $\{\sigma \in 2^{<\omega} : \sigma \leq_{lex} \Omega^R\}$ is c.e. in R but not left-c.e (because left-c.e. sets have c.e. degree). □

3.5. Listable classes. The author started looking at this property with the intention of characterizing the low property.

Definition 3.18. A class $\mathbb{K}$ is *listable* if there exists a Turing functional which, for every oracle X, produces an X-computable sequence of structures listing all the X-computable structures in $\mathbb{K}$ (allowing repetitions).

This definition appeared first in [49], but the underlying idea of considering classes whose computable models can be listed computably is much older. However, the uniformity in Definition 3.18 is needed to get the consequences we want. Nurtazin [54], almost four decades ago, gave a sufficient condition for a class of structures to be listable which includes the classes of linear orderings, Boolean algebras, equivalence structures, Abelian p-groups, and algebraic fields of characteristic p. Nurtazin's result says that, if there exists a computable structure in the class such that any other structure can be embedded into it, and such that any subset of that structure generates a structure in the class, then the class is listable (see [24, Theorem 5.1]). Nurtazin's condition is not a necessary condition for a class to be listable, and for many of the cases we are interested in, it is too strong.

The more general way of proving that a class is listable is by a priority argument, where one monitors all computable functions and tries to list the ones that code structures in the class. In [49], the author developed a game, $\mathsf{G}^\infty(\mathbb{K})$, that captures the combinatorial argument behind these constructions. This game is played by two players, C and D. Along the game, player C builds an infinite list of structures in $\mathbb{K}$ via finite approximations, adding a new element to each structure in each move. Player D, however, builds only one structure and he is allowed to wait and only add elements to his structure when he thinks it is worth it. Player C's goal is to get one of his structures to be isomorphic to D's structure (i.e. "to copy"), while D's goal is to diagonalize. See [49] for a more detailed definition and for other related games. The classes $\mathbb{K}$ for which C has a winning strategy are said to be ∞-*copyable*.

The connection between listability, the low property, and the games was unexpected.

Theorem 3.19 ([49]). *Let $\mathbb{K}$ be a $\Pi^{\mathtt{in}}_2$ class of structures. The following are equivalent:*

(1) *$\mathbb{K}$ has the low property on a cone.*

(2) *$\mathbb{K}$ is Σ-small and $\mathbb{K}'$ is listable on a cone.*

(3) *$\mathbb{K}$ is Σ-small and $\mathbb{K}'$ is ∞-copyable.*

3.6. Computable categoricity. The objective of this subsection is to argue that computable categoricity is easier to analyze on Σ-small classes.

The notion of computable categoricity has been studied intensively for the past few decades. A feature of computable structure theory is that computational properties of presentations need not be invariant under isomorphism, and instead they are invariant under computable isomorphisms. In other words, a structure can have two isomorphic computable

presentations which have different computational properties. For instance, there are computable presentations of the countable, infinite-dimensional $\mathbb{Q}$-vector space, $\mathbb{Q}^\infty$, where all the finite-dimensional subspaces are computable, and computable presentations of $\mathbb{Q}^\infty$ where no finite-dimensional subspace is computable (see [12]). The computably categorical structures are exactly the ones where this does not happen:

Definition 3.20. A computable structure $\mathcal{A}$ is *computably categorical* if between any two computable copies of $\mathcal{A}$ there is a computable isomorphism.

There are many results classifying the computably categorical structures within certain classes. A linear order is computably categorical if and only if it has finitely many adjacencies (Dzgoev and Goncharov [23]); a Boolean algebra is computably categorical if and only if it has finitely many atoms (Goncharov, and independently La Roche [39]); a $\mathbb{Q}$-vector space is computably categorical if and only if it has finite dimension; a p-group is computably categorical if and only if it can be written in one of the following forms: (i) $(\mathbb{Z}(p^\infty))^\ell \oplus G$ for $\ell \in \omega \cup \{\infty\}$ and G finite, or (ii) $(\mathbb{Z}(p^\infty))^n \oplus (\mathbb{Z}_{p^k})^\infty \oplus G$ where G is finite, and $n, k \in \omega$ (Goncharov [26] and Smith [58]); a tree of finite height is computably categorical if and only if it is of finite type (Lempp, McCoy, R. Miller, and Solomon [38]); and so on.

There are also many classes where computably categoricity is quite difficult to describe. Indeed, it was recently proved by Downey, Kach, Lempp, Lewis, Montalbán and Turetsky [14] that the index set of computable categorical structures is Π^1_1-complete. The relativized notion is, however, much better behaved and usually easier to characterize (recall the notions of "relatively P" and "P on a cone" from Section 1). A nice characterization of the relatively computably categorical structures was given by Goncharov [25]: they are exactly the atomic models, over a finite set of parameters, where all the types are generated by $\exists$-formulas, and there is a c.e. listing of those formulas. The author [43] has recently found that there is an even nicer characterization of the structures which are computably categorical on a cone. They are exactly the ones that have a $\Sigma^{\mathtt{in}}_3$ Scott sentence. The classes where we have the best hope of characterizing computable categoricity are the ones where the three notions of computable categoricity – plain, relative, and on a cone – coincide.

Definition 3.21. We say that $\mathbb{K}$ has the *categoricity property* if every computably categorical structure in $\mathbb{K}$ is relatively computably categorical.

$\mathbb{Q}$-vectors spaces, algebraically closed fields, Boolean algebras, linear orderings, equivalence structures, trees (as posets), ordered abelian groups, and p-groups all have the categoricity property.

Conjecture 1. Every Σ-small $\Pi^{\mathtt{in}}_2$-class $\mathbb{K}$ satisfies the categoricity property on a cone.

One thing we know is that Σ-small $\Pi^{\mathtt{in}}_2$-classes always contain structures which are categorical on a cone [43].

3.7. The back-and-forth ordinal. It is not hard to observe that $\Sigma^{\mathtt{c}}_2$ types over a structure $\mathcal{A}$ are equivalent to $\exists$-types over $\mathcal{A}'$, and $\Sigma^{\mathtt{c}}_3$-types to $\exists$-types over $\mathcal{A}''$. One can use this to get, for instance, that a structure is Δ^0_2-categorical on a cone if and only if it is $\mathcal{A}'$ is computably categorical on a cone; or that a a class $\mathbb{K}$ has the low_2 property on a cone if and only if both $\mathbb{K}$ and $\mathbb{K}'$ have the low property on a cone. Thus, understanding $\mathbb{K}'$ can be helpful for the understanding of these higher-level properties. If we have an effectively Σ-small class $\mathbb{K}$, we

have a nice and uniform notion of jump among the structures in $\mathbb{K}$ (Theorem 3.14). We then want to know if $\mathbb{K}'$ is effectively Σ-small. If it is, we can then consider $\mathbb{K}''$ and ask if it is Σ-small.

Definition 3.22. $\mathbb{K}$ is $\Sigma^{\mathtt{in}}_\alpha$-small if it realizes countably many $\Sigma^{\mathtt{in}}_\alpha$ types.

The effective notion of $\Sigma^{\mathtt{in}}_n$-smallness is considered in [46, 48] under the name "effective n-back-and-forth structure." We omit the definition here. In [46], we proved that on $\Sigma^{\mathtt{in}}_n$-small classes have nice $\Sigma^{\mathtt{c}}_n$-complete relations and that no set can be coded in the $(n-1)$st jump of their structures. An opposite behaviour happens, on a cone, for structures that are not $\Sigma^{\mathtt{in}}_n$-small. Thus, for a class $\mathbb{K}$, there is a qualitative jump in behavior from the α's at which $\mathbb{K}$ is $\Sigma^{\mathtt{in}}_\alpha$-small to the ones where it is not.

Definition 3.23. The *bf-ordinal* of a class of structures $\mathbb{K}$ is the least $\alpha \in \omega_1 + 1$ such that $\mathbb{K}$ is not $\Sigma^{\mathtt{in}}_\alpha$-small, and we let it be ∞ if there is no such α.

The notion of bf-ordinal is quite close to the notion of "Turing-ordinal" introduced by Jockusch and Soare [31].

Definition 3.24. A structure $\mathcal{A}$ has *βth jump Turing degree* $\mathbf{x}$ if the βth jump of every copy of $\mathcal{A}$ computes $\mathbf{x}$, and $\mathbf{x}$ computes the βth jump of some copy of $\mathcal{A}$.

A class $\mathbb{K}$ has *Turing-ordinal* τ if for all $\beta < \tau$, whenever a structure in $\mathbb{K}$ has βth-jump Turing degree, it is $0^{(\beta)}$, while for all $\mathbf{x} \geq_T 0^{(\tau)}$, there is a structure in $\mathbb{K}$ with τth-jump Truing degree $\mathbf{x}$.

Theorem 3.25. *Let $\mathbb{K}$ be a $\Pi^{\mathtt{in}}_2$-class with bf-ordinal $\tau < \omega_1$, and suppose that it has Turing ordinal on a cone. Then, on a cone, $\mathbb{K}$ has Turing ordinal either τ or $\tau + 1$.*

Sketch of the proof. This sketches assumes familiarity with either [46, Proof of theorem 3.1] or [48, Section 5]. Since for all $\gamma < \tau$, $\mathbb{K}$ is $\Sigma^{\mathtt{in}}_\gamma$-small, on a cone the greatest lower bound of each degree spectrum of structures in $\mathbb{K}$ is $0^{(\gamma)}$. On the other hand, since $\mathbb{K}$ is not $\Sigma^{\mathtt{in}}_\tau$-small, there is a perfect tree T of $\exists$-types over the language $\mathcal{L}_\tau$ as in [46, Proof of theorem 3.1] and also [48, Section 5]. Let us relativize to that tree, and hence assume T is computable. For every X above $0'$, there is a 1-generic G such that $G \oplus 0' \equiv_T G' \equiv_T X$. Then, there is a structure $\mathcal{A}_\tau \in \mathbb{K}$ computable in G realizing the $\exists$-type $T[G]$ (by [46, Lemma 2.9]). So X computes the jump of some copy of $\mathcal{A}_\tau$. On the other hand, G is left-c.e. in every copy of $\mathcal{A}_\tau$. So, the jump of every copy of $\mathcal{A}'_\tau$ computes G and $0'$, and hence X. It follows that $\mathcal{A}$ has $\tau + 1$-jump degree X.

Thus, the Turing ordinal on a cone of $\mathbb{K}$ is at least τ and at most $\tau + 1$. □

It is not hard to see that the bf-ordinal of $\mathbb{K}$ is ∞ if and only if it has countably many structures. If $\mathbb{K}$ is an $\mathcal{L}_{\omega_1,\omega}$-axiomatizable class, it follows from Morley's proof [51] that, for every $\alpha < \omega_1$, the number of $\Sigma^{\mathtt{in}}_\alpha$ types it realizes is either countable or continuum. Therefore, the bf-ordinal of $\mathbb{K}$ is less than ω_1 if and only if $\mathbb{K}$ has continuum many countable models. In the remaining case, when the bf-ordinal of $\mathbb{K}$ is ω_1, $\mathbb{K}$ must then have $\aleph_1$ many countable models. Vaught's conjecture [60] claims that this last case never occurs. Thus, if $\mathbb{K}$ is $\mathcal{L}_{\omega_1,\omega}$-axiomatizable and its bf-ordinal is ω_1, we say that $\mathbb{K}$ is a *counterexample to Vaught's conjecture*.

We actually do not know any example of a $\Pi^{\mathtt{in}}_\alpha$-class $\mathbb{K}$ whose bf-ordinal is greater than $\alpha + \omega$ unless it is ∞. The closest example we know is the class of Boolean algebras which is $\forall_2$ and has bf-ordinal ω.

Question 2. Is there a Π^{in}_2 class $\mathbb{K}$ with $2^{\aleph_0}$ many models whose bf-ordinal is greater than ω?

Let us remark that question 2 asks about a strengthening of Vaught's conjecture in the opposite direction as Martin's conjecture. Martin's model-theoretic conjecture is about complete first order theories which have less than $2^{\aleph_0}$ many countable models, and implies that they all have bounded Scott rank by at most $\omega + \omega$. Wagner had proposed a strengthening of Martin's conjecture which included theories with $2^{\aleph_0}$ many countable models, which turned out to be false (Gao [22]).

The author found the following connection between these examples and the iterates of the low-property mentioned above. Recall that $\mathbb{K}$ satisfies "hyperarithmetic-is-recursive" if it has the low_α property for all $\alpha < \omega_1^{CK}$, which is equivalent to saying that every hyperarithmetic structure in $\mathbb{K}$ has a computable copy.

Theorem 3.26 (ZFC+$\forall X$ ($X^\sharp$ exists)). *Let $\mathbb{K}$ be an $\mathcal{L}_{\omega_1,\omega}$-axiomatizable class with uncountably many models. The following are equivalent:*

(1) *$\mathbb{K}$ is a counterexample to Vaught's conjecture.*

(2) *$\mathbb{K}$ satisfies hyperarithmetic-is-recursive relative to all oracles on a cone.*

The proof in [48] used projective determinacy, but this was then improved to $\forall X$ ($X^\sharp$ exists) in [41]. Furthermore, in [41] the result above is extended to all analytic equivalence classes E: E has $\aleph_1$ equivalence classes if and only if it satisfies hyperarithmetic-is-recursive on a cone non-trivially.

The main theorem of [48] is actually stronger than Theorem 3.26. Assuming $\mathbf{\Sigma}^1_2$-determinacy and relative to all oracles on a cone, $\mathbb{K}$ has the *low-for-ω_1 property*, that is, if a structure in $\mathbb{K}$ has a presentation that is low-for-ω_1, then it has a computable copy. (We recall that X is low-for-ω_1 if $\omega_1^X = \omega_1^{CK}$.)

4. Comparing the complexity of classes

Reducibilities between classes allow us to classify structures in one class in terms of structures in another class. With this in mind, Friedman and Stanley [20] defined the notion of Borel reducibility. Since then, the study of Borel reducibility on arbitrary Borel and analytic equivalence relations has been extremely active in descriptive set theory. We concentrate here on the isomorphism relation.

Definition 4.1 (H. Friedman and L. Stanley [20]). A class of structures $\mathbb{K}$ is *Borel reducible* to a class $\mathbb{S}$, and we write $\mathbb{K} \leq_B \mathbb{S}$, if there is a Borel function $f\colon 2^\omega \to 2^\omega$ that maps presentations of structures in $\mathbb{K}$ to structures in $\mathbb{S}$ and preserves isomorphism. That is, for all $\mathcal{A} \in \mathbb{K}$, $f(D(\mathcal{A})) = D(\mathcal{B})$ for some $\mathcal{B} \in \mathbb{S}$, and if $\tilde{\mathcal{A}} \in \mathbb{K}$ with $f(D(\tilde{\mathcal{A}})) = D(\tilde{B})$, then

$$\mathcal{A} \cong \tilde{\mathcal{A}} \iff \mathcal{B} \cong \tilde{\mathcal{B}}.$$

(Recall that $D(\mathcal{A})$ is the atomic diagram of $\mathcal{A}$ coded as a subset of ω.)

A class $\mathbb{K}$ is *on top for Borel reducibility* if for every language $\mathcal{L}$, the class of $\mathcal{L}$-structures is Borel-reducible to $\mathbb{K}$.[4]

[4] In the literature these classes are sometimes called *Borel complete*, but we want to avoid that notation here.

They first observed that it is enough to use the language with only one binary relation (i.e. directed graphs) in the definition above. Then, they built Borel reductions to show that the classes of trees, linear orderings, 2-step nilpotent groups and fields are all on top for Borel reducibility. Camerlo and Gao [9] added Boolean algebras to that list. Friedman and Stanley observed that if a class is on top, then its isomorphism problem must be $\mathbf{\Sigma}^1_1$-complete, giving them a whole range of examples which are not on top for Borel reducibility. Torsion abelian groups are an interesting class: their isomorphism problem is Σ^1_1-complete, but they are not on top for Borel reducibility [20, Theorem 5]. (The reason is that their isomorphism problem can be reduced to countable subsets of ordinals via de Ulm invariants in a constructible way, and hence E_0 does not reduce to it.) Whether torsion-free abelian groups are on top was stated as an open question then and remains so. Since then, Downey and Montalbán showed that their isomorphism problem is Σ^1_1-complete, using ideas of Hjorth [27]. It is also open if abelian groups are on top.

In this paper, we are interested in effective versions of this reducibility.

4.1. Effective reducibility. One way of effectivizing the notion of Borel reducibility is by considering computable reductions that act on indices of computable structures. There has been some recent interest on this reducibility which has turned out to be much more interesting than expected [11, 16, 17, 42].

Definition 4.2. We say that a class of structures $\mathbb{K}$ is *effectively reducible* to a class $\mathbb{S}$ if there is a computable function $f\colon \omega \to \omega$ which maps indices of computable structures in $\mathbb{K}$ to indices of computable structures in $\mathbb{S}$ preserving isomorphism. A class of structures $\mathbb{K}$ is said to be *on top for effective reducibility* if for any computable language $\mathcal{L}$, the class of $\mathcal{L}$-structures effectively reduces to it.

(One could also consider hyperarithmetic reductions, but the author has recently shown that, on a cone, being on top for effective reducibility is equivalent to being on top for hyperarithmetic reducibility [42, Theorem 1.6], and hence does not make a difference for natural classes.)

E. Fokina, S. Friedman, V. Harizanov, J. Knight, C. MaCoy and A. Montalbán [17] gave proofs that linear orderings, trees, fields, p-groups and torsion-free abelian groups are all on top. Note that this is different from the Borel-reducibility case where p-groups are not on top, and where it is open if abelian groups are.

It is not hard to see that if a class is on top, its isomorphism-index-set,

$$E(\mathbb{K}) = \{\langle n, m\rangle \in \omega^2 : \ n \text{ and } m \text{ are indices for isomorphic computable structures in } \mathbb{K}\},$$

must be Σ^1_1-complete. Thus, $\mathbb{Q}$-vector spaces, equivalence structures, torsion-free abelian groups of finite rank, etc. cannot be on top because they have arithmetic isomorphism problems. So far, this is the only way we know to produce examples of classes which are not on top.

Definition 4.3. A class $\mathbb{K}$ is *intermediate for effective reducibility* if it is not on top for effective reducibility, and its isomorphism-index-set is not hyperarithmetic.

The reason is that when we say that $\mathbb{K}$ is $\mathbf{\Sigma}^1_1$-complete we mean that there is a continuous reduction from any $\mathbf{\Sigma}^1_1$ subset of 2^ω to the isomorphism problem of $\mathbb{K}$ as a set, and not as an equivalence relation. Reductions that preserve equivalence relations are quite different.

No specific example of an intermediate class is known. Becker [5], and independently Knight and Montalbán [unpublished], showed that such a class of structures exists under the assumption that Vaught's conjecture fails (relative to some oracle). The question now is whether such examples can be built without using a counterexample to Vaught's conjecture:

Question 3. Are the following statements equivalent?

- Vaught's conjecture.
- No $L_{\omega_1,\omega}$-axiomatizable class of structures is intermediate for effective reducibility, relative to every oracle on a cone.

Recent work by the author [42] gives a partial reversal, showing that the second statement follows from a strengthening of Vaught's conjecture (which might turn out to be equivalent to Vaught's conjecture too).

4.2. Turing-computable reducibility. The notion of Turing computable reducibility between classes of structures was introduced by Calvert, Cummins, Knight and S. Miller [8]. It is defined exactly as Borel reducibility (Definition 4.1) except that the function f is required to be a computable operator.

Definition 4.4. A class $\mathbb{K}$ is *Turing computable reducible (tc-reducible)* to $\mathbb{S}$, and we write $\mathbb{K} \leq_{tc} \mathbb{S}$, if there is a Turing operator Φ such that for every presentation $\mathcal{A} \in \mathbb{K}$, $\Phi^{D(\mathcal{A})}$ is the characteristic function of $D(\mathcal{B})$ for some $\mathcal{B} \in \mathbb{S}$ in a way that, if also $\Phi^{D(\tilde{\mathcal{A}})} = D(\tilde{\mathcal{B}})$, then

$$\mathcal{A} \cong \tilde{\mathcal{A}} \iff \mathcal{B} \cong \tilde{\mathcal{B}}.$$

So, instead of working on indices, these operators act on the atomic diagrams given as reals. This makes more of a difference than it seems. It is not hard to see that tc-reducibility implies effective reducibility. This implication does not reverse, as tc-reducibility also implies Borel-reducibility, which is not implied by effective-reducibility (e.g. abelian p-groups).

A class $\mathbb{K}$ is then *on top for tc-reducibility* if for every computable language $\mathcal{L}$, the class of $\mathcal{L}$-structures tc-reduces to $\mathbb{K}$. All the reducibilities produced in [20] are not just Borel but also effective, showing that trees, linear orderings, nilpotent groups and fields are actually on top for tc-reducibility. However, the fact that tc-reduction is finer than Borel reduction allows it to get finer comparabilities between certain classes of structures. For instance, any two classes of structures with countably infinitely many models are Borel-equivalent – this is not the case for tc-reductions, and an interesting structure can be found among these (see [34]). There are even classes of finite structures that are not trivial under tc-reducibility. But the most interesting fact about tc-reducibility is that it preserves the back-and-forth structure:

Theorem 4.5 (Pull Back theorem). *(Knight, S. Miller and Vanden Boom [34]) Let Φ be a Turing computable embedding from $\mathbb{K}$ to $\mathbb{S}$. Then, for every Π^c_α-formula φ, there is a Π^c_α formula φ^* such that for all $\mathcal{A} \in \mathbb{K}$,*

$$\mathcal{A} \models \varphi^* \quad \iff \quad \Phi(\mathcal{A}) \models \varphi$$

(where $\Phi(\mathcal{A})$ is the presentation $\mathcal{B}$ such that $\Phi^{D(\mathcal{A})} = D(\mathcal{B})$). It follows that if $\mathcal{A} \leq_\alpha \tilde{\mathcal{A}}$, then $\Phi(\mathcal{A}) \leq_\alpha \Phi(\mathcal{B})$ (where $\leq_\alpha$ is the α-back-and-forth relation as in [1, Chapter 15]).

This theorem allowed Knight, S. Miller and Vanden Boom to characterize the classes $\mathbb{K}$ such that $\mathbb{K} \leq_{tc} \mathbb{S}$ for certain fixed classes $\mathbb{S}$, like $\mathbb{Q}$-vector spaces.

4.3. Completeness for degree spectrum. A stronger notion of completeness was analyzed by Hirschfeldt, Khoussainov, Shore and Slinko [28]. The idea is that these complete classes of structures contains structures exhibiting all the possible computability theoretic behaviors that structures can have. Their objective was to show that certain nice classes of structures are indeed complete in this sense. Their definition is rather cumbersome, but we include it here for completeness.

Definition 4.6 ([28], Definition 1.21). A class of structures $\mathbb{K}$ is *complete with respect to degree spectra of nontrivial structures, effective dimensions, expansion by constants, and degree spectra of relations* (which we will write as HKSS-complete) if for every non-trivial structure $\mathcal{G}$ over a computable language $\mathcal{L}$, there is a structure $\mathcal{A} \in \mathbb{K}$ with the following properties:

(1) $DgSp(\mathcal{A}) = DgSp(\mathcal{G})$.

(2) If $\mathcal{G}$ is computably presentable, then the following holds:

  (i) For any degree $\mathbf{d}$, $\mathcal{A}$ has the same $\mathbf{d}$-computable dimension as $\mathcal{G}$.

  (ii) If $x \in G$, there is an $a \in A$ such that $(\mathcal{A}, a)$ has the same computable dimension as $(\mathcal{G}, x)$.

  (iii) If $S \subseteq G$, there exists $U \subseteq A$ such that $DgSp_{\mathcal{A}}(U) = DgSp_{\mathcal{G}}(S)$ and if S is intrinsically c.e., then so is U.

We recall that the *degree spectrum* of a structure $\mathcal{A}$ is

$$DgSp(\mathcal{A}) = \{X \in 2^\omega : X \text{ computes a copy of } \mathcal{A}\}.$$

They did not talk about a reducibility, but their notion can be easily be made into a reduction.

They showed that undirected graphs, partial orderings, lattices, integral domains of arbitrary characteristic (and in particular rings), commutative semigroups, and 2-step nilpotent groups are all HKSS-complete. Σ-small classes cannot be HKSS-complete. This is because the degree spectrum of a structure in a Σ-small class can never be the upper cone over a base which is higher than the complexity of all the $\exists$-types realized in the structure.

We suspect that if a nice class is not on top for tc-reducibility, it should not be HKSS-complete either.

5. Complete classes for effective-bi-interpretability

In this section, we introduce a notion of completeness much stronger than the ones above. This new notion is more structural, as its definition does not involve presentations of structures. Its main attraction is that it preserves a whole range of computational properties. The notion as defined here is new, although it is composed of a few already-well-known concepts. In [28], one can already see the idea of having interpretations which are somewhat effective. However, the properties they require use presentations and are not as clean cut or as general as the one here.

We start by introducing the notion of *effective-bi-interpretability* which is a variation of the classical model theoretic notion of bi-interpretability.

5.1. Effective-bi-interpretability. Before looking at effective-bi-interpretability, let us consider effective-interpretability in just one direction. Informally, a structure $\mathcal{A}$ is *effectively-interpretable* in a structure $\mathcal{B}$ if there is an interpretation of $\mathcal{A}$ in $\mathcal{B}$ as in model theory, but where the domain of the interpretation is allowed to be a subset of $\mathcal{B}^{<\omega}$, and where all sets in the interpretation are required to be uniformly r.i. computable, except for the domain which is allowed to be uniformly r.i.c.e.[5]

Before giving the formal definition, we need to review one more concept. A relation R on $\mathcal{A}^{<\omega}$ is said to be *uniformly r.i.c.e.* if there is a c.e. operator W such that for every copy $(\mathcal{B}, R^{\mathcal{B}})$ or $(\mathcal{A}, R)$, $R^B = W^{D(\mathcal{B})}$. These are exactly the Σ^{c}_1-definable relations without parameters. We can then extend this definition and define *uniformly r.i. computable* in the obvious way.

Definition 5.1. Let $\mathcal{A}$ be an $\mathcal{L}$-structure, and $\mathcal{B}$ be any structure. Let us assume that $\mathcal{L}$ is a relational language $\mathcal{L} = \{P_0, P_1, P_2, ...\}$ where P_i has arity $a(i)$; so $\mathcal{A} = (A; P_0^{\mathcal{A}}, P_1^{\mathcal{A}}, ...)$ and $P_i^{\mathcal{A}} \subseteq A^{a(i)}$.

We say that $\mathcal{A}$ is *effectively-interpretable* in $\mathcal{B}$ if, in $\mathcal{B}$, there is

- a uniformly r.i.c.e. set $D^{\mathcal{B}}_{\mathcal{A}} \subseteq \mathcal{B}^{<\omega}$ (the domain of the interpretation),
- a uniformly r.i. computable relation $\eta \subseteq \mathcal{B}^{<\omega} \times \mathcal{B}^{<\omega}$ which is an equivalence relation on $D^{\mathcal{B}}_{\mathcal{A}}$ (interpreting equality),
- a uniformly r.i. computable sequence of relations $R_i \subseteq (B^{<\omega})^{a(i)}$, closed under the equivalence η within $D^{\mathcal{B}}_{\mathcal{A}}$ (interpreting the relations P_i),
- and a function $f^{\mathcal{B}}_{\mathcal{A}} \colon D^{\mathcal{B}}_{\mathcal{A}} \to \mathcal{A}$ which induces an isomorphism:

$$(D^{\mathcal{B}}_{\mathcal{A}}/\eta; R_0, R_1, ...) \cong (A; P_0^{\mathcal{A}}, P_1^{\mathcal{A}}, ...).$$

Let us clarify that: The sets R_i do not need to be subsets of $(D^{\mathcal{B}}_{\mathcal{A}})^{a(i)}$, and, when we refer to the structure $(D^{\mathcal{B}}_{\mathcal{A}}/\eta; R_0, R_1, ...)$, we of course mean

$$(D^{\mathcal{B}}_{\mathcal{A}}/\eta; (R_0 \cap (D^{\mathcal{B}}_{\mathcal{A}})^{a(0)})/\eta, (R_1 \cap (D^{\mathcal{B}}_{\mathcal{A}})^{a(1)})/\eta, \ldots).$$

By *uniformly r.i. computable sequence* we mean that $\bigoplus_i R_{i\in I}$ is uniformly r.i. computable.

If we add parameters, this notion is equivalent to that of Σ-definability introduced by Ershov [15] and widely studied in Russia. Ershov's definition is quite different in format: it uses $HF(\mathcal{B})$ instead of $\mathcal{B}^{<\omega}$, and $\exists$-definable sets with parameters instead of uniformly r.i.c.e. ones. (It is known that r.i.c.e. subsets of $\mathcal{B}^{<\omega}$ are equivalent to Σ-definable (with parameters) subsets of $HF(\mathcal{B})$; see [47, Section 4].) Another well-known notion is that of Σ-*equivalence* between two structures, which just means that the structures are Σ-definable in each other. This is, indeed, quite a strong notion of equivalence, but the one we consider below is stronger, as we also require the composition of the isomorphisms to be computable in the respective structures. Here is the formal definition:

Definition 5.2. Two structures $\mathcal{A}$ and $\mathcal{B}$ are *effectively-bi-interpretable* if there are effective-interpretations of each structure in the other as in Definition 5.1 such that the compositions

$$f^{\mathcal{A}}_{\mathcal{B}} \circ \tilde{f}^{\mathcal{B}}_{\mathcal{A}} \colon D^{(D^{\mathcal{B}}_{\mathcal{A}})}_{\mathcal{B}} \to \mathcal{B} \quad \text{and} \quad f^{\mathcal{B}}_{\mathcal{A}} \circ \tilde{f}^{\mathcal{A}}_{\mathcal{B}} \colon D^{(D^{\mathcal{A}}_{\mathcal{B}})}_{\mathcal{A}} \to \mathcal{A}$$

[5] We remark that this definition is slightly different from what the author called effective-interpretability in [50, Definition 1.7], as we now allow the domain to be a subset of $\mathcal{B}^{<\omega}$ rather than $\mathcal{B}^n$ for some n, and we do not allow parameters in the definitions.

are uniformly r.i. computable in $\mathcal{B}$ and $\mathcal{A}$ respectively. (Here $\tilde{f}^{\mathcal{B}}_{\mathcal{A}}\colon (D^{\mathcal{B}}_{\mathcal{A}})^{<\omega} \to \mathcal{A}^{<\omega}$ is the obvious extension of $f^{\mathcal{B}}_{\mathcal{A}}\colon D^{\mathcal{B}}_{\mathcal{A}} \to \mathcal{A}$.)

In the next lemma, we see how effective-bi-interpretability preserves most computability theoretic properties.

Lemma 5.3. *Let $\mathcal{A}$ and $\mathcal{B}$ be effectively-bi-interpretable.*

(1) *$\mathcal{A}$ and $\mathcal{B}$ have the same degree spectrum.*

(2) *$\mathcal{A}$ is computably categorical if and only if $\mathcal{B}$ is.*

(3) *$\mathcal{A}$ and $\mathcal{B}$ have the same computable dimension.*

(4) *$\mathcal{A}$ is rigid if and only if $\mathcal{B}$ is.*

(5) *$\mathcal{A}$ and $\mathcal{B}$ have the same Scott rank.*

(6) *For every $\bar{a} \in \mathcal{A}^{<\omega}$, there is a $\bar{b} \in \mathcal{B}^{<\omega}$ such that $(\mathcal{A}, \bar{a})$ and $(\mathcal{B}, \bar{b})$ have the same computable dimension, and vice-versa.*

(7) *For every $R \subseteq \mathcal{A}^{<\omega}$, there is a $Q \subseteq \mathcal{B}^{<\omega}$ which has the same degree spectrum, and vice-versa.*

(8) *$\mathcal{A}$ has the c.e. extendibility condition if and only if $\mathcal{B}$ does.*

(9) *The index sets of $\mathcal{A}$ and $\mathcal{B}$ are Turing equivalent, assuming $\mathcal{A}$ and $\mathcal{B}$ are infinite structures.*

(10) *The jumps of $\mathcal{A}$ and $\mathcal{B}$ are effectively-bi-interpretable.*

(Of course, items 2-10 assume $\mathcal{A}$ and $\mathcal{B}$ are computable.)

Sketch of the proof. Throughout this proof, assume that $\mathcal{A}$ is the presentation that is coded inside $\mathcal{B}^{<\omega}$, i.e. with domain $D^{\mathcal{B}}_{\mathcal{A}}$, and $\tilde{\mathcal{B}}$ is the copy of $\mathcal{B}$ coded inside $\mathcal{A}^{<\omega}$, i.e. with domain $D^{\mathcal{A}}_{\mathcal{B}} = D^{\mathcal{D}^{\mathcal{B}}_{\mathcal{A}}}_{B}$. We let f be the isomorphism from $\tilde{\mathcal{B}}$ to $\mathcal{B}$.

For part 1, just observe that via the Σ-interpretation, given a copy of $\mathcal{B}$, we can enumerate a copy of $\mathcal{A}$, and hence compute one.

For part 2, we need the following observation: Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be copies of $\mathcal{B}$, and let $\mathcal{A}_1$ and $\mathcal{A}_2$ be the presentations of $\mathcal{A}$ coded inside $\mathcal{B}_1^{<\omega}$ and $\mathcal{B}_2^{<\omega}$ respectively. The point we need to make here is that if $\mathcal{A}_1$ and $\mathcal{A}_2$ are computably isomorphic, then so are $\mathcal{B}_1$ and $\mathcal{B}_2$: A computable isomorphism between $\mathcal{A}_1$ and $\mathcal{A}_2$ induces a computable isomorphism between $\tilde{\mathcal{B}}_1$ and $\tilde{\mathcal{B}}_2$, which each are computably-isomorphic to $\mathcal{B}_1$ and $\mathcal{B}_2$ respectively. Thus, if $\mathcal{A}$ is computably categorical, so is $\mathcal{B}$. For 3, we have that if $\mathcal{B}$ has k non-computably isomorphic copies $\mathcal{B}_1, ..., \mathcal{B}_k$, then the respective structures $\mathcal{A}_1, ..., \mathcal{A}_k$ cannot be computably isomorphic either. So the effective dimension of $\mathcal{A}$ is at least that of $\mathcal{B}$, and hence, by symmetry, they must be equal.

For part 4, suppose $\mathcal{B}$ is not rigid. Let h be a nontrivial automorphism of $\mathcal{B}$. It then induces an automorphism of $\mathcal{B}^{<\omega}$, which then induces an automorphism g of $\mathcal{A}$, which then induces an automorphism h_1 of $\tilde{\mathcal{B}}$. Since f is invariant, we have $h \circ f = f \circ h_1$ and, since h is nontrivial, h_1 is not trivial either. It follows that the automorphism g of $\mathcal{A}$ cannot be trivial either.

For part 5, suppose that $SR(\mathcal{A}) = \alpha$, that is, that every automorphism orbit in $\mathcal{A}$ is $\Sigma^{\mathtt{in}}_\alpha$ definable. Take a tuple $\bar{b} \in \mathcal{B}^{<\omega}$; we will show its orbit is also $\Sigma^{\mathtt{in}}_\alpha$ definable. Let $\bar{c} \in \tilde{\mathcal{B}}^{<\omega} \subseteq \mathcal{B}^{<\omega}$ be such that $f(\bar{c}) = \bar{b}$. The orbit of $\bar{c}$ is $\Sigma^{\mathtt{in}}_\alpha$ definable inside $\mathcal{A}$, and since

$\mathcal{A}$ is $\Sigma^{\mathtt{c}}_1$-definable in $\mathcal{B}$, the orbit of $\bar{c}$ is also $\Sigma^{\mathtt{in}}_\alpha$ definable in $\mathcal{B}$. Since f is $\Sigma^{\mathtt{c}}_1$-definable in $\mathcal{B}$, the orbit of $\bar{b}$ is also $\Sigma^{\mathtt{in}}_\alpha$ definable. It follows that $SR(\mathcal{B}) \leq \alpha$, and, by symmetry, that $SR(\mathcal{B}) = \alpha$.

For part 6, think of $\bar{a}$ as a tuple in $(D^B_{\mathcal{A}})^{<\omega} \subseteq \mathcal{B}^{<\omega}$ and call it $\bar{b}$. It is not hard to show that $(\mathcal{A}, \bar{a})$ and $(\mathcal{B}, \bar{b})$ are effectively-bi-interpretable.

For part 7, think of R as a subset of $(D^B_{\mathcal{A}})^{<\omega} \subseteq \mathcal{B}^{<\omega}$ and call it Q. Clearly, for every copy of $\mathcal{B}$, R and Q have the same degree. Conversely, for each copy of $\mathcal{A}$, if we look at the copy of $\mathcal{B}$ inside and then at the one of $\mathcal{A}$ inside it, we get that R and Q have the same degree too.

For part 8, all we have to notice is that each $\exists$-type in $\mathcal{A}$ is 1-1 reducible to a $\Sigma^{\mathtt{c}}_1$-type in $\mathcal{B}$, and vice-versa.

For part 9, given an index of a structure that we want to know if it is isomorphic to $\mathcal{B}$, we can produce an index for the structure that is then supposed to be isomorphic to $\mathcal{A}$. If it is not, then we know the original structure was not isomorphic to $\mathcal{B}$. If it is, we need to check that the bi-interpretability does produce an isomorphism, which $0''$ can check. One has to notice that all index sets compute $0''$, as their domain must be infinite.

Last, for part 10, it is not hard to interpret the complete r.i.c.e. relations from one structure into the other by interpreting $\Sigma^{\mathtt{c}}_1$-formulas in one by $\Sigma^{\mathtt{c}}_1$ formulas in the other. □

5.2. Reduction via effective-bi-interpretability. As we mentioned before, uniform r.i.c.e. sets are $\Sigma^{\mathtt{c}}_1$ definable. So, an effective-bi-interpretation is given by a list of $\Sigma^{\mathtt{c}}_1$ formulas defining all the relations involved. When we fix these formulas, we obtain a map from one kind of structure into another (which might not always define a bi-interpretation). We can use this to define a reducibility between classes:

Definition 5.4. A class $\mathbb{K}$ is *reducible to* $\mathbb{S}$ *via effective-bi-interpretability* if there are $\Sigma^{\mathtt{c}}_1$ formulas such that for every $\mathcal{A} \in \mathbb{K}$, there is a $\mathcal{B} \in \mathbb{S}$ such that $\mathcal{A}$ and $\mathcal{B}$ are effectively-bi-interpretable using those formulas. A class $\mathbb{K}$ is *on top for effective-bi-interpretability* if for every computable language $\mathcal{L}$, the class of $\mathcal{L}$-structures is reducible to $\mathbb{K}$ via effective-bi-interpretability.

Not much is known about this definition. Classes that are Σ-small are not on top for effective-bi-interpretability for the same reason they are not HKSS-complete. Classes that have bounded Scott rank cannot be on top because effective-bi-interpretability preserves Scott ranks. Using the interpretations defined by Hirschfeldt, Khoussainov, Shore and Slinko [28], we get the following: undirected graphs, partial orderings, and lattices are on top for effective-bi-interpretability; if we add a finite set of constants to the languages of integral domains, commutative semigroups, or 2-step nilpotent groups, they become on top for effective-bi-interpretability too. A recent result by R. Miller, J. Park, B. Poonen, H. Schoutens, and A. Shlapentokh [53] shows that fields are on top for effective-bi-interpretability.

Acknowldgements. The author was partially supported the Packard Fellowship. The author would like to thank Matthew Harrison-Trainor and Noah Schweber for useful comments on an earlier draft of this paper.

References

[1] C.J. Ash and J. Knight, *Computable Structures and the Hyperarithmetical Hierarchy*, Elsevier Science, 2000.

[2] Chris Ash, Julia Knight, Mark Manasse, and Theodore Slaman, *Generic copies of countable structures*, Ann. Pure Appl. Logic, **42**(3):195–205, 1989.

[3] P.E. Alaev, J. J. Thurber, and A. N. Frolov, *Computability on linear orders enriched by predicates*, Algebra Logika, **48**(5):549–563, 677, 680, 2009.

[4] V. Baleva, *The jump operation for structure degrees*, Arch. Math. Logic, **45**(3):249–265, 2006.

[5] Howard Becker, *Isomorphism of computable structures and Vaught's conjecture*, To appear.

[6] Howard Becker and Alexander S. Kechris, *The descriptive set theory of Polish group actions*, volume **232** of London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, 1996.

[7] Wesley Calvert, Douglas Cenzer, Valentina Harizanov, and Andrei Morozov, *Effective categoricity of equivalence structures*, Ann. Pure Appl. Logic, **141**(1-2):61–78, 2006.

[8] W. Calvert, D. Cummins, J. F. Knight, and S. Miller, *Comparison of classes of finite structures*, Algebra Logika, **43**(6):666–701, 759, 2004.

[9] Riccardo Camerlo and Su Gao, *The completeness of the isomorphism relation for countable Boolean algebras*, Trans. Amer. Math. Soc., **353**(2):491–518, 2001.

[10] John Chisholm, *Effective model theory vs. recursive model theory*, J. Symbolic Logic, **55**(3):1168–1191, 1990.

[11] Samuel Coskey, Joel David Hamkins, and Russell Miller, *The hierarchy of equivalence relations on the natural numbers under computable reducibility*, Computability, **1**(1):15–38, 2012.

[12] Downey, Hirschfeldt, Kach, Lempp, A. Montalbán, and Mileti, *Subspaces of computable vector spaces*, Journal of Algebra, **314**(2):888–894, August 2007.

[13] Rod Downey and Carl G. Jockusch, *Every low Boolean algebra is isomorphic to a recursive one*, Proc. Amer. Math. Soc., **122**(3):871–880, 1994.

[14] R. Downey, A. Kach, S. Lempp, A.E.M. Lewis-Pye, A. Montalbán, and D. Turetsky, *The complexity of computable categoricity*, Submitted for publication.

[15] Yuri L. Ershov, *Definability and computability*. Siberian School of Algebra and Logic. Consultants Bureau, New York, 1996.

[16] Ekaterina B. Fokina and Sy-David Friedman, *Equivalence relations on classes of computable structures*, In Mathematical theory and computational practice, volume **5635** of Lecture Notes in Comput. Sci., pages 198–207. Springer, Berlin, 2009.

[17] E. B. Fokina, S. Friedman, V. Harizanov, J. F. Knight, C. McCoy, and A. Montalbán, *Isomorphism and bi-embeddability relations on computable structures*, Journal of Symbolic Logic, **77**(1):122–132, 2012.

[18] Andrey N. Frolov, *Linear orderings of low degree*, Sibirsk. Mat. Zh., **51**(5):1147–1162, 2010.

[19] ———, *Low linear orderings*, J. Logic Comput., **22**(4):745–754, 2012.

[20] Harvey Friedman and Lee Stanley, *A Borel reducibility theory for classes of countable structures*, J. Symbolic Logic, **54**(3):894–914, 1989.

[21] Jörg Flum and Martin Ziegler, *Topological model theory*, volume **769** of Lecture Notes in Mathematics. Springer, Berlin, 1980.

[22] Su Gao, *Some dichotomy theorems for isomorphism relations of countable models*, J. Symbolic Logic, **66**(2):902–922, 2001.

[23] S. S. Gončarov and V. D. Dzgoev, *Autostability of models*, Algebra i Logika, **19**(1):45–58, 132, 1980.

[24] S. S. Goncharov and Dzh. Naĭt, *Computable structure and antistructure theorems*, Algebra Logika, **41**(6):639–681, 757, 2002.

[25] S. S. Gončarov, *The number of nonautoequivalent constructivizations*, Algebra i Logika, **16**(3):257–282, 377, 1977.

[26] Sergey S. Goncharov, *Autostability of models and abelian groups*, Algebra i Logika, **19**(1):23–44, 132, 1980.

[27] Greg Hjorth, *The isomorphism relation on countable torsion free abelian groups*, Fund. Math., **175**(3):241–257, 2002.

[28] Denis R. Hirschfeldt, Bakhadyr Khoussainov, Richard A. Shore, and Arkadii M. Slinko, *Degree spectra and computable dimensions in algebraic structures*, Ann. Pure Appl. Logic, **115**(1-3):71–113, 2002.

[29] Kenneth Harris and Antonio Montalbán, *On the n-back-and-forth types of Boolean algebras*, Trans. Amer. Math. Soc., **364**(2):827–866, 2012.

[30] Carl G. Jockusch, Jr. and Robert I. Soare, *Degrees of orderings not isomorphic to recursive linear orderings*, Ann. Pure Appl. Logic, **52**(1-2):39–64, 1991, International Symposium on Mathematical Logic and its Applications (Nagoya, 1988).

[31] ———, *Boolean algebras, Stone spaces, and the iterated Turing jump*, J. Symbolic Logic, **59**(4):1121–1138, 1994.

[32] A. N. Khisamiev, *On the Ershov upper semilattice L_E*, Sibirsk. Mat. Zh., **45**(1):211–228, 2004.

[33] A. Kach and A. Montalbán, *Cuts of linear orders*, Order, **28**(3):593–600, 2011.

[34] Julia F. Knight, Sara Miller, and M. Vanden Boom, *Turing computable embeddings*, J. Symbolic Logic, **72**(3):901–918, 2007.

[35] J. B. Kruskal, *Well-quasi-ordering, the Tree Theorem, and Vazsonyi's conjecture*, Trans. Amer. Math. Soc., **95**:210–225, 1960.

[36] Julia F. Knight and Michael Stob, *Computable Boolean algebras*, J. Symbolic Logic, **65**(4):1605–1623, 2000.

[37] E. G. K. Lopez-Escobar, *An interpolation theorem for denumerably long formulas*, Fund. Math., **57**:253–272, 1965.

[38] Steffen Lempp, Charles McCoy, Russell Miller, and Reed Solomon, *Computable categoricity of trees of finite height*, J. Symbolic Logic, **70**(1):151–215, 2005.

[39] Peter E. La Roche, *Contributions to Recursive Algebra*, ProQuest LLC, Ann Arbor, MI, 1978, Thesis (Ph.D.)–Cornell University.

[40] David Marker, *Bounds on Scott rank for various nonelementary classes*, Arch. Math. Logic, **30**(2):73–82, 1990.

[41] Antonio Montalbán, *Analytic equivalence relations satisfying hyperarithmetic - is - recursive*, Submitted for publication.

[42] _____, *Classes of structures with no intermediate isomorphism problems*, Submitted for publication.

[43] _____, *A robuster Scott rank*, Submitted for publication.

[44] _____, *Notes on the jump of a structure*, Mathematical Theory and Computational Practice, pages 372–378, 2009.

[45] _____, *Coding and definability in computable structures*, Notes from a course at Notre Dame University to be published in the NDJFL, 2010.

[46] _____, *Counting the back-and-forth types*, Journal of Logic and Computability, page doi: 10.1093/logcom/exq048, 2010.

[47] _____, *Rice sequences of relations*, Philosophical Transactions of the Royal Society A, **370**:3464–3487, 2012.

[48] _____, *A computability theoretic equivalent to Vaught's conjecture*, Adv. Math., **235**:56–73, 2013.

[49] _____, *Copyable structures*, Journal of Symbolic Logic, **78**(4):1025–1346, 2013.

[50] _____, *A fixed point for the jump operator on structures*, Journal of Symbolic Logic, **78**(2):425–438, 2013.

[51] Michael Morley, *The number of countable models*, J. Symbolic Logic, **35**:14–18, 1970.

[52] Yiannis N. Moschovakis, *Descriptive set theory*, volume 100 of Studies in Logic and the Foundations of Mathematics, North-Holland Publishing Co., Amsterdam, 1980.

[53] R. Miller, J. Park, B. Poonen, H. Schoutens, and A. Shlapentokh, *Fields are complete for isomorphisms*, To appear.

[54] A. T. Nurtazin, *Computable classes and algebraic criteria for autostability*, PhD thesis, Institute of Mathematics and Mechanics, Alma-Ata, 1974.

[55] Linda Richter, *Degrees of unsolvability of models*, PhD thesis, University of Illinois at Urbana-Champaign, 1977.

[56] Linda Jean Richter, *Degrees of structures*, J. Symbolic Logic, **46**(4):723–731, 1981.

[57] Gerald E. Sacks, *Bounds on weak scattering*, Notre Dame J. Formal Logic, **48**(1):5–31, 2007.

[58] Rick L. Smith, *Two theorems on autostability in p-groups*, In Logic Year 1979–80 (Proc. Seminars and Conf. Math. Logic, Univ. Connecticut, Storrs, Conn., 1979/80), volume **859** of Lecture Notes in Math., pages 302–311. Springer, Berlin, 1981.

[59] John J. Thurber, *Every* low_2 *Boolean algebra has a recursive copy*, Proc. Amer. Math. Soc., **123**(12):3859–3866, 1995.

[60] R. L. Vaught, *Denumerable models of complete theories*, In Infinitistic Methods (Proc. Sympos. Foundations of Math., Warsaw, 1959), pages 303–321. Pergamon, Oxford, 1961.

[61] M. Vanden Boom, *The effective Borel hierarchy*, Fund. Math., **195**(3):269–289, 2007.

Department of Mathematics, University of California, Berkeley, CA, USA
E-mail: antonio@math.berkeley.edu

Recent developments in finite Ramsey theory: foundational aspects and connections with dynamics

Sławomir Solecki

Abstract. We survey some recent results in Ramsey theory. We indicate their connections with topological dynamics. On the foundational side, we describe an abstract approach to finite Ramsey theory. We give one new application of the abstract approach through which we make a connection with the theme of duality in Ramsey theory. We finish with some open problems.

Mathematics Subject Classification (2010). 03E15, 05D10, 22F50.

Keywords. Ramsey theory, extreme amenability, duality in Ramsey theory.

1. Ramsey theory and topological dynamics

Recent years have seen a renewed interest in Ramsey theory that lead to advances both in proving new concrete Ramsey results and in developing the foundational aspects of the theory. To a large extent this interest in Ramsey theory was sparked by the discovery of its close connections with topological dynamics and especially with the notion of extreme amenability and related to it problem of computing universal minimal flows of topological groups. A topological group is called **extremely amenable** if each continuous action of it on a compact (always assumed Hausdorff) space has a fixed point. First such groups were discovered by Herer and Christensen [13] using functional analytic methods. It was then shown by Veech [40] that extremely amenable groups cannot be locally compact. It turned out, however, that some very interesting groups are extremely amenable; for example, Gromov and Milman [12] showed that the unitary group of a separable infinite dimensional Hilbert space, taken with the strong operator topology and with composition as the group operation, is extremely amenable. The proof in [12] of this theorem used probabilistic methods of concentration of measure through the notion of Lévy group. (Lévy groups are topological groups possessing an increasing sequence of compact subgroups with dense union and with concentration of measure exhibited by the sequence of the normalized Haar measures on the compact subgroups.) Concentration of measure grew to be one of the two main methods used in proving extreme amenability.

It was not until Pestov's paper [27] that the second general method—Ramsey theory—was discovered. Pestov showed that the group of all increasing bijections from $\mathbb{Q}$ to itself, with pointwise convergence topology and composition as the group operation, is extremely amenable. His proof used the classical Ramsey theorem in a way that appeared, as it turned

Proceedings of the International Congress of Mathematicians, Seoul, 2014

out correctly, fundamental. Pestov's article was followed by two papers by Glasner and Weiss [7] and [8], one of which [8] used the dual Ramsey theorem of Graham and Rothschild, see Theorem 3.3 below, to determine the universal minimal flow of the group of all homeomorphisms of the Cantor set. The full and unexpectedly tight connection between extreme amenability and Ramsey theory was then established by Kechris, Pestov, and Todorcevic in [15].

Theorem 1.1 below is the main result of the theory found in [15]. (Paper [25] contains some further developments.) Recall that a topological group is **non-Archimedean** if it has a basis at the identity consisting of open subgroups. A structure, which we understand in the sense of Model Theory, is **ultrahomogeneous** if each isomorphism between two finite substructures extends to an automorphism of the whole structure, and it is **locally finite** if its finitely generated substructures are finite. A class $\mathcal{F}$ of finite structures is said to have the **Ramsey property** if for any positive integer c any two structures A and B in $\mathcal{F}$ there is a structure C in $\mathcal{F}$ such that for each coloring with c colors of all substructures of C isomorphic to A there is a substructure B' of C isomorphic to B such that substructures of B' isomorphic to A get the same color.

Theorem 1.1 (Kechris–Pestov–Todorcevic [15]). *Let G be a non-Archimedean, second countable, completely metrizable group. Then G is extremely amenable if and only if G is isomorphic to the group of all automorphisms (taken with the pointwise convergence topology and composition as the group operation) of a countable, ultra-homogeneous, locally finite structure A such that*

- *A is linearly ordered and*
- *the class of all finite substructures of A has the Ramsey property.*

For example, $\mathbb{Q}$ taken with its linear order is ultrahomogeneous, locally finite, linearly ordered, and the class of its finite substructures consists of finite linear orders, which has the Ramsey property by the classical theorem of Ramsey [30]. This gives back Pestov's theorem [27] mentioned above. More broadly, Theorem 1.1 related topological dynamics to Ramsey theory for finite structures, the latter having been developed by Nešetřil, Rödl [22–24], Abramson and Harrington [1], and others.

As it turned out, the connection from Theorem 1.1 suggested new Ramsey results. One, although not the only one, way this took place was through comparisons with the concentration of measure method. Given a group whose extreme amenability was proved using concentration of measure, one could sometimes formulate a Ramsey statement that would yield the extreme amenability, and then ask if the Ramsey statement itself held. On the other hand, one could also ask for Ramsey statements that gave extreme amenability in situations to which concentration of measure did not apply. We give below two examples, one on either side.

The following theorem for finite linearly ordered metric spaces was proved by Nešetřil [21]. In its statement by an **order isometry** from a linearly ordered metric space A to a linearly ordered metric space B we understand a bijection from A to B that preserves the metric and the linear order.

Theorem 1.2 (Nešetřil [21]). *Given a positive integer c and two finite linearly ordered metric spaces A and B, there exits a finite linearly ordered metric space C such that for each coloring with c colors of all subspaces of C order isometric to A there exists a subspace B'*

of C order isometric to B such that all subspaces of B' order isometric to A have the same color.

The theorem above implies, as shown in [15], that the group of all isometries of the separable Urysohn metric space taken with the pointwise convergence topology is extremely amenable. Extreme amenability of this group was earlier established by Pestov in [28] with concentration of measure methods.

To state the second theorem, also resulting from analyzing connections between Ramsey theory, concentration of measure, and extreme amenability, consider the following notions. Let $[n]$ stand for the set $\{1, 2, \dots, n\}$. Given a prime number p, let $(\mathbb{Z}/p)^{n:l}$ be the set of all partial functions from $[n]$ to $\mathbb{Z}/p$ whose domains have at least $n-l$ elements, and let $(\mathbb{Z}/p)^n$ be the set of all functions from $[n]$ to $\mathbb{Z}/p$. A set $L \subseteq (\mathbb{Z}/p)^{n:l}$ is called **full** if there exists $h \in (\mathbb{Z}/p)^n$ and $a \subseteq [n]$ with $n - l$ elements such that for each $r \in \mathbb{Z}/p$

$$(r + h) \restriction a_r \in L$$

for some $a \subseteq a_r \subseteq [n]$.

We now have the following Ramsey theorem. We will come back to it in the last section of the paper when discussing open problems.

Theorem 1.3 (Farah–Solecki [4]). *Let $p_1, \dots, p_k$ be prime numbers, and let c be a positive integer. Then*

$$\exists l_1 \forall n_1 \geq l_1 \cdots \exists l_k \forall n_k \geq l_k \textit{ for each coloring of } \textstyle\prod_{i=1}^{k} (\mathbb{Z}/p)^{n_i:l_i} \textit{ with } c \textit{ colors}$$

there exist full sets $L_1 \subseteq (\mathbb{Z}/p_1)^{n_1:l_1}, \dots, L_k \subseteq (\mathbb{Z}/p_k)^{n_k:l_k}$ with $L_1 \times \cdots \times L_k$ monochromatic.

The proof of the above result uses Lovasz's method for calculating the chromatic numbers of the Kneser graphs, see [17]. The theorem above implies that, for example, $L_0(\phi, A)$ is extremely amenable. The group $L_0(\phi, A)$ is the completion of the group of all continuous functions (with pointwise addition) from the Cantor set $2^{\mathbb{N}}$ to a finite abelian group A with respect to convergence in ϕ, where ϕ is a diffuse *submeasure* on all closed-and-open subsets of $2^{\mathbb{N}}$. (These groups are related to the ones considered by Herer and Christensen [13].) On the other hand, it is shown in [4] that extreme amenability of $L_0(\phi, A)$ as above cannot be proved using the concentration of measure method—such groups are not Lévy despite possessing sequences of compact subgroups with dense unions.

There are many other examples of recently found Ramsey theorems with application to topological dynamics; for a sample, see [14, 25], or [32].

2. Finite Ramsey theory—abstract approach

The Kechris–Pestov–Todorcevic theory lead indirectly to rethinking of the foundations of finite Ramsey theory. In this section, we present an abstract approach to finite Ramsey theory from [35]. This approach recovers most of the core Ramsey theory and makes it possible to prove new results. At the same time, it reveals the formal algebraic structure underlying finite Ramsey theorems: there exist a single type of structure, called Ramsey domain over a normed composition space, that underlies Ramsey theorems. One formulates within this

algebraic setting an abstract pigeonhole principle and an abstract Ramsey statement, and proves, as the main theorem, that the pigeonhole principle implies the Ramsey statement. This abstract Ramsey theorem, which we state at the end of this section as Theorem 2.1, gives particular Ramsey theorems as instances, or iterative instances, for particular Ramsey domains.

We outline the general approach in this section. For details and proofs the reader should consult [35]. We give one new concrete application in the next section, which will allow us to illustrate the abstract notions in a concrete situation and also to discuss the theme of duality in Ramsey theory. We ask the reader to consult [35, 36], and [42] for more concrete applications. Let us only mention here that the following theorems can be obtained as particular instances of the abstract approach to Ramsey theory, see [35, 36], and [42]:

— the classical Ramsey theorem, see [20];

— the Hales–Jewett theorem, see [20];

— the Graham–Rothschild theorem, [9], see also [20];

— the versions of the two results directly above for partial rigid surjections due to Voigt, [41], see also [20];

— a self-dual Ramsey theorem, [35];

— the Milliken Ramsey theorem for finite trees, [18], see also [31];

— a common generalization of Deuber's and Jasiński's Ramsey theorems for finite trees, [2, 14];

— Spencer's generalization of the Graham–Rothschild theorem and the Ramsey theorem for affine subspaces, [38];

— dual Ramsey theorem for trees, [36].

2.1. Normed composition spaces. The algebraic structure is initially defined at the level of points and it is lifted later to the level of sets. We describe first the point level structure. Let A and X be sets. Assume we are given a *partial* function from $A \times A$ to A,

$$(a, b) \to a \cdot b \in A,$$

and a *partial* function from $A \times X$ to X,

$$(a, x) \to a \,.\, x \in X,$$

such that for $a, b \in A$ and $x \in X$ if $a \,.\, (b \,.\, x)$ and $(a \cdot b) \,.\, x$ are both defined, then

$$a \,.\, (b \,.\, x) = (a \cdot b) \,.\, x.$$

The above equation is just the usual action condition. We assume we also have a function $\partial\colon X \to X$ and a function $|\cdot|\colon X \to L$, where L is equipped with a partial order $\leq$. The operations $\cdot$ and $\,.\,$ are called a **multiplication** and an **action** (of A on X), respectively. We call ∂ a **truncation** and $|\cdot|$ a **norm**.

A structure $(A, X, \,.\,, \cdot, \partial, |\cdot|)$ as above is called a **normed composition space** if the following conditions hold for $a \in A$ and $x, y \in X$:

(i) if $a \,.\, x$ and $a \,.\, \partial x$ are defined, then

$$\partial(a \,.\, x) = a \,.\, \partial x;$$

(ii) $|\partial x| \leq |x|$;

(iii) if $|x| \leq |y|$ and $a \,.\, y$ is defined, then $a \,.\, x$ is defined and $|a \,.\, x| \leq |a \,.\, y|$.

The conditions above record the interactions between pairs of objects among $\,.\,$, ∂, and $|\cdot|$. So the action is done by homomorphisms with respect to the truncation, by (i), the truncation does not increase the norm, by (ii), and the action respects the norm, by (iii).

We isolate one notion that will turn out to be useful later on. Given $a, b \in A$, we say that b **extends** a if for each x for which $a \,.\, x$ is defined, $b \,.\, x$ is defined as well and is equal to $a \,.\, x$.

For $t \in \mathbb{N}$, we write ∂^t for the t-th iteration of ∂. For a subset P of X, we write $\partial P = \{\partial x \colon x \in P\}$.

2.2. Ramsey domains. Here we lift the algebraic structure from points to subsets of A and X. Let $\mathcal{F}$ and $\mathcal{P}$ be families of non-empty subsets of A and X, respectively. Assume we have a *partial* function from $\mathcal{F} \times \mathcal{F}$ to $\mathcal{F}$,

$$(F, G) \to F \bullet G \in \mathcal{F},$$

with the property that if $F \bullet G$ is defined, then it is given point-wise, that is, $f \cdot g$ is defined for all $f \in F$ and $g \in G$ and

$$F \bullet G = \{f \cdot g \colon f \in F,\ g \in G\}.$$

Assume we also have a *partial* function from $\mathcal{F} \times \mathcal{P}$ to $\mathcal{P}$,

$$(F, P) \to F \bullet P \in \mathcal{P},$$

such that if $F \bullet P$ is defined, then $f \,.\, x$ is defined for all $f \in F$ and $x \in P$ and

$$F \bullet P = \{f \,.\, x \colon f \in F,\ x \in P\}.$$

The structure $(\mathcal{F}, \mathcal{P}, \bullet, \bullet)$ as above is called a **Ramsey domain** over the normed composition space $(A, X, \,.\,, \cdot, \partial, |\cdot|)$ if the following conditions hold:

(a) if $F, G \in \mathcal{F}$, $P \in \mathcal{P}$, and $F \bullet (G \bullet P)$ is defined, then so is $(F \bullet G) \bullet P$;

(b) if $P \in \mathcal{P}$, then $\partial P \in \mathcal{P}$;

(c) if $F \in \mathcal{F}$, $P \in \mathcal{P}$, and $F \bullet \partial P$ is defined, then there is $G \in \mathcal{F}$ such that $G \bullet P$ is defined and for each $f \in F$ there is $g \in G$ extending f.

The following two conditions on Ramsey domains are crucial in running inductive arguments. A Ramsey domain is called **vanishing** if for each $P \in \mathcal{P}$ there is $t \in \mathbb{N}$ such that the set $\partial^t P$ has one element. It is called **linear** if for each $P \in \mathcal{P}$, the set $\{|x| \colon x \in P\}$ is a linearly ordered subset of L. The first one of these conditions makes it possible to start inductive arguments, the second one is used to organize induction.

2.3. Ramsey theorem. Using the structure described earlier, we state here the abstract Ramsey theorem—Theorem 2.1. The theorem will say that an appropriate pigeonhole principle implies an appropriate Ramsey condition. The following statement is our Ramsey condition for a Ramsey domain $(\mathcal{F}, \mathcal{P}, \bullet, \bullet)$.

(R) Given a positive integer c, for each $P \in \mathcal{P}$, there is an $F \in \mathcal{F}$ such that $F \bullet P$ is defined, and for every coloring with c colors of $F \bullet P$ there is an $f \in F$ such that $f \,.\, P$ is monochromatic.

For $P \subseteq X$ and $y \in X$, put

$$P_y = \{x \in P \colon \partial x = y\}.$$

For $F \subseteq A$ and $a \in A$, let

$$F_a = \{f \in F \colon f \text{ extends } a\}.$$

The following criterion is our pigeonhole principle, which we called local pigeonhole principle in [35] and denoted it there by (LP). We keep this notation here.

(LP) Given a positive integer c, for all $P \in \mathcal{P}$ and $y \in \partial P$, there are $F \in \mathcal{F}$ and $a \in A$ such that $F \bullet P$ is defined, $a \,.\, y$ is defined, and for every coloring with c colors of $F_a \,.\, P_y$ there is an $f \in F_a$ such that $f \,.\, P_y$ is monochromatic.

The following is the abstract Ramsey theorem.

Theorem 2.1 (Solecki [36]). *Let $(\mathcal{F}, \mathcal{P}, \bullet, \bullet)$ be a linear, vanishing Ramsey domain over a normed composition space. Assume that each set in $\mathcal{P}$ is finite. Then (LP) implies (R).*

3. Duality and the dual Ramsey theorem for trees

In this section, we touch on the theme of duality. In Ramsey theory for finite structures, Abramson–Harrington, Nešetřil–Rödl's theorem [1, 24] has a dual counterpart due to Prömel [29]. This duality was made precise and shown to extend to proofs in [33] and [34]. In the unstructured Ramsey theory, the classical theorem of Ramsey [30] has a dual counterpart due to Graham and Rothschild [9]. We will extend here this last instance of duality to trees and we will relate it to the concept of Galois connection. This new concrete Ramsey result will also allow us to give an illustration of the abstract notions presented in the previous section.

3.1. The context for duality among trees–Galois connections. Let $(S, \sqsubseteq_S)$ and $(T, \sqsubseteq_T)$ be two partial orders. A pair (f, e) is called a **Galois connection** if $f \colon T \to S$, $e \colon S \to T$, and

$$e \circ f \sqsubseteq_T \mathrm{id}_T \text{ and } f \circ e \sqsupseteq_S \mathrm{id}_S, \tag{3.1}$$

that is, $e(f(w)) \sqsubseteq_T w$ and $v \sqsubseteq_S f(e(v))$ for all $w \in T$ and $v \in S$. Usually the functions e and f in a Galois connection are assumed to be monotone. It is crucial for us, however, to use the more relaxed notion given above. Galois connections in their abstract form were first defined by Ore in [26]; for a comprehensive treatment see [5]. As already noticed by Ore, of particular importance are Galois connections fulfilling a strengthening of (3.1) consisting of assuming that equality holds in one of the two inequalities in (3.1). (Ore called such

connections perfect.) Since the situation we consider, when the partial orders are trees, is asymmetric, only one of these strengthenings is interesting—the one with equality holding in the second formula in (3.1), in which case (3.1) becomes

$$e \circ f \sqsubseteq_T \mathrm{id}_T \text{ and } f \circ e = \mathrm{id}_S. \tag{3.2}$$

Galois connections with (3.2) are sometimes called embedding–projection pairs, and are important in denotational semantics of programming languages, see for example [3].

3.2. The notion of rigid surjection and the dual Ramsey theorem for trees. By a **tree** we understand a finite, partially ordered set with a smallest element, called **root**, and such that the set of predecessors of each element is linearly ordered. So below, *all trees are non-empty and finite*. Maximal elements of the tree order are called **leaves**. We always denote the tree order on T by $\sqsubseteq_T$.

Each tree T carries a binary function $\wedge_T$ that assigns to each $v, w \in T$ the largest with respect to $\sqsubseteq_T$ element $v \wedge_T w$ of T that is a predecessor of both v and w. By convention, we regard every node of a tree as one of its own predecessors and as one of its own successors.

For a tree T and $v \in T$, let $\mathrm{im}_T(v)$ be the set of all **immediate successors** of v, and we do not regard v as one of them. A tree T is called **ordered** if for each $v \in T$ we have a fixed linear order on $\mathrm{im}_T(v)$. Such an assignment of linear orders defines the lexicographic linear order $\leq_T$ on all the nodes of T by stipulating that $v \leq_T w$ if v is a predecessor of w and, in case v is not a predecessor of w and w is not a predecessor of v, that $v \leq_T w$ if the predecessor of v in $\mathrm{im}_T(v \wedge w)$ is less than or equal to the predecessor of w in $\mathrm{im}_T(v \wedge w)$ in the given order on $\mathrm{im}_T(v \wedge w)$.

Let S and T be ordered trees. A function $e\colon S \to T$ is called a **morphism** if the following conditions hold:

(i) $e(v \wedge_S w) = e(v) \wedge_T e(w)$, for all $v, w \in S$;

(ii) e is monotone between $\leq_S$ and $\leq_T$, that is, $v \leq_S w$ implies $e(v) \leq_T e(w)$, for all $v, w \in S$;

(iii) e maps the root of S to the root of T.

Now we give the definition of functions appearing in the dual Ramsey theorem for trees. Let S, T be ordered trees. A function $f\colon T \to S$ is called a **rigid surjection** provided there exists a morphism $e\colon S \to T$ such that equation (3.2) holds. It is not difficult to see that in this situation f determines e uniquely, so the definition above could be stated without invoking e.

Here is the dual Ramsey theorem for trees.

Theorem 3.1 (Solecki [37]). *Let c be a positive integer. Let S, T be ordered trees. There exists an ordered tree U such that for each coloring with c colors of all rigid surjections from U to S there is a rigid surjection $g_0\colon U \to T$ such that*

$$\{f \circ g_0\colon f\colon T \to S \text{ a rigid surjection}\}$$

is monochromatic.

Ramsey theorems for trees proved so far were usually stated in terms of injective morphisms e; see [36] for a survey. (An exception here is the dual Ramsey theorem of Graham–Rothschild, which we discuss below.) Each such injective morphism e is an element of a

unique pair (f, e) with the pair fulfilling (3.2) and with f being a surjective morphism. In this situation, when both e and f are morphisms, e determines f and f determines e. So Ramsey theorems formulated in terms of e can be equivalently stated in terms of pairs (f, e) or in terms of f. One could call the formulation in terms of f dual. Now, it turns out, that on the dual side, surjective morphisms f are part of a much richer family of functions—rigid surjections; one abandons the assumption that f is a morphism and obtains a Ramsey theorem for this larger class of functions. In fact, the statement for the larger class easily implies the statements for morphisms. We discuss it briefly below.

An injective morphism between ordered trees is called an **embedding**. An image of a tree S under an embedding from S to T is called a **copy** of S in T. The following theorem is due to Leeb, see [10].

Theorem 3.2 (Leeb). *Given a positive integer c and ordered trees S and T, there is an ordered tree U such that for each coloring with c colors of all copies of S in U there is a copy T' of T in U such that all copies of S in T' get the same color.*

An embedding uniquely determines a copy which is the image of the embedding, but also vice versa, a copy uniquely determines an embedding of which it is the image. So the theorem above can be restated in terms of embeddings and can be easily seen to be a particular case of Theorem 3.1 by viewing an embedding e as an element of pairs (f, e) fulfilling (3.2).

Theorem 3.1 also generalizes the dual Ramsey theorem of Graham–Rothschild, as we indicate below. A k-**partition** of a set X is a family of k non-empty pairwise disjoint subsets of X whose union is X. A k-partition $\mathcal{P}$ is an k-**subpartition** of an l-partition $\mathcal{Q}$ if each element of $\mathcal{P}$ is the union of some elements of $\mathcal{Q}$. For $m \in \mathbb{N}$, let $[m]$ be the set $\{1, \dots, m\}$. The following is the dual Ramsey theorem of Graham and Rothschild [9]. (We come back to it in the last section of the paper.)

Theorem 3.3 (Graham–Rothschild [9]). *Let c be a positive integer. For each k, l, there exists m such that for each coloring with c colors of all k-partitions of $[m]$ there exists an l-partition $\mathcal{Q}$ of $[m]$ such that all k-subpartitions of $\mathcal{Q}$ get the same color.*

If $\mathcal{P}$ a k-partition of $[m]$, then we can write $\mathcal{P} = \{p_1, \dots, p_k\}$ with $\min p_i < \min p_{i+1}$, for $1 \leq i < k$, and define $f_{\mathcal{P}} \colon [m] \to [k]$ by

$$f_{\mathcal{P}}(x) = \text{ the unique } i \text{ such that } x \in p_i.$$

Note that $[m]$, for $m \in \mathbb{N}$, is an ordered tree if we take $[m]$ with its natural order relation and with the unique trivial ordering of the immediate successors of each vertex. If $[m]$ is treated as a tree, $f_{\mathcal{P}} \colon [m] \to [k]$ is a rigid surjection. This observation leads to a restatement of the Graham–Rothschild theorem in terms of rigid surjections. This restatement follows easily from Theorem 3.1 by considering ordered trees $S = [k]$ and $T = [l]$ and viewing the resulting tree U with its linear order $\leq_U$ only (and forgetting its tree order $\sqsubseteq_U$).

3.3. Description of algebraic structures for the dual Ramsey theorem for trees. We describe here concrete examples of the general structures defined in Section 2 that are used to prove Theorem 3.1. For technical reasons, we consider only a restricted class of rigid surjections. One proves Theorem 3.1 for this restricted class and then derives the full version of the theorem from the restricted one. For ordered trees S, T, a rigid surjection $f \colon T \to S$

is called **sealed** if $f^{-1}(v) = \{w\}$, where v is $\leq_S$-largest in S and w is $\leq_T$-largest in T. For $w \in T$, let

$$T^w = \{v \in T \colon v \leq_T w\},$$

and for $f \colon T \to S$ and $v \in S$, let

$$f^v = f \restriction T^{e(v)},$$

where $e \colon S \to T$ is the unique morphism with (f, e) fulfilling (3.2).

Now we define a normed composition space. Let $\mathcal{L}$ be a family of ordered trees such that for $T \in \mathcal{L}$ and $w \in T$, we have $T^w \in \mathcal{L}$. We will specify $\mathcal{L}$ later. The sets A and X will be equal to each other, as will be the operations $.$ and $\cdot$. We let $A = X$ be the set of all sealed rigid surjections $g \colon T_2 \to T_1$ for $T_1, T_2 \in \mathcal{L}$. Let $f, g \in A = X$. We let $g \cdot f = g \,.\, f$ be defined precisely when $f \colon T^y \to S$ and $g \colon V \to T$ for some ordered trees $S, T, V \in \mathcal{L}$ and a vertex y in T. We let

$$g \cdot f = g \,.\, f = f \circ g^y.$$

For $f \in X$ whose image is a tree S define ∂f as follows. If S consists only of its root, let

$$\partial f = f.$$

If S has a vertex that is not a root, let v be the second $\leq_S$-largest vertex in S, and let

$$\partial f = f^v.$$

Consider $\mathcal{L}$ as a partial order with the partial order relation given by requiring that T_1 be less than T_2 if and only if there exists $w \in T_2$ with $T_1 = T_2^w$. For $f \in X$, let

$$|f| = (\text{domain of } f) \in \mathcal{L}.$$

It is easy to check that the structure $(A, X, \ .\ , \cdot, \partial, |\cdot|)$ defined above is a normed composition space.

Now we define a Ramsey domain over this normed composition space. To specify $\mathcal{L}$, fix a family $\mathcal{T}$ of ordered trees such that each ordered tree has an isomorphic copy in $\mathcal{T}$ and such that $T_1 \cap T_2 = \emptyset$, for $T_1, T_2 \in \mathcal{T}$, and let

$$\mathcal{L} = \{T^w \colon T \in \mathcal{T},\ w \in T\}.$$

We consider non-empty sets $K \subseteq A = X$ for which there exist ordered trees T_1, T_2 such that each element of K has its domain included in T_2 and its image equal to T_1. We require that $T_2 \in \mathcal{T}$. Since the trees in $\mathcal{T}$ are pairwise disjoint, each element of K determines T_2. We define $d(K) = T_2$ and $r(K) = T_1$. Now, let $\mathcal{F}$ consist of all such sets K with $r(K) \in \mathcal{T}$, and let $\mathcal{P}$ consist of all such sets K with $r(K) \in \mathcal{L}$. For $F_1, F_2, F \in \mathcal{F}$ and $P \in \mathcal{P}$, let $F_1 \bullet F_2$ and $F \bullet P$ be defined precisely when $d(F_2) = r(F_1)$ and $d(P) = r(F)$, respectivly. In these cases, we let

$$F_1 \bullet F_2 = F_1 \cdot F_2 \ \text{ and } \ F \bullet P = F \,.\, P.$$

Again one checks that the structure defined above is a Ramsey domain that is linear and vanishing. Condition (R) for it gives the statement of the dual Ramsey theorem for trees (for sealed surjections); condition (LP) for it is proved using a version of the Hales–Jewett theorem, but we will not describe this argument here. For all the proofs the reader may consult [37].

4. Further developments and problems

We present below two groups of problems. Both of them aim at extending, in two different ways, the point of view from Section 2 beyond its original context. The first problem has to do with unifying the approach to finite Ramsey theory of [35], which was described in Section 2, with Todorcevic's infinite Ramsey theory of [39]. Issues in the second group center around proving certain analogous or finding a better understanding of Theorems 1.1, 1.3 and 3.3 presented earlier.

First, there exists a general approach to *infinite* Ramsey theory given by Todorcevic in [39] that incorporates earlier work of Nash-Williams, Ellentuck, and Carlson, among others. Roughly speaking, this is a theory of finding infinite sequences (x_n) such that the set of all infinite sequences formed from (x_n) by, for example, amalgamating or taking subsequences or acting by a semigroup, is monochromatic. The question arises whether one can view the approach to finite Ramsey theory outlined in Section 2 as a starting point, or as the underlying layer, of the infinite Ramsey theory. For example, given a normed composition space $(A, X, \, .\, , \cdot, \partial, |\cdot|)$ as in Section 2, it is natural to consider the space of sequences

$$\varprojlim(X, \partial) = \{(x_n) \in X^{\mathbb{N}} \colon x_n = \partial x_{n+1} \text{ for each } n \in \mathbb{N}\}$$

with the induced partial action of A. It seems plausible that Todorcevic's theory can be recovered in spaces of the form $\varprojlim(X, \partial)$, which would unify the two approaches.

Second, there exist certain Ramsey statements that point to a possible relationship of Ramsey theory with combinatorial tools coming from algebraic topology as in [17] or from fixed point theorems in convex analysis. (A similar view is expressed by Gromov in [11, Introduction to Section 1].) We may recall that Theorem 1.3 is proved using methods that originated with Lovasz's proof of Kneser's conjecture, which is done with the aid of insights coming from algebraic topology around the Lefschetz fixed point theorem. Below, we describe two other purely Ramsey theoretic statements with some intriguing additional features. Both of them merit attention in their own right. It would also be very interesting to see if the combinatorial methods stemming from algebraic topology as in [17] can be incorporated into the approach outlined in Section 2 to shed light on these or similar statements.

Moore [19] carried out an analysis, analogous to the Kechris–Pestov–Todorcevic [15] analysis described in Section 1, of amenability among non-Archimedean groups. As a by-product, he uncovered a Ramsey statement relevant to amenability of well known Thompson's group F. This is the group, under composition, of all piecewise linear increasing homeomorphisms of the interval $[0, 1]$ whose non-differentiability points are dyadic rationals and whose slopes are integer powers of 2. Moore found a Ramsey statement equivalent to amenability of F (establishing which is a major problem concerning this group). This Ramsey statement has a new feature—it involves convex combinations. We reproduce it below.

By a **binary tree** we understand an ordered tree T, as in Section 3.2, with the property that for each vertex $v \in T$ the set of its immediate successors $\mathrm{im}_T(v)$ has size 0 or 2. For $n \in \mathbb{N}$, $n > 0$, let $\mathbb{T}_n$ denote the set of all binary trees with n leaves. Given a sequence of binary trees $\vec{U} = (U_1, \dots, U_m)$ such that the number of leaves in all of them totals n and given a tree T in $\mathbb{T}_m$, let $T(\vec{U})$ be the tree in $\mathbb{T}_n$ that results from T by attaching U_i to the i-th leaf of T, where the leaves of T are numbered according to the linear order $\leq_T$ on T. The root of U_i is identified with the i-th leave of T in the resulting tree. Here is the Ramsey statement formulated by Moore.

For every m there exists $n \geq m$ such that for each coloring $c\colon \mathbb{T}_n \to \{0,1\}$ there exist non-negative numbers $\alpha_{\vec{U}}$, where $\vec{U}$ ranges over all m-tuples $\vec{U} = (U_1, \ldots, U_m)$ of binary trees with a total of n leaves, such that $\sum_{\vec{U}} \alpha_{\vec{U}} = 1$ and

$$\sum_{\vec{U}} \alpha_{\vec{U}} c(T(\vec{U}))$$

is constant as T varies over $\mathbb{T}_m$.

Theorem 4.1 (Moore [19]). *The above Ramsey statement is equivalent to amenability of Thompson's group F.*

More broadly, Moore's analysis of amenability parallel to the analysis of extreme amenability for non-Archimedean groups lead him to a general class of Ramsey statements phrased in terms of convex combinations. At this point, no statements of this form appear to be known that do not follow from ordinary Ramsey statements.

There is another Ramsey statement that seems to fit here. It was formulated by Kechris, Sokić and Todorcevic [16], and was motivated by the desire to give a Ramsey theoretic proof of the theorem of Giordano–Pestov [6] that the group of all measure preserving transformations of the interval $[0,1]$ with Lebesgue measure, taken with the weak topology, is extremely amenable. The original proof in [6] used concentration of measure. What appears to be a minor modification of the Graham–Rothschild theorem, Theorem 3.3 above, yields a Ramsey statement that would imply Giordano–Pestov's result. The statement, which was formulated in [16] and which we reproduce below, is not known to be true.

We say that a partition $\mathcal{Q}$ of a finite set X is **homogeneous** if any two sets in $\mathcal{Q}$ contain the same number of elements of X.

Given a positive integer c, for each k and l there exists m such that for each coloring with c colors of all homogeneous k-partitions of $[m]$ there exists a homogeneous l-partition $\mathcal{Q}$ of $[m]$ such that all homogeneous k-subpartitions of $\mathcal{Q}$ get the same color.

Acknowledgements. This research was supported by NSF grant DMS-1266189.

References

[1] F.G. Abramson and L.A. Harrington, *Models without indiscernibles*, J. Symb. Logic **43** (1978), 572–600.

[2] W. Deuber, *A generalization of Ramsey's theorem for regular trees*, J. Combin. Theory, Ser. B **18** (1975), 18–23.

[3] M. Droste and R. Göbel, *Universal domains and the amalgamation property*, Math. Structures Comput. Sci. **3** (1993), 137–159.

[4] I. Farah and S. Solecki, *Extreme amenability of L_0, a Ramsey theorem, and Lévy groups*, J. Funct. Anal. **255** (2008), 471–493.

[5] G. Gierz, K.H. Hofmann, K. Keimel, J.D. Lawson, M. Mislove, and D.S. Scott, *Continuous Lattices and Domains*, Encyclopedia of Mathematics and its Applications, **93**, Cambridge University Press, 2003.

[6] T. Giordano and V. Pestov, *Some extremely amenable groups*, C. R. Math. Acad. Sci. Paris **334** (2002), 273–278.

[7] E. Glasner and B. Weiss, *Minimal actions of the group $\mathbb{S}(\mathbb{Z})$ of permutations of the integers*, Geom. Funct. Anal. **12** (2002), 964–988.

[8] ———, *The universal minimal system for the group of homeomorphisms of the Cantor set*, Fund. Math. **176** (2003), 277–289.

[9] R.L. Graham, B.L. Rothschild, *Ramsey's theorem for n-parameter sets*, Trans. Amer. Math. Soc. **159** (1971), 257–292.

[10] ———, *Some recent developments in Ramsey theory*, in Combinatorics, Mathematical Centre, Amsterdam, 1975, pp. 261–276.

[11] M. Gromov, *Number of questions*, preprint 2014.

[12] M. Gromov and V.D. Milman, *A topological application of the isoperimetric inequality*, Amer. J. Math. **105** (1983), 843–854.

[13] W. Herer and J.P.R. Christensen, *On the existence of pathological submeasures and the construction of exotic topological groups*, Math. Ann. **213** (1975), 203–210.

[14] J. Jasiński, *Ramsey degrees of boron tree structures*, Combinatorica **33** (2013), 23–44.

[15] A.S. Kechris, V.G. Pestov, and S. Todorcevic, *Fraissé limits, Ramsey theory, and topological dynamics of automorphism groups*, Geom. Funct. Anal. **15** (2005), 106–189.

[16] A.S. Kechris, M. Sokić, and S. Todorcevic, *Ramsey properties of finite measure algebras and topological dynamics of the group of measure preserving automorphisms: some results and an open problem*, preprint, 2012.

[17] J. Matoušek, *Using the Borsuk-Ulam theorem*, Lectures on topological methods in combinatorics and geometry, Universitext, Springer-Verlag, 2003.

[18] K. Milliken, *A Ramsey theorem for tress*, J. Combin. Theory, Ser. A **26** (1979), 137–148.

[19] J.T. Moore, *Amenability and Ramsey theory*, Fund. Math. **220** (2013), 263–280.

[20] J. Nešetřil, *Ramsey theory*, in Handbook of Combinatorics, eds. R. Graham, M. Grötschel, L. Lovász, Elsevier Science, 1995, pp. 1331–1403.

[21] ———, *Metric spaces are Ramsey*, European J. Combin. **28** (2007), 457–468.

[22] J. Nešetřil and V. Rödl, *Partition for relational and set systems*, J. Combin. Theory Ser. A **22** (1977), 289–312.

[23] ———, *Ramsey classes of set systems*, J. Combin. Theory Ser. A **34** (1983), 183–201.

[24] ———, *The partite construction and Ramsey systems*, Discrete Math. **74** (1989), 327–334.

[25] L. Nguyen Van The, *More on the Kechris-Pestov-Todorcevic correspondence: precompact expansions*, Fund. Math. **222** (2013), 19–47.

[26] O. Ore, *Galois connexions*, Trans. Amer. Math. Soc. **55** (1944), 493–513.

[27] V.G. Pestov, *On free actions, minimal flows, and a problem by Ellis*, Trans. Amer. Math. Soc. **350** (1998), 4149–4165.

[28] ———, *Ramsey - Milman phenomenon, Urysohn metric spaces, and extremely amenable groups*, Israel J. Math. **127** (2002), 317–357; *A corrigendum to: "Ramsey-Milman phenomenon, Urysohn metric spaces, and extremely amenable groups"*, Israel J. Math. **145** (2005), 375–379.

[29] H.J. Prömel, *Induced partition properties of combinatorial cubes*, J. Combin. Theory Ser. A **39** (1985), 177–208.

[30] F.P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. **30** (1930), 264–286.

[31] M. Sokić, *Bounds on trees*, Discrete Math. **311** (2011), 398–407.

[32] ———, *Ramsey property, ultrametric spaces, finite posets, and universal minimal flows*, Israel J. Math. **194** (2013), 609–640.

[33] S. Solecki, *A Ramsey theorem for structures with both relations and functions*, J. Combin. Theory, Ser. A **117** (2010), 704–714.

[34] ———, *Direct Ramsey theorem for structures involving relations and functions*, J. Combin. Theory, Ser. A **119** (2012), 440–449.

[35] ———, *Abstract approach to finite Ramsey theory and a self-dual Ramsey theorem*, Adv. Math. **248** (2013), 1156–1198.

[36] ———, *Abstract approach to Ramsey theory and Ramsey theorems for finite trees*, in Asymptotic Geometric Analysis, Fields Institute Communications, Springer, 2013, pp. 313–340.

[37] ———, *Dual Ramsey theorem for trees*, preprint 2014.

[38] J. Spencer, *Ramsey's theorem for spaces*, Trans. Amer. Math. Soc. **249** (1979), 363–371.

[39] S. Todorcevic, *Introduction to Ramsey Spaces*, Annals of Mathematics Studies, **174**, Princeton University Press, 2010.

[40] W. Veech, *Topological dynamics*, Bull. Amer. Math. Soc. **83** (1977), 775–830.

[41] B. Voigt, *The partition problem for finite Abelian groups*, J. Combin. Theory, Ser. A **28** (1980), 257–271.

[42] M. Zhao, *A self-dual Ramsey theorem for parameter systems*, preprint 2013.

Department of Mathematics, University of Illinois, 1409 W. Green St., Urbana, IL 61801, USA
E-mail: ssolecki@math.uiuc.edu

2. Algebra

On finite-dimensional Hopf algebras

Dedicado a Biblioco 34

Nicolás Andruskiewitsch

Abstract. This is a survey on the state-of-the-art of the classification of finite-dimensional complex Hopf algebras. This general question is addressed through the consideration of different classes of such Hopf algebras. Pointed Hopf algebras constitute the class best understood; the classification of those with abelian group is expected to be completed soon and there is substantial progress in the non-abelian case.

Mathematics Subject Classification (2010). 16T05, 16T20, 17B37, 16T25, 20G42.

Keywords. Hopf algebras, quantum groups, Nichols algebras.

1. Introduction

Hopf algebras were introduced in the 1950's from three different perspectives: algebraic groups in positive characteristic, cohomology rings of Lie groups, and group objects in the category of von Neumann algebras. The study of non-commutative non-cocommutative Hopf algebras started in the 1960's. The fundamental breakthrough is Drinfeld's report [25]. Among many contributions and ideas, a systematic construction of solutions of the quantum Yang-Baxter equation (qYBE) was presented. Let V be a vector space. The qYBE is equivalent to the braid equation:

$$(c \otimes \mathrm{id})(\mathrm{id} \otimes c)(c \otimes \mathrm{id}) = (\mathrm{id} \otimes c)(c \otimes \mathrm{id})(\mathrm{id} \otimes c), \qquad c \in GL(V \otimes V). \tag{1.1}$$

If c satisfies (1.1), then (V, c) is called a braided vector space; this is a down-to-the-earth version of a braided tensor category [54]. Drinfeld introduced the notion of quasi-triangular Hopf algebra, meaning a pair (H, R) where H is a Hopf algebra and $R \in H \otimes H$ is invertible and satisfies the appropppriate conditions, so that every H-module V becomes a braided vector space, with c given by the action of R composed with the usual flip. Furthermore, every finite-dimensional Hopf algebra H gives rise to a quasi-triangular Hopf algebra, namely the Drinfeld double $D(H) = H \otimes H^*$ as vector space. If H is not finite-dimensional, some precautions have to be taken to construct $D(H)$, or else one considers Yetter-Drinfeld modules, see §2.2. In conclusion, every Hopf algebra is a source of solutions of the braid equation. Essential examples of quasi-triangular Hopf algebras are the quantum groups $U_q(\mathfrak{g})$ [25, 53] and the finite-dimensional variations $\mathfrak{u}_q(\mathfrak{g})$ [59, 60].

In the approach to the classification of Hopf algebras exposed in this report, braided vector spaces and braided tensor categories play a decisive role; and the finite quantum groups are the main actors in one of the classes that splits off.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

By space limitations, there is a selection of the topics and references included. Particularly, we deal with finite-dimensional Hopf algebras over an algebraically closed field of characteristic zero with special emphasis on description of examples and classifications. Interesting results on Hopf algebras either infinite-dimensional, or over other fields, unfortunately can not be reported. There is no account of the many deep results on tensor categories, see [30]. Various basic fundamental results are not explicitly cited, we refer to [1, 62, 66, 75, 79, 83] for them; classifications of Hopf algebras of fixed dimensions are not evoked, see [21, 71, 86].

2. Preliminaries

Let $\theta \in \mathbb{N}$ and $\mathbb{I} = \mathbb{I}_\theta = \{1, 2, \dots, \theta\}$. The base field is $\mathbb{C}$. If X is a set, then $|X|$ is its cardinal and $\mathbb{C}X$ is the vector space with basis $(x_i)_{i\in X}$. Let G be a group: we denote by $\operatorname{Irr} G$ the set of isomorphism classes of irreducible representations of G and by $\widehat{G}$ the subset of those of dimension 1; by G^x the centralizer of $x \in G$; and by $\mathcal{O}_x^G$ its conjugacy class. More generally we denote by $\operatorname{Irr} \mathfrak{C}$ the set of isomorphism classes of simple objects in an abelian category $\mathfrak{C}$. The group of n-th roots of 1 in $\mathbb{C}$ is denoted $\mathbb{G}_n$; also $\mathbb{G}_\infty = \bigcup_{n\geq 1} \mathbb{G}_n$. The group presented by $(x_i)_{i\in I}$ with relations $(r_j)_{j\in J}$ is denoted $\langle (x_i)_{i\in I} | (r_j)_{j\in J}\rangle$. The notation for Hopf algebras is standard: Δ, ε, $\mathcal{S}$, denote respectively the comultiplication, the counit, the antipode (always assumed bijective, what happens in the finite-dimensional case). We use Sweedler's notation: $\Delta(x) = x_{(1)} \otimes x_{(2)}$. Similarly, if C is a coalgebra and V is a left comodule with structure map $\delta : V \to C \otimes V$, then $\delta(v) = v_{(-1)} \otimes v_{(0)}$. If D, E are subspaces of C, then $D \wedge E = \{c \in C : \Delta(c) \in D \otimes C + C \otimes E\}$; also $\wedge^0 D = D$ and $\wedge^{n+1} D = (\wedge^n D) \wedge D$ for $n > 0$.

2.1. Basic constructions and results. The first examples of finite-dimensional Hopf algebras are the group algebra $\mathbb{C}G$ of a finite group G and its dual, the algebra of functions $\mathbb{C}^G$. Indeed, the dual of a finite-dimensional Hopf algebra is again a Hopf algebra by transposing operations. By analogy with groups, several authors explored the notion of extension of Hopf algebras at various levels of generality; in the finite-dimensional context, every extension $\mathbb{C} \to A \to C \to B \to \mathbb{C}$ can be described as C with underlying vector space $A \otimes B$, via a heavy machinery of actions, coactions and non-abelian cocycles, but actual examples are rarely found in this way (extensions from a different perspective are in [9]). Relevant exceptions are the so-called *abelian extensions* [56] (rediscovered by Takeuchi and Majid): here the input is a matched pair of groups (F, G) with mutual actions $\triangleright, \triangleleft$ (or equivalently, an exact factorization of a finite group). The actions give rise to a Hopf algebra $\mathbb{C}^G \# \mathbb{C}F$. The multiplication and comultiplication can be further modified by compatible cocycles (σ, κ), producing to the abelian extension $\mathbb{C} \to \mathbb{C}^G \to \mathbb{C}^G {}^\kappa\#_\sigma \mathbb{C}F \to \mathbb{C}F \to \mathbb{C}$. Here (σ, κ) turns out to be a 2-cocycle in the total complex associated to a double complex built from the matched pair; the relevant H^2 is computed via the so-called Kac exact sequence.

It is natural to approach Hopf algebras by considering algebra or coalgebra invariants. There is no preference in the finite-dimensional setting but coalgebras and comodules are locally finite, so we privilege the coalgebra ones to lay down general methods. The basic coalgebra invariants of a Hopf algebra H are:

- The group $G(H) = \{g \in H - 0 : \Delta(g) = x \otimes g\}$ of group-like elements of H.

- The space of skew-primitive elements $\mathcal{P}_{g,h}(H)$, $g, h \in G(H)$; $\mathcal{P}(H) := \mathcal{P}_{1,1}(H)$.
- The coradical H_0, that is the sum of all simple subcoalgebras.
- The coradical filtration $H_0 \subset H_1 \subset \dots$, where $H_n = \wedge^n H_0$; then $H = \bigcup_{n\geq 0} H_n$.

2.2. Modules. The category ${}_H\mathcal{M}$ of left modules over a Hopf algebra H is monoidal with tensor product defined by the comultiplication; ditto for the category ${}^H\mathcal{M}$ of left comodules, with tensor product defined by the multiplication. Here are two ways to deform Hopf algebras without altering one of these categories.

- Let $F \in H\otimes H$ be invertible such that $(1\otimes F)(\mathrm{id}\otimes\Delta)(F) = (F\otimes 1)(\Delta\otimes\mathrm{id})(F)$ and $(\mathrm{id}\otimes\varepsilon)(F) = (\varepsilon\otimes\mathrm{id})(F) = 1$. Then H^F (the same algebra with comultiplication $\Delta^F := F\Delta F^{-1}$) is again a Hopf algebra, named the *twisting* of H by F [26]. The monoidal categories ${}_H\mathcal{M}$ and ${}_{H^F}\mathcal{M}$ are equivalent. If H and K are finite-dimensional Hopf algebras with ${}_H\mathcal{M}$ and ${}_K\mathcal{M}$ equivalent as monoidal categories, then there exists F with $K \simeq H^F$ as Hopf algebras (Schauenburg, Etingof-Gelaki). Examples of twistings not mentioned elsewhere in this report are in [31, 65].
- Given a linear map $\sigma : H\otimes H \to \mathbb{C}$ with analogous conditions, there is a Hopf algebra H_σ (same coalgebra, multiplication twisted by σ) such that the monoidal categories ${}^H\mathcal{M}$ and ${}^{H_\sigma}\mathcal{M}$ are equivalent [24].

A Yetter-Drinfeld module M over H is left H-module and left H-comodule with the compatibility $\delta(h.m) = h_{(1)}m_{(-1)}\mathcal{S}(h_{(3)})\otimes h_{(2)}\cdot m_{(0)}$, for all $m \in M$ and $h \in H$. The category ${}^H_H\mathcal{YD}$ of Yetter-Drinfeld modules is braided monoidal. That is, for every $M, N \in {}^H_H\mathcal{YD}$, there is a natural isomorphism $c : M\otimes N \to N\otimes M$ given by $c(m\otimes n) = m_{(-1)}\cdot n\otimes m_{(0)}$, $m \in M$, $n \in N$. When H is finite-dimensional, the category ${}^H_H\mathcal{YD}$ is equivalent, as a braided monoidal category, to ${}_{D(H)}\mathcal{M}$.

The definition of Hopf algebra makes sense in any braided monoidal category. Hopf algebras in ${}^H_H\mathcal{YD}$ are interesting because of the following facts–discovered by Radford and interpreted categorically by Majid, see [62, 75]:

- If R is a Hopf algebra in ${}^H_H\mathcal{YD}$, then $R\#H := R\otimes H$ with semidirect product and coproduct is a Hopf algebra, named the *bosonization* of R by H.
- Let π, ι be Hopf algebra maps as in $K \underset{\pi}{\overset{\iota}{\rightleftarrows}} H$ with $\pi\iota = \mathrm{id}_H$. Then $R = H^{\mathrm{co}\pi} := \{x \in K : (\mathrm{id}\otimes\pi)\Delta(x) = 1\otimes x\}$ is a Hopf algebra in ${}^H_H\mathcal{YD}$ and $K \simeq R\#H$.

For instance, if $V \in {}^H_H\mathcal{YD}$, then the tensor algebra $T(V)$ is a Hopf algebra in ${}^H_H\mathcal{YD}$, by requiring $V \hookrightarrow \mathcal{P}(T(V))$. If $c : V\otimes V \to V\otimes V$ satisfies $c = -\tau$, τ the usual flip, then the exterior algebra $\Lambda(V)$ is a Hopf algebra in ${}^H_H\mathcal{YD}$.

There is a braided adjoint action of a Hopf algebra R in ${}^H_H\mathcal{YD}$ on itself, see e.g. [12, (1.26)]. If $x \in \mathcal{P}(R)$ and $y \in R$, then $\mathrm{ad}_c(x)(y) = xy - \mathrm{mult}\, c(x\otimes y)$.

2.2.1. Triangular Hopf algebras. A quasitriangular Hopf algebra (H, R) is *triangular* if the braiding induced by R is a symmetry: $c_{V\otimes W}c_{W\otimes V} = \mathrm{id}_{W\otimes V}$ for all $V, W \in {}_H\mathcal{M}$. A finite-dimensional triangular Hopf algebra is a twisting of a bosonization $\Lambda(V)\#\mathbb{C}G$, where G is a finite group and $V \in {}^G_G\mathcal{YD}$ has $c = -\tau$ [6]. This lead eventually to the classification of triangular finite-dimensional Hopf algebras [29]; previous work on the semisimple case culminated in [28].

2.3. Semisimple Hopf algebras. The algebra of functions $\mathbb{C}^G$ on a finite group G admits a Haar measure, i.e., a linear function $\int : \mathbb{C}^G \to \mathbb{C}$ invariant under left and right translations, namely $\int =$ sum of all elements in the standard basis of $\mathbb{C}G$. This is adapted as follows: a right integral *on* a Hopf algebra H is a linear function $\int : H \to \mathbb{C}$ which is invariant under the left regular coaction: analogously there is the notion of left integral. The notion has various applications. Assume that H is finite-dimensional. Then an integral *in* H is an integral on H^*; the subspace of left integrals in H has dimension one, and there is a generalization of Maschke's theorem for finite groups: H is semisimple if and only if $\varepsilon(\Lambda) \neq 0$ for any integral $0 \neq \Lambda \in H$. This characterization of semisimple Hopf algebras, valid in any characteristic, is one of several, some valid only in characteristic 0. See [79]. Semisimple Hopf algebras can be obtained as follows:

- ◇ A finite-dimensional Hopf algebra H is semisimple if and only if it is cosemisimple (that is, H^* is semisimple).
- ◇ Given an extension $\mathbb{C} \to K \to H \to L \to \mathbb{C}$, H is semisimple iff K and L are. Notice that there are semisimple extensions that are not abelian [40, 69, 74].
- ◇ If H is semisimple, then so are H^F and H_σ, for any twist F and cocycle σ. If G is a finite simple group, then any twisting of $\mathbb{C}G$ is a simple Hopf algebra (i.e., not a non-trivial extension) [73], but the converse is not true [37].
- ◇ A bosonization $R\#H$ is semisimple iff R and H are.

To my knowledge, all examples of semisimple Hopf algebras arise from group algebras by the preceding constructions; this was proved in [68, 70] in low dimensions and in [32] for dimensions $p^a q^b$, pqr, where p, q and r are primes. See [1, Question 2.6]. An analogous question in terms of fusion categories: is any semisimple Hopf algebra weakly group-theoretical? See [32, Question 2].

There are only finitely many isomorphism classes of *semisimple* Hopf algebras in each dimension [81], but this fails in general [13, 20].

Conjecture 2.1 (Kaplansky). *Let H be a semisimple Hopf algebra. The dimension of every $V \in \operatorname{Irr}{}_H\mathcal{M}$ divides the dimension of H.*

The answer is affirmative for iterated extensions of group algebras and duals of group algebras [67] and notably for semisimple quasitriangular Hopf algebras [27].

3. Lifting methods

3.1. Nichols algebras. Nichols algebras are a special kind of Hopf algebras in braided tensor categories. We are mainly interested in Nichols algebras in the braided category ${}^H_H\mathcal{YD}$, where H is a Hopf algebra, see page 3. In fact, there is a functor $V \mapsto \mathfrak{B}(V)$ from ${}^H_H\mathcal{YD}$ to the category of Hopf algebras in ${}^H_H\mathcal{YD}$. Their first appearence is in the precursor [72]; they were rediscovered in [85] as part of a "quantum differential calculus", and in [61] to present the positive part of $U_q(\mathfrak{g})$. See also [76, 78].

There are several, unrelated at the first glance, alternative definitions. Let $V \in {}^H_H\mathcal{YD}$. The first definition uses the representation of the braid group $\mathbb{B}_n$ in n strands on $V^{\otimes n}$, given by $\varsigma_i \mapsto \mathrm{id}^{\otimes}{}^{n-1} \otimes c \otimes \mathrm{id}^{\otimes}{}^{n-i-1}$, c in $(i, i+1)$ tensorands; here recall that

$$\mathbb{B}_n = \langle \varsigma_1, \dots, \varsigma_{n-1} | \varsigma_i\varsigma_j = \varsigma_j\varsigma_i, |i-j| > 1, \varsigma_i\varsigma_j\varsigma_i = \varsigma_j\varsigma_i\varsigma_j, |i-j| = 1 \rangle.$$

Let $M : \mathbb{S}_n \to \mathbb{B}_n$ be the Matsumoto section and let $\mathcal{Q}_n : V^{\otimes n} \to V^{\otimes n}$ be the quantum symmetrizer, $\mathcal{Q}_n = \sum_{s\in\mathbb{S}_n} M(s) : V^{\otimes n} \to V^{\otimes n}$. Then define

$$\mathfrak{J}^n(V) = \ker \mathcal{Q}_n, \qquad \mathfrak{J}(V) = \oplus_{n\geq 2}\mathfrak{J}^n(V), \qquad \mathfrak{B}(V) = T(V)/\mathfrak{J}(V). \tag{3.1}$$

Hence $\mathfrak{B}(V) = \oplus_{n\geq 0}\mathfrak{B}^n(V)$ is a graded Hopf algebra in ${}^H_H\mathcal{YD}$ with $\mathfrak{B}^0(V) = \mathbb{C}$, $\mathfrak{B}^1(V) \simeq V$; by (3.1) the algebra structure depends only on c. To explain the second definition, let us observe that the tensor algebra $T(V)$ is a Hopf algebra in ${}^H_H\mathcal{YD}$ with comultiplication determined by $\Delta(v) = v\otimes 1 + 1\otimes v$ for $v \in V$. Then $\mathfrak{J}(V)$ coincides with the largest homogeneous ideal of $T(V)$ generated by elements of degree ≥ 2 that is also a coideal. Let now $T = \oplus_{n\geq 0}T^n$ be a graded Hopf algebra in ${}^H_H\mathcal{YD}$ with $T^0 = \mathbb{C}$. Consider the conditions

$$T^1 \text{ generates } T \text{ as an algebra,} \tag{3.2}$$

$$T^1 = \mathcal{P}(T). \tag{3.3}$$

These requirements are dual to each other: if T has finite-dimensional homogeneous components and $R = \oplus_{n\geq 0}R^n$ is the graded dual of T, i.e., $R^n = (T^n)^*$, then T satisfies (3.2) if and only if R satisfies (3.3). These conditions determine $\mathfrak{B}(V)$ up to isomorphisms, as the unique graded connected Hopf algebra T in ${}^H_H\mathcal{YD}$ that satisfies $T^1 \simeq V$, (3.2) and (3.3). There are still other characterizations of $\mathfrak{J}(V)$, e.g. as the radical of a suitable homogeneous bilinear form on $T(V)$, or as the common kernel of some suitable skew-derivations. See [15] for more details.

Despite all these different definitions, Nichols algebras are extremely difficult to deal with, e.g. to present by generators and relations, or to determine when a Nichols algebra has finite dimension or finite Gelfand-Kirillov dimension. It is not even known a priori whether the ideal $\mathfrak{J}(V)$ is finitely generated, except in a few specific cases. For instance, if c is a symmetry, that is $c^2 = \mathrm{id}$, or satisfies a Hecke condition with generic parameter, then $\mathfrak{B}(V)$ is quadratic. By the efforts of various authors, we have some understanding of finite-dimensional Nichols algebras of braided vector spaces either of diagonal or of rack type, see §3.5, 3.6.

3.2. Hopf algebras with the (dual) Chevalley property. We now explain how Nichols algebras enter into our approach to the classification of Hopf algebras. Recall that a Hopf algebra has the *dual Chevalley property* if the tensor product of two simple comodules is semisimple, or equivalently if its coradical is a (cosemisimple) Hopf subalgebra. For instance, a *pointed* Hopf algebra, one whose simple comodules have all dimension one, has the dual Chevalley property and its coradical is a group algebra. Also, a *copointed* Hopf algebra (one whose coradical is the algebra of functions on a finite group) has the dual Chevalley property. The Lifting Method is formulated in this context [13]. Let H be a Hopf algebra with the dual Chevalley property and set $K := H_0$. Under this assumption, the graded coalgebra $\operatorname{gr} H = \oplus_{n\in\mathbb{N}_0} \operatorname{gr}^n H$ associated to the coradical filtration becomes a Hopf algebra and considering the homogeneous projection π as in $R = H^{\operatorname{co}\pi} \hookrightarrow \operatorname{gr} H \underset{\pi}{\rightleftarrows} K$ we see that $\operatorname{gr} H \simeq R\#K$. The subalgebra of coinvariants R is a graded Hopf algebra in ${}^K_K\mathcal{YD}$ that inherits the grading with $R^0 = \mathbb{C}$; it satisfies (3.3) since the grading comes from the coradical filtration. Let R' be the subalgebra of R generated by $V := R^1$; then $R' \simeq \mathfrak{B}(V)$. The braided vector space V is a basic invariant of H called its *infinitesimal braiding*. Let us fix then a semisimple Hopf algebra K. To classify all finite-dimensional Hopf algebras H with $H_0 \simeq K$ as Hopf algebras, we have to address the following questions.

(a) Determine those $V \in {}^K_K\mathcal{YD}$ such that $\mathfrak{B}(V)$ is finite-dimensional, and give an efficient defining set of relations of these.
(b) Investigate whether any *finite-dimensional* graded Hopf algebra R in ${}^K_K\mathcal{YD}$ satisfying $R^0 = \mathbb{C}$ and $P(R) = R^1$, is a Nichols algebra.
(c) Compute all Hopf algebras H such that $\operatorname{gr} H \simeq \mathfrak{B}(V)\#K$, V as in (a).

Since the Nichols algebra $\mathfrak{B}(V)$ depends as an algebra (and as a coalgebra) only on the braiding c, it is convenient to restate Question (a) as follows:

(a_1) Determine those braided vector spaces (V,c) in a suitable class such that $\dim \mathfrak{B}(V) < \infty$, and give an efficient defining set of relations of these.
(a_2) For those V as in (a_1), find in how many ways, if any, they can be realized as Yetter-Drinfeld modules over K.

For instance, if $K = \mathbb{C}\Gamma$, Γ a finite abelian group, then the suitable class is that of braided vector spaces of diagonal type. In this context, Question (a_2) amounts to solve systems of equations in Γ. The answer to (a) is instrumental to attack (b) and (c). Question (b) can be rephrased in two equivalent statements:

(b_1) Investigate whether any *finite-dimensional* graded Hopf algebra T in ${}^K_K\mathcal{YD}$ with $T^0 = \mathbb{C}$ and generated as algebra by T^1, is a Nichols algebra.
(b_2) Investigate whether any *finite-dimensional* Hopf algebra H with $H_0 = K$ is generated as algebra by H_1.

We believe that the answer to (b) is affirmative at least when K is a group algebra. In other words, by the reformulation (b_2):

Conjecture 3.1 ([14]). *Every finite-dimensional pointed Hopf algebra is generated by group-like and skew-primitive elements.*

As we shall see in §3.7, the complete answer to (a) is needed in the approach proposed in [14] to attack Conjecture 3.1. It is plausible that the answer of (b_2) is affirmative for every semisimple Hopf algebra K. Question (c), known as *lifting* of the relations, also requires the knowledge of the generators of $\mathfrak{J}(V)$, see §3.8.

3.3. Generalized lifting method. Before starting with the analysis of the various questions in §3.2, we discuss a possible approach to more general Hopf algebras [5]. Let H be a Hopf algebra; we consider the following invariants of H:

- The *Hopf coradical* $H_{[0]}$ is the subalgebra generated by H_0.
- The *standard filtration* $H_{[0]} \subset H_{[1]} \subset \dots$, $H_{[n]} = \wedge^{n+1} H_{[0]}$; then $H = \bigcup_{n\geq 0} H_{[n]}$.

If H has the dual Chevalley property, then $H_{[n]} = H_n$ for all $n \in \mathbb{N}_0$. In general, $H_{[0]}$ is a Hopf subalgebra of H with coradical H_0 and we may consider the graded Hopf algebra $\operatorname{gr} H = \oplus_{n\geq 0} H_{[n]}/H_{[n-1]}$. As before, if $\pi : \operatorname{gr} H \to H_{[0]}$ is the homogeneous projection, then $R = (\operatorname{gr} H)^{\operatorname{co}\pi}$ is a Hopf algebra in ${}^{H_{[0]}}_{H_{[0]}}\mathcal{YD}$ and $\operatorname{gr} H \cong R\#H_{[0]}$. Furthermore, $R = \oplus_{n\geq 0} R^n$ with grading inherited from $\operatorname{gr} H$. This discussion raises the following questions.

(A) Let C be a finite-dimensional cosemisimple coalgebra and $S : C \to C$ a bijective anti-coalgebra map. Classify all finite-dimensional Hopf algebras L generated by C, such that $\mathcal{S}|_C = S$.

(B) Given L as in the previous item, classify all finite-dimensional connected graded Hopf algebras R in ${}^L_L\mathcal{YD}$.

(C) Given L and R as in previous items, classify all deformations or liftings, that is, classify all Hopf algebras H such that $\operatorname{gr} H \cong R\#L$.

Question (A) is largely open, except for the remarkable [82, Theorem 1.5]: if H is a Hopf algebra generated by an $\mathcal{S}$-invariant 4-dimensional simple subcoalgebra C, such that $1 < \operatorname{ord}(\mathcal{S}^2|_C) < \infty$, then H is a Hopf algebra quotient of the quantized algebra of functions on SL_2 at a root of unity ω. Nichols algebras enter into the picture in Question (B); if $V = R^1$, then $\mathfrak{B}(V)$ is a subquotient of R. Question (C) is completely open, as it depends on the previous Questions.

3.4. Generalized root systems and Weyl groupoids. Here we expose two important notions introduced in [51].

Let $\theta \in \mathbb{N}$ and $\mathbb{I} = \mathbb{I}_\theta$. A *basic datum* of type $\mathbb{I}$ is a pair $(\mathcal{X}, \rho)$, where $\mathcal{X} \neq \emptyset$ is a set and $\rho : \mathbb{I} \to \mathbb{S}_\mathcal{X}$ is a map such that $\rho_i^2 = \mathrm{id}$ for all $i \in \mathbb{I}$. Let $\mathcal{Q}_\rho$ be the quiver $\{\sigma_i^x := (x, i, \rho_i(x)) : i \in \mathbb{I}, x \in \mathcal{X}\}$ over $\mathcal{X}$, with $t(\sigma_i^x) = x$, $s(\sigma_i^x) = \rho_i(x)$ (here t means target, s means source). Let $F(\mathcal{Q}_\rho)$ be the free groupoid over $\mathcal{Q}_\rho$; in any quotient of $F(\mathcal{Q}_\rho)$, we denote

$$\sigma_{i_1}^x \sigma_{i_2} \cdots \sigma_{i_t} = \sigma_{i_1}^x \sigma_{i_2}^{\rho_{i_1}(x)} \cdots \sigma_{i_t}^{\rho_{i_{t-1}} \cdots \rho_{i_1}(x)}; \tag{3.4}$$

i.e., the implicit superscripts are those allowing compositions.

3.4.1. Coxeter groupoids. A *Coxeter datum* is a triple $(\mathcal{X}, \rho, \mathbf{M})$, where $(\mathcal{X}, \rho)$ is a basic datum of type $\mathbb{I}$ and $\mathbf{M} = (\mathbf{m}^x)_{x \in \mathcal{X}}$ is a family of Coxeter matrices $\mathbf{m}^x = (m_{ij}^x)_{i,j \in \mathbb{I}}$ with

$$s((\sigma_i^x \sigma_j)^{m_{ij}^x}) = x, \qquad i, j \in \mathbb{I}, \qquad x \in \mathcal{X}. \tag{3.5}$$

The *Coxeter groupoid* $\mathcal{W}(\mathcal{X}, \rho, \mathbf{M})$ associated to $(\mathcal{X}, \rho, \mathbf{M})$ [51, Definition 1] is the groupoid presented by generators $\mathcal{Q}_\rho$ with relations

$$(\sigma_i^x \sigma_j)^{m_{ij}^x} = \mathrm{id}_x, \qquad i, j \in \mathbb{I}, x \in \mathcal{X}. \tag{3.6}$$

3.4.2. Generalized root system. A generalized root system (GRS for short) is a collection $\mathcal{R} := (\mathcal{X}, \rho, \mathcal{C}, \Delta)$, where $\mathcal{C} = (C^x)_{x \in \mathcal{X}}$ is a family of generalized Cartan matrices $C^x = (c_{ij}^x)_{i,j \in \mathbb{I}}$, cf. [57], and $\Delta = (\Delta^x)_{x \in \mathcal{X}}$ is a family of subsets $\Delta^x \subset \mathbb{Z}^\mathbb{I}$. We need the following notation: Let $\{\alpha_i\}_{i \in \mathbb{I}}$ be the canonical basis of $\mathbb{Z}^\mathbb{I}$ and define $s_i^x \in GL(\mathbb{Z}^\mathbb{I})$ by $s_i^x(\alpha_j) = \alpha_j - c_{ij}^x \alpha_i$, $i, j \in \mathbb{I}$, $x \in \mathcal{X}$. The collection should satisfy the following axioms:

$$c_{ij}^x = c_{ij}^{\rho_i(x)} \qquad \text{for all } x \in \mathcal{X},\ i, j \in \mathbb{I}. \tag{3.7}$$

$$\Delta^x = \Delta_+^x \cup \Delta_-^x, \qquad \Delta_\pm^x := \pm(\Delta^x \cap \mathbb{N}_0^\mathbb{I}) \subset \pm\mathbb{N}_0^\mathbb{I}; \tag{3.8}$$

$$\Delta^x \cap \mathbb{Z}\alpha_i = \{\pm\alpha_i\}; \tag{3.9}$$

$$s_i^x(\Delta^x) = \Delta^{\rho_i(x)}; \tag{3.10}$$

$$(\rho_i\rho_j)^{m_{ij}^x}(x) = (x), \qquad m_{ij}^x := |\Delta^x \cap (\mathbb{N}_0\alpha_i + \mathbb{N}_0\alpha_j)|, \tag{3.11}$$

for all $x \in \mathcal{X}$, $i \neq j \in \mathbb{I}$. We call Δ_+^x, respectively Δ_-^x, the set of *positive*, respectively *negative*, roots. Let $\mathcal{G} = \mathcal{X} \times GL_\theta(\mathbb{Z}) \times \mathcal{X}$, $\varsigma_i^x = (x,\, s_i^x,\, \rho_i(x))$, $i \in \mathbb{I}$, $x \in \mathcal{X}$, and $\mathcal{W} = \mathcal{W}(\mathcal{X}, \rho, \mathcal{C})$ the subgroupoid of $\mathcal{G}$ generated by all the ς_i^x, i.e., by the image of the morphism of quivers $\mathcal{Q}_\rho \to \mathcal{G}$, $\sigma_i^x \mapsto \varsigma_i^x$. There is a Coxeter matrix $\mathbf{m}^x = (m_{ij}^x)_{i,j\in\mathbb{I}}$, where m_{ij}^x is the smallest natural number such that $(\varsigma_i^x\varsigma_j)^{m_{ij}^x} = \mathrm{id}\, x$. Then $\mathbf{M} = (\mathbf{m}^x)_{x\in\mathcal{X}}$ fits into a Coxeter datum $(\mathcal{X}, \rho, \mathbf{M})$, and there is an isomorphism of groupoids $\mathcal{W}(\mathcal{X}, \rho, \mathbf{M}) \twoheadrightarrow \mathcal{W} = \mathcal{W}(\mathcal{X}, \rho, \mathcal{C})$ [51]; this is called the *Weyl groupoid* of $\mathcal{R}$. If $w \in \mathcal{W}(x, y)$, then $w(\Delta^x) = \Delta^y$, by (3.10). The sets of *real* roots at $x \in \mathcal{X}$ are $(\Delta^{\mathrm{re}})^x = \bigcup_{y\in\mathcal{X}}\{w(\alpha_i) : i \in \mathbb{I},\, w \in \mathcal{W}(y, x)\}$; correspondingly the *imaginary* roots are $(\Delta^{\mathrm{im}})^x = \Delta^x - (\Delta^{\mathrm{re}})^x$. Assume that $\mathcal{W}$ is connected. Then the following conditions are equivalent [22, Lemma 2.11]:

- Δ^x is finite for some $x \in \mathcal{X}$,
- Δ^x is finite for all $x \in \mathcal{X}$,
- $(\Delta^{\mathrm{re}})^x$ is finite for all $x \in \mathcal{X}$,
- $\mathcal{W}$ is finite.

If these hold, then all roots are real [22]; we say that $\mathcal{R}$ is *finite*. We now discuss two examples of GRS, central for the subsequent discussion.

Example 3.2 ([2]). Let $\mathbf{k}$ be a field of characteristic $\ell \geq 0$, $\theta \in \mathbb{N}$, $\mathbf{p} \in \mathbb{G}_2^\theta$ and $A = (a_{ij}) \in \mathbf{k}^{\theta\times\theta}$. We assume $\ell \neq 2$ for simplicity. Let $\mathfrak{h} = \mathbf{k}^{2\theta - \mathrm{rk}\, A}$. Let $\mathfrak{g}(A, \mathbf{p})$ be the Kac-Moody Lie superalgebra over $\mathbf{k}$ defined as in [57]; it is generated by $\mathfrak{h}$, e_i and f_i, $i \in \mathbb{I}$, and the parity is given by $|e_i| = |f_i| = p_i$, $i \in \mathbb{I}$, $|h| = 0$, $h \in \mathfrak{h}$. Let $\Delta^{A,\mathbf{p}}$ be the root system of $\mathfrak{g}(A, \mathbf{p})$. We make the following technical assumptions:

$$a_{jk} = 0 \implies a_{kj} = 0, \qquad j \neq k; \tag{3.12}$$

$$\mathrm{ad}\, f_i \text{ is locally nilpotent in } \mathfrak{g}(A, \mathbf{p}), \qquad i \in \mathbb{I}. \tag{3.13}$$

The matrix A is *admissible* if (3.13) holds [80]. Let $C^{A,\mathbf{p}} = (c_{ij}^{A,\mathbf{p}})_{i,j\in\mathbb{I}}$ be given by

$$c_{ij}^{A,\mathbf{p}} := -\min\{m \in \mathbb{N}_0 : (\mathrm{ad}\, f_i)^{m+1} f_j = 0\}, i \neq j \in \mathbb{I}, \qquad c_{ii}^{A,\mathbf{p}} := 2. \tag{3.14}$$

We need the following elements of $\mathbf{k}$:

$$\text{if } p_i = 0, \qquad d_m = m\, a_{ij} + \binom{m}{2} a_{ii}; \tag{3.15}$$

$$\text{if } p_i = 1; \qquad d_m = \begin{cases} k\, a_{ii}, & m = 2k, \\ k\, a_{ii} + a_{ij}, & m = 2k+1; \end{cases} \tag{3.16}$$

$$\nu_{j,0} = 1, \qquad \nu_{j,n} = \prod_{t=1}^{n} (-1)^{p_i((t-1)p_i + p_j)} d_t; \tag{3.17}$$

$$\mu_{j,0} = 0, \qquad \mu_{j,n} = (-1)^{p_i p_j} n \left(\prod_{t=2}^{n} (-1)^{p_i((t-1)p_i + p_j)} d_t \right) a_{ji}. \tag{3.18}$$

With the help of these scalars, we define a reflection $r_i(A,\mathbf{p}) = (r_iA, r_i\mathbf{p})$, where $r_i\mathbf{p} = (\overline{p}_j)_{j\in\mathbb{I}}$, with $\overline{p}_j = p_j - c_{ij}^{A,\mathbf{p}}p_i$, and $r_iA = (\overline{a}_{jk})_{j,k\in\mathbb{I}}$, with

$$\overline{a}_{jk} = \begin{cases} -c_{ik}^{A,\mathbf{p}}\mu_{j,-c_{ij}^{A,\mathbf{p}}}a_{ii} + \mu_{j,-c_{ij}^{A,\mathbf{p}}}a_{ik} \\ \qquad -c_{ik}^{A,\mathbf{p}}\nu_{j,-c_{ij}^{A,\mathbf{p}}}a_{ji} + \nu_{j,-c_{ij}^{A,\mathbf{p}}}a_{jk}, & j,k\neq i; \\ c_{ik}^{A,\mathbf{p}}a_{ii} - a_{ik}, & j = i \neq k; \\ -\mu_{j,-c_{ij}^{A,\mathbf{p}}}a_{ii} - \nu_{j,-c_{ij}^{A,\mathbf{p}}}a_{ji}, & j \neq k = i; \\ a_{ii}, & j = k = i. \end{cases} \tag{3.19}$$

Theorem 3.3. *There is an isomorphism $T_i^{A,\mathbf{p}} : \mathfrak{g}(r_i(A,\mathbf{p})) \to \mathfrak{g}(A,\mathbf{p})$ of Lie superalgebras given (for an appropppriate basis (h_i) of $\mathfrak{h}$) by*

$$T_i^{A,\mathbf{p}}(e_j) = \begin{cases} (\operatorname{ad} e_i)^{-c_{ij}^{A,\mathbf{p}}}(e_j), & i\neq j\in\mathbb{I}, \\ f_i, & j = i \end{cases}$$
$$T_i^{A,\mathbf{p}}(f_j) = \begin{cases} (\operatorname{ad} f_i)^{-c_{ij}^{A,\mathbf{p}}}f_j, & j\in\mathbb{I}, j\neq i, \\ (-1)^{p_i}e_i, & j = i, \end{cases} \tag{3.20}$$
$$T_i^{A,\mathbf{p}}(h_j) = \begin{cases} \mu_{j,-c_{ij}^{A,\mathbf{p}}}h_i + \nu_{j,-c_{ij}^{A,\mathbf{p}}}h_j, & i\neq j\in\mathbb{I} \\ -h_i, & j = i, \\ h_j, & \theta+1\leq j\leq 2\theta - \operatorname{rk} A. \end{cases}$$

Assume that $\dim\mathfrak{g}(A,\mathbf{p}) < \infty$; then (3.12) and (3.13) hold. Let

$$\mathcal{X} = \{r_{i_1}\cdots r_{i_n}(A,\mathbf{p}) \,|\, n\in\mathbb{N}_0, i_1,\dots,i_n\in\mathbb{I}\}.$$

Then $(\mathcal{X}, r, \mathcal{C}, \Delta)$, where $\mathcal{C} = (C^{(B,\mathbf{q})})_{(B,\mathbf{q})\in\mathcal{X}}$ and $\Delta = (\Delta^{(B,\mathbf{q})})_{(B,\mathbf{q})\in\mathcal{X}}$, is a finite GRS, an invariant of $\mathfrak{g}(A,\mathbf{p})$.

Example 3.4. Let H be a Hopf algebra, assumed semisimple for easiness. Let $M\in {}^H_H\mathcal{YD}$ be finite-dimensional, with a fixed decomposition $M = M_1\oplus\cdots\oplus M_\theta$, where $M_1,\dots,M_\theta\in \operatorname{Irr}{}^H_H\mathcal{YD}$. Then $T(M)$ and $\mathfrak{B}(M)$ are $\mathbb{Z}^\theta$-graded, by $\deg x = \alpha_i$ for all $x\in M_i$, $i\in\mathbb{I}_\theta$. Recall that $\mathbb{Z}^\theta_{\geq 0} = \sum_{i\in\mathbb{I}_\theta}\mathbb{Z}_{\geq 0}\alpha_i$.

Theorem 3.5 ([46, 48]). *If $\dim\mathfrak{B}(M) < \infty$, then M has a finite GRS.*

We discuss the main ideas of the proof. Let $i\in\mathbb{I} = \mathbb{I}_\theta$. We define $M_i' = V_i^*$,

$$c_{ij}^M = -\sup\{h\in\mathbb{N}_0 : \operatorname{ad}_c^h(M_i)(M_j)\neq 0 \text{ in } \mathfrak{B}(M)\}, \qquad i\neq j, \quad c_{ii}^M = 2;$$
$$M_j' = \operatorname{ad}_c^{-c_{ij}}(M_i)(M_j), \qquad \rho_i(M) = M_1'\oplus\cdots\oplus M_\theta'.$$

Then $\dim\mathfrak{B}(M) = \dim\mathfrak{B}(\rho_i(M))$ and $C^M = (c_{ij}^M)_{i,j\in\mathbb{I}}$ is a generalized Cartan matrix [12]. Also, M_j' is irreducible [12, 3.8], [46, 7.2]. Let $\mathcal{X}$ be the set of objects in ${}^H_H\mathcal{YD}$ with fixed decomposition (up to isomorphism) of the form

$$\{\rho_{i_1}\cdots\rho_{i_n}(M) \,|\, n\in\mathbb{N}_0, i_1,\dots,i_n\in\mathbb{I}\}.$$

Then $(\mathcal{X}, \rho, \mathcal{C})$, where $\mathcal{C} = (C^N)_{N\in\mathcal{X}}$, satisfies (3.7). Next we need:

- [46, Theorem 4.5]; [42] There exists a totally ordered index set $(L, \leq)$ and families $(W_l)_{l\in L}$ in $\operatorname{Irr} {}^H_H\mathcal{YD}$, $(\beta_l)_{l\in L}$ such that $\mathfrak{B}(M) \simeq \otimes_{l\in L}\mathfrak{B}(W_l)$ as $\mathbb{Z}^\theta$-graded objects in ${}^H_H\mathcal{YD}$, where $\deg x = \beta_l$ for all $x \in W_l$, $l \in L$.

Let $\Delta^M_\pm = \{\pm\beta_l : l \in L\}$, $\Delta^M = \Delta^M_+ \cup \Delta^M_-$, $\Delta = (\Delta^N)_{N\in\mathcal{X}(M)}$. Then $\mathcal{R} = (\mathcal{X}, \rho, \mathcal{C}, \Delta)$ is a finite GRS.

Theorem 3.6 ([23]). *The classification of all finite GRS is known.*

The proof is a combinatorial tour-de-force and requires computer calculations. It is possible to recover from this result the classification of the finite-dimensional contragredient Lie superalgebras in arbitrary characteristic [2]. However, the list of [23] is substantially larger than the classifications of the alluded Lie superalgebras or the braidings of diagonal type with finite-dimensional Nichols algebra.

3.5. Nichols algebras of diagonal type. Let G be a finite group. We denote ${}^G_G\mathcal{YD} = {}^H_H\mathcal{YD}$ for $H = \mathbb{C}G$. So $M \in {}^G_G\mathcal{YD}$ is a left G-module with a G-grading $M = \oplus_{g\in G}M_g$ such that $t \cdot M_g = M_{tgt^{-1}}$, for all $g, t \in G$. If $M, N \in {}^G_G\mathcal{YD}$, then the braiding $c : M \otimes N \to N \otimes M$ is given by $c(m \otimes n) = g \cdot n \otimes m$, $m \in M_g$, $n \in N$, $g \in G$. Now assume that $G = \Gamma$ is a finite abelian group. Then every $M \in {}^\Gamma_\Gamma\mathcal{YD}$ is a Γ-graded Γ-module, hence of the form $M = \oplus_{g\in\Gamma,\chi\in\widehat{\Gamma}}M^\chi_g$, where M^χ_g is the χ-isotypic component of M_g. So ${}^\Gamma_\Gamma\mathcal{YD}$ is just the category of $\Gamma \times \widehat{\Gamma}$-graded modules, with the braiding $c : M \otimes N \to N \otimes M$ given by $c(m \otimes n) = \chi(g)n \otimes m$, $m \in M^\eta_g$, $n \in N^\chi_t$, $g, t \in G$, $\chi, \eta \in \widehat{\Gamma}$. Let $\theta \in \mathbb{N}$, $\mathbb{I} = \mathbb{I}_\theta$.

Definition 3.7. Let $\mathbf{q} = (q_{ij})_{i,j\in\mathbb{I}}$ be a matrix with entries in $\mathbb{C}^\times$. A braided vector space (V, c) is of *diagonal type* with matrix $\mathbf{q}$ if V has a basis $(x_i)_{i\in\mathbb{I}}$ with

$$c(x_i \otimes x_j) = q_{ij}x_j \otimes x_i, \qquad i, j \in \mathbb{I}. \tag{3.21}$$

Thus, every finite-dimensional $V \in {}^\Gamma_\Gamma\mathcal{YD}$ is a braided vector space of diagonal type. Question (a), more precisely (a_1), has a complete answer in this setting. First we can assume that $q_{ii} \neq 1$ for $i \in \mathbb{I}$, as otherwise $\dim \mathfrak{B}(V) = \infty$. Also, let $\mathbf{q}' = (q'_{ij})_{i,j\in\mathbb{I}} \in (\mathbb{C}^\times)^{\mathbb{I}\times\mathbb{I}}$ and V' a braided vector space with matrix $\mathbf{q}'$. If $q_{ii} = q'_{ii}$ and $q_{ij}q_{ji} = q'_{ij}q'_{ji}$ for all $j \neq i \in \mathbb{I}_\theta$, then $\mathfrak{B}(V) \simeq \mathfrak{B}(V')$ as braided vector spaces.

Theorem 3.8 ([44]). *The classification of all braided vector spaces of diagonal type with finite-dimensional Nichols algebra is known.*

The proof relies on the Weyl groupoid introduced in [43], a particular case of Theorem 3.5. Another fundamental ingredient is the following result, generalized at various levels in [42, 46, 48].

Theorem 3.9 ([58]). *Let V be a braided vector space of diagonal type. Every Hopf algebra quotient of $T(V)$ has a PBW basis.*

The classification in Theorem 3.8 can be organized as follows:

- ◇ For most of the matrices $\mathbf{q} = (q_{ij})_{i,j\in\mathbb{I}_\theta}$ in the list of [44] there is a field $\mathbf{k}$ and a pair $(A, \mathbf{p})$ as in Example 3.2 such that $\dim \mathfrak{g}(A, \mathbf{p}) < \infty$, and $\mathfrak{g}(A, \mathbf{p})$ has the same GRS as the Nichols algebra corresponding to $\mathbf{q}$ [2].
- ◇ Besides these, there are 12 (yet) unidentified examples.

We believe that Theorem 3.8 can be proved from Theorem 3.6, via Example 3.2.

Theorem 3.10 ([18, 19]). *An efficient set of defining relations of each finite-dimensional Nichols algebra of a braided vector space of diagonal type is known.*

The proof uses most technical tools available in the theory of Nichols algebras; of interest in its own is the introduction of the notion of convex order in Weyl groupoids. As for other classifications above, it is not possible to state precisely the list of relations. We just mention different types of relations that appear.

- Quantum Serre relations, i.e., $\mathrm{ad}_c(x_i)^{1-a_{ij}}(x_j)$ for suitable $i \neq j$.
- Powers of root vectors, i.e., $x_\beta^{N_\beta}$, where the x_β's are part of the PBW basis.
- More exotic relations; they involve 2, 3, or at most 4 i's in $\mathbb{I}$.

3.6. Nichols algebras of rack type. We now consider Nichols algebras of objects in ${}^G_G\mathcal{YD}$, where G is a finite not necessarily abelian group. The category ${}^G_G\mathcal{YD}$ is semisimple and the simple objects are parametrized by pairs $(\mathcal{O}, \rho)$, where $\mathcal{O}$ is a conjugacy class in G and $\rho \in \operatorname{Irr} G^x$, for a fixed $x \in \mathcal{O}$; the corresponding simple Yetter-Drinfeld module $M(\mathcal{O}, \rho)$ is $\operatorname{Ind}_{G^x}^G \rho$ as a module. The braiding c is described in terms of the conjugation in $\mathcal{O}$. To describe the related suitable class, we recall that a rack is a set $X \neq \emptyset$ with a map $\triangleright : X \times X \to X$ satisfying

- $\varphi_x := x \triangleright \_\_$ is a bijection for every $x \in X$.
- $x \triangleright (y \triangleright z) = (x \triangleright y) \triangleright (x \triangleright z)$ for all $x, y, z \in X$ (self-distributivity).

For instance, a conjugacy class $\mathcal{O}$ in G with the operation $x \triangleright y = xyx^{-1}$, $x, y \in \mathcal{O}$ is a rack; actually we only consider racks realizable as conjugacy classes. Let X be a rack and $\mathfrak{X} = (X_k)_{k\in I}$ a decomposition of X, i.e., a disjoint family of subracks with $X_l \triangleright X_k = X_k$ for all $k, l \in I$.

Definition 3.11. [10] A *2-cocycle of degree* $\mathbf{n} = (n_k)_{k\in I}$, associated to $\mathfrak{X}$, is a family $\mathbf{q} = (q_k)_{k\in I}$ of maps $q_k : X \times X_k \to \mathbf{GL}(n_k, \mathbb{C})$ such that

$$q_k(i, j \triangleright h) q_k(j, h) = q_k(i \triangleright j, i \triangleright h) q_k(i, h), \qquad i, j \in X,\ h \in X_k,\ k \in I. \tag{3.22}$$

Given such $\mathbf{q}$, let $V = \oplus_{k\in I} \mathbb{C}X_k \otimes \mathbb{C}^{n_k}$ and let $c^{\mathbf{q}} \in \mathbf{GL}(V \otimes V)$ be given by

$$c^{\mathbf{q}}(x_i v \otimes x_j w) = x_{i \triangleright j} q_k(i, j)(w) \otimes x_i v, \qquad i \in X_l,\ j \in X_k,\ v \in \mathbb{C}^{n_l},\ w \in \mathbb{C}^{n_k}.$$

Then $(V, c^{\mathbf{q}})$ is a braided vector space called of *rack type*; its Nichols algebra is denoted $\mathfrak{B}(X, \mathbf{q})$. If $\mathfrak{X} = (X)$, then we say that $\mathbf{q}$ is *principal*.

Every finite-dimensional $V \in {}^G_G\mathcal{YD}$ is a braided vector space of rack type [10, Theorem 4.14]. Question (a_1) in this setting has partial answers in three different lines: computation of some finite-dimensional Nichols algebras, Nichols algebras of reducible Yetter-Drinfeld modules and collapsing of racks.

3.6.1. Finite-dimensional Nichols algebras of rack type. The algorithm to compute a Nichols algebra $\mathfrak{B}(V)$ is as follows: compute the space $\mathfrak{J}^i(V) = \ker \mathcal{Q}_i$ of relations of degree i, for $i = 2, 3, \dots, m$; then compute the m-th partial Nichols algebra $\widehat{\mathfrak{B}_m(V)} =$

$T(V)/\langle \oplus_{2\le i\le m}\mathfrak{J}^i(V)\rangle$, say with a computer program. If lucky enough to get $\dim \widehat{\mathfrak{B}_m}(V) < \infty$, then check whether it is a Nichols algebra, e.g. via skew-derivations; otherwise go to $m+1$. The description of $\mathfrak{J}^2(V) = \ker(\mathrm{id}+c)$ is not difficult [38] but for higher degrees it turns out to be very complicated. We list all known examples of finite-dimensional Nichols algebras $\mathfrak{B}(X,\mathbf{q})$ with X indecomposable and $\mathbf{q}$ principal and abelian ($n_1 = 1$).

Example 3.12. Let $\mathcal{O}_d^m$ be the conjugacy class of d-cycles in $\mathbb{S}_m$, $m \ge 3$. We start with the rack of transpositions in $\mathbb{S}_m$ and the cocycles -1, χ that arise from the $\rho \in \operatorname{Irr}\mathbb{S}_m^{(12)}$ with $\rho(12) = -1$, see [64, (5.5), (5.9)]. Let V be a vector space with basis $(x_{ij})_{(ij)\in\mathcal{O}_2^m}$ and consider the relations

$$x_{ij}^2 = 0, \qquad (ij) \in \mathcal{O}_2^m; \tag{3.23}$$

$$x_{ij}x_{kl} + x_{kl}x_{ij} = 0, \qquad (ij),(kl) \in \mathcal{O}_2^m,\ |\{i,j,k,l\}| = 4; \tag{3.24}$$

$$x_{ij}x_{kl} - x_{kl}x_{ij} = 0, \qquad (ij),(kl) \in \mathcal{O}_2^m,\ |\{i,j,k,l\}| = 4; \tag{3.25}$$

$$x_{ij}x_{ik} + x_{jk}x_{ij} + x_{ik}x_{jk} = 0, \qquad (ij),(ik),(jk) \in \mathcal{O}_2^m,\ |\{i,j,k\}| = 3; \tag{3.26}$$

$$x_{ij}x_{ik} - x_{jk}x_{ij} - x_{ik}x_{jk} = 0, \qquad (ij),(ik),(jk) \in \mathcal{O}_2^m,\ |\{i,j,k\}| = 3. \tag{3.27}$$

The quadratic algebras $\mathfrak{B}_m := \widehat{\mathfrak{B}_2}(\mathcal{O}_2^m, -1) = T(V)/\langle (3.23), (3.24), (3.26)\rangle$ and $\mathcal{E}_m := \widehat{\mathfrak{B}_2}(\mathcal{O}_2^m, \chi) = T(V)/\langle (3.23), (3.25), (3.27)\rangle$ were considered in [64], [36] respectively; $\mathcal{E}_m$ are named the *Fomin-Kirillov algebras*. It is known that

- The Nichols algebras $\mathfrak{B}(\mathcal{O}_2^m, -1)$ and $\mathfrak{B}(\mathcal{O}_2^m, \chi)$ are twist-equivalent, hence have the same Hilbert series. Ditto for the algebras $\mathfrak{B}_m$ and $\mathcal{E}_m$ [84].
- If $3 \le m \le 5$, then $\mathfrak{B}_m = \mathfrak{B}(\mathcal{O}_2^m, -1)$ and $\mathcal{E}_m = \mathfrak{B}(\mathcal{O}_2^m, \chi)$ are finite-dimensional [36, 38, 64] (for $m = 5$ part of this was done by Graña). In fact

$$\dim \mathfrak{B}_3 = 12, \qquad \dim \mathfrak{B}_4 = 576, \qquad \dim \mathfrak{B}_5 = 8294400.$$

  But for $m \ge 6$, it is not known whether the Nichols algebras $\mathfrak{B}(\mathcal{O}_2^m, -1)$ and $\mathfrak{B}(\mathcal{O}_2^m, \chi)$ have finite dimension or are quadratic.

Example 3.13 ([10]). The Nichols algebra $\mathfrak{B}(\mathcal{O}_4^4, -1)$ is quadratic, has the same Hilbert series as $\mathfrak{B}(\mathcal{O}_2^4, -1)$ and is generated by $(x_\sigma)_{\sigma\in\mathcal{O}_4^4}$ with defining relations

$$x_\sigma^2 = 0, \tag{3.28}$$

$$x_\sigma x_{\sigma^{-1}} + x_{\sigma^{-1}}x_\sigma = 0, \tag{3.29}$$

$$x_\sigma x_\kappa + x_\nu x_\sigma + x_\kappa x_\nu = 0, \qquad \sigma\kappa = \nu\sigma,\ \kappa \ne \sigma \ne \nu \in \mathcal{O}_4^4. \tag{3.30}$$

Example 3.14 ([41]). Let A be a finite abelian group and $g \in \operatorname{Aut} A$. The *affine* rack (A, g) is the set A with product $a \triangleright b = g(b) + (\mathrm{id} - g)(a)$, $a, b \in A$. Let $p \in \mathbb{N}$ be a prime, $q = p^{v(q)}$ a power of p, $A = \mathbb{F}_q$ and g the multiplication by $N \in \mathbb{F}_q^\times$; let $X_{q,N} = (A, g)$. Assume that $q = 3$, 4, 5, or 7, with $N = 2$, $\omega \in \mathbb{F}_4 - \mathbb{F}_2$, 2 or 3, respectively. Then $\dim \mathfrak{B}(X_{q,N}, -1) = q\varphi(q)(q-1)^{q-2}$, φ being the Euler function, and $\mathfrak{J}(X_{q,N}, -1) = \langle \mathfrak{J}^2 + \mathfrak{J}^{v(q)(q-1)}\rangle$, where $\mathfrak{J}^2$ is generated by

$$x_i^2, \qquad \text{always} \tag{3.31}$$

$$x_i x_j + x_{-i+2j}x_i + x_j x_{-i+2j}, \qquad \text{for } q = 3, \tag{3.32}$$

$$x_ix_j + x_{(\omega+1)i+\omega j}x_i + x_jx_{(\omega+1)i+\omega j}, \qquad \text{for } q = 4, \tag{3.33}$$
$$x_ix_j + x_{-i+2j}x_i + x_{3i-2j}x_{-i+2j} + x_jx_{3i-2j}, \qquad \text{for } q = 5, \tag{3.34}$$
$$x_ix_j + x_{-2i+3j}x_i + x_jx_{-2i+3j}, \qquad \text{for } q = 7, \tag{3.35}$$

with $i, j \in \mathbb{F}_q$; and $\mathfrak{J}^{v(q)(q-1)}$ is generated by $\sum_h T^h(V)\mathfrak{J}^2T^{v(q)(q-1)-h-2}(V)$ and

$$(x_\omega x_1x_0)^2 + (x_1x_0x_\omega)^2 + (x_0x_\omega x_1)^2, \qquad \text{for } q = 4, \tag{3.36}$$
$$(x_1x_0)^2 + (x_0x_1)^2, \qquad \text{for } q = 5, \tag{3.37}$$
$$(x_2x_1x_0)^2 + (x_1x_0x_2)^2 + (x_0x_2x_1)^2, \qquad \text{for } q = 7. \tag{3.38}$$

Of course $X_{3,2} = \mathcal{O}_2^3$; also $\dim \mathfrak{B}(X_{4,\omega}, -1) = 72$. By duality, we get

$$\dim \mathfrak{B}(X_{5,3}, -1) = \dim \mathfrak{B}(X_{5,2}, -1) = 1280,$$
$$\dim \mathfrak{B}(X_{7,5}, -1) = \dim \mathfrak{B}(X_{7,3}, -1) = 326592.$$

Example 3.15 ([45]). There is another finite-dimensional Nichols algebra associated to $X_{4,\omega}$ with a cocycle $\mathbf{q}$ with values $\pm\xi$, where $1 \neq \xi \in \mathbb{G}_3$. Concretely, $\dim \mathfrak{B}(X_{4,\omega}, \mathbf{q}) = 5184$ and $\mathfrak{B}(X_{4,\omega}, \mathbf{q})$ can be presented by generators $(x_i)_{i\in\mathbb{F}_4}$ with defining relations

$$x_0^3 = x_1^3 = x_\omega^3 = x_{\omega^2}^3 = 0,$$
$$\xi^2x_0x_1 + \xi x_1x_\omega - x_\omega x_0 = 0, \quad \xi^2x_0x_\omega + \xi x_\omega x_{\omega^2} - x_{\omega^2}x_0 = 0,$$
$$\xi x_0x_{\omega^2} - \xi^2x_1x_0 + x_{\omega^2}x_1 = 0, \quad \xi x_1x_{\omega^2} + \xi^2x_\omega x_1 + x_{\omega^2}x_\omega = 0,$$
$$x_0^2x_1x_\omega x_1^2 + x_0x_1x_\omega x_1^2x_0 + x_1x_\omega x_1^2x_0^2 + x_\omega x_1^2x_0^2x_1 + x_1^2x_0^2x_1x_\omega + x_1x_0^2x_1x_\omega x_1$$
$$+x_1x_\omega x_1x_0^2x_\omega + x_\omega x_1x_0x_1x_0x_\omega + x_\omega x_1^2x_0x_\omega x_0 = 0.$$

3.6.2. Nichols algebras of decomposable Yetter-Drinfeld modules over groups. The ideas of Example 3.4 in the context of decomposable Yetter-Drinfeld modules over groups were pushed further in a series of papers culminating with a remarkable classification result [50]. Consider the groups

$$\Gamma_n = \langle a, b, \nu | ba = \nu ab, \quad \nu a = a\nu^{-1}, \quad \nu b = b\nu, \quad \nu^n = 1\rangle, \quad n \geq 2; \tag{3.39}$$
$$T = \langle \zeta, \chi_1, \chi_2 | \zeta\chi_1 = \chi_1\zeta, \quad \zeta\chi_2 = \chi_2\zeta, \quad \chi_1\chi_2\chi_1 = \chi_2\chi_1\chi_2, \quad \chi_1^3 = \chi_2^3\rangle. \tag{3.40}$$

- [47] Let G be a suitable quotient of Γ_2. Then there exist $V_1, W_1 \in \operatorname{Irr}{}^G_G\mathcal{YD}$ such that $\dim V_1 = \dim W_1 = 2$ and $\dim \mathfrak{B}(V_1 \oplus W_1) = 64 = 2^6$.
- [50] Let G be a suitable quotient of Γ_3. Then there exist V_2, V_3, V_4, $W_2, W_3, W_4 \in \operatorname{Irr}{}^G_G\mathcal{YD}$ such that $\dim V_2 = 1$, $\dim V_3 = \dim V_4 = 2$, $\dim W_2 = \dim W_3 = \dim W_4 = 3$ and $\dim \mathfrak{B}(V_2 \oplus W_2) = \dim \mathfrak{B}(V_3 \oplus W_3) = 10368 = 2^7 3^4$, $\dim \mathfrak{B}(V_4 \oplus W_4) = 2304 = 2^{18}$.
- [49] Let G be a suitable quotient of Γ_4. Then there exist V_5, $W_5 \in \operatorname{Irr}{}^G_G\mathcal{YD}$ such that $\dim V_5 = 2$, $\dim W_5 = 4$ and $\dim \mathfrak{B}(V_5 \oplus W_5) = 262144 = 2^{18}$.
- [49] Let G be a suitable quotient of T. Then there exist V_6, $W_6 \in \operatorname{Irr}{}^G_G\mathcal{YD}$ such that $\dim V_6 = 1$, $\dim W_6 = 4$ and $\dim \mathfrak{B}(V_6 \oplus W_6) = 80621568 = 2^{12}3^9$.

Theorem 3.16 ([50]). *Let G be a non-abelian group and $V, W \in \operatorname{Irr}{}^G_G\mathcal{YD}$ such that G is generated by the support of $V \oplus W$. Assume that $c^2_{|V\otimes W} \neq \mathrm{id}$ and that $\dim \mathfrak{B}(V \oplus W) < \infty$. Then $V \oplus W$ is one of $V_i \oplus W_i$, $i \in \mathbb{I}_6$, above, and correspondingly G is a quotient of either Γ_n, $2 \leq n \leq 4$, or T.*

3.6.3. Collapsing racks. Implicit in Question (a_1) in the setting of racks is the need to compute all non-principal 2-cocycles for a fixed rack X. Notably, there exist criteria that dispense of this computation. To state them and explain their significance, we need some terminology. All racks below are finite.

- A rack X is *abelian* when $x \triangleright y = y$, for all $x, y \in X$.
- A rack is *indecomposable* when it is not a disjoint union of two proper subracks.
- A rack X with $|X| > 1$ is *simple* when for any projection of racks $\pi : X \to Y$, either π is an isomorphism or Y has only one element.

Theorem 3.17 ([10, 3.9, 3.12], [55]). *Every simple rack is isomorphic to one of:*

(1) *Affine racks $(\mathbb{F}_p^t, T)$, where p is a prime, $t \in \mathbb{N}$, and T is the companion matrix of a monic irreducible polynomial $f \in \mathbb{F}_p[\mathtt{X}]$ of degree t, $f \neq \mathtt{X}, \mathtt{X} - 1$.*

(2) *Non-trivial (twisted) conjugacy classes in simple groups.*

(3) *Twisted conjugacy classes of type (G, u), where $G = L^t$, with L a simple non-abelian group and $1 < t \in \mathbb{N}$; and $u \in \operatorname{Aut}(L^t)$ acts by $u(\ell_1, \ell_2, \dots, \ell_t) = (\theta(\ell_t), \ell_1, \ell_2, \dots, \ell_{t-1})$, where $\theta \in \operatorname{Aut}(L)$.*

Definition 3.18 ([7, 3.5]). We say that a finite rack X is *of type D* when there are a decomposable subrack $Y = R \coprod S$, $r \in R$ and $s \in S$ such that $r \triangleright (s \triangleright (r \triangleright s)) \neq s$.

Also, X is *of type F* [4] if there are a disjoint family of subracks $(R_a)_{a\in\mathbb{I}_4}$ and a family $(r_a)_{a\in\mathbb{I}_4}$ with $r_a \in R_a$, such that $R_a \triangleright R_b = R_b$, $r_a \triangleright r_b \neq r_b$, for all $a \neq b \in \mathbb{I}_4$.

An indecomposable rack X *collapses* when $\dim \mathfrak{B}(X, \mathbf{q}) = \infty$ for every finite *faithful* 2-cocycle $\mathbf{q}$ (see [7] for the definition of faithful).

Theorem 3.19 ([7, 3.6]; [4, 2.8]). *If a rack is of type D or F, then it collapses.*

The proofs use results on Nichols algebras from [12, 23, 46].

If a rack projects onto a rack of type D (or F), then it is also of type D (or F), hence it collapses by Theorem 3.19. Since every indecomposable rack X, $|X| > 1$, projects onto a simple rack, it is natural to ask for the determination of all *simple* racks of type D or F. A rack is *cthulhu* if it is neither of type D nor F; it is *sober* if every subrack is either abelian or indecomposable [4]. Sober implies cthulhu.

- Let $m \geq 5$. Let $\mathcal{O}$ be either $\mathcal{O}^{\mathbb{S}_m}_\sigma$, if $\sigma \in \mathbb{S}_m - \mathbb{A}_m$, or else $\mathcal{O}^{\mathbb{A}_m}_\sigma$ if $\sigma \in \mathbb{A}_m$. The type of σ is formed by the lengths of the cycles in its decomposition.
  - ◇ [7, 4.2] If the type of σ is (3^2), $(2^2, 3)$, $(1^n, 3)$, (2^4), $(1^2, 2^2)$, $(2, 3)$, (2^3), or $(1^n, 2)$, then $\mathcal{O}$ is cthulhu. If the type of σ is $(1, 2^2)$, then $\mathcal{O}$ is sober.
  - ◇ [33] Let $p \in \mathbb{N}$ be a prime. Assume the type of σ is (p). If $p = 5, 7$ or not of the form $(r^k - 1)/(r - 1)$, r a prime power, then $\mathcal{O}$ is sober; otherwise $\mathcal{O}$ is of type D. Assume the type of σ is $(1, p)$. If $p = 5$ or not of the form $(r^k - 1)/(r - 1)$, r a prime power, then $\mathcal{O}$ is sober; otherwise $\mathcal{O}$ is of type D.

Table 3.1. Classes in sporadic simple groups not of type D

Group	Classes	Group	Classes
T	2A	Co_3	23A, 23B
M_{11}	8A, 8B, 11A, 11B	J_1	15A, 15B, 19A, 19B, 19C
M_{12}	11A, 11B	J_2	2A, 3A
M_{22}	11A, 11B	J_3	5A, 5B, 19A, 19B
M_{23}	23A, 23B	J_4	29A, 43A, 43B, 43C
M_{24}	23A, 23B	Ly	37A, 37B, 67A, 67B, 67C
Ru	29A, 29B	$O'N$	31A, 31B
Suz	3A	Fi_{23}	2A
HS	11A, 11B	Fi_{22}	2A, 22A, 22B
McL	11A, 11B	Fi'_{24}	29A, 29B
Co_1	3A	B	2A, 46A, 46B, 47A, 47B
Co_2	2A, 23A, 23B		

- ◇ [7, 4.1] For all other types, $\mathcal{O}$ is of type D, hence it collapses.
- ○ [4] Let $n \geq 2$ and q be a prime power. Let $x \in \mathbf{PSL}_n(q)$ not semisimple and $\mathcal{O} = \mathcal{O}_x^{\mathbf{PSL}_n(q)}$. The type of a unipotent element are the sizes of its Jordan blocks.
- ◇ Assume x is unipotent. If x is either of type (2) and q is even or not a square, or of type (3) and $q = 2$, then $\mathcal{O}$ is sober. If x is either of type $(2,1)$ and q is even, or of type $(2,1,1)$ and $q = 2$ then $\mathcal{O}$ is cthulhu. If x is of type $(2,1,1)$ and $q > 2$ is even, then $\mathcal{O}$ is not of type D, but it is open if it is of type F.
- ◇ Otherwise, $\mathcal{O}$ is either of type D or of type F, hence it collapses.
- ○ [8, 35] Let $\mathcal{O}$ be a conjugacy class in a sporadic simple group G. If $\mathcal{O}$ appears in Table 3.1, then $\mathcal{O}$ is not of type D. If $G = M$ is the Monster and $\mathcal{O}$ is one of 32A, 32B, 41A, 46A, 46B, 47A, 47B, 59A, 59B, 69A, 69B, 71A, 71B, 87A, 87B, 92A, 92B, 94A, 94B, then it is open whether $\mathcal{O}$ is of type D. Otherwise, $\mathcal{O}$ is of type D.

3.7. Generation in degree one. Here is the scheme of proof proposed in [14] to attack Conjecture 3.1: Let T be a finite-dimensional graded Hopf algebra in ${}^K_K\mathcal{YD}$ with $T^0 = \mathbb{C}$ and generated as algebra by T^1. We have a commutative diagram of Hopf algebra maps $\begin{array}{ccc} T & \xrightarrow{\pi} & \mathfrak{B}(V) \\ & {\scriptstyle p}\nwarrow \quad \nearrow & \\ & T(V) & \end{array}$. To show that π is injective, take a generator r (or a family of generators) of $\mathfrak{J}(V)$ such that $r \in \mathcal{P}(T(V))$ and consider the Yetter-Drinfeld submodule $U = \mathbb{C}r \oplus V$ of $T(V)$; if $\dim \mathfrak{B}(U) = \infty$, then $p(r) = 0$. Then p factorizes through $T(V)/\mathfrak{J}_1(V)$, where $\mathfrak{J}_1(V)$ is the ideal generated by primitive generators of $\mathfrak{J}(V)$, and so on.

The Conjecture has been verified in all known examples in characteristic 0 (it is false in positive characteristic or for infinite-dimensional Hopf algebras).

Theorem 3.20. *A finite-dimensional pointed Hopf algebra H is generated by group-like and skew-primitive elements if either of the following holds:*

- ◇ [19] *The infinitesimal braiding is of diagonal type, e. g. $G(H)$ is abelian.*
- ◇ [11, 38]. *The infinitesimal braiding of H is any of $(\mathcal{O}_2^m, -1)$, $(\mathcal{O}_2^m, \chi)$ $(m = 3, 4, 5)$, $(X_{4,\omega}, -1)$, $(X_{5,2}, -1)$, $(X_{5,3}, -1)$, $(X_{7,3}, -1)$, $(X_{7,5}, -1)$.*

3.8. Liftings. We address here Question (c) in §3.2. Let X be a finite rack and $q : X \times X \to \mathbb{G}_\infty$ a 2-cocycle. A Hopf algebra H is a *lifting* of (X, q) if H_0 is a Hopf subalgebra, H is generated by H_1 and its infinitesimal braiding is a realization of $(\mathbb{C}X, c^q)$. See [39] for liftings in the setting of copointed Hopf algebras.

We start discussing realizations of braided vector spaces as Yetter-Drinfeld modules. Let $\theta \in \mathbb{N}$ and $\mathbb{I} = \mathbb{I}_\theta$. First, a *YD-datum of diagonal type* is a collection

$$\mathcal{D} = ((q_{ij})_{i,j\in\mathbb{I}}, G, (g_i)_{i\in\mathbb{I}}, (\chi_i)_{i\in\mathbb{I}}), \tag{3.41}$$

where $q_{ij} \in \mathbb{G}_\infty$, $q_{ii} \neq 1$, $i, j \in \mathbb{I}$; G is a finite group; $g_i \in Z(G)$; $\chi_i \in \widehat{G}$, $i \in \mathbb{I}$; such that $q_{ij} := \chi_j(g_i)$, $i, j \in \mathbb{I}$. Let (V, c) be the braided vector space of diagonal type with matrix (q_{ij}) in the basis $(x_i)_{i\in\mathbb{I}_\theta}$. Then $V \in {}^G_G\mathcal{YD}$ by declaring $x_i \in V^{\chi_i}_{g_i}$, $i \in \mathbb{I}$. More generally, a *YD-datum of rack type* [11, 64] is a collection

$$\mathcal{D} = (X, q, G, \cdot, g, \chi), \tag{3.42}$$

where X is a finite rack; $q : X \times X \to \mathbb{G}_\infty$ is a 2-cocycle; G is a finite group; $\cdot$ is an action of G on X; $g : X \to G$ is equivariant with respect to the conjugation in G; and $\chi = (\chi_i)_{i\in X}$ is a family of 1-cocycles $\chi_i : G \to \mathbb{C}^\times$ (that is, $\chi_i(ht) = \chi_i(t)\chi_{t\cdot i}(h)$, for all $i \in X$, $h, t \in G$) such that $g_i \cdot j = i \triangleright j$ and $\chi_i(g_j) = q_{ij}$ for all $i, j \in X$. Let $(V, c) = (\mathbb{C}X, c^q)$ be the associated braided vector space. Then V becomes an object in ${}^G_G\mathcal{YD}$ by $\delta(x_i) = g_i \otimes x_i$ and $t \cdot x_i = \chi_i(t)x_{t\cdot i}$, $t \in G$, $i \in X$.

Second, let $\mathcal{D}$ be a YD-datum of either diagonal or rack type and $V \in {}^G_G\mathcal{YD}$ as above; let $\mathcal{T}(V) := T(V)\#\mathbb{C}G$. The desired liftings are quotients of $\mathcal{T}(V)$; write a_i in these quotients instead of x_i to distinguish them from the elements in $\mathfrak{B}(V)\#\mathbb{C}G$. Let $\mathcal{G}$ be a minimal set of generators of $\mathfrak{J}(V)$, assumed homogeneous both for the $\mathbb{N}$- and the G-grading. Roughly speaking, the deformations will be defined by replacing the relations $r = 0$ by $r = \phi_r$, $r \in \mathcal{G}$, where $\phi_r \in \mathcal{T}(V)$ belongs to a lower term of the coradical filtration, and the ideal $\mathfrak{J}_\phi(V)$ generated by ϕ_r, $r \in \mathcal{G}$, is a Hopf ideal. The problem is to describe the ϕ_r's and to check that $\mathcal{T}(V)/\mathfrak{J}_\phi(V)$ has the right dimension. If $r \in \mathcal{P}(T(V))$ has G-degree g, then $\phi_r = \lambda(1 - g)$ for some $\lambda \in \mathbb{C}$; depending on the action of G on r, it may happen that λ should be 0. In some cases, all $r \in \mathcal{G}$ are primitive, so all deformations can be described; see [13] for quantum linear spaces (their liftings can also be presented as Ore extensions [20]) and the Examples 3.21 and 3.22. But in most cases, not all $r \in \mathcal{G}$ are primitive and some recursive construction of the deformations is needed. This was achieved in [15] for diagonal braidings of Cartan type A_n, with explicit formulae, and in [16] for diagonal braidings of finite Cartan type, with recursive formulae. Later it was observed that the so obtained liftings are cocycle deformations of $\mathfrak{B}(V)\#\mathbb{C}G$, see e.g. [63]. This led to the strategy in [3]: pick an adapted stratification $\mathcal{G} = \mathcal{G}_0 \cup \mathcal{G}_1 \cup \cdots \cup \mathcal{G}_N$ [3, 5.1]; then construct recursively the deformations of $T(V)/\langle \mathcal{G}_0 \cup \mathcal{G}_1 \cup \cdots \cup \mathcal{G}_{k-1}\rangle$ by determining the cleft extensions of the deformations in the previous step and applying the theory of Hopf bi-Galois extensions [77]. In the Examples below, $\chi_i = \chi \in \widehat{G}$ for all $i \in X$ by [11, 3.3 (d)].

Example 3.21 ([11, 39]). Let $\mathcal{D} = (\mathcal{O}_2^3, -1, G, \cdot, g, \chi)$ be a YD-datum. Let $\lambda \in \mathbb{C}^2$ be such that

$$\lambda_1 = \lambda_2 = 0, \qquad \text{if } \chi^2 \neq \varepsilon; \tag{3.43}$$

$$\lambda_1 = 0, \qquad \text{if } g_{12}^2 = 1; \qquad \lambda_2 = 0, \qquad \text{if } g_{12}g_{13} = 1. \tag{3.44}$$

Let $\mathfrak{u} = \mathfrak{u}(\mathcal{D}, \lambda)$ be the quotient of $\mathcal{T}(V)$ by the relations

$$a_{12}^2 = \lambda_1(1 - g_{12}^2), \tag{3.45}$$

$$a_{12}a_{13} + a_{23}a_{12} + a_{13}a_{23} = \lambda_2(1 - g_{12}g_{13}). \tag{3.46}$$

Then $\mathfrak{u}$ is a pointed Hopf algebra, a cocycle deformation of $\operatorname{gr}\mathfrak{u} \simeq \mathfrak{B}(V)\#\mathbb{C}G$ and $\dim \mathfrak{u} = 12|G|$; $\mathfrak{u}(\mathcal{D}, \lambda) \simeq \mathfrak{u}(\mathcal{D}, \lambda')$ iff $\lambda = c\lambda'$ for some $c \in \mathbb{C}^\times$. Conversely, any lifting of $(\mathcal{O}_2^3, -1)$ is isomorphic to $\mathfrak{u}(\mathcal{D}, \lambda)$ for some YD-datum $\mathcal{D} = (\mathcal{O}_2^3, -1, G, \cdot, g, \chi)$ and $\lambda \in \mathbb{C}^2$ satisfying (3.43), (3.44).

Example 3.22 ([11]). Let $\mathcal{D} = (\mathcal{O}_2^4, -1, G, \cdot, g, \chi)$ be a YD-datum. Let $\lambda \in \mathbb{C}^3$ be such that

$$\lambda_i = 0, \quad i \in \mathbb{I}_3, \qquad \text{if } \chi^2 \neq \varepsilon; \tag{3.47}$$

$$\lambda_1 = 0, \quad \text{if } g_{12}^2 = 1; \qquad \lambda_2 = 0, \quad \text{if } g_{12}g_{34} = 1; \quad \lambda_3 = 0, \quad \text{if } g_{12}g_{13} = 1. \tag{3.48}$$

Let $\mathfrak{u} = \mathfrak{u}(\mathcal{D}, \lambda)$ be the quotient of $\mathcal{T}(V)$ by the relations

$$a_{12}^2 = \lambda_1(1 - g_{12}^2), \tag{3.49}$$

$$a_{12}a_{34} + a_{34}a_{12} = \lambda_2(1 - g_{12}g_{34}), \tag{3.50}$$

$$a_{12}a_{13} + a_{23}a_{12} + a_{13}a_{23} = \lambda_3(1 - g_{12}g_{13}). \tag{3.51}$$

Then $\mathfrak{u}$ is a pointed Hopf algebra, a cocycle deformation of $\mathfrak{B}(V)\#\mathbb{C}G$ and $\dim \mathfrak{u} = 576|G|$; $\mathfrak{u}(\mathcal{D}, \lambda) \simeq \mathfrak{u}(\mathcal{D}, \lambda')$ iff $\lambda = c\lambda'$ for some $c \in \mathbb{C}^\times$. Conversely, any lifting of $(\mathcal{O}_2^4, -1)$ is isomorphic to $\mathfrak{u}(\mathcal{D}, \lambda)$ for some YD-datum $\mathcal{D} = (\mathcal{O}_2^4, -1, G, \cdot, g, \chi)$ and $\lambda \in \mathbb{C}^3$ satisfying (3.47), (3.48).

Example 3.23 ([39]). Let $\mathcal{D} = (X_{4,\omega}, -1, G, \cdot, g, \chi)$ be a YD-datum. Let $\lambda \in \mathbb{C}^3$ be such that

$$\lambda_1 = \lambda_2 = 0, \qquad \text{if } \chi^2 \neq \varepsilon; \qquad \lambda_3 = 0, \text{ if } \chi^6 \neq \varepsilon; \tag{3.52}$$

$$\lambda_1 = 0, \text{ if } g_0^2 = 1, \qquad \lambda_2 = 0, \text{ if } g_0g_1 = 1, \qquad \lambda_3 = 0, \text{ if } g_0^3g_1^3 = 1. \tag{3.53}$$

Let $\mathfrak{u} = \mathfrak{u}(\mathcal{D}, \lambda)$ be the quotient of $\mathcal{T}(V)$ by the relations

$$x_0^2 = \lambda_1(1 - g_0^2), \tag{3.54}$$

$$x_0x_1 + x_\omega x_0 + x_1x_\omega = \lambda_2(1 - g_0g_1) \tag{3.55}$$

$$(x_\omega x_1 x_0)^2 + (x_1x_0x_\omega)^2 + (x_0x_\omega x_1)^2 = \zeta_6 - \lambda_3(1 - g_0^3g_1^3), \qquad \text{where} \tag{3.56}$$

$$\zeta_6 = \lambda_2(x_\omega x_1x_0x_\omega + x_1x_0x_\omega x_1 + x_0x_\omega x_1x_0) - \lambda_2^3(g_0g_1 - g_0^3g_1^3)$$

$$
\begin{aligned}
&+\lambda_1^2 g_0^2\big(g_{1+\omega}^2(x_\omega x_3 + x_0 x_\omega) + g_1 g_{1+\omega}(x_\omega x_1 + x_1 x_3) + g_1^2(x_1 x_0 + x_0 x_3)\big)\\
&-2\lambda_1^2 g_0^2(x_0 x_3 - x_\omega x_3 - x_1 x_\omega + x_1 x_0) - 2\lambda_1^2 g_\omega^2(x_\omega x_3 - x_1 x_3 + x_0 x_\omega - x_0 x_1)\\
&-2\lambda_1^2 g_1^2(x_\omega x_1 + x_1 x_3 + x_1 x_\omega - x_0 x_3 + x_0 x_1)\\
&+\lambda_2\lambda_1(g_\omega^2 x_0 x_3 + g_1^2 x_\omega x_3 + g_0^2 x_1 x_3) + \lambda_2^2 g_0 g_1(x_\omega x_1 + x_1 x_0 + x_0 x_\omega - \lambda_1)\\
&-\lambda_2\lambda_1^2(3g_0^3 g_{1+\omega} - 2g_0 g_1^3 - g_0^2 g_\omega^2 - 2g_0^3 g_1 + g_\omega^2 - g_1^2 + g_0^2)\\
&-\lambda_2(\lambda_1 - \lambda_2)\big(\lambda_1\, g_0^2(g_{1+\omega}^2 + g_1 g_{1+\omega} + g_1^2 + 2g_0 g_1^3) + x_\omega x_1 + x_1 x_0 + x_0 x_\omega\big).
\end{aligned}
$$

Then $\mathfrak{u}$ is a pointed Hopf algebra, a cocycle deformation of $\operatorname{gr}\mathfrak{u} \simeq \mathfrak{B}(V)\#\mathbb{C}G$ and $\dim \mathfrak{u} = 72|G|$; $\mathfrak{u}(\mathcal{D},\lambda) \simeq \mathfrak{u}(\mathcal{D},\lambda')$ iff $\lambda = c\lambda'$ for some $c \in \mathbb{C}^\times$. Conversely, any lifting of $(X_{4,\omega}, -1)$ is isomorphic to $\mathfrak{u}(\mathcal{D},\lambda)$ for some YD-datum $\mathcal{D} = (X_{4,\omega}, -1, G, \cdot, g, \chi)$ and $\lambda \in \mathbb{C}^3$ satisfying (3.52), (3.53).

Example 3.24 ([39]). Let $\mathcal{D} = (X_{5,2}, -1, G, \cdot, g, \chi)$ be a YD-datum. Let $\lambda \in \mathbb{C}^3$ be such that

$$
\lambda_1 = \lambda_2 = 0, \qquad \text{if } \chi^2 \neq \varepsilon; \qquad \lambda_3 = 0, \text{ if } \chi^4 \neq \varepsilon; \tag{3.57}
$$

$$
\lambda_1 = 0, \text{ if } g_0^2 = 1, \qquad \lambda_2 = 0, \text{ if } g_0 g_1 = 1, \qquad \lambda_3 = 0, \text{ if } g_0^2 g_1 g_2 = 1. \tag{3.58}
$$

Let $\mathfrak{u} = \mathfrak{u}(\mathcal{D},\lambda)$ be the quotient of $\mathcal{T}(V)$ by the relations

$$
x_0^2 = \lambda_1(1 - g_0^2), \tag{3.59}
$$

$$
x_0 x_1 + x_2 x_0 + x_3 x_2 + x_1 x_3 = \lambda_2(1 - g_0 g_1), \tag{3.60}
$$

$$
(x_1 x_0)^2 + (x_0 x_1)^2 = \zeta_4 - \lambda_3(1 - g_0^2 g_1 g_2), \quad \text{where} \tag{3.61}
$$

$\zeta_4 = \lambda_2\,(x_1 x_0 + x_0 x_1) + \lambda_1\, g_1^2(x_3 x_0 + x_2 x_3) - \lambda_1\, g_0^2(x_2 x_4 + x_1 x_2) + \lambda_2\lambda_1\, g_0^2(1 - g_1 g_2)$.
Then $\mathfrak{u}$ is a pointed Hopf algebra, a cocycle deformation of $\operatorname{gr}\mathfrak{u} \simeq \mathfrak{B}(V)\#\mathbb{C}G$ and $\dim \mathfrak{u} = 1280|G|$; $\mathfrak{u}(\mathcal{D},\lambda) \simeq \mathfrak{u}(\mathcal{D},\lambda')$ iff $\lambda = c\lambda'$ for some $c \in \mathbb{C}^\times$. Conversely, any lifting of $(X_{5,2}, -1)$ is isomorphic to $\mathfrak{u}(\mathcal{D},\lambda)$ for some YD-datum $\mathcal{D} = (X_{5,2}, -1, G, \cdot, g, \chi)$ and $\lambda \in \mathbb{C}^3$ satisfying (3.57), (3.58).

Example 3.25 ([39]). Let $\mathcal{D} = (X_{5,3}, -1, G, \cdot, g, \chi)$ be a YD-datum. Let $\lambda \in \mathbb{C}^3$ be such that

$$
\lambda_1 = \lambda_2 = 0, \qquad \text{if } \chi^2 \neq \varepsilon; \qquad \lambda_3 = 0, \text{ if } \chi^4 \neq \varepsilon; \tag{3.62}
$$

$$
\lambda_1 = 0, \text{ if } g_0^2 = 1, \qquad \lambda_2 = 0, \text{ if } g_1 g_0 = 1, \qquad \lambda_3 = 0, \text{ if } g_0^2 g_1 g_3 = 1. \tag{3.63}
$$

Let $\mathfrak{u} = \mathfrak{u}(\mathcal{D},\lambda)$ be the quotient of $\mathcal{T}(V)$ by the relations

$$
x_0^2 = \lambda_1(1 - g_0^2), \tag{3.64}
$$

$$
x_1 x_0 + x_0 x_2 + x_2 x_3 + x_3 x_1 = \lambda_2(1 - g_1 g_0) \tag{3.65}
$$

$$
x_0 x_2 x_3 x_1 + x_1 x_4 x_3 x_0 = \zeta_4' - \lambda_3(1 - g_0^2 g_1 g_3), \quad \text{where} \tag{3.66}
$$

$\zeta_4' = \lambda_2\,(x_0 x_1 + x_1 x_0) - \lambda_1\, g_1^2(x_3 x_2 + x_0 x_3) - \lambda_1\, g_0^2(x_3 x_4 + x_1 x_3) + \lambda_1\lambda_2(g_1^2 + g_0^2 - 2g_0^2 g_1 g_3)$.
Then $\mathfrak{u}$ is a pointed Hopf algebra, a cocycle deformation of $\operatorname{gr}\mathfrak{u} \simeq \mathfrak{B}(V)\#\mathbb{C}G$ and $\dim \mathfrak{u} = 1280|G|$; $\mathfrak{u}(\mathcal{D},\lambda) \simeq \mathfrak{u}(\mathcal{D},\lambda')$ iff $\lambda = c\lambda'$ for some $c \in \mathbb{C}^\times$. Conversely, any lifting of $(X_{5,3}, -1)$ is isomorphic to $\mathfrak{u}(\mathcal{D},\lambda)$ for some YD-datum $\mathcal{D} = (X_{5,3}, -1, G, \cdot, g, \chi)$ and $\lambda \in \mathbb{C}^3$ satisfying (3.62), (3.63).

4. Pointed Hopf algebras

4.1. Pointed Hopf algebras with abelian group. Here is a classification from [16]. Let $\mathcal{D} = ((q_{ij})_{i,j\in\mathbb{I}_\theta}, \Gamma, (g_i)_{i\in\mathbb{I}_\theta}, (\chi_i)_{i\in\mathbb{I}_\theta})$ be a YD-datum of diagonal type as in (3.41) with Γ a finite abelian group and let $V \in {}^\Gamma_\Gamma\mathcal{YD}$ be the corresponding realization. We say that $\mathcal{D}$ is a *Cartan datum* if there is a Cartan matrix (of finite type) $\mathbf{a} = (a_{ij})_{i,j\in\mathbb{I}_\theta}$ such that $q_{ij}q_{ji} = q_{ii}^{a_{ij}}$, $i \neq j \in \mathbb{I}_\theta$.

Let Φ be the root system associated to $\mathbf{a}$, $\alpha_1, \dots, \alpha_\theta$ a choice of simple roots, $\mathcal{X}$ the set of connected components of the Dynkin diagram of Φ and set $i \sim j$ whenever α_i, α_j belong to the same $J \in \mathcal{X}$. We consider two classes of parameters:

- $\lambda = (\lambda_{ij})_{i<j\in\mathbb{I}_\theta,\ i\not\sim j}$, is a family in $\{0,1\}$ with $\lambda_{ij} = 0$ when $g_ig_j = 1$ or $\chi_i\chi_j \neq \varepsilon$.
- $\mu = (\mu_\alpha)_{\alpha\in\Phi^+}$ is a family in $\mathbb{C}$ with $\mu_\alpha = 0$ when $\operatorname{supp}\alpha \subset J$, $J \in \mathcal{X}$, and $g_\alpha^{N_J} = 1$ or $\chi_\alpha^{N_J} \neq \varepsilon$. Here $N_J = \operatorname{ord} q_{ii}$ for an arbitrary $i \in J$.

We attach a family $(u_\alpha(\mu))_{\alpha\in\Phi^+}$ in $\mathbb{C}\Gamma$ to the parameter μ , defined recursively on the length of α, starting by $u_{\alpha_i}(\mu) = \mu_{\alpha_i}(1 - g_i^{N_i})$. From all these data we define a Hopf algebra $\mathfrak{u}(\mathcal{D}, \lambda, \mu)$ as the quotient of $\mathcal{T}(V) = T(V)\#\mathbb{C}\Gamma$ by the relations

$$ga_ig^{-1} = \chi_i(g)a_i, \tag{4.1}$$

$$\operatorname{ad}_c(a_i)^{1-a_{ij}}(a_j) = 0, \qquad i \neq j, i \sim j, \tag{4.2}$$

$$\operatorname{ad}_c(a_i)(a_j) = \lambda_{ij}(1 - g_ig_j), \qquad i < j, i \not\sim j, \tag{4.3}$$

$$a_\alpha^{N_J} = u_\alpha(\mu). \tag{4.4}$$

Theorem 4.1. *The Hopf algebra $\mathfrak{u}(\mathcal{D}, \lambda, \mu)$ is pointed, $G(\mathfrak{u}(\mathcal{D}, \lambda, \mu)) \simeq \Gamma$ and $\dim \mathfrak{u}(\mathcal{D}, \lambda, \mu) = \prod_{J\in\mathcal{X}} N_J^{|\Phi_J^+|}|\Gamma|$. Let H be a pointed finite-dimensional Hopf algebra and set $\Gamma = G(H)$. Assume that the prime divisors of $|\Gamma|$ are > 7. Then there exists a Cartan datum $\mathcal{D}$ and parameters λ and μ such that $H \simeq \mathfrak{u}(\mathcal{D}, \lambda, \mu)$. It is known when two Hopf algebras $\mathfrak{u}(\mathcal{D}, \lambda, \mu)$ and $\mathfrak{u}(\mathcal{D}', \lambda', \mu')$ are isomorphic.*

The proof offered in [16] relies on [14, 43, 59, 60]. Some comments: the hypothesis on $|\Gamma|$ forces the infinitesimal braiding V of H to be of Cartan type, and the relations of $\mathfrak{B}(V)$ to be just quantum Serre and powers of root vectors. The quantum Serre relations are not deformed in the liftings, except those linking different components of the Dynkin diagram; the powers of the root vectors are deformed to the $u_\alpha(\mu)$ that belong to the coradical. All this can fail without the hypothesis, see [52] for examples in rank 2.

4.2. Pointed Hopf algebras with non-abelian group. We present some classification results of pointed Hopf algebras with non-abelian group. We say that a finite group G *collapses* whenever any finite-dimensional pointed Hopf algebra H with $G(H) \simeq G$ is isomorphic to $\mathbb{C}G$.

- [7, 8] Let G be either $\mathbb{A}_m$, $m \geq 5$, or a sporadic simple group, different from Fi_{22}, the Baby Monster B or the Monster M. Then G collapses.

  The proof uses §3.6.3; the remaining Yetter-Drinfeld modules are discarded considering abelian subracks of the supporting conjugacy class and the list in [44].

- [12] Let $V = M(\mathcal{O}_2^3, \operatorname{sgn})$ and let $\mathcal{D}$ be the corresponding YD-datum. Let H be a finite-dimensional pointed Hopf algebra with $G(H) \simeq \mathbb{S}_3$. Then H is isomorphic either to $\mathbb{CS}_3$, or to $\mathfrak{u}(\mathcal{D}, 0) = \mathfrak{B}(V)\#\mathbb{CS}_3$, or to $\mathfrak{u}(\mathcal{D}, (0,1))$, cf. Example 3.21.
- [38] Let $H \not\simeq \mathbb{CS}_4$ be a finite-dimensional pointed Hopf algebra with $G(H) \simeq \mathbb{S}_4$. Let $V_1 = M(\mathcal{O}_2^4, \operatorname{sgn} \otimes \operatorname{id})$, $V_2 = M(\mathcal{O}_2^4, \operatorname{sgn} \otimes \operatorname{sgn})$, $W = M(\mathcal{O}_4^4, \operatorname{sgn} \otimes \operatorname{id})$, with corresponding data $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$. Then H is isomorphic to one of

$$\mathfrak{u}(\mathcal{D}_1, (0,\mu)), \quad \mu \in \mathbb{C}^2; \qquad \mathfrak{u}(\mathcal{D}_2, t), \quad t \in \{0,1\}; \qquad \mathfrak{u}(\mathcal{D}_3, \lambda), \quad \lambda \in \mathbb{C}^2.$$

  Here $\mathfrak{u}(\mathcal{D}_1, (0,\mu))$ is as in Example 3.22; $\mathfrak{u}(\mathcal{D}_2, t)$ is the quotient of $\mathcal{T}(V_2)$ by the relations $a_{12}^2 = 0$, $a_{12}a_{34} - a_{34}a_{12} = 0$, $a_{12}a_{23} - a_{13}a_{12} - a_{23}a_{13} = t(1 - g_{(12)}g_{(23)})$ and $\mathfrak{u}(\mathcal{D}_3, \lambda)$ is the quotient of $\mathcal{T}(W)$ by the relations

$$a_{(1234)}^2 = \lambda_1(1 - g_{(13)}g_{(24)}); \qquad a_{(1234)}a_{(1432)} + a_{(1432)}a_{(1234)} = 0;$$
$$a_{(1234)}a_{(1243)} + a_{(1243)}a_{(1423)} + a_{(1423)}a_{(1234)} = \lambda_2(1 - g_{(12)}g_{(13)}).$$

  Clearly $\mathfrak{u}(\mathcal{D}_1, 0) = \mathfrak{B}(V_1)\#\mathbb{CS}_4$, $\mathfrak{u}(\mathcal{D}_2, 0) = \mathfrak{B}(V_2)\#\mathbb{CS}_4$, $\mathfrak{u}(\mathcal{D}_3, 0) = \mathfrak{B}(W)\#\mathbb{CS}_4$. Also $\mathfrak{u}(\mathcal{D}_1, (0,\mu)) \simeq \mathfrak{u}(\mathcal{D}_1, (0,\nu))$ iff $\mu = c\nu$ for some $c \in \mathbb{C}^\times$, and $\mathfrak{u}(\mathcal{D}_3, \lambda) \simeq \mathfrak{u}(\mathcal{D}_3, \kappa)$ iff $\lambda = c\kappa$ for some $c \in \mathbb{C}^\times$.
- [7, 38] Let H be a finite-dimensional pointed Hopf algebra with $G(H) \simeq \mathbb{S}_5$, but $H \not\simeq \mathbb{CS}_5$. It is not known whether $\dim \mathfrak{B}(\mathcal{O}_{2,3}^5, \operatorname{sgn} \otimes \varepsilon) < \infty$. Let $\mathcal{D}_1$, $\mathcal{D}_2$ be the data corresponding to $V_1 = M(\mathcal{O}_2^5, \operatorname{sgn} \otimes \operatorname{id})$, $V_2 = M(\mathcal{O}_2^5, \operatorname{sgn} \otimes \operatorname{sgn})$. If the infinitesimal braiding of H is not $M(\mathcal{O}_{2,3}^5, \operatorname{sgn} \otimes \varepsilon)$, then H is isomorphic to one of $\mathfrak{u}(\mathcal{D}_1, (0,\mu))$, $\mu \in \mathbb{C}^2$ (defined as in Example 3.22), or $\mathfrak{B}(V_2)\#\mathbb{CS}_5$, or $\mathfrak{u}(\mathcal{D}_2, 1)$ (defined as above).
- [7] Let $m > 6$. Let $H \not\simeq \mathbb{CS}_m$ be a finite-dimensional pointed Hopf algebra with $G(H) \simeq \mathbb{S}_m$. Then the infinitesimal braiding of H is $V = M(\mathcal{O}, \rho)$, where the type of σ is $(1^{m-2}, 2)$ and $\rho = \rho_1 \otimes \operatorname{sgn}$, $\rho_1 = \operatorname{sgn}$ or ε; it is an open question whether $\dim \mathfrak{B}(V) < \infty$, see Example 3.12. If $m = 6$, there are two more Nichols algebras with unknown dimension corresponding to the class of type (2^3), but they are conjugated to those of type $(1^4, 2)$ by the outer automorphism of $\mathbb{S}_6$.
- [34] Let $m \geq 12$, $m = 4h$ with $h \in \mathbb{N}$. Let $G = \mathbb{D}_m$ be the dihedral group of order $2m$. Then there are infinitely many finite-dimensional Nichols algebras in ${}^G_G\mathcal{YD}$; all of them are exterior algebras as braided Hopf algebras. Let H be a finite-dimensional pointed Hopf algebra with $G(H) \simeq \mathbb{D}_m$, but $H \not\simeq \mathbb{CD}_m$. Then H is a lifting of an exterior algebra, and there infinitely many such liftings.

4.2.1. Copointed Hopf algebras. We say that a semisimple Hopf algebra K *collapses* if any finite-dimensional Hopf algebra H with $H_0 \simeq K$ is isomorphic to K. Thus, if G collapses, then $\mathbb{C}^G$ and $(\mathbb{C}G)^F$ collapse, for any twist F. Next we state the classification of the finite-dimensional copointed Hopf algebras over $\mathbb{S}_3$ [17]. Let $V = M(\mathcal{O}_2^3, \operatorname{sgn})$ as a Yetter-Drinfeld module over $\mathbb{C}^{\mathbb{S}_3}$. Let $\lambda \in \mathbb{C}^{\mathcal{O}_2^3}$ be such that $\sum_{(ij) \in \mathcal{O}_2^3} \lambda_{ij} = 0$. Let $\mathfrak{v} = \mathfrak{v}(V, \lambda)$ be the quotient of $T(V)\#\mathbb{C}^{\mathbb{S}_3}$ by the relations $a_{(13)}a_{(23)} + a_{(12)}a_{(13)} + a_{(23)}a_{(12)} = 0$, $a_{(23)}a_{(13)} + a_{(13)}a_{(12)} + a_{(12)}a_{(23)} = 0$, $a_{(ij)}^2 = \sum_{g \in \mathbb{S}_3} (\lambda_{ij} - \lambda_{g^{-1}(ij)g})\delta_g$, for $(ij) \in \mathcal{O}_2^3$. Then $\mathfrak{v}$ is a Hopf algebra of dimension 72 and $\operatorname{gr} \mathfrak{v} \simeq \mathfrak{B}(V)\#\mathbb{C}^{\mathbb{S}_3}$. Any finite-dimensional

copointed Hopf algebra H with $H_0 \simeq \mathbb{C}^{\mathbb{S}_3}$ is isomorphic to $\mathfrak{v}(V, \lambda)$ for some λ as above; $\mathfrak{v}(V, \lambda) \simeq \mathfrak{v}(V, \lambda')$ iff λ and λ' are conjugated under $\mathbb{C}^\times \times \operatorname{Aut} \mathbb{S}_3$.

Acknowledgements. This work was partially supported by ANPCyT-Foncyt, CONICET, Secyt (UNC). I thank all my coauthors for pleasant and fruitful collaborations. I am particularly in debt with Hans-Jürgen Schneider, Matías Graña and Iván Angiono for sharing their ideas on pointed Hopf algebras with me.

References

[1] Andruskiewitsch, N., *About finite dimensional Hopf algebras*, Contemp. Math. **294** (2002), 1–54.

[2] Andruskiewitsch, N. and Angiono, I., *Weyl groupoids, contragredient Lie superalgebras and Nichols algebras*, in preparation.

[3] Andruskiewitsch, N., Angiono, I., García Iglesias, A., Masuoka, A., and Vay, C., *Lifting via cocycle deformation*, J. Pure Appl. Alg. **218** (2014), 684–703.

[4] Andruskiewitsch, N., Carnovale, G., and García, G. A., *Finite-dimensional pointed Hopf algebras over finite simple groups of Lie type I*, arXiv:1312.6238.

[5] Andruskiewitsch, N. and Cuadra, J., *On the structure of (co-Frobenius) Hopf algebras*, J. Noncommut. Geom. **7** (2013), 83–104.

[6] Andruskiewitsch, N., Etingof, P., and Gelaki, S., *Triangular Hopf algebras with the Chevalley property*, Michigan Math. J. **49** (2001), 277–298.

[7] Andruskiewitsch, N., Fantino, F., Graña, M., and Vendramin, L., *Finite-dimensional pointed Hopf algebras with alternating groups are trivial*, Ann. Mat. Pura Appl. (4), **190** (2011), 225–245.

[8] ———, *Pointed Hopf algebras over the sporadic simple groups*, J. Algebra **325** (2011), 305–320.

[9] Andruskiewitsch, N. and García, G. A., *Finite subgroups of a simple quantum group*, Compositio Math. **145** (2009), 476–500.

[10] Andruskiewitsch, N. and Graña, M., *From racks to pointed Hopf algebras*, Adv. Math. **178** (2003), 177–243.

[11] ———, *Examples of liftings of Nichols algebras over racks*, AMA Algebra Montp. Announc. (electronic), Paper **1**, (2003).

[12] Andruskiewitsch, N., Heckenberger, I., and Schneider, H.-J., *The Nichols algebra of a semisimple Yetter-Drinfeld module*, Amer. J. Math. **132** (2010), 1493–1547.

[13] Andruskiewitsch, N. and Schneider, H.-J., *Lifting of quantum linear spaces and pointed Hopf algebras of order p^3*, J. Algebra **209** (1998), 658–691.

[14] ———, *Finite quantum groups and Cartan matrices*, Adv. Math. **154** (2000), 1–45.

[15] ———, *Pointed Hopf algebras, in New directions in Hopf algebras*, MSRI series, Cambridge Univ. Press (2002), 1–68.

[16] ———, *On the classification of finite-dimensional pointed Hopf algebras*, Ann. of Math. **171** (2010), 375–417.

[17] Andruskiewitsch, N. and Vay, C., *Finite dimensional Hopf algebras over the dual group algebra of the symmetric group in 3 letters*, Comm. Algebra **39** (2011), 4507–4517.

[18] Angiono, I., *A presentation by generators and relations of Nichols algebras of diagonal type and convex orders on root systems*, J. Europ. Math. Soc., to appear.

[19] ———, *On Nichols algebras of diagonal type*, J. Reine Angew. Math. **683** (2013), 189–251.

[20] Beattie, M., Dăscălescu, S., and Grünenfelder, L., *On the number of types of finite-dimensional Hopf algebras*, Invent. Math. **136** (1999), 1–7.

[21] Beattie, M. and García, G. A., *Classifying Hopf algebras of a given dimension*, Contemp. Math. **585** (2013), 125–152.

[22] Cuntz, M. and Heckenberger, I., *Weyl groupoids with at most three objects*, J. Pure Appl. Algebra **213** (2009), 1112–1128.

[23] ———, *Finite Weyl groupoids*. J. Reine Angew. Math., to appear.

[24] Doi, Y. and Takeuchi, M., *Multiplication alteration by two-cocycles. The quantum version*, Comm. Algebra **22** (1994), 5715–5732.

[25] Drinfeld, V. G., *Quantum groups*, Proc. Int. Congr. Math., Berkeley/Calif. 1986, Vol. 1, 798–820 (1987).

[26] ———, *Quasi-Hopf algebras*, Leningrad Math. J. **1** (1990), 1419–1457.

[27] Etingof, P. and Gelaki, S., *Some properties of finite-dimensional semisimple Hopf algebras*, Math. Res. Lett. **5** (1998), 191–197.

[28] ———, *The classification of triangular semisimple and cosemisimple Hopf algebras over an algebraically closed field*, Int. Math. Res. Not. **2000**, No. 5 (2000), 223–234.

[29] ———, *The classification of finite-dimensional triangular Hopf algebras over an algebraically closed field of characteristic* 0, Mosc. Math. J. **3** (2003), 37–43.

[30] Etingof, P., Gelaki, S., Nikshych, D., and Ostrik, V., *Tensor categories*, book to appear.

[31] Etingof, P. and Nikshych, D., *Dynamical quantum groups at roots of* 1, Duke Math. J. **108** (2001), 135–168.

[32] Etingof, P., Nikshych, D., and Ostrik, V., *Weakly group-theoretical and solvable fusion categories*, Adv. Math. **226** (2011), 176–205.

[33] Fantino, F., *Conjugacy classes of p-cycles of type D in alternating groups*, Comm. Algebra, to appear.

[34] Fantino, F. and García, G. A., *On pointed Hopf algebras over dihedral groups*, Pacific J. Math. **252** (2011), 69–91.

[35] Fantino, F. and Vendramin, L., *On twisted conjugacy classes of type D in sporadic simple groups*, Contemp. Math. **585** (2013), 247–259.

[36] Fomin, S. and Kirillov, K. N., *Quadratic algebras, Dunkl elements, and Schubert calculus*, Progr. Math. **172** (1999), 146–182.

[37] Galindo, C. and Natale, S., *Simple Hopf algebras and deformations of finite groups*, Math. Res. Lett. **14** (2007), 943–954.

[38] García, G. A. and García Iglesias, A., *Finite dimensional pointed Hopf algebras over* $\mathbb{S}_4$, Israel J. Math. **183** (2011), 417–444.

[39] García Iglesias, A. and Vay, C., *Finite-dimensional pointed or copointed Hopf algebras over affine racks*, J. Algebra **397** (2014), 379–406.

[40] Gelaki, S., Naidu, D., and Nikshych, D., *Centers of graded fusion categories*, Algebra Number Theory **3** (2009), 959–990.

[41] Graña, M., *On Nichols algebras of low dimension*, Contemp. Math. **267** (2000), 111–134.

[42] Graña, M. and Heckenberger, I., *On a factorization of graded Hopf algebras using Lyndon words*, J. Algebra **314** (2007), 324–343.

[43] Heckenberger, I., *The Weyl groupoid of a Nichols algebra of diagonal type*, Invent. Math. **164** (2006), 175–188.

[44] ———, *Classification of arithmetic root systems*, Adv. Math. **220** (2009), 59–124.

[45] Heckenberger, I., Lochmann, A., and Vendramin, L., *Braided racks, Hurwitz actions and Nichols algebras with many cubic relations*, Transform. Groups **17** (2012), 157–194.

[46] Heckenberger, I. and Schneider, H.-J., *Root systems and Weyl groupoids for Nichols algebras*, Proc. Lond. Math. Soc. **101** (2010), 623–654.

[47] ———, *Nichols algebras over groups with finite root system of rank two I*, J. Algebra **324** (2010), 3090–3114.

[48] ———, *Right coideal subalgebras of Nichols algebras and the Duflo order on the Weyl groupoid*, Israel J. Math. **197** (2013), 139–187.

[49] Heckenberger, I. and Vendramin, L., *Nichols algebras over groups with finite root system of rank two III*, arXiv:1309.4634.

[50] ———, *The classification of Nichols algebras with finite root system of rank two*, arXiv:1311.2881.

[51] Heckenberger, I. and Yamane, H., *A generalization of Coxeter groups, root systems, and Matsumoto's theorem*, Math. Z. **259** (2008), 255–276.

[52] Helbig, M., *On the lifting of Nichols algebras*, Comm. Algebra **40** (2012), 3317–3351.

[53] Jimbo, M., *A q-difference analogue of $U(\mathfrak{g})$ and the Yang-Baxter equation*, Lett. Math. Phys. **10** (1985), 63–69.

[54] Joyal, A. and Street, R., *Braided tensor categories*, Adv. Math. **102** (1993), 20–78.

[55] Joyce, D., *Simple quandles*, J. Algebra **79** (1982), 307–318.

[56] Kac, G., *Extensions of groups to ring groups*, Math. USSR. Sb. **5** (1968), 451–474.

[57] Kac, V., *Infinite-dimensional Lie algebras*. 3rd edition. Cambridge Univ. Press, 1990.

[58] Kharchenko, V., *A quantum analogue of the Poincaré–Birkhoff–Witt theorem*, Algebra and Logic **38** (1999), 259–276.

[59] Lusztig, G., *Finite dimensional Hopf algebras arising from quantized universal enveloping algebras*, J. Amer. Math. Soc. **3** (1990), 257–296.

[60] ———, *Quantum groups at roots of* 1, Geom. Dedicata **35** (1990), 89–114.

[61] ———, *Introduction to quantum groups*, Birkhäuser, 1993.

[62] Majid, S., *Foundations of quantum group theory*, Cambridge Univ. Press, 1995.

[63] Masuoka, *A., Abelian and non-abelian second cohomologies of quantized enveloping algebras*, J. Algebra **320** (2008), 1–47.

[64] Milinski, A. and Schneider, H.-J., *Pointed indecomposable Hopf algebras over Coxeter groups*, Contemp. Math. **267** (2000), 215–236.

[65] Mombelli, M., *Families of finite-dimensional Hopf algebras with the Chevalley property*, Algebr. Represent. Theory **16** (2013), 421–435.

[66] Montgomery, S., *Hopf algebras and their actions on rings*, CMBS **82**, Amer. Math. Soc. 1993.

[67] Montgomery, S. and Whiterspoon, S., *Irreducible representations of crossed products*, J. Pure Appl. Algebra **129** (1998), 315–326.

[68] Natale, S., *Semisolvability of semisimple Hopf algebras of low dimension*, Mem. Amer. Math. Soc. **186**, 123 pp. (2007).

[69] ———, *Hopf algebra extensions of group algebras and Tambara-Yamagami categories*, Algebr. Represent. Theory **13** (2010), 673–691.

[70] ———, *Semisimple Hopf algebras of dimension* 60, J. Algebra **324** (2010), 3017–3034.

[71] Ng, S.-H., *Non-semisimple Hopf algebras of dimension p^2*, J. Algebra **255** (2002), 182–197.

[72] Nichols, W. D., *Bialgebras of type one*, Comm. Algebra **6** (1978), 1521–1552.

[73] Nikshych, D., *K_0-rings and twisting of finite-dimensional semisimple Hopf algebras*, Comm. Algebra **26** (1998), 321–342; Corrigendum, **26** (1998), 2019.

[74] ———, *Non-group-theoretical semisimple Hopf algebras from group actions on fusion categories*, Selecta Math. (N. S.) **14** (2008), 145–161.

[75] Radford, D. E., *Hopf algebras*, Series on Knots and Everything **49**, Hackensack, NJ: World Scientific. xxii, 559 p. (2012).

[76] Rosso, M., *Quantum groups and quantum shuffles*, Invent. Math. **133** (1998), 399–416.

[77] Schauenburg, P., *Hopf bi-Galois extensions*, Comm. Algebra **24** (1996), 3797–3825.

[78] ———, *A characterization of the Borel-like subalgebras of quantum enveloping algebras*, Comm. Algebra **24** (1996), 2811–2823.

[79] Schneider, H.-J., *Lectures on Hopf algebras*, Trab. Mat. 31/95 (FaMAF), 1995.

[80] Serganova, V., *Kac-Moody superalgebras and integrability*, Progr. Math. **288** (2011), 169–218.

[81] Ştefan, D., *The set of types of n-dimensional semisimple and cosemisimple Hopf algebras is finite*, J. Algebra **193** (1997), 571–580.

[82] ———, *Hopf algebras of low dimension*, J. Algebra **211** (1999), 343–361.

[83] Sweedler, M. E., *Hopf algebras*, Benjamin, New York, 1969.

[84] Vendramin, L., *Nichols algebras associated to the transpositions of the symmetric group are twist-equivalent*, Proc. Amer. Math. Soc. **140** (2012), 3715–3723.

[85] Woronowicz, S. L., *Differential calculus on compact matrix pseudogroups (quantum groups)*, Comm. Math. Phys. **122** (1989), 125–170.

[86] Zhu, Y., *Hopf algebras of prime dimension*, Intern. Math. Res. Not. **1** (1994), 53–59.

FaMAF, Universidad Nacional de Córdoba. CIEM – CONICET. , Medina Allende s/n (5000) Ciudad Universitaria, Córdoba, Argentina

E-mail: andrus@famaf.unc.edu.ar

Excision, descent, and singularity in algebraic K-theory

Guillermo Cortiñas

Abstract. Algebraic K-theory is a homology theory that behaves very well on sufficiently nice objects such as stable C^*-algebras or smooth algebraic varieties, and very badly in singular situations. This survey explains how to exploit this to detect singularity phenomena using K-theory and cyclic homology.

Mathematics Subject Classification (2010). Primary 19D55; Secondary 19D50, 19E08.

Keywords. Algebraic K-theory, cyclic homology, topological algebras, singular varieties.

1. Introduction

A homology theory of rings associates groups $H_n(R)$ ($n \in \mathbb{Z}$) depending covariantly on the ring R. We say that H is (polynomially) *homotopy invariant* if the map $H_n(R) \to H_n(R[t])$ is an isomorphism. A homology theory also associates groups $H_n(R : I)$ to every two-sided ideal $I \lhd R$, which fit into a long exact sequence

$$H_{n+1}(R/I) \to H_n(R : I) \to H_n(R) \to H_n(R/I).$$

We say that H is *nilinvariant* if $H_*(R : I) = 0$ when I is nilpotent. The homology theory is said to satisfy *excision* if whenever $f : R \to S$ is a ring homomorphism sending an ideal $I \lhd R$ isomorphically onto an ideal of S, the map $H_*(R : I) \to H_*(S : f(I))$ is an isomorphism. Similarly, a cohomology theory of algebraic varieties over a field (schemes of finite type) associates groups $H_n(X : Y)$ with every closed immersion $Y \to X$, depending contravariantly on $(X : Y)$. In particular if $\tilde{X}$ is the blowup of X along Y and $\tilde{Y}$ is the exceptional fiber, we have a map $H_*(X : Y) \to H_*(\tilde{X} : \tilde{Y})$; this map is an isomorphism whenever H satisfies *cdh-descent*. For example, Quillen's *algebraic K-theory* K is a (co)homology theory of rings and schemes which has none of the aforementioned properties. For another example, Weibel's *homotopy algebraic K-theory* KH is also defined for rings and schemes, and satisfies all of them. There is a natural map $K_n(R) \to KH_n(R)$ which is an isomorphism when R is K_n*-regular*; this means that the map $K_n(R) \to K_n(R[t_1, \dots, t_p])$ is an isomorphism for all p. For instance Noetherian regular rings such as rings of polynomial functions on smooth varieties, and stable C^*-algebras such as the algebra $\mathcal{K}$ of compact operators, are K_n-regular for all n. In general the map $K \to KH$ is part of a long exact sequence

$$KH_{n+1}(R) \to K_n^{\mathrm{nil}}(R) \to K_n(R) \to KH_n(R).$$

Proceedings of International Congress of Mathematicians, 2014, Seoul

In this article we survey a series of results that help describe the groups K_*^{nil} in terms of cyclic homology, and explain how this description has helped make significant progress in several long standing problems, ranging from the comparison of algebraic and topological K-theory of topological algebras to the relation between K-regularity and nonsingularity of algebraic varieties.

The article is organized as follows. In Section 2 we recall (from Goodwillie's paper [29] and from [6]) two key properties of the Chern character $ch_* : K_*(R) \to HN_*(R)$ to negative cyclic homology of $\mathbb{Q}$-algebras. They can be summarized by saying that *infinitesimal K-theory* is nilinvariant and satisfies excision (Theorem 2.4); the infinitesimal K-groups fit into a long exact sequence

$$HN_{n+1}(R) \to K_n^{\mathrm{inf}}(R) \to K_n(R) \xrightarrow{ch_n} HN_n(R).$$

In Section 3 we explain how the results of the previous one were used in [15] to compare the algebraic and the topological K-theory of a stable locally convex algebra L. The main result reviewed in this section is Theorem 3.2, which says that for such L, $KH_*(L) = K_*^{\mathrm{top}}(L)$ and $K_n^{\mathrm{nil}}(L) = HC_{n-1}(L)$ is cyclic homology. In Section 4 we recall the notions of descent and Mayer-Vietories properties. We review Thomason's theorems that K-theory satisfies Nisnevich descent and has the Mayer-Vietoris property for blow-ups along regularly embedded closed subschemes (Theorem 4.1), and their analogues for cyclic homology and infinitesimal K-theory of schemes of finite type over a field of characteristic zero (Theorem 4.2). In Section 5 we review Haesemeyer's theorem on cdh-descent and its generalization (Theorems 5.1 and 5.2) and derive from them a long exact sequence (Theorem 5.3)

$$KH_{n+1}(X) \to \mathcal{F}_{n-1}^{HC}(X) \to K_n(X) \to KH_n(X), \tag{1.1}$$

for every scheme X of finite type over a field of characteristic zero. Up to extension, the groups $\mathcal{F}_*^{HC}(X)$ are computed from the cyclic homology groups of X and of an array of smooth schemes that appear in its desingularization process. In Section 6 we show how these results were used to prove Weibel's dimension conjecture for schemes of finite type over a field of characteristic zero (Theorem 6.3). In Section 7 we begin by explaining how the sequence 1.1 can be used to compute the obstruction to K_n-regularity (Proposition 7.1). Then we review several results on K-regularity, including that Vorst's regularity conjecture and Gubeladze's nilpotence conjecture hold for algebras and coefficient rings containing a field of characteristic zero (Theorems 7.3 and 7.8), and the answer to Bass' question on whether $K_n(R) = K_n(R[t])$ implies that $K_n(R) = K_n(R[t_1, t_2])$ (Theorem 7.5). Versions of some of these results in characteristic $p > 0$ are reviewed in Section 8. They are based on the good properties of the Bökstedt-Hsiang-Madsen cyclotomic trace $\mathrm{tr} : K \to TC$, which takes values in topological cyclic homology [3]. Theorems of McCarthy and Geisser-Hesselholt (8.1 and 8.2) say that tr computes the obstruction groups to nilinvariance and excision up to p-adic completion. The results of Geisser-Hesselholt extending Haesemeyer's theorem (Theorem 8.4) and proving Weibel's dimension conjecture (Theorem 8.5) and Vorst's regularity conjecture (Theorem 8.6) over a perfect field of characteristic $p > 0$ which admits strong resolution of singularities are reviewed, as well as the solution of Gubeladze's nilpotence conjecture over a field of charactristic p (Theorem 8.8).

2. The obstructions to excision and nilinvariance

Quillen's algebraic K-theory associates groups $K_n(R)$ ($n \in \mathbb{Z}$) to each associative, not necessarily unital ring R. These groups are defined as the stable homotopy groups of a functorial spectrum $K(R)$. If $I \lhd R$ is an ideal, the relative K-theory spectrum is $K(R:I) = \mathrm{hofiber}(K(R) \to K(R/I))$. The spectrum $K(R:I)$ is defined so as to fit into a *homotopy fibration sequence*

$$K(R:I) \to K(R) \to K(R/I).$$

Thus taking homotopy groups we obtain a long exact sequence

$$K_{n+1}(R/I) \to K_n(R:I) \to K_n(R) \to K_n(R/I).$$

Observe that the failure of the map $K_*(R) \to K_*(R/I)$ to be an isomorphism is measured by the relative groups $K_*(R:I)$. If I is nilpotent, the groups $K_n(R:I)$ vanish for $n \leq 0$. Thus $K_n(R) \to K_n(R/I)$ is an isomorphism for $n \leq 0$; because of this, we say that K-theory is *nilinvariant* in nonpositive degrees. For $n \geq 1$ the groups $K_n(R:I)$ may be nonzero for nilpotent I; if R is a $\mathbb{Q}$-algebra they are $\mathbb{Q}$-vector spaces ([49, Consequence 1.4]). For general R, the groups $K_n(R:I) \otimes \mathbb{Q}$ are computed by means of the Chern character [29]

$$ch : K_*(R) \to HN_*(R \otimes \mathbb{Q}). \tag{2.1}$$

Negative cyclic homology is connected with cyclic and periodic cyclic homology by means of Connes' SBI-sequence

$$HP_{n+1}(R) \xrightarrow{S} HC_{n-1}(R) \xrightarrow{B} HN_n(R) \xrightarrow{I} HP_n(R). \tag{2.2}$$

Theorem 2.1 (Goodwillie, [28, Theorem II.5.1], [29, Main Theorem]). *Let R be a ring and $I \lhd R$ a nilpotent ideal. Then*

$$HP_n(R \otimes \mathbb{Q} : I \otimes \mathbb{Q}) = 0,$$

and the Chern character induces an isomorphism

$$K_n(R:I) \otimes \mathbb{Q} \cong HN_n(R \otimes \mathbb{Q} : I \otimes \mathbb{Q}) \cong HC_{n-1}(R \otimes \mathbb{Q} : I \otimes \mathbb{Q}).$$

In particular, $K_n(R:I) \otimes \mathbb{Q} \cong K_n(R \otimes \mathbb{Q} : I \otimes \mathbb{Q})$.

Remark 2.2. Goodwillie states his results for unital R; the nonunital case follows from the unital case (see [5, Lemma 6.1], e.g.).

If $I \lhd R$ is an ideal and $f : R \to S$ is a ring homomorphism such that $f(I) \lhd S$ is an ideal and $f : I \to f(I)$ is bijective, we put

$$K(R, S : I) = \mathrm{hofiber}(K(R:I) \to K(S : f(I))).$$

The groups $K_n(R, S : I)$ are zero for $n \leq -1$. Thus for $n \leq 0$ the groups $K_n(R:I)$ depend only on I and not on the ideal embedding $I \lhd R$. Because of this, we say that K-theory satisfies *excision* (or that it is *excisive*) in nonpositive degrees. If R is a $\mathbb{Q}$-algebra, the groups $K_n(R, S : I)$ are $\mathbb{Q}$-vector spaces ([49, Consequence 1.5]). For general R, the

groups $K_n(R, S : I) \otimes \mathbb{Q}$ are again computed by means of the Chern character; this is the content of Theorem 2.3 below. First let us recall the Cuntz-Quillen excision theorem [18, Theorem 5.3], which says that HP *satisfies excision* in the category of $\mathbb{Q}$-algebras; this means that if in the situation just described f is a $\mathbb{Q}$-algebra homomorphism, then

$$HP_*(R, S : I) = 0. \tag{2.3}$$

It follows that the B-map in the SBI sequence (2.2) induces an isomorphism

$$HC_{*-1}(R, S : I) \cong HN_*(R, S : I).$$

Theorem 2.3 ([6, Theorem 0.1]). *Let $I \lhd R$ be an ideal and let $f : R \to S$ be a ring homomorphism. Assume that $f(I) \lhd S$ is an ideal and that $f : I \to f(I)$ is bijective. Then the Chern character induces an isomorphism*

$$\begin{aligned} K_*(R, S : I) \otimes \mathbb{Q} &\cong HN_*(R \otimes \mathbb{Q}, S \otimes \mathbb{Q} : I \otimes \mathbb{Q}) \\ &\cong HC_{*-1}(R \otimes \mathbb{Q}, S \otimes \mathbb{Q} : I \otimes \mathbb{Q}). \end{aligned}$$

In particular, $K_*(R, S : I) \otimes \mathbb{Q} \cong K_*(R \otimes \mathbb{Q}, S \otimes \mathbb{Q} : I \otimes \mathbb{Q})$.

In the case of $\mathbb{Q}$-algebras Theorem 2.3 confirmed a conjecture formulated in the mid 1980's by Geller-Reid-Weibel ([24, 25]). Its proof combines the techniques developed by Cuntz-Quillen to prove their excision theorem with those used by Suslin-Wodzicki to characterize those nonunital rings I such that $K_*(-, - : I) \otimes \mathbb{Q} = 0$ [44, Theorem A].

In the applications considered in the next sections, all the rings involved are $\mathbb{Q}$-algebras. In this case we will find it useful to formulate the results above in terms of infinitesimal K-theory. The Chern character (2.1) comes from a map of spectra $K \to HN$; *infinitesimal K-theory* is the homotopy fiber of this map:

$$K^{\inf}(R) = \operatorname{hofiber}(K(R) \to HN(R)).$$

In terms of $K^{\inf}$, Theorems 2.1 and 2.3 can be expressed as follows.

Theorem 2.4. *Infinitesimal K-theory of $\mathbb{Q}$-algebras is nilinvariant and satisfies excision.*

3. $K^{\inf}$-regularity and algebraic K-theory of topological algebras.

Let A be a ring and $F : \mathrm{Rings} \to \mathfrak{Ab}$ a functor. We say that A is *F-regular* if the map induced by the canonical inclusion

$$F(A) \to F(A[t_1, \dots, t_n])$$

is an isomorphism for all $n \geq 1$. The functor F is called *invariant under polynomial homotopy* (homotopy invariant, for short) on a subcategory $\mathfrak{C} \subset \mathrm{Rings}$ if every ring $A \in \mathfrak{C}$ is F-regular. A *homology theory* of rings is a covariant functor E from rings to spectra; we write $E_n(R) = \pi_n E(R)$ for the stable homotopy group. If E is a homology theory of rings, we call a ring *E-regular* if it is E_n-regular for all n. We call E *homotopy invariant* if each E_n is homotopy invariant. There is a well-known procedure $E \to EH$, the *homotopization procedure* that transforms E to a homotopy invariant homology

theory EH. This construction comes with a natural map $E(A) \to EH(A)$ and setting $E^{\mathrm{nil}}(A) = \mathrm{hofiber}(E(A) \to EH(A))$ one obtains a long exact sequence

$$EH_{n+1}(A) \to E_n^{\mathrm{nil}}(A) \to E_n(A) \to EH_n(A). \tag{3.1}$$

If A is E_m-regular for all $m \leq n$, then $E_m(A) \to EH_m(A)$ is an isomorphism for $m \leq n$ and is onto for $m = n+1$. Applying this procedure to Quillen's K-theory yields Weibel's homotopy K-theory KH [50]. The following theorem summarizes two key properties of KH.

Theorem 3.1 (Weibel, [50, Theorems 2.1 and 2.3]). *Homotopy algebraic K-theory KH is nilinvariant and satisfies excision.*

If A is a $\mathbb{Q}$-algebra, then A is HP-regular (e.g. by [28, Corollary II.4.4] or by [38, Equation 3.13]). Geller and Weibel showed in [26, Theorem 4.1] that if A is a $\mathbb{Q}$-algebra then $HNH_*(A) \cong HP_*(A)$, and (3.1) identifies with Connes' SBI-sequence (2.2). If A is a $\mathbb{Q}$-algebra, set

$$K^{\mathrm{ninf}}(A) = K^{\mathrm{inf,nil}}(A).$$

The homotopization procedure preserves fibration sequences. Hence we have a homotopy commutative diagram whose rows and columns are homotopy fibration sequences:

$$\begin{array}{ccccc} K^{\mathrm{ninf}}(A) & \longrightarrow & K^{\mathrm{inf}}(A) & \longrightarrow & K^{\mathrm{inf}}H(A) \\ \downarrow & & \downarrow & & \downarrow \\ K^{\mathrm{nil}}(A) & \longrightarrow & K(A) & \longrightarrow & KH(A) \\ \downarrow & & \downarrow & & \downarrow \\ HC(A)[-1] & \longrightarrow & HN(A) & \longrightarrow & HP(A). \end{array} \tag{3.2}$$

By the discussion above,

$$A \;\; K^{\mathrm{inf}} - \text{regular} \;\Rightarrow K_*^{\mathrm{nil}}(A) \xrightarrow{\cong} HC_{*-1}(A). \tag{3.3}$$

A *locally convex algebra* is a complete locally convex $\mathbb{C}$-vector space L equipped with a jointly continuous, associative multiplication $L \otimes_{\mathbb{C}} L \to L$. The following theorem concerns the K-theory of a large class of stable locally convex algebras. It computes the obstruction for the map $K \to K^{\mathrm{top}}$ to be an isomorphism. Here $K_*^{\mathrm{top}}(L) = kk_*^{lc}(\mathbb{C}, L)$ is the covariant part of Cuntz' bivariant K-theory for locally convex algebras [17]. As expected of a bivariant K-theory of complex topological algebras, it is periodic of period two:

$$kk_*^{lc}(L, M) = kk_{*+2}^{lc}(L, M). \tag{3.4}$$

Thus there are only two covariant topological K-groups, K_0^{top} and K_1^{top}.

Let $\mathcal{B} = \mathcal{B}(\ell^2(\mathbb{N}))$ be the C^*-algebra of bounded operators in an infinite dimensional, separable Hilbert space. An ideal $I \lhd \mathcal{B}$ is a *Fréchet operator ideal* if it carries a complete metrizable locally convex topology such that the inclusion $I \to \mathcal{B}$ is continuous.

Theorem 3.2 ([15, Theorems 6.2.1 and 6.3.1]). *Let L be a locally convex $\mathbb{C}$-algebra and let I be a Fréchet operator ideal. Write $\hat{\otimes}$ for the projective tensor product. Then:*

(i) $L\hat{\otimes}I$ *is* $K^{\inf}$*-regular.*

(ii) $KH_*(L\hat{\otimes}I) = K_*^{\mathrm{top}}(L\hat{\otimes}I)$.

(iii) *For each* $n \in \mathbb{Z}$*, there is a natural* 6*-term exact sequence of abelian groups as follows:*

$$\begin{array}{ccccc} K_1^{\mathrm{top}}(L\hat{\otimes}I) & \longrightarrow & HC_{2n-1}(L\hat{\otimes}I) & \longrightarrow & K_{2n}(L\hat{\otimes}I) \\ \uparrow & & & & \downarrow \\ K_{2n-1}(L\hat{\otimes}I) & \longleftarrow & HC_{2n-2}(L\hat{\otimes}I) & \longleftarrow & K_0^{\mathrm{top}}(L\hat{\otimes}I). \end{array}$$

Sketch of the proof of Theorem 3.2. Call a Banach ideal $I \lhd \mathcal{B}$ *harmonic* if it contains an operator whose sequence of singular values is the harmonic sequence $\{1/n\}$. Results of Cuntz-Thom [19, Theorems 4.2.1 and 5.1.2] imply that if H is a homology theory which satisfies excision and $I \lhd \mathcal{B}$ is harmonic, then $H(-\hat{\otimes}I)$ is invariant under C^∞ homotopies, also called *diffotopies*. It is shown in [15, Theorem 6.16] that if in addition H is nilinvariant, then $H(-\hat{\otimes}I)$ is diffotopy invariant for any Fréchet ideal I. In particular this applies to $K^{\inf}$ and KH by Theorems 2.4 and 3.1. Part i) of the theorem follows from the fact that $K^{\inf}(-\hat{\otimes}I)$ is diffotopy invariant, using the argument of [42, Theorem 3.4]. By [15, Lemma 3.2.1], $K^{\inf}$-regular $\mathbb{Q}$-algebras are K_n-regular for $n \leq 0$. Hence we have $KH_n(L\hat{\otimes}I) = K_n(L\hat{\otimes}I)$ for $n \leq 0$. Cuntz and Thom proved in [19, Theorem 6.2.1] that $K_0(L\hat{\otimes}I) = K_0^{\mathrm{top}}(L\hat{\otimes}I)$ when I is harmonic; by the argument of [15, Theorem 6.16], this extends to all Fréchet ideals. Summing up $KH(-\hat{\otimes}I)$ is an excisive and diffotopy invariant homology theory in $\mathfrak{Loc}\mathfrak{Alg}$ which agrees with $K^{\mathrm{top}}(-\hat{\otimes}I)$ in dimension 0. A standard argument now shows that they must agree in all dimensions ([16, Abschnitt 6]). This proves ii). Part iii) follows from what we have already done, using (3.3), (3.4), and diagram (3.2). □

The following theorem can be proved using Theorem 3.2.

Theorem 3.3 ([15, Theorem 8.3.3]). *Let* $\mathcal{K} \lhd \mathcal{B}$ *be the ideal of compact operators and let* L *be a Fréchet algebra whose topology is generated by a countable family of submultiplicative seminorms and which has a uniformly bounded, one-sided approximate unit. Then the map*

$$K_*(L\hat{\otimes}\mathcal{K}) \to K_*^{\mathrm{top}}(L\hat{\otimes}\mathcal{K})$$

is an isomorphism.

The particular case of the theorem above when L is a Banach algebra with a one-sided approximate unit confirms a conjecture formulated by Karoubi in [37]. Wodzicki announced a proof of the latter case in [54, Theorem 1]; he told us he has also proved the general case of Theorem 3.3. The proof of Theorem 3.3 given in [15] consists of showing that $HC_*(L\hat{\otimes}\mathcal{K}) = 0$, and then using Theorem 3.2. For a different proof see [7, Theorem 12.1.1]. Karoubi also proved that if $\mathfrak{A}$ is a C^*-algebra and $\otimes$ is the spatial tensor product, then the comparison map

$$K_*(\mathfrak{A}\otimes\mathcal{K}) \to K_*^{\mathrm{top}}(\mathfrak{A}\otimes\mathcal{K}) \tag{3.5}$$

is an isomorphism for $* \leq 0$, and conjectured that it is also an isomorphism for $* \geq 1$. This conjecture was proved by Suslin-Wodzicki ([44, Theorem 10.9]). We remark that the analog of Theorem 3.2 for $\mathfrak{A}\otimes\mathcal{K}$ also holds. Indeed a theorem of Higson ([42, Theorem 20]) says that $\mathfrak{A}\otimes\mathcal{K}$ is K-regular, so $KH_*(\mathfrak{A}\otimes\mathcal{K}) = K_*^{\mathrm{top}}(\mathfrak{A}\otimes\mathcal{K})$, since (3.5) is an isomorphism; a

similar argument as that of [42, Theorem 20] shows that $\mathfrak{A}\underset{\sim}{\otimes}\mathcal{K}$ is also $K^{\inf}$-regular. Finally it is not hard to see that $HC_*(\mathfrak{A}\underset{\sim}{\otimes}\mathcal{K}) = 0$ (the argument is similar to that of [15, Lemma 8.2.3]).

Remark 3.4. No nonzero commutative unital $\mathbb{Q}$-algebra is $K^{\inf}$-regular. Indeed [15, Proposition 3.3.1] says that if a unital algebra R is $K_n^{\inf}$-regular for some $n \geq 1$, then every element $a \in R$ can be written as a finite linear combination

$$a = \sum_{i=1}^{m} b_i(x_i y_i - y_i x_i)c_i \qquad (b_i, c_i, x_i, y_i \in R).$$

4. Mayer-Vietoris sequences and descent

Fix a field k. Let Sch/k be the category of schemes essentially of finite type over $\mathrm{Spec}(k)$. A *k-scheme* is an object of Sch/k. A *cohomology theory* (also called *presheaf of spectra*) is a functor $(\mathrm{Sch}/k)^{op} \to \mathrm{Spt}$. Let

$$\begin{array}{ccc} Y' & \longrightarrow & X' \\ \downarrow & & \downarrow{\scriptstyle p} \\ Y & \xrightarrow[i]{} & X \end{array} \tag{4.1}$$

be a cartesian square in Sch/k. A cohomology theory E of k-schemes is said to have the *Mayer-Vietoris* property with respect to (4.1) if it sends (4.1) to a homotopy cartesian diagram of spectra. This implies that we have a long exact sequence of Mayer-Vietoris type:

$$E_{n+1}(Y') \to E_n(X) \to E_n(Y) \oplus E_n(X') \to E_n(Y').$$

In the sequence above, we have used subscript notation; we will switch to superscripts when needed, following the usual convention $E_* = E^{-*}$.

We consider classes of cartesian squares in Sch/k which are closed under isomorphism. We call such a class a *cd-structure*. Each such class generates a Grothendieck topology on Sch/k. We shall consider the following cd-structures; note that each of them is contained in the next one:

- The cd-structure of elementary *Zariski* squares. A square (4.1) is an elementary Zariski square if i and p are open immersions and $X = Y' \cup X'$. We write *zar* for the Grothendieck topology generated by this structure.
- The cd-structure of elementary *Nisnevich* squares. A square (4.1) is an elementary Nisnevich square if i is an open embedding, p is étale, and $p : (X' \setminus Y')_{\mathrm{red}} \to (X \setminus Y)_{\mathrm{red}}$ is an isomorphism. Here X_{red} is the reduced scheme. We remark that in the particular case when p is an open immersion, the latter condition is equivalent to the condition that $Y' \cup X' = X$; thus every elementary Zariski square is an elementary Nisnevich square. We write *nis* for the Nisnevich topology.
- Voevodsky's *combined cd-structure*. It consists of the elementary Nisnevich squares and the *abstract blow-up squares*. A square (4.1) is an abstract blow-up if p is proper, i is a closed embedding and $p : (X' \setminus Y') \to (X \setminus Y)$ is an isomorphism. We write *cdh* for the corresponding topology.

Let c be one of the above cd-structures on Sch/k and let t be the Grothendieck topology it generates. A cohomology theory E of k-schemes satisfies *descent* with respect to t if it has the Mayer-Vietoris property for every square in c. There is a construction $(E,t)\mapsto \mathbb{H}_t(-,E)$ which given a topology t and a cohomology theory E produces a cohomology theory $\mathbb{H}_t(-,E)$ which satisfies t-descent, and a natural transformation ([36])

$$E\to \mathbb{H}_t(-,E). \tag{4.2}$$

The case when t comes from a cd-structure was studied in depth by Voevodsky, who obtained several key results in his fundamental article [47]. Using Voevodsky's results, one can show that if t comes from a cd-structure which meets some technical conditions [8, Theorem 3.4], which are satisfied in all the examples we consider here, then E satisfies t-descent if and only if the map (4.2) is a *global weak equivalence*; this means that

$$E_*(X)\to \mathbb{H}_t^{-*}(X,E)$$

is an isomorphism for all X.

A cohomology theory of schemes E which satisfies Zariski descent is determined by its value on affine schemes, and

$$E(R):=E(\operatorname{Spec} R) \tag{4.3}$$

is a homology theory of commutative rings which satisfies Zariski descent. Each of the homology theories of rings considered in the previous sections extends to a cohomology theory of schemes which satisfies Zariski descent so that (4.3) holds. Moreover, we have the following landmark results of Thomason. Recall that a closed immersion is *regular* if it is locally defined by a regular sequence. A closed subscheme $Y\subset X$ is of *pure codimension* when all its irreducible components have the same codimension. The square (4.1) is a *regular blow-up* if i is a regular closed immersion of pure codimension and p is the blow-up along i.

Theorem 4.1 (Thomason, [45, Theorem 10.8], [46, Théorème 2.1]). *K-theory of schemes satisfies Nisnevich descent and has the Mayer-Vietoris property for regular blow-up squares.*

Thomason defined the K-theory of a quasi-compact and separated scheme X in terms of the complicial category of perfect complexes of quasi-coherent sheaves; the main technical tool for proving Theorem 4.1 is the Thomason-Waldhausen localization theorem ([45, Theorems 1.8.2 and 1.9.8]) which roughly says that K-theory maps sequences of complicial categories which induce exact sequences of derived categories into homotopy fibration sequences of spectra. Weibel introduced cyclic homology of schemes as the hyperhomology of the sheafified Connes' complex [52]. He and Geller showed in [27, Theorem 4.8] that HC satisfies Nisnevich descent. Keller proved a version of the Thomason-Waldhausen localization theorem for cyclic homology [40, Theorem 2.4] and showed [39, Section 5.2] that for quasi-compact separated schemes the cyclic homology of schemes introduced by Weibel in [52] admits a categorical description analogous to that of K-theory. The proof given in [8] of the following theorem uses the latter result of Keller and the argument of Thomason's proof of Theorem 4.1.

Theorem 4.2 ([27, Theorem 4.8], [8, Theorems 2.9 and 2.10]). *Cyclic, negative cyclic and periodic cyclic homology satisfy Nisnevich descent and have the Mayer-Vietoris property for regular blow-ups.*

Corollary 4.3. *Let $E \in \{HC, HN, HP, K\}$ and let $X \in \mathrm{Sch}/k$. Assume that X is smooth. Then the map $E_*(X) \to H^{-*}_{\mathrm{cdh}}(X, E)$ is an isomorphism.*

Proof. By [8, Corollary 3.9], any cohomology theory on Sch/k which satisfies Nisnevich descent and has the Mayer-Vietoris property for regular blow-ups also satisfies this. □

A homology theory of k-schemes E is *invariant under infinitesimal thickenings* if $E_*(X) \to E_*(X_{\mathrm{red}})$ is an isomorphism for all X. An abstract blow-up square is a *finite blow-up* if p is a finite morphism. If E satisfies Zariski descent then E is invariant under infinitesimal thickenings (resp. has the Mayer-Vietoris property for finite blow-ups) if and only if the associated homology of commutative algebras (4.3) is nilinvariant (resp. satisfies excision).

The Chern character (2.1) extends to schemes. If $X \in \mathrm{Sch}/k$ and k is of characteristic zero it is given by a map of spectra $ch : K(X) \to HN(X)$; infinitesimal K-theory is defined as $K^{\mathrm{inf}}(X) = \mathrm{hofiber}(K(X) \to HN(X))$.

The following is an immediate corollary of Theorems 2.4, 4.1 and 4.2.

Corollary 4.4 ([8, Theorems 4.3 and 4.4]). *Let k be a field of characteristic zero. Then the infinitesimal K-theory of k-schemes is invariant under infinitesimal thickenings, satifies Nisnevich descent, and has the Mayer-Vietoris property for regular blow-ups and for finite blow-ups.*

5. Haesemeyer's theorem

In this section k is a field of characteristic zero. If E is a cohomology theory of k-schemes, we write $\mathcal{F}^E(X) = \mathrm{hofiber}(E(X) \to H_{cdh}(X, E))$. Thus we have a homotopy fibration sequence

$$\mathcal{F}^E(X) \to E(X) \to \mathbb{H}_{\mathrm{cdh}}(X, E). \tag{5.1}$$

Both $\mathbb{H}_{\mathrm{cdh}}(X, -)$ and $\mathcal{F}^-(X)$ preserve homotopy fibration sequences; this fact is used in diagram 5.2 below.

Theorem 5.1 (Haesemeyer, [34, Theorem 6.4]). *Let k be a field of characteristic zero and $X \in \mathrm{Sch}/k$. Then $KH(X) \xrightarrow{\sim} \mathbb{H}_{\mathrm{cdh}}(X, K)$.*

It follows from the theorem above that for $E = K$ and $X = \mathrm{Spec}\, A$, the sequence (5.1) is equivalent to the middle row of diagram (3.2). For $X \in \mathrm{Sch}/k$ we have a homotopy commutative diagram whose rows and columns are homotopy fibration sequences:

$$\begin{array}{ccccc}
\mathcal{F}^{K^{\mathrm{inf}}}(X) & \longrightarrow & K^{\mathrm{inf}}(X) & \longrightarrow & H_{\mathrm{cdh}}(X, K^{\mathrm{inf}}) \\
\downarrow & & \downarrow & & \downarrow \\
\mathcal{F}^{K}(X) & \longrightarrow & K(X) & \longrightarrow & KH(X) \\
\downarrow & & \downarrow & & \downarrow \\
\mathcal{F}^{HN}(X) & \longrightarrow & HN(X) & \longrightarrow & H_{\mathrm{cdh}}(X, HN).
\end{array} \tag{5.2}$$

Theorem 5.1 implies that KH satisfies cdh-descent. On the other hand we know from [50, Theorem 1.2] that it is also *homotopy invariant*, that is, it sends the projection $X \times \mathbb{A}^1 \to X$

to an isomorphism $KH_*(X) \cong KH_*(X \times \mathbb{A}^1)$. For the proof of Theorem 5.1, Haesemayer first showed that KH has the Mayer-Vietoris property for regular blow-ups ([34, Theorem 3.6]). In almost all the rest of the proof, the only other properties of KH that are used are excision and invariance under infinitesimal thickenings. Homotopy invariance is used only once, in [34, Proposition 3.7]. It was found later that homotopy invariance is not needed; this led to the following generalization of Haesemeyer's theorem.

Theorem 5.2 ([8, Theorem 3.12]). *Let k be a field of characteristic zero and let E be a cohomology theory of k-schemes. Assume that E satisfies excision, is invariant under infinitesimal thickenings, satisfies Nisnevich descent, and has the Mayer-Vietoris property for regular blow-ups. Then E satisfies* cdh*-descent.*

By Corollary 4.4, Theorem 5.2 applies to K^{inf}. It follows that $\mathcal{F}_*^{K^{\mathrm{inf}}}(X) = 0$ in diagram (5.2), and therefore we have a weak equivalence

$$\mathcal{F}^K(X) \xrightarrow{\sim} \mathcal{F}^{HN}(X).$$

By Cuntz-Quillen's theorem (2.3) and by Theorems 2.1 and 4.2, HP also satisfies the hypothesis of Theorem 5.2. It follows from this and the SBI-sequence that there is a weak equivalence

$$\mathcal{F}^{HC}(X)[-1] \xrightarrow{\sim} \mathcal{F}^{HN}(X).$$

Summing up, we have proved the following.

Theorem 5.3 ([14, Theorem 1.6], see also [8, Corollary 3.13 and Theorem 4.6]). *Let k be a field of characteristic zero and $X \in \mathrm{Sch}/k$. Then there is a homotopy fibration sequence*

$$\mathcal{F}^{HC}(X)[-1] \to K(X) \to KH(X).$$

Theorems 5.1 and 5.3 together say that the obstruction to cdh-descent in algebraic K-theory is measured by $\mathcal{F}^{HC}$. By Corollary 4.3 and Hironaka's desingularization theorem [35] the groups $\mathcal{F}_*^{HC}(X)$ can be computed in terms of the cyclic homology of X and of the cyclic homology of a finite array (called a hyperresolution) of smooth schemes that appear in the desingularization process. In this sense we can say that the fiber of $K(X) \to KH(X)$ can be computed in terms of cyclic homology. Note that this represents a lot of progress from our starting point. Indeed regarding KH as K-theory made homotopic lead us to a map $K_*^{\mathrm{nil}} \to HC_{*-1}$ (see (3.2)) that is never an isomorphism in the commutative case (see Remark 3.4); regarding it instead as a version of K-theory with cdh-descent lead us to a character $K_*^{\mathrm{nil}} = \mathcal{F}_*^K \to \mathcal{F}_{*-1}^{HC}$ which is always an isomorphism. Haesemeyer's theorem made all the difference.

Theorems 5.1 and 5.3 are powerful tools for studying the K-theory of singular schemes. In the next sections we review a number of results whose proofs used these tools.

Remark 5.4. Building on work of Ayoub and ideas of Voevodsky, Cisinski proved in [4, Théorème 3.9] that KH satisfies cdh-descent in the category of Noetherian schemes of finite dimension. Since the latter category includes the schemes of finite type over a field of any characteristic, Cisinki's result is far more general than Haesemeyer's Theorem 5.1. However, Cisinski's proof relies heavily on homotopy invariance, and therefore no analog of Theorem 5.2 can be derived from his argument. A version of Theorem 5.2 for schemes of finite type over a perfect field which admits strong resolution of singularities, due to Geisser-Hesselholt, is given in Theorem 8.4 below.

6. Weibel's dimension conjecture

Conjecture 6.1 (Weibel's dimension conjecture [51, Questions 2.9]). *Let X be a Noetherian scheme of dimension d. Then $K_m(X) = 0$ for $m < -d$ and X is K_{-d}-regular.*

The KH-version of the conjecture for schemes over a field of characteristic zero was settled by Haesemeyer:

Theorem 6.2 (Haesemeyer, [34, Theorem 7.1]). *Let k be a field of characteristic zero and let $X \in \mathrm{Sch}/k$. Then $KH_m(X) = 0$ for $m < -\dim X$.*

Proof. By Theorem 5.1 and [34, Theorem 2.8] there is a spectral sequence

$$E_2^{p,q} = H^p_{\mathrm{cdh}}(X, aK_{-q}) \Rightarrow KH_{-p-q}(X).$$

The E_2-term is the cohomology of the cdh sheaf associated to the K-theory groups. We have $aK_{-q} = 0$ for $q > 0$ by Hironaka's desingularization theorem and the fact that negative K-theory vanishes on smooth schemes. By a result of Suslin-Voevodsky [43, Theorem 5.13], if $n > d = \dim X$ then $H^n_{\mathrm{cdh}}(X, \mathcal{S}) = 0$ for every cdh sheaf $\mathcal{S}$; thus $E_2^{p,q} = 0$ if either $p > d$ or $q > 0$. The theorem is immediate from this. □

Theorem 6.3 ([8, Theorem 6.2]). *Weibel's conjecture holds for schemes essentially of finite type over a field of characteristic zero.*

The vanishing part of Theorem 6.3 can be proved using Theorems 5.3 and 6.2 and a similar spectral sequence argument; the regularity part requires more work (see [8, Section 6]).

7. Singularity and the obstruction to K-regularity

Let $F : \mathrm{Rings} \to \mathfrak{Ab}$ be a functor from rings to abelian groups. If A is a ring, then the inclusion $A \subset A[t]$ is split by evaluation at zero. Hence for

$$NF(A) = \mathrm{coker}(F(A) \to F(A[t]))$$

we have a direct sum decomposition

$$F(A[t]) = F(A) \oplus NF(A).$$

Similarly,

$$F(A[t_1, t_2]) = F(A) \oplus NF(A) \oplus NF(A) \oplus N^2F(A),$$

where $N^2F = N(NF)$. Iterating this process one obtains an expression for $F(A[t_1, \ldots, t_n])$ in terms of $F(A)$ and of the groups $N^jF(A)$ for $j \geq 1$. Hence A is F-regular if and only if $N^jF(A) = 0$ for all $j \geq 1$.

Kassel proved ([38, Example 3.3]) that if A is a unital associative $\mathbb{Q}$-algebra, then

$$NHC_n(A) = HH_n(A) \otimes t\mathbb{Q}[t]. \tag{7.1}$$

Here HH is *Hochschild homology*; it is related to cyclic homology by Connes' SBI-sequence

$$HC_{n-1}(A) \xrightarrow{B} HH_n(A) \xrightarrow{I} HC_n(A) \xrightarrow{S} HC_{n-2}(A).$$

For example if A is commutative, then $HH_0(A) = A$ and

$$HH_1(A) = \Omega^1_A \tag{7.2}$$

is the module of absolute *Kähler* 1-*differential forms*. Using (7.1), (7.2), and the Künneth formula for Hochschild homology [53, Proposition 9.4.1], we obtain

$$\begin{aligned} N^2HC_n(A) =& NHH_n(A) \otimes t\mathbb{Q}[t] \\ =& HH_n(A) \otimes \mathbb{Q}[t] \otimes \mathbb{Q}[t] \oplus HH_{n-1}(A) \otimes \Omega^1_{\mathbb{Q}[t]} \otimes \mathbb{Q}[t]. \end{aligned} \tag{7.3}$$

Iterating this process one obtains a formula for $N^jHC_n(A)$ for all j in terms of $HH_{n-p}(A)$ $(p \leq j)$ (see [9, Formula 1.5]). Like cyclic homology, Hochschild homology is defined for schemes (see [52]) and has all the properties stated for cyclic homology in Theorem 4.2 and Corollary 4.3 ([8, Theorem 2.9]). Formulas (7.1) and (7.3) generalize to schemes. Moreover if k is a field of characteristic zero and $X \in \mathrm{Sch}/k$, then for $V = t\mathbb{Q}[t]$ and $dV = \Omega^1_{\mathbb{Q}[t]}$, we have ([9, Lemma 3.2 and Corollary 3.3]

$$\begin{aligned} N\mathcal{F}^{HC}_n(X) =& \mathcal{F}^{HH}_n(X) \otimes V, \\ N^2\mathcal{F}^{HC}_n(X) =& \big(\mathcal{F}^{HH}_n(X) \otimes V \otimes V\big) \oplus \big(\mathcal{F}^{HH}_{n-1}(X) \otimes V \otimes dV\big). \end{aligned}$$

In view of Theorem 5.3 and of the fact that KH is homotopy invariant, we obtain the following formulas:

$$\begin{aligned} NK_n(X) =& \mathcal{F}^{HH}_{n-1}(X) \otimes V \\ N^2K_n(X) =& \big(\mathcal{F}^{HH}_{n-1}(X) \otimes V \otimes V\big) \oplus \big(\mathcal{F}^{HH}_{n-2}(X) \otimes V \otimes dV\big). \end{aligned}$$

Of course we can iterate this to produce formulas for $N^jK_n(X)$ in terms of $\mathcal{F}^{HH}_{n-p}(X)$ $(p \leq j-1)$. Summing up, we obtain the following.

Proposition 7.1. *Let k be a field of characteristic zero, let $X \in \mathrm{Sch}/k$ and let $n \in \mathbb{Z}$. Then X is K_n-regular if and only if $\mathcal{F}^{HH}_m(X) = 0$ for all $m \leq n-1$.*

In view of (5.1), the equivalent conditions of the proposition are also equivalent to the assertion that the map

$$HH_m(X) \to \mathbb{H}^{-m}_{\mathrm{cdh}}(X, HH)$$

is an isomorphism for $m \leq n-1$ and a surjection for $m = n$. By [14, Proposition 2.1 and Theorem 2.2], $\mathbb{H}^*_{\mathrm{cdh}}(X, HH)$ decomposes as follows:

$$\mathbb{H}^{-m}_{\mathrm{cdh}}(X, HH) = \bigoplus_{i \geq 0} H^{i-m}_{\mathrm{cdh}}(X, a\Omega^i).$$

Here $H^*_{\mathrm{cdh}}(X, a\Omega^i)$ is the cohomology of the cdh-sheaf associated to Kähler i-differential forms. Summing up, K-regularity questions translate into comparing Hochschild homology with cdh cohomology of differential forms. This point of view has been useful for solving the following old questions about K-regularity.

Conjecture 7.2 (Vorst's regularity conjecture [48]). *Let R be a commutative unital ring of dimension d, essentially of finite type over a field k. If R is K_{d+1}-regular then R is a regular ring.*

For $d = 0$, the conjecture is trivial; the case $d = 1$ was proved by Vorst [48, Theorem 3.6]. Recall that a commutative unital ring R is *regular in codimension* d if the local ring $R_{\mathfrak{p}}$ is regular for each prime $\mathfrak{p}$ of codimension d. A Noetherian ring of dimension d is regular if and only if it is regular in codimension d.

Theorem 7.3 ([14, Theorem 0.1]). *Let R be a commutative unital ring which is essentially of finite type over a field k of characteristic zero. Assume that R is K_n-regular. Then R is regular in codimension $< n$. In particular, if R is $K_{\dim R+1}$-regular, then it is regular.*

Question 7.4 (Bass' question [1, Question $(VI)_n$]). *Does $NK_n(R) = 0$ imply $N^2K_n(R) = 0$?*

Theorem 7.5 ([9, Corollary 6.7], [10, Theorem 4.1]).

a) *For any field F algebraic over $\mathbb{Q}$, the 2-dimensional normal algebra*

$$R = F[x, y, z]/(z^2 + y^3 + x^{10} + x^7 y)$$

has $NK_0(R) = 0$ but $N^2K_0(R) \neq 0$.

b) *Suppose R is essentially of finite type over a field of infinite transcendence degree over $\mathbb{Q}$. Then $NK_n(R) = 0$ implies that R is K_n-regular.*

The proof in *loc.cit.* of the theorem above employs the method described in the paragraph after Proposition 7.1. Gubeladze gave a different proof of part b) in [33, Theorem 1].

The next conjecture concerns abelian monoids; we shall use multiplicative notation. An abelian monoid is called *cancellative* if $ac = bc$ implies $a = b$, and *torsion-free* if $a^n = b^n$ with $n \in \mathbb{Z}_{\geq 1}$ implies $a = b$. An element $u \in M$ is called a *unit* if it has an inverse in M. If k is a commutative ring and M a commutative monoid, then the *monoid k-algebra* $k[M]$ is the set of finitely supported functions $M \to k$ equipped with pointwise addition and convolution product. Any element of $k[M]$ can be written uniquely as a finite k-linear combination $\sum_a \lambda_a \chi_a$ where χ_a is the characteristic function $\chi_a(b) = \delta_{a,b}$. Each integer $c \geq 2$ defines a k-algebra homomorphism

$$\theta_c : k[M] \to k[M], \qquad \theta_c(\chi_a) = \chi_{a^c}.$$

The map θ_c is called the *dilation* of ratio c. If $F : \mathrm{Rings} \to \mathfrak{Ab}$ is a functor and $\mathfrak{c} = (c_1, c_2, \dots)$ is a sequence of integers $c_i \geq 2$, then we have an inductive system $\{\theta_{c_n} : F(k[M]) \to F(k[M]) : n \geq 1\}$. We write

$$F(k[M])^{\mathfrak{c}} = \operatorname*{colim}_{\theta_{c_n}} F(k[M])$$

for its colimit.

Conjecture 7.6 (Gubeladze's nilpotence conjecture, [30, Conjecture 2.1]). *Let k be a commutative ring, M an abelian monoid and $\mathfrak{c} = (c_1, c_2, \dots)$ a sequence of integers $c_i \geq 2$. Assume that k is regular Noetherian and that M is cancellative and torsion-free and has no non-trivial units. Then*

$$K_*(k[M])^{\mathfrak{c}} = K_*(k).$$

Remark 7.7. Under the conditions of the conjecture, the map $k[M] \to k$ is a polynomial homotopy equivalence (see the proof of [13, Corollary 8.4]). Thus Gubeladze's conjecture is a homotopy invariance statement.

Theorem 7.8 (Gubeladze, [31, Theorem 1.2], [32, Theorem 1]). *Conjecture 7.6 holds if k contains a field of characteristic zero.*

Gubeladze's proof uses relies on Theorem 2.3 and on his previous work on special cases of the conjecture. A different proof, using Theorem 5.2, was given in [11, Corollary 6.10] for the particular case when k is a field. The general case when k is a regular ring containing a field (of any characteristic) was proved in [13, Theorem 0.2]; this is discussed in Section 8.

8. A glimpse at characteristic $p > 0$.

We have seen in Section 2 that the obstructions to nilinvariance and excision with rational coefficients are computed by the Chern character to negative cyclic homology. To compute these obstructions in the p-adically complete case for a prime p, one has to replace the Chern character by the *cyclotomic trace* of Bökstedt-Hsiang-Madsen [3]

$$\mathrm{tr} : K(R) \to TC(R),$$

which takes values in *topological cyclic homology*.

Theorem 8.1 (McCarthy, [41, Main Theorem]). *Let R be a unital ring, $I \lhd R$ a nilpotent ideal and $p > 0$ a prime. Then $\mathrm{tr} : K(R : I) \to TC(R : I)$ becomes a weak equivalence after p-completion.*

The topological cyclic homology spectrum TC is the homotopy limit of a pro-spectrum TC^n; the following theorem is formulated in terms of the latter pro-spectrum. It implies that $K(R, S : I) \to TC(R, S : I)$ becomes a weak equivalence after p-completion.

Theorem 8.2 (Geisser-Hesselholt [21, Theorem 1]). *Let $f : R \to S$ be a homomorphism of unital associative rings, let $I \lhd R$ be a two-sided ideal and assume that $f : I \to f(I)$ is an isomorphism onto a two-sided ideal of S. Then the map induced by the cyclotomic trace map*

$$K_q(R, S : I, \mathbb{Z}/p^\nu) \to TC^n_q(R, S : I, \mathbb{Z}/p^\nu)$$

is an isomorphism of pro-abelian groups, for all integers q, all primes p, and all positive integers ν.

Remark 8.3. In their recent article [20], Dundas and Kittang have shown that Goodwillie's global cyclotomic trace $K \to \mathcal{TC}$ induces an integral isomorphism $K_*(R, S : I) \to \mathcal{TC}_*(R, S : I)$.

The following is a version of Theorem 5.2 for perfect fields which admit *strong resolution of singularities*. This means that for every integral scheme X separated and of finite type over k, there exists a sequence of blow-ups

$$X_r \to X_{r-1} \to \cdots \to X_1 \to X_0 = X$$

such that the reduced scheme X_r^{red} is smooth over k; the center Y_i of the blow-up $X_{i+1} \to X_i$ is connected and smooth over k; the closed embedding of Y_i in X_i is normally flat; and Y_i is nowhere dense in X_i.

Theorem 8.4 (Geisser-Hesselholt, [22, Theorem 1.1]). *Let k be an infinite perfect field such that strong resolution of singularities holds over k, and let $\{F^n(-)\}$ be a presheaf of pro-spectra on the category of schemes essentially of finite type over k. Assume that $\{F^n(-)\}$ takes infinitesimal thickenings to weak equivalences and finite abstract blow-up squares to homotopy cartesian squares. Assume further that each $F^n(-)$ takes elementary Nisnevich squares and squares associated with blow-ups along regular embeddings to homotopy cartesian squares. Then the canonical map defines a weak equivalence of pro-spectra*

$$\{F^n(X)\} \xrightarrow{\sim} \{\mathbb{H}^{\cdot}_{\mathrm{cdh}}(X, F^n(-))\}$$

for every scheme X essentially of finite type over k.

By Theorems 3.1, 8.1, and 8.2, the theorem above applies to KH and to the fiber of the cyclotomic trace. Using this, Geisser and Hesselholt obtained the following result about Weibel's dimension conjecture.

Theorem 8.5 (Geisser-Hesselholt, [22, Theorem A]). *Let k be an infinite perfect field of characteristic $p > 0$ such that strong resolution of singularities holds over k, and let X be a d-dimensional scheme essentially of finite type over k. Then $K_q(X)$ vanishes for $q < -d$.*

Geisser and Hesselholt also obtained the following result about Vorst's conjecture.

Theorem 8.6 (Geisser-Hesselholt, [23, Theorem A]). *Let k be an infinite perfect field of characteristic $p > 0$ such that strong resolution of singularities holds over k. Let R be a localization of a d-dimensional commutative k-algebra of finite type and suppose that R is K_{d+1}-regular. Then R is a regular ring.*

If we restrict our attention to toric varieties, the assumption that the ground field admits strong resolution of singularities can be dropped. The Bierstone-Milman theorem [2, Thm. 1.1] provides a chacteristic-free resolution of singularities for such varieties that is sufficient to prove the following toric version of Haesemeyer's theorem.

Theorem 8.7 ([12, Theorem 1.1]). *Assume k is an infinite field and let $\mathcal{G}$ be a cohomology theory on Sch/k. If $\mathcal{G}$ satisfies the Mayer-Vietoris property for Zariski covers, finite abstract blow-up squares, and blow-ups along regularly embedded closed subschemes, then $\mathcal{G}$ satisfies the Mayer-Vietoris property for all abstract blow-up squares of toric k-varieties.*

Theorems 8.2 and 8.7 were used to prove the following.

Theorem 8.8 ([13, Theorem 0.2]). *Gubeladze's conjecture 7.6 holds if k contains a field of characteristic $p > 0$.*

As we saw in Theorem 7.8, the case when $k \supset \mathbb{Q}$ of the theorem above had already been proved by Gubeladze in [32]. A different proof of Gubeladze's theorem, using Theorem 8.7, was also given in [13, Theorem 7.5].

Acknowledgements. This work was supported by CONICET and partially supported by grants UBACyT 20020100100386, PIP 11220110100800 and MTM2012-36917-C03-02. I wish to thank the following people for helpful comments on previous versions of this article: Christian Haesemeyer, Lars Hesselholt, Emanuel Rodríguez Cirone, Andreas Thom and Bruce Williams.

References

[1] Hyman Bass, *Some problems in "classical" algebraic K-theory*, Algebraic K-theory, II: "Classical" algebraic K-theory and connections with arithmetic (Proc. Conf., Battelle Memorial Inst., Seattle, Wash., 1972), Springer, Berlin, 1973, pp. 3–73. Lecture Notes in Math., Vol. 342. MR0409606 (53 #13358)

[2] Edward Bierstone and Pierre D. Milman, *Desingularization of toric and binomial varieties*, J. Algebraic Geom. **15** (2006), no. 3, 443–486, DOI 10.1090/S1056-3911-06-00430-9. MR2219845 (2007e:14025)

[3] M. Bökstedt, W. C. Hsiang, and I. Madsen, *The cyclotomic trace and algebraic K-theory of spaces*, Invent. Math. **111** (1993), no. 3, 465–539, DOI 10.1007/BF01231296. MR1202133 (94g:55011)

[4] Denis-Charles Cisinski, *Descente par éclatements en K-théorie invariante par homotopie*, Ann. of Math. (2) **177** (2013), no. 2, 425–448, DOI 10.4007/annals.2013.177.2.2 (French, with English and French summaries). MR3010804

[5] Guillermo Cortiñas, *On the derived functor analogy in the Cuntz-Quillen framework for cyclic homology*, Algebra Colloq. **5** (1998), no. 3, 305–328. MR1679567 (2000a:19003)

[6] ———, *The obstruction to excision in K-theory and in cyclic homology*, Invent. Math. **164** (2006), no. 1, 143–173, DOI 10.1007/s00222-005-0473-9. MR2207785 (2006k:19006)

[7] ———, *Algebraic v. topological K-theory: a friendly match*, Topics in algebraic and topological K-theory, Lecture Notes in Math., vol. 2008, Springer, Berlin, 2011, pp. 103–165, DOI 10.1007/978-3-642-15708-0_3. MR2762555 (2012c:19001)

[8] G. Cortiñas, C. Haesemeyer, M. Schlichting, and C. Weibel, *Cyclic homology, cdh-cohomology and negative K-theory*, Ann. of Math. (2) **167** (2008), no. 2, 549–573, DOI 10.4007/annals.2008.167.549. MR2415380 (2009c:19006)

[9] G. Cortiñas, C. Haesemeyer, Mark E. Walker, and C. Weibel, *Bass' NK groups and cdh-fibrant Hochschild homology*, Invent. Math. **181** (2010), no. 2, 421–448, DOI 10.1007/s00222-010-0253-z. MR2657430 (2011g:19003)

[10] ———, *A negative answer to a question of Bass*, Proc. Amer. Math. Soc. **139** (2011), no. 4, 1187–1200, DOI 10.1090/S0002-9939-2010-10728-1. MR2748413 (2011m:19001)

[11] ———, *The K-theory of toric varieties*, Trans. Amer. Math. Soc. **361** (2009), no. 6, 3325–3341, DOI 10.1090/S0002-9947-08-04750-8. MR2485429 (2010b:19001)

[12] ———, *Toric varieties, monoid schemes and cdh descent*, J. reine angew. Math., posted on 2013, DOI 10.1515/crelle-2012-0123.

[13] ———, *The K-theory of toric varieties in positive characteristic*, J. Topol. **7** (2014), no. 1, 247–286, DOI 10.1112/jtopol/jtt026.

[14] G. Cortiñas, C. Haesemeyer, and C. Weibel, *K-regularity, cdh-fibrant Hochschild homology, and a conjecture of Vorst*, J. Amer. Math. Soc. **21** (2008), no. 2, 547–561, DOI 10.1090/S0894-0347-07-00571-1. MR2373359 (2008k:19002)

[15] Guillermo Cortiñas and Andreas Thom, *Comparison between algebraic and topological K-theory of locally convex algebras*, Adv. Math. **218** (2008), no. 1, 266–307, DOI 10.1016/j.aim.2007.12.007. MR2409415 (2009h:46136)

[16] Joachim Cuntz, *Bivariante K-Theorie für lokalkonvexe Algebren und der Chern-Connes-Charakter*, Doc. Math. **2** (1997), 139–182 (electronic) (German, with English summary). MR1456322 (98h:19006)

[17] ———, *Bivariant K-theory and the Weyl algebra*, K-Theory **35** (2005), no. 1-2, 93–137, DOI 10.1007/s10977-005-3464-0. MR2240217 (2008a:46068)

[18] Joachim Cuntz and Daniel Quillen, *Excision in bivariant periodic cyclic cohomology*, Invent. Math. **127** (1997), no. 1, 67–98, DOI 10.1007/s002220050115. MR1423026 (98g:19003)

[19] Joachim Cuntz and Andreas Thom, *Algebraic K-theory and locally convex algebras*, Math. Ann. **334** (2006), no. 2, 339–371, DOI 10.1007/s00208-005-0722-7. MR2207702 (2006j:46070)

[20] Bjørn Ian Dundas and Harald Øyen Kittang, *Integral excision for K-theory*, Homology Homotopy Appl., to appear.

[21] Thomas Geisser and Lars Hesselholt, *Bi-relative algebraic K-theory and topological cyclic homology*, Invent. Math. **166** (2006), no. 2, 359–395, DOI 10.1007/s00222-006-0515-y. MR2249803 (2008a:19003)

[22] ———, *On the vanishing of negative K-groups*, Math. Ann. **348** (2010), no. 3, 707–736, DOI 10.1007/s00208-010-0500-z. MR2677901 (2011j:19004)

[23] ———, *On a conjecture of Vorst*, Math. Z. **270** (2012), no. 1-2, 445–452, DOI 10.1007/s00209-010-0806-2. MR2875843 (2012k:19003)

[24] S. Geller, L. Reid, and C. Weibel, *The cyclic homology and K-theory of curves*, J. Reine Angew. Math. **393** (1989), 39–90, DOI 10.1090/S0273-0979-1986-15474-1. MR972360 (89m:14006)

[25] S. Geller and C. Weibel, *$K(A, B, I)$. II*, K-Theory **2** (1989), no. 6, 753–760, DOI 10.1007/BF00538431. MR1010981 (90h:18013)

[26] ———, *Hochschild and cyclic homology are far from being homotopy functors*, Proc. Amer. Math. Soc. **106** (1989), no. 1, 49–57, DOI 10.2307/2047373. MR965242 (89i:18010)

[27] Susan C. Geller and Charles A. Weibel, *Étale descent for Hochschild and cyclic homology*, Comment. Math. Helv. **66** (1991), no. 3, 368–388, DOI 10.1007/BF02566656. MR1120653 (92e:19006)

[28] Thomas G. Goodwillie, *Cyclic homology, derivations, and the free loopspace*, Topology **24** (1985), no. 2, 187–215, DOI 10.1016/0040-9383(85)90055-2. MR793184 (87c:18009)

[29] _____, *Relative algebraic K-theory and cyclic homology*, Ann. of Math. (2) **124** (1986), no. 2, 347–402, DOI 10.2307/1971283. MR855300 (88b:18008)

[30] Joseph Gubeladze, *K-theory of affine toric varieties*, Homology Homotopy Appl. **1** (1999), 135–145 (electronic). MR1693095 (2000e:14009)

[31] _____, *The nilpotence conjecture in K-theory of toric varieties*, Invent. Math. **160** (2005), no. 1, 173–216, DOI 10.1007/s00222-004-0410-3. MR2129712 (2006d:14057)

[32] _____, *Global coefficient ring in the nilpotence conjecture*, Proc. Amer. Math. Soc. **136** (2008), no. 2, 499–503 (electronic), DOI 10.1090/S0002-9939-07-09106-X. MR2358489 (2008j:19009)

[33] _____, *On Bass' question for finitely generated algebras over large fields*, Bull. Lond. Math. Soc. **41** (2009), no. 1, 36–40, DOI 10.1112/blms/bdn101. MR2481986 (2009m:19001)

[34] Christian Haesemeyer, *Descent properties of homotopy K-theory*, Duke Math. J. **125** (2004), no. 3, 589–620, DOI 10.1215/S0012-7094-04-12534-5. MR2166754 (2006g:19002)

[35] Heisuke Hironaka, *Resolution of singularities of an algebraic variety over a field of characteristic zero. I, II*, Ann. of Math. (2) 79 (1964), 109–203; ibid. (2) **79** (1964), 205–326. MR0199184 (33 #7333)

[36] J. F. Jardine, *Generalized étale cohomology theories*, Progress in Mathematics, vol. 146, Birkhäuser Verlag, Basel, 1997. MR1437604 (98c:55013)

[37] Max Karoubi, *K-théorie algébrique de certaines algèbres d'opérateurs*, Algèbres d'opérateurs (Sém., Les Plans-sur-Bex, 1978), Lecture Notes in Math., vol. 725, Springer, Berlin, 1979, pp. 254–290 (French). MR548119 (81i:46095)

[38] Christian Kassel, *Cyclic homology, comodules, and mixed complexes*, J. Algebra **107** (1987), no. 1, 195–216, DOI 10.1016/0021-8693(87)90086-X. MR883882 (88k:18019)

[39] Bernhard Keller, *On the cyclic homology of ringed spaces and schemes*, Doc. Math. **3** (1998), 231–259 (electronic). MR1647519 (99i:16018)

[40] _____, *On the cyclic homology of exact categories*, J. Pure Appl. Algebra **136** (1999), no. 1, 1–56, DOI 10.1016/S0022-4049(97)00152-7. MR1667558 (99m:18012)

[41] Randy McCarthy, *Relative algebraic K-theory and topological cyclic homology*, Acta Math. **179** (1997), no. 2, 197–222, DOI 10.1007/BF02392743. MR1607555 (99e:19006)

[42] Jonathan Rosenberg, *Comparison between algebraic and topological K-theory for Banach algebras and C^*-algebras*, Handbook of K-theory. Vol. 1, 2, Springer, Berlin, 2005, pp. 843–874, DOI 10.1007/978-3-540-27855-9_16. MR2181834 (2006f:46071)

[43] Andrei Suslin and Vladimir Voevodsky, *Bloch-Kato conjecture and motivic cohomology with finite coefficients*, The arithmetic and geometry of algebraic cycles (Banff, AB, 1998), NATO Sci. Ser. C Math. Phys. Sci., vol. 548, Kluwer Acad. Publ., Dordrecht, 2000, pp. 117–189. MR1744945 (2001g:14031)

[44] Andrei A. Suslin and Mariusz Wodzicki, *Excision in algebraic K-theory*, Ann. of Math. (2) **136** (1992), no. 1, 51–122.

[45] R. W. Thomason and Thomas Trobaugh, *Higher algebraic K-theory of schemes and of derived categories*, The Grothendieck Festschrift, Vol. III, Progr. Math., vol. 88, Birkhäuser Boston, Boston, MA, 1990, pp. 247–435, DOI 10.1007/978-0-8176-4576-2_10. MR1106918 (92f:19001)

[46] R. W. Thomason, *Les K-groupes d'un schéma éclaté et une formule d'intersection excédentaire*, Invent. Math. **112** (1993), no. 1, 195–215, DOI 10.1007/BF01232430 (French). MR1207482 (93k:19005)

[47] Vladimir Voevodsky, *Homotopy theory of simplicial sheaves in completely decomposable topologies*, J. Pure Appl. Algebra **214** (2010), no. 8, 1384–1398, DOI 10.1016/j.jpaa.2009.11.004. MR2593670 (2011a:55022)

[48] Ton Vorst, *Localization of the K-theory of polynomial extensions*, Math. Ann. **244** (1979), no. 1, 33–53, DOI 10.1007/BF01420335. With an appendix by Wilberd van der Kallen. MR550060 (80k:18016)

[49] C. A. Weibel, *Mayer-Vietoris sequences and mod p K-theory*, Algebraic K-theory, Part I (Oberwolfach, 1980), Lecture Notes in Math., vol. 966, Springer, Berlin, 1982, pp. 390–407. MR689385 (84f:18026)

[50] Charles A. Weibel, *Homotopy algebraic K-theory*, Algebraic K-theory and algebraic number theory (Honolulu, HI, 1987), Contemp. Math., vol. 83, Amer. Math. Soc., Providence, RI, 1989, pp. 461–488, DOI 10.1090/conm/083/991991. MR991991 (90d:18006)

[51] ———, *K-theory and analytic isomorphisms*, Invent. Math. **61** (1980), no. 2, 177–197, DOI 10.1007/BF01390120. MR590161 (83b:13011)

[52] Charles Weibel, *Cyclic homology for schemes*, Proc. Amer. Math. Soc. **124** (1996), no. 6, 1655–1662, DOI 10.1090/S0002-9939-96-02913-9. MR1277141 (96h:19003)

[53] Charles A. Weibel, *An introduction to homological algebra*, Cambridge Studies in Advanced Mathematics, vol. 38, Cambridge University Press, Cambridge, 1994. MR1269324 (95f:18001)

[54] Mariusz Wodzicki, *Algebraic K-theory and functional analysis*, First European Congress of Mathematics, Vol. II (Paris, 1992), Progr. Math., vol. 120, Birkhäuser, Basel, 1994, pp. 485–496. MR1341858 (97f:46112)

Dep. Matemática-IMAS, FCEyN-UBA Ciudad Universitaria Pab 1, C1428EGA Buenos Aires, Argentina

E-mail: gcorti@dm.uba.ar

Applications of the classification of finite simple groups

Dedicated to the memory of my friend and mentor Robert Steinberg

Robert Guralnick

Abstract. We discuss how the classification of finite simple groups is used and mention some specific applications to various other fields of mathematics as well as to group theory.

Mathematics Subject Classification (2010). Primary 20D05; Secondary 20B15, 14H30, 12F05.

Keywords. Finite groups, finite simple groups, applications of simple groups, Brauer groups, Riemann surfaces, polynomials, function fields.

1. Introduction

Recall that a subgroup N of a group G is normal in G if it is closed under conjugation. A *simple group* is a nontrivial group in which the only normal subgroups are the group itself and the identity subgroup. The first step in understanding groups is to understand the simple groups. This is much too hard for general groups and so one wants to focus on special types of groups. The classification of simple Lie groups turned out to have a beautiful theory (and has been amazingly fruitful). The classification of simple affine algebraic groups turns out to have essentially the same description albeit with a more difficult proof.

The classification of finite simple groups is one of the most amazing theorems in mathematics. See [55] for a very nice history. It is probably the longest and most complicated single proof in mathematics and involved contributions by many different mathematicians. Perhaps the most significant breakthrough came in John Thompson's series of papers on N-groups [56]. This built upon the earlier groundbreaking work of Feit and Thompson proving that all groups of odd order are solvable (recall that a finite group is solvable if and only if for K a normal subgroup of H with $K < H < G$ and H/K simple, then H/K is cyclic of prime order – this is equivalent to saying the derived series of G terminates in the identity subgroup). Thompson classified all finite simple groups such that the normalizer of any nontrivial prime power order subgroup is solvable. In particular, he classified all the minimal finite simple groups.

Much of the structure of the proof of the classification was based on the strategy used in the Thompson papers. In the early 1970's, Gorenstein realized that based on Thompson's ideas (and others) a proof of the classification was feasible. As Gorenstein has remarked, Aschbacher came along and started proving one amazing result after another and as a result

Proceedings of the International Congress of Mathematicians, Seoul, 2014

sped up the process significantly. Although the result was announced in the early 1980's, it was only completed with the publication of the books of Aschbacher and Smith in 2004 [6]. The original proof is scattered throughout the literature and there is a second generation proof that is being worked on. See the article [3] for a summary of the ideas. This second generation proof is being published by the AMS. Several volumes (plus some preliminary material and the two Aschbacher-Smith volumes) have already been published.

One major advantage of the second generation proof is that one wants to prove that there is no minimal counterexample to the theorem. Thus, one can assume that all proper subgroups of a possible counterexample only involve the known simple groups. This allows one to prove properties of simple groups and use them in the proof.

Indeed, there are other approaches being worked as on well – including one by Meiefrankenfeld et al (see [49]). Also, recently an approach using fusion systems which allows more topological tools to be used has been intensively studied and shown to be a useful way of thinking. Some parts of the classification have been recast in this language.

The tools and techniques in the proof are an amazing array of deep and clever ideas. Results about generators and relations, representation theory, amalgamated products and local subgroup structure are just a few of the ideas involved in the proof. One key result which finally settled a problem which many thought might not be resolved for quite a while was obtained by Bombieri [8] following work of Thompson. Bombieri used elimination theory to show that groups of Ree type are actually Ree groups.

The classification allows one to reduce many problems to questions of simple groups. Some of these reduction theorems could have been proved much earlier but there was not much point as the state of knowledge about arbitrary finite simple groups was quite limited before Thompson came on the scene.

There has also been some thought that with the completion of the classification, finite group theory does not have much of a future. To the contrary, with the tools we now have, one can prove some beautiful results in group theory and in other areas as well. When Richard Brauer was asked whether the classification of finite simple groups would be the end of finite group theory, he answered that it was only the beginning.

Recall that the classification of finite simple groups essentially says:

Theorem 1.1. *Let G be a finite simple group. Then one of the following holds:*

(1) *G is cyclic of prime order;*

(2) *G is an alternating group of degree $n \geq 5$;*

(3) *G is one of 26 sporadic groups; or*

(4) *G is a finite group of Lie type.*

The 26 sporadic groups are a fascinating subject and there have been thousands of pages written about them (and some amazing computations). The smallest of these is the Mathieu group of degree 11 and order 7920 while the largest is the Monster of order approximately 8×10^{53} and has some very interesting connections with modular functions. The first five of these groups were constructed by Mathieu during latter part of the nineteenth century. The remaining sporadic groups were exhibited in the 1960's and 70's.

The simple algebraic groups over an algebraically closed field k of characteristic p are classified (up to isogeny) by the irreducible Dynkin diagrams. The ones of types A, B, C and D correspond to classical groups (i.e. linear, orthogonal and symplectic groups) while the groups of type G_2, F_4 and $E_n, n = 6, 7, 8$ are called exceptional groups. If F is an

endomorphism of a simple algebraic group X such that the fixed set $G := X^F$ is finite, then G is close to being a finite simple (with just a few exceptions, the derived group of G modulo its center is simple). The simplest example is to let F be the q-Frobenius map on k (i.e. it is the map $x \to x^q$ for $q = p^e$). Then F is an automorphism of k and we can extend this to matrices and so to any group of matrices that is invariant under F. For example, if $X = \mathrm{SL}_n(k)$ and F is the q-Frobenius, then $X^F = \mathrm{SL}_n(q)$ and this modulo its center (of order $\gcd(n, q-1)$) is simple for $n \geq 3$ or $n = 2$ and $q \geq 4$. If the Dynkin diagram has graph automorphisms, then we get more examples. For example if we take $F = \sigma\tau$ on $X := \mathrm{SL}_n(k), n \geq 3$ where σ is the q-Frobenius and τ is a graph automorphism (which we can take as transpose inverse), then $X^F = \mathrm{SU}_n(q)$, the special unitary group. Again, almost always this is simple modulo its center (of order $\gcd(n, q+1)$). If $p > 3$, it turns out that these are the only such endomorphisms (note these maps are actually bijections but the inverses are not morphisms in the category of algebraic groups).

In characteristic 2 for the algebraic groups of type B_2 and F_4 and in characteristic 3 for the group G_2, one gets more exotic endomorphisms. This gives the complete list of the finite simple groups of Lie type. Special cases of these constructions were observed but Chevalley, Steinberg and Ree proved the general cases.

See the wonderful book [17] for many more properties of the finite simple groups. One crucial fact about the classification of finite simple groups is not just that gives the full list but that the typical finite simple group is of Lie type and this often allows one to use the theory of algebraic groups and its rich structure to deduce results about finite simple groups. We give some examples of results which use the classification below.

2. Using the classification

The classification of finite simple groups has had many consequences in different areas of mathematics. Many problems in number theory and algebraic geometry can be translated to problems in group theory and very often these involve permutation groups. Typically, one reduces the problem first to transitive groups and then to primitive permutation groups (recall we say that a group G acts primitively on a set S of cardinality larger than 1 if G preserves no nontrivial equivalence relation on S – this is equivalent to G being transitive on S and the stabilizer of a point being a maximal subgroup; these are the simple objects in the category of G-sets).

Note that even if one is trying to prove a result about a known simple group, often one needs to know about subgroups. The classification of simple groups is typically needed for this.

There is a very useful theorem that describes the structure of finite primitive permutation groups due to Aschbacher, O'Nan and Scott [4]. See also [21, 45]. We will not state the theorem but just note that finite primitive permutation groups have a very restricted structure. In particular:

Theorem 2.1. *If G is a finite primitive permutation group, then either G has a unique minimal normal subgroup or G has exactly 2 minimal normal subgroups which are isomorphic and act regularly on the set.*

The Aschbacher-O'Nan-Scott theorem reduces the study of primitive permutation groups to questions about (almost) simple groups and irreducible faithful representations of finite

groups (and the cohomology of such).

Aschbacher [1] gives a very nice description of the subgroups of the classical (almost) simple groups. Basically, the result says that any subgroup of a classical group either preserves some natural structure on the space or is almost simple (modulo the center). Similarly, Liebeck and Seitz [47] give a description of the maximal subgroups of the finite exceptional groups of Lie type. The maximal subgroups of symmetric and alternating groups are described in [46].

Linear representations of finite groups (both complex and over finite fields) also come into various applications. Fortunately, there have been major advances in recent years in both cases. In particular, the Deligne-Lusztig theory has played an important role in many applications.

There is a beautiful paper of Larsen and Pink [42] (see also an earlier related paper by Nori [51]) which asserts in particular that there are only finitely many sporadic groups which have a linear representation of fixed dimension. This can be used as a replacement for the classification for certain types of asymptotic results.

3. Coverings of Riemann Surfaces

Zariski in his thesis answered a conjecture of Enrique and proved:

Theorem 3.1. *Let X be a generic Riemann surface of genus $g > 6$. Let $n > 1$ be a positive integer. There are no solvable rational maps $f : X \to Y$ of degree n.*

If $f : X \to Y$ is a nonconstant map of Riemann surfaces, let $\mathrm{Mon}(f)$ denote the monodromy group of f (i.e the Galois group of the Galois closure of the field extension $\mathbb{C}(X)/\mathbb{C}(Y)$). A solvable map is one whose monodromy group is solvable.

One needs to define generic appropriately (basically, it means that the set of Riemann surfaces of genus g which admit a degree n solvable map to another Riemann surface is contained in a proper closed subvariety of the moduli space of genus g Riemann surfaces). Note that any Riemann surface of genus $g \leq 6$ admits a map of degree at most 4 to $\mathbb{P}^1$ and in particular a solvable map and so $g > 6$ is necessary in the theorem.

In fact, Zariski's proof was mostly about finite primitive solvable groups. In [31], the hypothesis of n being fixed was removed (for $n > 1$, one obtains a proper subvariety but in fact the countable union of these subvarieties is again contained in a proper subvariety).

Using the classification of finite simple groups, one can extend Zariski's result considerably. The following is proved in a series of papers [27, 31, 35]:

Theorem 3.2. *Let X be a generic Riemann surface of genus $g > 3$. Suppose that there exists an indecomposable map f of degree n from X to another Riemann surface Y. Then $Y = \mathbb{P}^1$ and one of the following holds:*

(1) *The monodromy group of f is S_n with $n \geq (g+2)/2$; or*

(2) *The monodromy group of f is A_n with $n > 2g$.*

The indecomposable case is the critical case to consider (any generic map is a composition of an indecomposable map and a rational function on $\mathbb{P}^1$). The key fact about primitive permutation groups that is required is the classification of primitive permutation groups

containing an element that fixes at least $1/2$ the points (a problem that was considered extensively already in the late 1800's).

It is a classical result that the first case in the previous theorem could arise. It was only relatively recently [48] that it was shown the second case can occur as well. There is a similar result for $g = 3$ [27]. The possibilities for $g \leq 2$ have not been worked out. Note that for $g = 0$, we just saying that f is a rational function. However, we do have the following theorem that is a result of the work of a dozen or so authors. This project was completed in [15].

Theorem 3.3. *Let g be a positive integer with X and Y Riemann surfaces with X of genus g. Suppose that $f : X \to Y$ is a rational map. Let S be any simple composition factor of the monodromy group of f. Then there exists a positive integer $N(g)$ such that one of the following holds:*

(1) *S is cyclic of prime order;*

(2) *S is an alternating group; or*

(3) *$|S| < N(g)$.*

Of course, the first two cases do occur. If $g = 0$ then essentially the complete list of possible composition factors of monodromy groups of rational functions has been worked out (indeed, Frohardt, Guralnick and Magaard are working on a project to describe all indecomposable rational functions of degree n whose monodromy groups are not alternating or symmetric groups of degree n).

4. Division Algebras

One of the earliest consequences of the classification of finite simple groups was pointed out by Fein, Kantor and Schacher [12]. Recall that a number field is a finite extension of the rational field.

Theorem 4.1. *Let L/K be a finite nontrivial extension of number fields. There are infinitely many non-isomorphic finite dimensional division algebras D with center K such that $L \otimes D$ is a full matrix ring over L*

The Brauer group $\mathrm{Br}(L)$ of a field is the set of equivalence classes of finite dimensional central division algebras over L (with the multiplication coming from tensor product). The theorem says that the relative Brauer group $\mathrm{Br}(L|K)$ is infinite ($\mathrm{Br}(L|K)$ is the kernel of the natural map from $\mathrm{Br}(K) \to \mathrm{Br}(L)$ given by extension of scalars). If L/K is Galois this follows easily from the description of division algebras over a number field. In the general case, it turns out [12] that one has the following equivalences:

Theorem 4.2. *The following are equivalent:*

(1) *$\mathrm{Br}(L|K)$ is infinite for any nontrivial extension of number fields;*

(2) *$\mathrm{Br}(L|K)$ is nonzero for any nontrivial extension of number fields;*

(3) *If G is a finite group acting transitively on a set S of cardinality $n > 1$, then there exists an element $x \in G$ of prime power order not fixing any points of S.*

Another equivalent condition is that the image of the norm map from L^* to K^* has infinite index. The fact there always exists a fixed point free element in a finite transitive group of degree $n > 1$, is quite elementary (going back to Jordan) but the only known proof to date of the existence of an element of prime power order with no fixed points depends on the classification of finite simple groups.

A proof of a more general result (again with a number theoretic consequence) is given in [19]. We state the number theoretic result.

Theorem 4.3. *Let L/K be a degree $n \geq 1$ extension of number fields. Assume that the Galois group of the Galois closure of L/K is S_n. If L'/K is an extension of number fields (of any degree) and $\mathrm{Br}(L|K)$ is commensurable with $\mathrm{Br}(L'|K)$, then L' and L are K-conjugate fields.*

Recall that two subgroups of a group are said to be commensurable if their intersection has finite index in each subgroup.

An easy corollary of the previous result also gives:

Corollary 4.4. *Let L/K be a degree $n \geq 1$ extension of number fields. Assume that the Galois group of the Galois closure of L/K is S_n. If L'/K is an extension of number fields (of any degree) and the set of degree 1 primes of K with respect to L and L' are the same (up to a finite set), then L' and L are K-conjugate fields.*

We say that a prime P of K (i.e. a prime ideal of the ring $\mathcal{O}_K$ of algebraic integers of K) is a degree 1 prime with respect to L, if some prime ideal of $\mathcal{O}_L$ laying over P has the same residue field as P. For $n = 1$, the result follows from the Cebatorev density theorem and the elementary fact that for any finite transitive group there is a fixed point free element For $n = 2$, this was a question posed by Fried and Jarden.

5. Polynomials

Here are a few results about polynomials. It seems surprising that the classification of simple groups is required to prove some of these results. See [34] for the first result.

Theorem 5.1. *Let F be a field of characteristic p and $f \in F[x]$ of degree n. Assume that f is indecomposable over F but decomposes over some extension field. Then one of:*

(1) $p = 11, n = 55$*; or*

(2) $p = 7, n = 21$*; or*

(3) $n = p^a > p$.

In characteristic zero, this is classical and does not require the classification (or more generally if $\gcd(n, p) = 1$). In [39], the polynomials themselves are actually classified. This involved some interesting curve theory as well as the group theory. There are many examples of such polynomials in the final case. The fact that there are two exceptions suggests that there cannot be a simple proof of the result.

If $f \in F[x]$ is a separable polynomial of degree n with F a finite field of size q, then f is called exceptional if f is bijective over the field of size q^a for some large a (all we need is that $q^a > n^4$). There is another (geometric) description of exceptionality over any field.

Moreover, one can also extend the definition to morphisms between varieties. This idea goes back to the thesis of Dickson in the 1890's.

The following result is obtained in [33, 39] following earlier work in [14].

Theorem 5.2. *Let F be a finite field of characteristic p. Let $f \in F[x]$ be exceptional and indecomposable of degree n with n not a power of p.*

(1) *If $p > 3$, then f is has prime degree and is either a Dickson polynomial or essentially a power of x;*

(2) *if $p \leq 3$, then one of the following holds:*

   (a) *f is a Dickson polynomial of prime degree;*
   (b) *f is a power of x; or*
   (c) *$n = p^a(p^a - 1)/2$ with $a > 1$, a odd.*

Moreover, all such f are known.

Recall that the Dickson polynomials satisfy $D_{n,a}(x + a/x) = x^n + a/x^n$ and have monodromy group the dihedral group of order $2n$. They are generalizations of certain Chebyshev polynomials.

The previous theorem requires a nontrivial reduction to the monodromy group, a difficult group theory problem and then a quite interesting determination of which group theoretic solutions actually led to solutions of the original problem (and then to use that information to actually classify the polynomials). The monodromy groups occurring in the last case when $p \leq 3$ are $\mathrm{PSL}_2(p^a)$.

Here is another result which asserts that many polynomials over finite fields either are bijective or miss quite a lot of points [38].

Theorem 5.3. *Suppose that $f(x) \in \mathbb{F}_q[x]$ of degree n prime to p. Then for any a, f is either bijective on $\mathbb{F}_{q^a}$ or $|f(\mathbb{F}_{q^a})| \leq (5/6)q^a + O(q^{a/2})$.*

If the assumption that n is coprime to the characteristic is dropped, then one can construct examples where f is close to being surjective (the $5/6$ needs to be replaced by $1 - 1/n$ with n the degree of f).

6. Generation

Here is a group theoretic application (although these results do have some applications in number theory). Let $d(G)$ denote the minimal size of a generating set of G. Let $P_2(G)$ denote the probability that a random pair of elements of G generate G. We say that two normal subsets S_1, S_2 of G invariably generate G if $G = \langle s_1, s_2 \rangle$ for any pair $s_i \in S_i$.

Theorem 6.1. *Let G be a finite simple group:*

(1) *$d(G) \leq 2$;*

(2) *$P_2(G) \to 1$ as $|G| \to \infty$*

(3) *for $1 \neq y \in G$, there is $x \in G$ with $G = \langle x, y \rangle$.*

(4) *There exist conjugacy classes C_1 and C_2 of G so that G can be invariably generated by C_1, C_2.*

Remarks
(2) only implies that (1) holds with possibly finitely many exceptions.
Thompson proved (1) for G minimal simple and showed that a finite group is solvable if and only if every 2-generated subgroup is.
(1) was proved by Miller, Steinberg and finished (using the classification) by Aschbacher-Guralnick.
(2) was proved in three papers by Dixon, Kantor-Lubotzky and Liebeck-Shalev.
(4) was proved by Guralnick and Malle. The property of invariable generation is quite useful in the computation of Galois groups.

These types of generation results can give characterizations of various properties of finite groups. We mention just one. Recall that the solvable radical $R(G)$ of a finite group G is the largest normal solvable subgroup. We can characterize the elements of $R(G)$ [32].

Theorem 6.2. *Let G be a finite group. $g \in R(G)$ if and only if $\langle g, x\rangle$ is solvable for all $x \in G$.*

7. Cohomology

The computation of cohomology groups (particularly of low degree) has important consequences in many areas. The first result is obtained by reducing the problem to simple groups and then proving the result for simple groups. See [22] for the first part and [24] for the second.

Theorem 7.1. *Let G be a finite group with V an absolutely irreducible faithful G-module over a field k.*

(1) $\dim H^1(G, V) \leq (1/2)\dim V$;
(2) $\dim H^2(G, V) < 20 \dim V$.

The 20 should be replaced by $1/2$ as well, but in fact most of the time these constants can be reduced considerably. Until 2012, the largest dimension of any $H^1(G, V)$ with V absolutely irreducible and faithful was 3. There are now examples of dimension over $10,000,000$ (using Kazhdan-Lusztig polynomials and computations by Frank Lübeck as well as knowing that the Lusztig conjecture holds for p sufficiently large depending on the root system). This certainly suggests that there is no upper bound on $\dim H^1(G, V)$ with G acting absolutely irreducibly and faithfully on V.

The previous result was used in [24] to prove the following result about profinite presentations of the finite simple groups (and also gives consequences for profinite presentations of arbitrary finite groups). Recall that a profinite presentation for a group G is a short exact sequence:

$$1 \to R \to F \to G \to 1,$$

where F is a free profinite group and R is a closed normal subgroup of F. The number of relations required is the minimal number of elements of R which generate R as a closed normal subgroup of F.

Corollary 7.2. *Let G be a finite simple group. Then G has a profinite presentation with 2 generators and at most 20 relations.*

The correct answer should be 4 (and 2 if we replace the simple group by is universal central cover). For ordinary presentations, one can replace 20 by 50 with the possible exception of ${}^2G_2(3^{2a+1})$; see [23, 25].

8. Linear Groups

Even results on Lie theory sometimes require the classification.

Answering a question of Kollar-Larsen [41], it was shown in [36] that:

Theorem 8.1. *Let G be a closed subgroup of $\mathrm{GL}_n(\mathbb{C}) = \mathrm{GL}(V), n > 2$. Suppose that G acts irreducibly on $\mathrm{Sym}^d(V)$ for $d \geq 4$. Then either G contains $\mathrm{Sp}(V)$ (if $\dim V$ is even) or $\mathrm{SL}(V)$ unless $n = 6$ or 12.*

This has applications to holomony of vector bundles of smooth projective varieties. For $n = 6, 12$, sporadic groups lead to counterexamples. A similar result on subgroups of the classical groups (in characteristic 0) which leave invariant the same subspaces in small tensor powers was obtained by the same authors (answering a conjecture of Nick Katz).

The following results were obtained in [28, 30] answering two conjectures of Peter Neumann from 1966.

Theorem 8.2. *Let k be a field and G an irreducible subgroup of $GL(V)$.*

(1) *If G is finite (or compact), then the average dimension of the fixed point space of an element of G is at most $(1/2)\dim V$.*

(2) *If G is arbitrary, there exists $g \in G$ with the fixed points of g having dimension at most $(1/3)\dim V$.*

See [30] for a proof of the first result (and generalizations and applications). See [28] for the second (and generalizations). Note that both results are best possible (in the first case take G to be cyclic of order 2 – in fact this is the only example where equality holds; in the second case take G to be any irreducible subgroup of $\mathrm{SO}(3, k)$).

There is a classical result of Jordan.

Theorem 8.3. *There is a function f on the natural numbers such that if G is a finite subgroup of $\mathrm{GL}_n(\mathbb{C})$, then G contains an abelian normal subgroup of index at most $f(n)$.*

The estimates for $f(n)$ have been rather large. Using the classification, Weisfeiler, in unpublished work before his death, immensely improved the bounds. Finally, Collins [9] showed that for $n \geq 71$, one could take $f(n) = (n + 1)!$ (and of course this is best possible since S_{n+1} has an irreducible representation of dimension n).

If $\mathbb{C}$ is replaced by an algebraically closed field of characteristic p, then of course there can be arbitrarily large finite subgroups with tiny abelian normal subgroups (e.g. $\mathrm{SL}_n(p^q)$). However, one can prove an analog. This is done by Larsen and Pink [42] without the classification. Collins [10] proves essentially the best possible analog (using the classification).

9. Mashke's Theorem

Maschke's Theorem says that all finite dimensional representations of a finite group in characteristic 0 are completely reducible. The same is true for (rational) representations of Lie groups. Of course, this fails for positive characteristic (both for finite groups and for simple algebraic groups). However, we can recover some version of this. There are asymptotic versions of the following theorem which can be proved without the classification.

Theorem 9.1. *Let p be a prime and G a finite subgroup of $\mathrm{GL}(V)$ with p the characteristic of V. If $\dim V \leq p-2$, then V is completely reducible if and only if G has no nontrivial normal p-subgroup.*

See [20] for the proof. Note that the $p-2$ is sharp since A_p has indecomposable modules of dimensional $p-1$ which are not irreducible. The result depends upon essentially reducing to the case of simple groups and showing that in most cases, one only needs to consider finite groups of Lie type in the same characteristic as V. We then apply a result of Jantzen [40]. The result is also valid for algebraic groups.

10. Word Maps and Waring's Problem

Suppose that w is a nontrivial word in a free group of rank r. Then w defines a map from $G^r \to G$. Let $w(G)$ denote this image. Some particular interesting examples are powers and the commutator word. This can be viewed as an analog of Waring's problem: determine k so that every positive integer is a sum a k dth powers.

The best result along this lines is a result of Larsen, Shalev and Tiep [43]:

Theorem 10.1. *If G is a finite simple group and $|G| \geq N_w$, then $G = w(G)w(G)$.*

Of course, w may vanish on G. Even if w does not vanish, one has examples where many copies of $w(G)$ are required. Of course, power words are not surjective if the power is not relatively prime to the order of $|G|$. For some words, we do have $G = w(G)$ with no exceptions.

Theorem 10.2. *Let $w = x^{-1}y^{-1}xy$ or x^ny^n where $n = p^aq^b$ for primes p and q. If G is a finite simple group, then $G = w(G)$.*

The case of the commutator word was conjectured by Ore [52] and follows by the work in [11, 44, 52]. Special cases for the second case were proved in various papers. See [26] for the complete result.

11. Galois Groups

The question of whether every finite group is the Galois group of a Galois extension of $\mathbb{Q}$ is still very open. There is been quite a lot of progress (using methods from algebraic geometry) in showing that many groups close to being simple are Galois groups. Even in that case, there are only very partial results (there are better results if one works over cyclotomic extensions). See [50] for a survey. Rigidity has proved to be a useful tool. More recently, other methods involving more serious algebraic geometry have been used.

Serre noted the following interesting invariant of the inverse Galois problem (which perhaps seems less likely to be true but is equivalent): every finite group is the Galois group of a totally real Galois extension of $\mathbb{Q}$ (recall that a number field is totally real if every embedding into $\mathbb{C}$ is contained in $\mathbb{R}$ – for Galois extensions, it just amounts to being a subfield of $\mathbb{R}$).

12. Lattices

Quillen [54] studied the poset $\mathcal{A}_p(G)$ of elementary abelian p-subgroups of a finite group G. He proved that if if G has a nontrivial normal p-subgroup, then $\mathcal{A}_p(G)$ is contractible and conjectured that a strong converse held (involving the cohomology of the associated simplicial complex). This conjecture is still open. In [5], the best results to date have been obtained proving Quillen's conjecture for $p > 5$ as long as certain unitary groups are not involved in the group.

Pálfy and Pudlák asked whether every finite lattice embeds in the subgroup lattice of some finite group. Certainly this is false but to date no particular lattice has been ruled out. Aschbacher reduced the question to a more complicated question about simple group. There has been significant progress on this. See [2, 53] for more details on the motivation for the question.

While the knowledge of maximal subgroups of finite groups is relatively well understood (but a full classification would require knowing the dimensions of the irreducible representations of the finite quasi-simple groups and this does not seem within reach at the moment). More generally, one would like more information about the entire subgroup structure.

13. Prime Power Index

Recall that if G is a finite solvable group, then every maximal subgroup has prime power index (this is an easy exercise). Upon learning of this result as a graduate student, the author immediately made the conjecture that this characterized solvable groups (by a theorem of Philip Hall, it is true if the every maximal subgroup index a prime or the square of a prime). Bob Steinberg pointed out that the simple group $\mathrm{SL}(3,2)$ provided a counterexample (the maximal subgroups all have index either 7 or 8). Years later, the next result was obtained (and has been used many many times). See [18]. This is another example where there is one special case. This suggests that the classification really is required.

Theorem 13.1. *Let G be a finite group and suppose that every maximal subgroup of G has prime power index in G. Then either G is solvable or G has a solvable normal subgroup with $G/N \cong \mathrm{SL}(3,2)$.*

14. Beauville Surfaces

A Beauville surface, first defined by F. Catanese, is a rigid compact complex surface of the form $(C_1 \times C_2)/G$ where C_1 and C_2 are curves of genus at least 2 and G is a finite group acting freely on $C_1 \times C_2$. In [7], it was shown that all sufficiently large alternating groups admit a Beauville structure and conjectured that all finite simple groups admit Beauville

structures (with the exception of A_5 which does not).

This was proved in [16] with finitely many possible exceptions. An independent proof giving full result was obtained in [29] (and also in [13]). The methods involve showing that for many triples of conjugacy classes C_1, C_2, C_3 we can find $x_i \in C_i$ with $x_1x_2x_3 = 1$ with the x_i generating the simple group.

Acknowledgements. The author is grateful for the support of the NSF grant DMS-1302886.

References

[1] Aschbacher, M., *On the maximal subgroups of the finite classical groups*, Invent. Math. **76** (1984), 469–514.

[2] ———, *On intervals in subgroup lattices of finite groups*, J. Amer. Math. Soc. **21** (2008), 809–830.

[3] Aschbacher, M., Lyons, R., Smith, S., and Solomon, R., *The classification of finite simple groups. Groups of characteristic 2 type*, Mathematical Surveys and Monographs **172**, American Mathematical Society, Providence, RI, 2011.

[4] Aschbacher, M. and Scott, L., *Maximal subgroups of finite groups*, J. Algebra **92** (1985), 44–80.

[5] Aschbacher, M. and Smith, S., *On Quillen's conjecture for the p-groups complex*, Ann. of Math. **137** (1993), 473–529.

[6] ———, *The classification of quasithin groups. I. Structure of strongly quasithin K-groups,* Mathematical Surveys and Monographs, **111**. American Mathematical Society, Providence, RI, 2004; *The classification of quasithin groups. II. Main theorems: the classification of simple QTKE-groups*, Mathematical Surveys and Monographs, **112**. American Mathematical Society, Providence, RI, 2004.

[7] Bauer, I., Catanese, F., and Grunewald, F., *Chebycheff and Belyi polynomials, dessins d'enfants, Beauville surfaces and group theory*, Mediterr. J. Math. **3** (2006), 121–146.

[8] Bombieri, E., Odlyzko, A., and Hunt, D., *Thompson's problem ($\sigma^2 = 3$)*, Appendices by A. Odlyzko and D. Hunt, Invent. Math. **58** (1980), no. 1, 77–100.

[9] Collins, M., *On Jordan's theorem for complex linear groups*, J. Group Theory **10** (2007), 411–423.

[10] ———, *Modular analogues of Jordan's theorem for finite linear groups*, J. Reine Angew. Math. **624** (2008), 143–171.

[11] Ellers, E. and Gordeev, N., *On the conjectures of J. Thompson and O. Ore*, Trans. Amer. Math. Soc. **350** (1998), 3657–3671.

[12] Fein, B., Kantor, W., and Schacher, M., *Relative Brauer groups. II*, J. Reine Angew. Math. **328** (1981), 39–57.

[13] Fairbairn, B., Magaard, K., and Parker, C., *Generation of finite quasisimple groups with an application to groups acting on Beauville surfaces*, Proc. Lond. Math. Soc. **107** (2013), 744–798.

[14] Fried, M., Guralnick, R., and Saxl, J., *Schur covers and Carlitz's conjecture*, Israel J. Math. **82** (1993), 157–225.

[15] Frohardt, D. and Magaard, K., *Composition factors of monodromy groups*, Ann. of Math. **154** (2001), 327–345.

[16] Garion, S., Larsen, M., and Lubotzky, A., *Beauville surfaces and finite simple groups*, J. Reine Angew. Math. **666** (2012), 225–243.

[17] Gorenstein, D., Lyons, R., and Solomon, R., *The classification of the finite simple groups. Number 3. Part I. Chapter A. Almost simple K-groups*, Mathematical Surveys and Monographs, **40.3**. American Mathematical Society, Providence, RI, 1998.

[18] Guralnick, R., *Subgroups of prime power index in a simple group*, J. Algebra **81** (1983), 304–311.

[19] ———, *Zeroes of permutation characters with applications to prime splitting and Brauer groups,* J. Algebra **131** (1990), 294–302.

[20] ———, *Small representations are completely reducible*, J. Algebra **220** (1999), 531–541.

[21] ———, *Monodromy groups of coverings of curves*, in Galois groups and fundamental groups, 1–46, Math. Sci. Res. Inst. Publ. **41**, Cambridge Univ. Press, Cambridge, 2003.

[22] Guralnick, R. and Hoffman, C., *The first cohomology group and generation of simple groups*, in Groups and geometries (Siena, 1996), 81–89, Trends Math., Birkhäuser, Basel, 1998.

[23] Guralnick, R., Kantor, W., Kassabov, M., and Lubotzky, A., *Presentations of finite simple groups: a quantitative approach*, J. Amer. Math. Soc. **21** (2008), 711–774.

[24] ———, *Presentations of finite simple groups: profinite and cohomological approaches*, Groups Geom. Dyn. **1** (2007), 469–523.

[25] ———, *Presentations of finite simple groups: a computational approach*, J. Eur. Math. Soc. **13** (2011), 391–458.

[26] Guralnick, R., Liebeck, M., O'Brien, E., Shalev, A., and Tiep, P. *Surjective word maps and Burnside's $p^a q^b$ theorem*, preprint.

[27] Guralnick, R. and Magaard, K., *On the minimal degree of a primitive permutation group*, J. Algebra **207** (1998), 127–145.

[28] Guralnick, R. and Malle, G., *Products of conjugacy classes and fixed point spaces*, J. Amer. Math. Soc. **25** (2012), 77–121.

[29] ———, *Simple groups admit Beauville structures*, J. Lond. Math. Soc. **85** (2012), 694–721.

[30] Guralnick, R. and Maróti, A., *Average dimension of fixed point spaces with applications*, Adv. Math. **226** (2011), 298–308.

[31] Guralnick, R. and Neubauer, M., *Monodromy groups of branched coverings: the generic case*, in Recent developments in the inverse Galois problem (Seattle, WA, 1993), 325–352, Contemp. Math. **186**, Amer. Math. Soc., Providence, RI, 1995.

[32] Guralnick, R., Kunyavskii, B., Plotkin, E., and Shalev, A., *Thompson-like characterizations of the solvable radical*, J. Algebra **300** (2006), 363–375.

[33] Guralnick, R., Rosenberg, J., and Zieve, M., *A new family of exceptional polynomials in characteristic two*, Ann. of Math. **172** (2010), 1361–1390.

[34] Guralnick, R. and Saxl, J. Monodromy groups of polynomials, in *Groups of Lie type and their geometries (Como, 1993)*, 125–150, London Math. Soc. Lecture Note Ser., 207, Cambridge Univ. Press, Cambridge, 1995.

[35] Guralnick, R. and Shareshian, J., *Symmetric and alternating groups as monodromy groups of Riemann surfaces. I*, Generic covers and covers with many branch points, With an appendix by Guralnick and R. Stafford, Mem. Amer. Math. Soc. **189** (2007), no. 886.

[36] Guralnick, R. and Tiep, P., *Symmetric powers and a problem of Kollár and Larsen*, Invent. Math. **174** (2008), 505–554.

[37] ———, *A problem of Kollár and Larsen on finite linear groups and crepant resolutions*, J. Eur. Math. Soc. **14** (2012), 605–657.

[38] Guralnick, R. and Wan, D., *Bounds for fixed point free elements in a transitive group and applications to curves over finite fields*, Israel J. Math. **101** (1997), 255–287.

[39] Guralnick, R. and Zieve, M., *Polynomials with PSL(2) monodromy*, Ann. of Math. **172** (2010), 1315–1359.

[40] Jantzen, J., *Low-dimensional representations of reductive groups are semisimple*, in Algebraic groups and Lie groups, 255–266, Austral. Math. Soc. Lect. Ser. 9, Cambridge Univ. Press, Cambridge, 1997.

[41] Kollär, J. and Larsen, M., *Quotients of Calabi-Yau varieties*, in Algebra, arithmetic, and geometry: in honor of Yu. I. Manin. Vol. II, 179–211, Progr. Math., 270, Birkhäuser Boston, Inc., Boston, MA, 2009.

[42] Larsen, M. and Pink, R., *Finite subgroups of algebraic groups*, J. Amer. Math. Soc. **24** (2011), 1105–1158.

[43] Larsen, M., Shalev, A., and Tiep, P., *The Waring problem for finite simple groups*, Ann. of Math. **174** (2011), 1885–1950.

[44] Liebeck, M., O'Brien, E., Shalev, A., and Tiep, P., *The Ore conjecture*, J. Eur. Math. Soc. **12** (2010), 939–1008.

[45] Liebeck, M., Praeger, C., and Saxl, J., *On the O'Nan-Scott theorem for finite primitive permutation groups*, J. Austral. Math. Soc. Ser. A **44** (1988), 389–396.

[46] ———, *A classification of the maximal subgroups of the finite alternating and symmetric groups*, J. Algebra **111** (1987), 365–383.

[47] Liebeck, M. and Seitz, G., *A survey of maximal subgroups of exceptional groups of Lie type*, in Groups, combinatorics & geometry (Durham, 2001), 139–146, World Sci. Publ., River Edge, NJ, 2003.

[48] Magaard, K. and Völklein, H., *The monodromy group of a function on a general curve*, Israel J. Math. **141** (2004), 355–368.

[49] Meierfrankenfeld, U., Stellmacher, B., and Stroth, G., *The structure theorem for finite groups with a large p-subgroup*, Mem. Amer. Math. Soc., to appear.

[50] Malle, G. and Matzat, B., *Inverse Galois theory*, Springer Monographs in Mathematics. Springer-Verlag, Berlin, 1999.

[51] Nori, M., *On subgroups of* $\mathrm{GL}_n(\mathbb{F}_p)$, Invent. Math. **88** (1987), 257–275.

[52] Ore, O., *Some remarks on commutators*, Proc. Amer. Math. Soc. **2** (1951), 307–314.

[53] Pálfy, P., Pudlák, P., *Congruence lattices of finite algebras and intervals in subgroup lattices of finite groups*, Algebra Universalis **11** (1980), 22–27.

[54] Quillen, D., *Homotopy properties of the poset of nontrivial p-subgroups of a group*, Adv. in Math. **28** (1978), 101–128.

[55] Solomon, R., *A brief history of the classification of the finite simple groups*, Bull. Amer. Math. Soc. (N.S.) **38** (2001), 315–352.

[56] Thompson, J. G., *Nonsolvable finite groups all of whose local subgroups are solvable, I, II, III, IV, V, VI*, Bull. Amer. Math. Soc. **74** (1968), 383–437; Pacific J. Math. **33** (1970), 451–536; Pacific J. Math. **39** (1971), 483–534; Pacific J. Math. **48** (1973), 511–592; Pacific J. Math. **50** (1974), 215–297; Pacific J. Math. **51** (1974), 573–630.

Department of Mathematics, University of Southern California, Los Angeles, CA 90089-2532, USA
E-mail: guralnic@usc.edu

Higher representation theory and quantum affine Schur-Weyl duality

Seok-Jin Kang

Abstract. In this article, we explain the main philosophy of 2-representation theory and quantum affine Schur-Weyl duality. The Khovanov-Lauda-Rouquier algebras play a central role in both themes.

Mathematics Subject Classification (2010). Primary 17B37; Secondary 16E99.

Keywords. 2-representation theory, Schur-Weyl duality, Khovanov-Lauda-Rouquier algebra, quantum group.

1. Introduction

The *Khovanov-Lauda-Rouquier algebras*, introduced by Khovanov-Lauda [27, 28] and Rouquier [32, 33], are a family of $\mathbf{Z}$-graded algebras that provide a fundamental framework for *2-representation theory* and *quantum affine Schur-Weyl duality*.

Let $H_k(\zeta)$ be the finite Hecke algebra with ζ a primitive n-th root of unity and let $U_q(A^{(1)}_{n-1})$ be the quantum affine algebra of type $A^{(1)}_{n-1}$. In [29], Lascoux-Leclerc-Thibon discovered a recursive algorithm of computing Kashiwara's lower global basis(=Lusztig's canonical basis) ([25, 30]) and conjectured that the coefficient polynomials, when evaluated at $q = 1$, give the composition multiplicities of simple $H_k(\zeta)$-modules inside Specht modules.

In [2], Ariki came up with a proof of the Lascoux-Leclerc-Thibon conjecture using the idea of *categorification*. More precisely, let Λ be a dominant integral weight associated with the affine Cartan datum of type $A^{(1)}_{n-1}$ and let $H^{\Lambda}_k(\zeta)$ be the corresponding cyclotomic Hecke algebra. Let proj $(H^{\Lambda}_k(\zeta))$ denote the category of finitely generated projective $H^{\Lambda}_k(\zeta)$-modules and let $K(\text{proj}\,(H^{\Lambda}_k(\zeta)))$ be the Grothendieck group of proj $(H^{\Lambda}_k(\zeta))$. Then Ariki proved

$$\bigoplus_{k=0}^{\infty} K(\text{proj}\,(H^{\Lambda}_k(\zeta)))_{\mathbf{C}} \cong V(\Lambda),$$

where $V(\Lambda)$ is the integrable highest weight module over $A^{(1)}_{n-1}$. Moreover, he showed that the isomorphism classes of projective indecomposable modules correspond to the lower global basis of $V(\Lambda)$ at $q = 1$, from which the Lascoux-Leclerc-Thibon conjecture follows.

The idea of categorification, which was originated from [7], can be explained as follows. In the classical representation theory, we study the properties of an algebra A that are reflected on various vector spaces V. That is, we investigate various algebra homomorphisms

Proceedings of the International Congress of Mathematicians, Seoul, 2014

$\phi : A \to \mathrm{End}(V)$. We identify A with a category having a single object and its elements as morphisms. Similarly, we consider $\mathrm{End}(V)$ as a category with V as its object and linear operators on V as morphisms. Then the classical representation theory can be understood as the study of functors from a category to another, whence the *1-representation theory*.

We now *categorify* the classical representation theory. Let $A = \bigoplus_{\alpha \in Q} A_\alpha$ be a graded algebra and let $V = \bigoplus_{\lambda \in P} V_\lambda$ be a graded A-module, where Q and P are appropriate abelian groups. We construct 2-categories $\mathfrak{A}$ and $\mathfrak{B}$ whose objects are certain categories $\mathcal{A}_\alpha$ $(\alpha \in Q)$ and $\mathcal{B}_\lambda$ $(\lambda \in P)$ such that

$$\bigoplus_{\alpha \in Q} K(\mathcal{A}_\alpha) \cong A, \qquad \bigoplus_{\lambda \in P} K(\mathcal{B}_\lambda) \cong V.$$

We now investigate the properties of 2-functors $R : \mathfrak{A} \to \mathfrak{B}$. That is, by categorifying the classical representation theory, we obtain the *2-representation theory*, the study of 2-functors from a 2-category to another.

So far, one of the most interesting developments in 2-representation theory is the one via Khovanov-Lauda-Rouquier algebras. The Khovanov-Lauda-Rouquier algebras categorify the negative half of quantum groups associated with *all* symmetrizable Cartan datum [27, 28, 32, 33]. Moreover, the cyclotomic Khovanov-Lauda-Rouquier algebras give a categorification of *all* integrable highest weight modules [16]. Hence Khovanov-Lauda-Rouquier's and Kang-Kashiwara's categorification theorems provide a vast generalization of Ariki's categorification theorem. (See also [38].)

When the Cartan datum is symmetric, as was conjectured by Khovanov-Lauda [28], Varagnolo-Vasserot proved that the isomorphism classes of simple modules (respectively, projective indecomposable modules) correspond to upper global basis(=dual canonical basis) (respectively, lower global basis) [36]. However, when the Cartan datum is not symmetric, the above statements do not hold in general. It is a very interesting problem to characterize the *perfect basis* and *dual perfect basis* that correspond to simple modules and projective indecomposable modules, respectively.

On the other hand, the Khovanov-Lauda-Rouquier algebras can be viewed as a huge generalization of affine Hecke algebras in the context of *Schur-Weyl duality*. The Schur-Weyl duality, established by Schur and others (see, for example, [34, 35]), reveals a deep connection between the representation theories of symmetric groups and general linear Lie algebras. Let $V = \mathbf{C}^n$ be the vector representation of the general linear Lie algebra gl_n and consider the k-fold tensor product of V. Then gl_n acts on $V^{\otimes k}$ by comultiplication and the symmetric group Σ_k acts on $V^{\otimes k}$ (from the right) by place permutation. Clearly, these actions commute with each other. The Schur-Weyl duality states that there exists a surjective algebra homomorphism

$$\phi_k : \mathbf{C}\Sigma_k \longrightarrow \mathrm{End}_{gl_n}(V^{\otimes k}),$$

where $\mathrm{End}_{gl_n}(V^{\otimes k})$ denotes the centralizer algebra of $V^{\otimes k}$ under the gl_n-action. Moreover, ϕ_k is an isomorphism whenever $k \leq n$.

The Schur-Weyl duality can be rephrased as follows. There is a functor $\mathcal{F}$ from the category of finite dimensional Σ_k-modules to the category of finite dimensional polynomial representations of gl_n given by

$$M \longmapsto V^{\otimes k} \otimes_{\mathbf{C}\Sigma_k} M,$$

where M is a finite dimensional Σ_k-module. The functor $\mathcal{F}$ is called the *Schur-Weyl duality functor* and it defines an equivalence of categories whenever $k \leq n$.

In [14], Jimbo extended the Schur-Weyl duality to the quantum setting: Σ_k is replaced by the finite Hecke algebra H_k and gl_n is replaced by the quantum group $U_q(gl_n)$. Then he obtained the *quantum Schur-Weyl duality functor* from the category of finite dimensional H_k-modules to the category of finite dimensional polynomial representations of $U_q(gl_n)$, which also defines an equivalence of categories whenever $k \leq n$.

In [4, 5, 10], Chari-Pressley, Cherednik and Ginzburg-Reshetikhin-Vasserot constructed a *quantum affine Schur-Weyl duality functor* which relates the category of finite dimensional representations of affine Hecke algebra H_k^{aff} and the category of finite dimensional integrable $U_q'(A_{n-1}^{(1)})$-modules. The main ingredients of their constructions are (i) the fundamental representation $V(\varpi_1)$, (ii) the R-matrices on the tensor products of $V(\varpi_1)$ satisfying the Yang-Baxter equation, (iii) the intertwiners in H_k^{aff} satisfying the braid relations.

Using Khovanov-Lauda-Rouquier algebras, one can construct quantum affine Schur-Weyl duality functors in much more generality. In [17], Kang, Kashiwara and Kim constructed such a functor which relates the category of finite dimensional modules over symmetric Khovanov-Lauda-Rouquuier algebras and the category of finite dimensional integrable modules over *all* quantum affine algebras. Roughly speaking, the basic idea can be explained as follows. Using a family of *good* modules and R-matrices, we determine a quiver Γ and construct a symmetric Khovanov-Lauda-Rouquier algebra $R^{\Gamma}(\beta)$ $(\beta \in Q_+)$. We then construct a $(U_q'(\mathfrak{g}), R^{\Gamma}(\beta))$-bimodule $\widehat{V}^{\otimes\beta}$, a completed tensor power arising from good modules, and define the quantum affine Schur-Weyl duality functor $\mathcal{F}$ by

$$M \mapsto \widehat{V}^{\otimes\beta} \otimes_{R^{\Gamma}(\beta)} M,$$

where M is an $R^{\Gamma}(\beta)$-module.

Various choices of quantum affine algebras and good modules would give rise to various quantum affine Schur-Weyl duality functors. We believe that our general approach will generate a great deal of exciting developments in the forthcoming years.

2. Quantum groups

We begin with a brief recollection of representation theory of quantum groups.

Let I be a finite index set. An integral matrix $A = (a_{ij})_{i,j\in I}$ is called a *symmetrizable Cartan matrix* if (i) $a_{ii} = 2$ for all $i \in I$, (ii) $a_{ij} \leq 0$ for $i \neq j$, (iii) $a_{ij} = 0$ if and only if $a_{ji} = 0$, (iv) there exists a diagonal matrix $D = \mathrm{diag}(d_i \in \mathbf{Z}_{>0} \mid i \in I)$ such that DA is symmetric.

A *Cartan datum* consists of :

(1) a symmetrizable Cartan matrix $A = (a_{ij})_{i,j\in I}$,

(2) a free abelian group P of finite rank, the *weight latice*,

(3) $\Pi = \{\alpha_i \in P \mid i \in I\}$, the set of *simple roots*,

(4) $P^{\vee} := \mathrm{Hom}(P, \mathbf{Z})$, the *dual weight lattice*,

(5) $\Pi^{\vee} = \{h_i \in P^{\vee} \mid i \in I\}$, the set of *simple coroots*

satisfying the following properties

i) $\langle h_i, \alpha_j \rangle = a_{ij}$ for all $i, j \in I$,

ii) Π is linearly independent,

iii) for each $i \in I$, there exists an element $\Lambda_j \in P$ such that

$$\langle h_i, \Lambda_j \rangle = \delta_{ij} \text{ for all } i, j \in I.$$

The Λ_i's $(i \in I)$ are called the *fundamental weights*.

We denote by

$$P^+ := \{\Lambda \in P \mid \langle h_i, \Lambda \rangle \geq 0 \text{ for all } i \in I\}$$

the set of *dominant integral weights*. The free abelian group $Q := \bigoplus_{i \in I} \mathbf{Z}\alpha_i$ is called the *root lattice*. Set $Q_+ = \sum_{i \in I} \mathbf{Z}_{\geq 0}\alpha_i$. For $\beta = \sum k_i\alpha_i \in Q_+$, we define its *height* to be $|\beta| := \sum k_i$.

Since A is symmetrizable, there exists a symmetric bilinear form $(\ ,\ )$ on $\mathfrak{h}^* := \mathbf{Q} \otimes_{\mathbf{Z}} P^\vee$ satisfying

$$(\alpha_i, \alpha_j) = d_i a_{ij}, \quad \langle h_i, \lambda \rangle = \frac{2(\alpha_i, \lambda)}{(\alpha_i, \alpha_i)} \quad \text{for all } \lambda \in \mathfrak{h}^*,\ i, j \in I.$$

Let q be an indeterminate and set $q_i = q^{d_i}$ $(i \in I)$. For $m, n \in \mathbf{Z}_{\geq 0}$, we define

$$[n]_i := \frac{q_i^n - q_i^{-n}}{q_i - q_i^{-1}}, \qquad [n]_i! := \prod_{k=1}^{n} [k]_i.$$

We write $e_i^{(k)} := e_i^k/[k]_i!$, $f_i^{(k)} := f_i^k/[k]_i!$ $(k \in \mathbf{Z}_{\geq 0},\ i \in I)$ for the *divided powers*.

Definition 2.1. The *quantum group* $U_q(\mathfrak{g})$ corresponding to a Cartan datum $(A, P, \Pi, P^\vee, \Pi^\vee)$ is the associative algebra over $\mathbf{Q}(q)$ generated by the elements e_i, f_i $(i \in I)$, q^h $(h \in P^\vee)$ with defining relations

$$\begin{aligned}
&q^0 = 1, \quad q^h q^{h'} = q^{h+h'} \ \ (h, h' \in P^\vee),\\
&q^h e_i q^{-h} = q^{\langle h, \alpha_i \rangle} e_i, \quad q^h f_i q^{-h} = q^{-\langle h, \alpha_i \rangle} f_i \ \ (h \in P^\vee,\ i \in I),\\
&e_i f_j - f_j e_i = \delta_{ij} \frac{K_i - K_i^{-1}}{q_i - q_i^{-1}} \ \ (K_i = q^{d_i h_i},\ i \in I),\\
&\sum_{k=0}^{1-a_{ij}} (-1)^k e_i^{(1-a_{ij}-k)} e_j e_i^{(k)} = 0 \quad (i \neq j),\\
&\sum_{k=0}^{1-a_{ij}} (-1)^k f_i^{(1-a_{ij}-k)} f_j f_i^{(k)} = 0 \quad (i \neq j).
\end{aligned} \tag{2.1}$$

Let $U_q^0(\mathfrak{g})$ be the subalgebra of $U_q(\mathfrak{g})$ generated by q^h $(h \in P^\vee)$ and let $U_q^+(\mathfrak{g})$ (respectively, $U_q^-(\mathfrak{g})$) be the subalgebra of $U_q(\mathfrak{g})$ generated by e_i (respectively, f_i) for all $i \in I$. Then the algebra $U_q(\mathfrak{g})$ has the *triangular decomposition*

$$U_q(\mathfrak{g}) \cong U_q^-(\mathfrak{g}) \otimes U_q^0(\mathfrak{g}) \otimes U_q^+(\mathfrak{g}).$$

Let $\mathbf{A} = \mathbf{Z}[q, q^{-1}]$. We define the *integral form* $U_{\mathbf{A}}(\mathfrak{g})$ of $U_q(\mathfrak{g})$ to be the $\mathbf{A}$-subalgebra of $U_q(\mathfrak{g})$ generated by $e_i^{(k)}$, $f_i^{(k)}$, q^h $(i \in I,\ h \in P^\vee,\ k \in \mathbf{Z}_{\geq 0})$. Let $U_{\mathbf{A}}^0(\mathfrak{g})$ be the $\mathbf{A}$-subalgebra of $U_q(\mathfrak{g})$ generated by q^h $(h \in P^\vee)$ and let $U_{\mathbf{A}}^+(\mathfrak{g})$ (respectively, $U_{\mathbf{A}}^-(\mathfrak{g})$) be the

$\mathbf{A}$-subalgebra of $U_q(\mathfrak{g})$ generated by $e_i^{(k)}$ (respectively, $f_i^{(k)}$) ($i \in I$, $k \in \mathbf{Z}_{\geq 0}$). Then we have

$$U_{\mathbf{A}}(\mathfrak{g}) \cong U_{\mathbf{A}}^-(\mathfrak{g}) \otimes U_{\mathbf{A}}^0(\mathfrak{g}) \otimes U_{\mathbf{A}}^+(\mathfrak{g}).$$

A $U_q(\mathfrak{g})$-module V is called a *highest weight module with highest weight* $\Lambda \in P$ if there exists a nonzero vector v_Λ in V, called the *highest weight vector*, such that

(i) $e_i\, v_\Lambda = 0$ for all $i \in I$,

(ii) $q^h\, v_\Lambda = q^{\langle h, \Lambda \rangle} v_\Lambda$ for all $h \in P^\vee$,

(iii) $V = U_q(\mathfrak{g})\, v_\Lambda$.

For each $\Lambda \in P$, there exists a unique irreducible highest weight module $V(\Lambda)$ with highest weight Λ. The *integral form* of $V(\Lambda)$ is defined to be

$$V_{\mathbf{A}}(\Lambda) := U_{\mathbf{A}}(\mathfrak{g})\, v_\Lambda,$$

where v_Λ is the highest weight vector.

Consider the anti-involution $\phi : U_q(\mathfrak{g}) \to U_q(\mathfrak{g})$ defined by

$$q^h \mapsto q^h, \qquad e_i \mapsto f_i, \qquad f_i \mapsto e_i \quad (h \in P^\vee,\ i \in I).$$

Then there exists a unique non-degenerate symmetric bilinear form $(\ ,\ )$ on $V(\Lambda)$ satisfying

$$(v_\Lambda, v_\Lambda) = 1, \quad (x\, u, v) = (u, \phi(x)\, v) \quad \text{for all } x \in U_q(\mathfrak{g}),\ u, v \in V(\Lambda). \tag{2.2}$$

The *dual* of $V_{\mathbf{A}}(\Lambda)$ is defined to be

$$V_{\mathbf{A}}(\Lambda)^\vee := \{v \in V(\Lambda) \mid (u, v) \in \mathbf{A} \ \text{ for all } u \in V_{\mathbf{A}}(\Lambda)\}.$$

Note that $V_{\mathbf{A}}(\Lambda)^\vee_\lambda = \mathrm{Hom}_{\mathbf{A}}(V_{\mathbf{A}}(\Lambda)_\lambda, \mathbf{A})$ for all $\lambda \in P$.

The *Category* $\mathcal{O}_{\text{int}}$ consists of $U_q(\mathfrak{g})$-modules M such that

i) $M = \bigoplus_{\mu \in P} M_\mu$, where $M_\mu := \{m \in M \mid q^h\, m = q^{\langle h, \mu \rangle}\, m \text{ for all } h \in P^\vee\}$,

ii) e_i, f_i ($i \in I$) are locally nilpotent on M,

iii) there exist finitely many elements $\lambda_1, \ldots, \lambda_s \in P$ such that

$$\mathrm{wt}(M) := \{\mu \in P \mid M_\mu \neq 0\} \subset \bigcup_{j=1}^{s} (\lambda_j - Q_+).$$

The following properties of the category $\mathcal{O}_{\text{int}}$ are well-known. (See, for example, [11, 15, 31].)

Proposition 2.2.

(a) *The category $\mathcal{O}_{int}$ is semisimple.*

(b) *The $U_q(\mathfrak{g})$-module $V(\Lambda)$ with $\Lambda \in P^+$ belongs to $\mathcal{O}_{int}$.*

(c) *Every simple object in $\mathcal{O}_{int}$ has the form $V(\Lambda)$ for some $\Lambda \in P^+$.*

3. Khovanov-Lauda-Rouquier algebras

Let $\mathbf{k}$ be a field and let $(A, P, \Pi, P^\vee, \Pi^\vee)$ be a Cartan datum.

For each $i \neq j$, set

$$S_{ij} := \{(p,q) \in \mathbf{Z}_{\geq 0} \times \mathbf{Z}_{\geq 0} \mid (\alpha_i, \alpha_i)p + (\alpha_j, \alpha_j)q = -2(\alpha_i, \alpha_j)\}.$$

Define a family of polynomials $\mathtt{Q} = (\mathtt{Q}_{ij})_{i,j \in I}$ in $\mathbf{k}[u,v]$ by

$$\mathtt{Q}_{ij}(u,v) := \begin{cases} 0 & \text{if } i = j, \\ \sum_{(p,q) \in S_{ij}} t_{i,j;p,q} u^p v^q & \text{if } i \neq j \end{cases} \tag{3.1}$$

for some $t_{i,j;p,q} \in \mathbf{k}$ such that $t_{i,j;p,q} = t_{j,i;q,p}$ and $t_{i,j;-a_{ij},0} \in \mathbf{k}^\times$. In particular,

$$\mathtt{Q}_{ii}(u,v) = 0, \quad \mathtt{Q}_{ij}(u,v) = \mathtt{Q}_{ji}(v,u) \ \ (i \neq j).$$

The symmetric group $\mathfrak{S}_n = \langle s_1, s_2, \dots, s_{n-1} \rangle$ acts on I^n by place permutation, where s_i denotes the transposition $(i, i+1)$.

Definition 3.1. The *Khovanov-Lauda-Rouquier algebra* $R(n)$ of degree $n \geq 0$ associated with $(A, \mathtt{Q})$ is the associative algebra over $\mathbf{k}$ generated by the elements $e(\nu)$ $(\nu \in I^n)$, x_k $(1 \leq k \leq n)$, τ_l $(1 \leq l \leq n-1)$ with defining relations

$$\begin{aligned}
&e(\nu)e(\nu') = \delta_{\nu,\nu'} e(\nu), \quad \sum_{\nu \in I^n} e(\nu) = 1, \\
&x_k x_l = x_l x_k, \quad x_k e(\nu) = e(\nu) x_k, \\
&\tau_l e(\nu) = e(s_l(\nu))\tau_l, \quad \tau_k \tau_l = \tau_l \tau_k \ \text{ if } |k-l| > 1, \\
&\tau_k^2 e(\nu) = \mathtt{Q}_{\nu_k, \nu_{k+1}}(x_k, x_{k+1}) e(\nu), \\
&(\tau_k x_l - x_{s_k(l)} \tau_k) e(\nu) = \begin{cases} -e(\nu) & \text{if } l = k,\ \nu_k = \nu_{k+1}, \\ e(\nu) & \text{if } l = k+1,\ \nu_k = \nu_{k+1}, \\ 0 & \text{otherwise}, \end{cases} \\
&(\tau_{k+1}\tau_k\tau_{k+1} - \tau_k\tau_{k+1}\tau_k) e(\nu) \\
&\quad = \begin{cases} \dfrac{\mathtt{Q}_{\nu_k,\nu_{k+1}}(x_k, x_{k+1}) - \mathtt{Q}_{\nu_k,\nu_{k+1}}(x_{k+2}, x_{k+1})}{x_k - x_{k+2}} e(\nu) & \text{if } \nu_k = \nu_{k+2}, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.2}$$

The algebra $R(n)$ has a $\mathbf{Z}$-grading by assigning the degrees as follows:

$$\deg e(\nu) = 0, \qquad \deg x_k e(\nu) = (\alpha_{\nu_k}, \alpha_{\nu_k}), \qquad \deg \tau_l e(\nu) = -(\alpha_{\nu_l}, \alpha_{\nu_{l+1}}).$$

We denote by q the *degree-shift functor* defined by

$$(qM)_k = M_{k-1},$$

where $M = \bigoplus_{k \in \mathbf{Z}} M_k$ is a graded $R(n)$-module. Also there is an algebra involution $\psi : R(n) \to R(n)$ given by

$$\begin{aligned}
&e(\nu) \mapsto e(\nu'), \qquad x_k \mapsto x_{n-k+1}, \\
&\tau_l e(\nu) \mapsto \begin{cases} -\tau_{n-l}\, e(\nu') & \text{if } \nu_l = \nu_{l+1}, \\ \tau_{n-l}\, e(\nu') & \text{if } \nu_l \neq \nu_{l+1}, \end{cases}
\end{aligned} \tag{3.3}$$

where $\nu = (\nu_1, \nu_2, \dots, \nu_n)$ and $\nu' = (\nu_n, \dots, \nu_2, \nu_1)$.

By the embedding $R(m) \otimes R(n) \hookrightarrow R(m+n)$, we may consider $R(m) \otimes R(n)$ as a subalgebra of $R(m+n)$. For an $R(m)$-module M and an $R(n)$-module N, we define their *convolution product* $M \circ N$ by

$$M \circ N := R(m+n) \otimes_{R(m)\otimes R(n)} (M \otimes N). \tag{3.4}$$

Since $R(m+n)$ is free over $R(m) \otimes R(n)$ ([27, Proposition 2.16]), the bifunctor $(M, N) \mapsto M \circ N$ is exact in M and N.

For $n \geq 0$ and $\beta \in Q_+$ with $|\beta| = n$, set

$$I^\beta := \{\nu = (\nu_1, \dots, \nu_n) \mid \alpha_{\nu_1} + \cdots + \alpha_{\nu_n} = \beta\}, \qquad e(\beta) := \sum_{\nu \in I^\beta} e(\nu).$$

Then $e(\beta)$ is a central idempotent in $R(n)$. We also define

$$e(\beta, \alpha_i) := \sum_{\substack{\nu \in I^{\beta+\alpha_i} \\ \nu_{n+1}=i}} e(\nu), \quad e(\alpha_i, \beta) := \sum_{\substack{\nu \in I^{\beta+\alpha_i} \\ \nu_1=i}} e(\nu).$$

The algebra

$$R(\beta) := R(n)e(\beta)$$

is called the *Khovanov-Lauda-Rouquier algebra at* β.

For a $\mathbf{k}$-agebra R, we denote by $\mathrm{mod}(R)$ (respectively, $\mathrm{proj}(R)$ and $\mathrm{rep}(R)$) the category of R-modules (respectively, the category of finitely generated projective R-modules and the category of finite dimensional R-modules).

If R is a graded $\mathbf{k}$-algebra, we will use $\mathrm{Mod}(R)$ (respectively, $\mathrm{Proj}(R)$ and $\mathrm{Rep}(R)$) for the category of graded R-modules (respectively, the category of finitely generated projective graded R-modules and the category of finite dimensional graded R-modules).

For each $i \in I$, define the functors

$$\begin{aligned} E_i &: \mathrm{Mod}\,(R(\beta+\alpha_i)) \longrightarrow \mathrm{Mod}\,(R(\beta)), \\ F_i &: \mathrm{Mod}\,(R(\beta)) \longrightarrow \mathrm{Mod}\,(R(\beta+\alpha_i)) \end{aligned} \tag{3.5}$$

by

$$\begin{aligned} E_i(N) &= e(\beta, \alpha_i)\, R(\beta+\alpha_i) \otimes_{R(\beta+\alpha_i)} N, \\ F_i(M) &= R(\beta+\alpha_i)\, e(\beta, \alpha_i) \otimes_{R(\beta)} M \end{aligned} \tag{3.6}$$

for $M \in \mathrm{Mod}\,(R(\beta))$, $N \in \mathrm{Mod}\,(R(\beta+\alpha_i))$.

By [27, Proposition 2.16], the functors E_i and F_i are exact and send finitely generated projective modules to finitely generated projective modules. Hence (3.5) restricts to the functors

$$\begin{aligned} E_i &: \mathrm{Proj}\,(R(\beta+\alpha_i)) \longrightarrow \mathrm{Proj}\,(R(\beta)), \\ F_i &: \mathrm{Proj}\,(R(\beta)) \longrightarrow \mathrm{Proj}\,(R(\beta+\alpha_i)). \end{aligned} \tag{3.7}$$

For $1 \leq k < n$, set $b_k := \tau_k x_{k+1}$ and $b'_k := x_{k+1}\tau_k$. Let $w_0 = s_{i_1} \cdots s_{i_r}$ be the longest element in S_n and set

$$\mathbf{b}(n) := b_{i_1} \cdots b_{i_r}, \quad \mathbf{b}'(n) := b'_{i_r} \cdots b'_{i_1}.$$

For each $n \geq 0$, we define the *divided powers* by

$$E_i^{(n)} := \mathbf{b}'(n)E_i^n, \qquad F_i^{(n)} := F_i^n \mathbf{b}(n).$$

In [27] and [32], Khovanov-Lauda and Rouquier proved the following *categorification theorem*.

Theorem 3.2 ([27, 32]). *There exists an* $\mathbf{A}$*-algebra isomorphism*

$$U_{\mathbf{A}}^-(\mathfrak{g}) \xrightarrow{\sim} K(\text{Proj}(R)) \text{ given by } f_i^{(n)} \longmapsto [F_i^{(n)}] \ (i \in I,\ n \geq 0),$$

where $K(\text{Proj}(R)) := \bigoplus_{\beta \in Q_+} K(\text{Proj}(R(\beta)))$.

Thus we have constructed a 2-category $\mathfrak{R}$ such that the objects are the categories $\text{Proj}\,(R(\beta))$ $(\beta \in Q_+)$ and the categories $\mathcal{H}om\,(\text{Proj}\,R(\alpha), \text{Proj}\,R(\beta))$ consist of the monomials $F_{i_1} \cdots F_{i_r}$ $(i_k \in I,\ r \geq 0)$ of functors satisfying

$$\alpha_{i_1} + \cdots + \alpha_{i_r} = \begin{cases} \alpha - \beta & \text{if } \alpha \geq \beta, \\ \beta - \alpha & \text{if } \beta \geq \alpha. \end{cases}$$

The morphisms in $\mathcal{H}om\,(\text{Proj}\,R(\alpha), \text{Proj}\,R(\beta))$ are the natural transformations generated by $x_i : F_i \to F_i$, $\tau_{ij} : F_iF_j \to F_jF_i$ $(i, j \in I)$ satisfying the relations

$$\begin{aligned}
&\tau_{ij} \circ \tau_{ji} = \mathsf{Q}_{ij}(F_j\, x_i, x_j F_i), \\
&\tau_{jk}F_i \circ F_j\, \tau_{ik} \circ \tau_{ij}F_k - F_k\, \tau_{ij} \circ \tau_{ik}F_j \circ F_i\, \tau_{jk} \\
&\qquad = \begin{cases} \dfrac{\mathsf{Q}_{ij}(x_iF_j, F_j\, x_i)F_i - F_i\mathsf{Q}_{ij}(F_j\, x_i, x_jF_i)}{x_iF_jF_i - F_iF_j\, x_i} F_i & \text{if } i = k, \\ 0 & \text{otherwise}, \end{cases} \\
&\tau_{ij} \circ x_iF_j - F_j\, x_i \circ \tau_{ij} = \delta_{ij}, \\
&\tau_{ij} \circ F_i\, x_j - x_jF_i \circ \tau_{ij} = -\delta_{ij}.
\end{aligned}$$

It is straightforward to verify that $\mathfrak{R}$ satisfies all the axioms for 2-categories [32, 33].

For the later use, we define a functor $\overline{F_i} : \text{Mod}\,(R(\beta)) \longrightarrow \text{Mod}\,(R(\beta + \alpha_i))$ by

$$\overline{F_i}(M) := R(\beta + \alpha_i)\, e(\alpha_i, \beta) \otimes_{R(\beta)} M \ \text{ for } i \in I,\ M \in \text{Mod}\,(R(\beta)).$$

The properties of the functors E_i, F_i and $\overline{F_i}$ $(i \in I)$ are given in the following proposition.

Proposition 3.3 ([16]).

(a) *We have an exact sequence in Mod* $(R(\beta))$

$$0 \longrightarrow \overline{F_i}\, E_i\, M \longrightarrow E_i\overline{F_i}\, M \longrightarrow q^{-(\alpha_i,\alpha_i)}M \otimes \mathbf{k}[t_i] \longrightarrow 0$$

which is functorial in $M \in$ *Mod* $(R(\beta))$.

(b) *There exist natural isomorphisms*

$$\begin{aligned}
&E_iF_j \xrightarrow{\sim} F_jE_i, \quad E_i\overline{F_j} \xrightarrow{\sim} \overline{F_j}E_i \ \text{ if } i \neq j, \\
&E_iF_i \xrightarrow{\sim} q^{-(\alpha_i,\alpha_i)}F_iE_i \oplus \mathbf{1} \otimes \mathbf{k}[t_i] \ \text{ if } i = j,
\end{aligned}$$

where t_i *is an indeterminate of degree* (α_i, α_i) *and*

$$\mathbf{1} \otimes \mathbf{k}[t_i] : \text{Mod}\,(R(\beta)) \to \text{Mod}\,(R(\beta))$$

is the degree-shift functor sending M *to* $M \otimes \mathbf{k}[t_i]$ *for* $M \in$ *Mod* $(R(\beta))$ $(\beta \in Q_+)$.

4. Cyclotomic categorification theorem

Let $\Lambda \in P^+$ and let

$$a^\Lambda(x_1) := \sum_{\nu \in I^\beta} x_1^{\langle h_{\nu_1}, \Lambda \rangle} e(\nu) \quad \in R(\beta).$$

Then the *cyclotomic Khovanov-Lauda-Rouquier algebra* $R^\Lambda(\beta)$ ($\beta \in Q_+$) is defined to be the quotient algebra

$$R^\Lambda(\beta) := R(\beta) \big/ R(\beta) a^\Lambda(x_1) R(\beta). \tag{4.1}$$

We would like to show that the cyclotomic Khovanov-Lauda-Rouquier algebras provide a categorification of irreducible highest weight $U_q(\mathfrak{g})$-modules in the category $\mathcal{O}_{\text{int}}$.

For each $i \in I$, define the functors

$$\begin{aligned} E_i^\Lambda &: \text{Mod}(R^\Lambda(\beta + \alpha_i)) \longrightarrow \text{Mod}(R^\Lambda(\beta)), \\ F_i^\Lambda &: \text{Mod}(R^\Lambda(\beta)) \longrightarrow \text{Mod}(R^\Lambda(\beta + \alpha_i)) \end{aligned} \tag{4.2}$$

by

$$\begin{aligned} E_i^\Lambda(N) &= e(\beta, \alpha_i) R^\Lambda(\beta + \alpha_i) \otimes_{R^\Lambda(\beta+\alpha_i)} N, \\ F_i^\Lambda(M) &= R^\Lambda(\beta + \alpha_i)\, e(\beta, \alpha_i) \otimes_{R^\Lambda(\beta)} M \end{aligned} \tag{4.3}$$

for $M \in \text{Mod}(R^\Lambda(\beta))$, $N \in \text{Mod}(R^\Lambda(\beta + \alpha_i))$. However, since $R^\Lambda(\beta + \alpha_i)$ is not free over $R^\Lambda(\beta)$, there is no guarantee that E_i^Λ and F_i^Λ send finitely generated projective modules to finitely generated projective modules. To prove this, we need to show that $R^\Lambda(\beta + \alpha_i)\, e(\beta, \alpha_i)$ is a projective right $R^\Lambda(\beta)$-module.

Let

$$\begin{aligned} F^\Lambda &:= R^\Lambda(\beta + \alpha_i)\, e(\beta, \alpha_i) \\ &= \frac{R(\beta + \alpha_i)\, e(\beta, \alpha_i)}{R(\beta + \alpha_i)\, a^\Lambda(x_1) R(\beta + \alpha_i)\, e(\beta, \alpha_i)}, \\ K_0 &:= R(\beta + \alpha_i)\, e(\beta, \alpha_i) \otimes_{R(\beta)} R^\Lambda(\beta) \\ &= \frac{R(\beta + \alpha_i)\, e(\beta, \alpha_i)}{R(\beta + \alpha_i)\, a^\Lambda(x_1) R(\beta)\, e(\beta, \alpha_i)}, \\ K_1 &:= R(\beta + \alpha_i)\, e(\alpha_i, \beta) \otimes_{R(\beta)} R^\Lambda(\beta) \\ &= \frac{R(\beta + \alpha_i)\, e(\alpha_i, \beta)}{R(\beta + \alpha_i)\, a^\Lambda(x_2) R^1(\beta)\, e(\alpha_i, \beta)}, \end{aligned}$$

where $R^1(\beta)$ is the subalgebra of $R(\beta + \alpha_i)$ generated by $e(\alpha_i, \nu)$ ($\nu \in I^\beta$), x_k ($2 \le k \le n+1$), τ_l ($2 \le l \le n$). Then F^Λ, K_0 and K_1 can be regarded as $(R(\beta + \alpha_i), R^\Lambda(\beta))$-bimodules.

Let t_i be an indeterminate of degree (α_i, α_i). Then $\mathbf{k}[t_i]$ acts on $R(\beta + \alpha_i)\, e(\alpha_i, \beta)$ and K_1 from the right by $t_i = x_1 e(\alpha_i, \beta)$. On the other hand, $\mathbf{k}[t_i]$ acts on K_0 and F^Λ from the right by $t_i = x_{n+1} e(\beta, \alpha_i)$. Hence all of them have a structure of $(R(\beta + \alpha_i), R(\beta) \otimes \mathbf{k}[t_i])$-bimodules. Moreover, F^Λ, K_0 and K_1 are in fact $(R(\beta + \alpha_i), R^\Lambda(\beta) \otimes \mathbf{k}[t_i])$-bimodules.

In [27], it was shown that K_0 and K_1 are finitely generated projective right $(R^\Lambda(\beta) \otimes \mathbf{k}[t_i])$-modules. Let $\pi : K_0 \to F^\Lambda$ be the canonical projection and let $P : K_1 \to K_0$ be the right multiplication by $a^\Lambda(x_1)\, \tau_1 \cdots \tau_n$.

The following theorem is one of the main results in [16].

Theorem 4.1 ([16]). *The sequence*

$$0 \longrightarrow K_1 \xrightarrow{P} K_0 \xrightarrow{\pi} F^\Lambda \longrightarrow 0 \tag{4.4}$$

is exact as $(R(\beta+\alpha_i, R^\Lambda(\beta)\otimes \mathbf{k}[t_i])$*-bimodules.*

Hence we get a projective resolution of F^Λ of length 1 as a right $R^\Lambda(\beta)[t_i]$-module. By the following lemma, we conclude that F^Λ is a finitely generated projective right $R^\Lambda(\beta)$-module.

Lemma 4.2 ([16]). Let R be a ring and let $f(t)$ be a monic polynomial in $R[t]$ with coefficients in the center of R.

If an $R[t]$-module M is annihilated by $f(t)$ and has projective dimension ≤ 1, then M is projective as an R-module.

Thus we obtain the following important theorem.

Theorem 4.3 ([16]).

(a) $R^\Lambda(\beta+\alpha_i)\, e(\beta,\alpha_i)$ *is a projective right* $R^\Lambda(\beta)$*-module.*

(b) $e(\beta,\alpha_i)R^\Lambda(\beta+\alpha_i)$ *is a projective left* R^Λ*-module.*

(c) *The functors* E_i^Λ *and* F_i^Λ *are exact.*

(d) *The functors* E_i^Λ *and* F_i^Λ *send finitely generated projective modules to finitely generated projective modules.*

Corollary 4.4 ([16]).

For all $i \in I$ *and* $\beta \in Q_+$*, we have an exact sequence of* $R(\beta+\alpha_i)$*-modules*

$$0 \longrightarrow q^{(\alpha_i, 2\Lambda-\beta)}\overline{F_i}\, M \longrightarrow F_i\, M \longrightarrow F_i^\Lambda\, M \longrightarrow 0$$

which is functorial in $M \in$ *Mod* $R^\Lambda(\beta)$.

To complete the construction of cyclotomic categorification, it remains to show that the adjoint pair $(F_i^\Lambda, E_i^\Lambda)$ gives an sl_2*-categorification* introduced by Chuang-Rouquier [6].

Theorem 4.5 ([16]).

(a) *For* $i \neq j$*, there exists a natural isomorphism*

$$q^{-(\alpha_i,\alpha_j)} F_j^\Lambda E_i^\Lambda \xrightarrow{\sim} E_i^\Lambda F_j^\Lambda.$$

(b) *Let* $\lambda = \Lambda - \beta$ $(\beta \in Q_+)$.

(i) *If* $\langle h_i, \Lambda\rangle \geq 0$*, there exists a natural isomorphism*

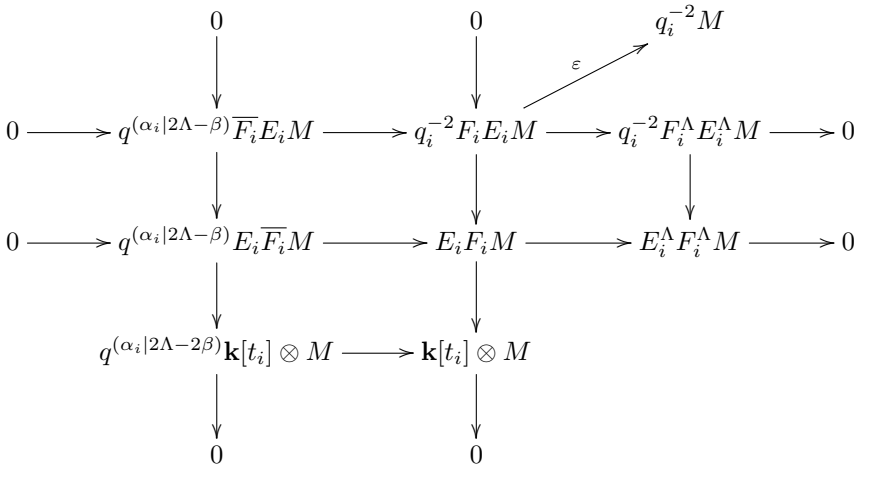
$$q_i^{-2} F_i^\Lambda E_i^\Lambda \oplus \bigoplus_{k=0}^{\langle h_i,\Lambda\rangle - 1} q_i^{2k}\mathbf{1} \xrightarrow{\sim} E_i^\Lambda F_i^\Lambda.$$

(ii) *If* $\langle h_i, \Lambda\rangle \leq 0$, *there exists a natural isomorphism*

$$q_i^{-2}F_i^{\Lambda}E_i^{\Lambda} \xrightarrow{\sim} E_i^{\Lambda}F_i^{\Lambda} \oplus \bigoplus_{k=0}^{-\langle h_i,\lambda\rangle-1} q_i^{2k-2}\mathbf{1}.$$

Proof. We will give a very rough sketch of the proof. The assertion (a) can be proved in a straightforward manner.

To prove (b), note that Theorem 4.1 and Corollary 4.4 yield the following commutative diagram.



Let $A : q^{2(\alpha_i,\Lambda-\beta)}\mathbf{k}[t_i]\otimes R^{\Lambda}(\beta) \longrightarrow \mathbf{k}[t_i]\otimes R^{\Lambda}(\beta)$ be the $R^{\Lambda}(\beta)$-bilinear map given by chasing the diagram. By a detailed analysis of the above commutative diagram at the kernel level, the Snake Lemma gives the following exact sequence of $R^{\Lambda}(\beta)$-bimodules

$$0 \longrightarrow \mathrm{Ker}\, A \longrightarrow q_i^{-2}F_i^{\Lambda}E_i^{\Lambda}R^{\Lambda}(\beta) \longrightarrow E_i^{\Lambda}F_i^{\Lambda}R^{\Lambda}(\beta) \longrightarrow \mathrm{Coker}A \longrightarrow 0.$$

If $\langle h_i, \lambda\rangle \geq 0$, we have $\mathrm{Ker}\, A = 0$, $\bigoplus_{k=0}^{a-1}\mathbf{k}\, t_i^k \otimes R^{\Lambda}(\beta) \xrightarrow{\sim} \mathrm{Coker}A$, and if $\langle h_i, \lambda\rangle \leq 0$, then $\mathrm{Coker}A = 0$, $\mathrm{Ker}(A) = q^{2(\alpha_i|\Lambda-\beta)}\bigoplus_{k=0}^{a-1}\mathbf{k}\, t_i^k \otimes R^{\Lambda}(\beta)$, from which our assertion(b) follows. □

Set

$$K(\mathrm{Proj}\,(R^{\Lambda})) := \bigoplus_{\beta\in Q_+} K(\mathrm{Proj}\, R^{\Lambda}(\beta)),$$
$$K(\mathrm{Rep}\,(R^{\Lambda})) := \bigoplus_{\beta\in Q_+} K(\mathrm{Rep}\, R^{\Lambda}(\beta)).$$

We define the endomorphisms E_i and F_i on $K(\mathrm{Proj}\,(R^{\Lambda}))$ by

$$\mathsf{E}_i = [q_i^{1-\langle h_i,\Lambda-\beta\rangle}E_i^{\Lambda}] : K(\mathrm{Proj}\, R^{\Lambda}(\beta+\alpha_i)) \longrightarrow K(\mathrm{Proj}\, R^{\Lambda}(\beta)),$$
$$\mathsf{F}_i = [F_i^{\Lambda}] : K(\mathrm{Proj}\, R^{\Lambda}(\beta)) \longrightarrow K(\mathrm{Proj}\, R^{\Lambda}(\beta+\alpha_i)).$$

On the other hand, we define E_i and F_i on $K(\text{Rep}\,(R^\Lambda))$ by

$$\mathsf{E}_i = [E_i^\Lambda] : K(\text{Rep}\, R^\Lambda(\beta+\alpha_i)) \longrightarrow K(\text{Rep}\, R^\Lambda(\beta)),$$
$$\mathsf{F}_i = [q_i^{1-\langle h_i, \Lambda-\beta\rangle} F_i^\Lambda] : K(\text{Rep}\, R^\Lambda(\beta)) \longrightarrow K(\text{Rep}\, R^\Lambda(\beta+\alpha_i)).$$

Let K_i be the endomorphism on $K(\text{Proj}\, R^\Lambda(\beta))$ and $K(\text{Rep}\, R^\Lambda(\beta))$ given by the multiplication by $q_i^{\langle h_i, \Lambda-\beta\rangle}$ for each $\beta \in Q_+$. Then we have

$$[\mathsf{E}_i, \mathsf{F}_j] = \delta_{ij} \frac{\mathsf{K}_i - \mathsf{K}_i^{-1}}{q_i - q_i^{-1}} \quad \text{for } i, j \in I.$$

Therefore, we obtain the *cyclotomic categrification theorem* for irreducible highest weight $U_q(\mathfrak{g})$-modules in the category $\mathcal{O}_{\text{int}}$.

Theorem 4.6 ([16]). *For each $\Lambda \in P^+$, there exist $U_{\mathbf{A}}(\mathfrak{g})$-module isomorphisms*

$$K(\textit{Proj}\, R^\Lambda) \xrightarrow{\sim} V_{\mathbf{A}}(\Lambda) \;\textit{ and }\; K(\textit{Rep}\, R^\Lambda) \xrightarrow{\sim} V_{\mathbf{A}}(\Lambda)^\vee.$$

Therefore, for each $\Lambda \in P^+$, we have constructed a 2-category $\mathfrak{R}^\Lambda$ consisting of Proj $(R^\Lambda(\beta))$ $(\beta \in Q_+)$, which gives an integrable 2-representation $\mathfrak{R}^\Lambda$ of $\mathfrak{R}$ in the sense of [32, 33]. (See also [38].)

Remark 4.7. There are several generalizations of Khovanov-Lauda-Rouquier algebras and categorification theorems. In [19, 22, 24], the Khovanov-Lauda-Rouquier algebras associated with Borcherds-Cartan data have been defined and their properties have been investigated including geometric realization, categorification and the connection with crystal bases. In [8, 9, 13, 20, 21, 23, 37], various versions of Khovanov-Lauda-Rouquier *super*-algebras have been introduced and the corresponding *super*-categorifications have been constructed.

5. Quantum affine algebras and R-matrices

In this section, we briefly review the finite dimensional representation theory of quantum affine algebras and the properties of R-matrices (see, for example, [1, 3, 4, 26]).

Let $(A, P, \Pi, P^\vee, \Pi^\vee)$ be a Cartan datum of affine type with $I = \{0, 1, \ldots, n\}$ the index set of simple roots. Let $0 \in I$ be the leftmost vertex in the affine Dynkin diagrams given in [15, Chapter 4]. Set $I_0 = I \setminus \{0\}$. Take relatively prime positive integers c_j's and d_j's $(j \in I)$ such that

$$\sum_{j\in I} c_j a_{ji} = 0, \quad \sum_{j\in I} a_{ij} d_j = 0 \;\text{ for all } i \in I.$$

Then the weight lattice can be written as

$$P = \bigoplus_{i\in I} \mathbf{Z}\Lambda_i \oplus \mathbf{Z}\delta,$$

where $\delta := \sum_{i\in I} d_i\alpha_i \in P$. We also define $c := \sum_{i\in I} c_i h_i \in P^\vee$.

We denote by $\mathfrak{g}$ the affine Kac-Moody algebra associated with $(A, P, P^\vee, \Pi, \Pi^\vee)$ and let $\mathfrak{g}_0$ be the finite dimensional simple Lie algebra inside $\mathfrak{g}$ generated by e_i, f_i, h_i $(i \in I_0)$. We will write W and W_0 for the Weyl group of $\mathfrak{g}$ and $\mathfrak{g}_0$, respectively.

Let $U_q(\mathfrak{g})$ be the corresponding quantum group and let $U'_q(\mathfrak{g})$ be the subalgebra of $U_q(\mathfrak{g})$ generated by e_i, f_i, $K_i^{\pm 1}$ $(i \in I)$. The algebra $U'_q(\mathfrak{g})$ will be called the *quantum affine algebra*.

Set $P_{\mathrm{cl}} := P\big/\mathbf{Z}\delta$ and let $\mathrm{cl} : P \to P_{\mathrm{cl}}$ be the canonical projection. Then we have

$$P_{\mathrm{cl}} = \bigoplus_{i\in I} \mathbf{Z}\mathrm{cl}(\Lambda_i) \ \text{ and } \ P_{\mathrm{cl}}^\vee := \mathrm{Hom}_{\mathbf{Z}}(P_{\mathrm{cl}}, \mathbf{Z}) = \bigoplus_{i\in I} \mathbf{Z}h_i.$$

A $U'_q(\mathfrak{g})$-module V is *integrable* if

(i) $V = \bigoplus_{\lambda\in P_{\mathrm{cl}}} V_\lambda$, where $V_\lambda = \{v \in V \mid K_i v = q_i^{\langle h_i,\lambda\rangle} v \text{ for all } i \in I\}$,

(ii) e_i, f_i $(i \in I)$ are locally nilpotent on V.

We denote by $\mathcal{C}_{\mathrm{int}}$ the category of finite dimensional integrable $U'_q(\mathfrak{g})$-modules.

Let M be an integrable $U'_q(\mathfrak{g})$-module. A weight vector $v \in M_\lambda$ $(\lambda \in P_{\mathrm{cl}})$ is called an *extremal weight vector* if there exists a family of nonzero vectors $\{v_{w\lambda} \mid w \in W\}$ such that

$$v_{s_i\lambda} = \begin{cases} f_i^{(\langle h_i,\lambda\rangle)} v_\lambda & \text{if } \langle h_i,\lambda\rangle \geq 0, \\ e_i^{(-\langle h_i,\lambda\rangle)} v_\lambda & \text{if } \langle h_i,\lambda\rangle \leq 0. \end{cases}$$

Let $P_{\mathrm{cl}}^0 := \{\lambda \in P_{\mathrm{cl}} \mid \langle c,\lambda\rangle = 0\}$ and set

$$\varpi_i := \Lambda_i - c_i\Lambda_0 \ \text{ for } i \in I_0.$$

Then there exists a unique finite dimensional integrable $U'_q(\mathfrak{g})$-module $V(\varpi_i)$ satisfying the following properties:

(i) all the weights of $V(\varpi_i)$ are contained in the convex hull of $W_0\,\mathrm{cl}(\varpi_i)$.

(ii) $\dim V(\varpi_i)_{\mathrm{cl}(\varpi_i)} = 1$,

(iii) for each $\mu \in W_0\,\mathrm{cl}(\varpi_i)$, there exists an extremal weight vector of weight μ,

(iv) $V(\varpi_i)$ is generated by $V(\varpi_i)_{\mathrm{cl}(\varpi_i)}$ as a $U'_q(\mathfrak{g})$-module.

The $U'_q(\mathfrak{g})$-module $V(\varpi_i)$ is called the *fundamental representation of weight* ϖ_i $(i \in I_0)$.

Let M be a $U'_q(\mathfrak{g})$-module. An involution on M is called a *bar involution* if $\overline{a\,v} = \overline{a}\,\overline{v}$ for all $a \in U'_q(\mathfrak{g})$, $v \in M$, where $\overline{e_i} = e_i$, $\overline{f_i} = f_i$, $\overline{K_i} = K_i^{-1}$ $(i \in I)$. A finite $U'_q(\mathfrak{g})$-crystal B is *simple* if (i) $\mathrm{wt}(B) \subset P_{\mathrm{cl}}^0$, (ii) there exists $\lambda \in \mathrm{wt}(B)$ such that $\#(B_\lambda) = 1$, (iii) the weight of every extremal vector of B is contained in $W_0\,\lambda$.

A finite dimensional integrable $U'_q(\mathfrak{g})$-module M is *good* if

(i) M has a bar involution,

(ii) M has a crystal basis with simple crystal,

(iii) M has a lower global basis.

For example, all the fundamental representations $V(\varpi_i)$ $(i \in I_0)$ are good. Every good module is irreducible. For any good module M, there exists an extremal weight vector v of weight λ such that $\mathrm{wt}(U'_q(\mathfrak{g})v) \subset \lambda - \sum_{i\in I_0} \mathbf{Z}_{\geq 0}\mathrm{cl}(\alpha_i)$. Such λ is called a *dominant extremal weight* and v is called a *dominant extremal weight vector*.

Take $\mathbf{k} = \overline{\mathbf{C}(q)} \subset \bigcup_{M>0} \mathbf{C}((q^{1/m}))$. Let $M_{\text{aff}} = \mathbf{k}[z, z^{-1}] \otimes_{\mathbf{k}} M$ be the *affinization* of M. For $v \in M$ and $k \in \mathbf{Z}$, the action of $U_q'(\mathfrak{g})$ on M_{aff} is given by

$$e_i(z^k \otimes v) = \begin{cases} z^{k+1} \otimes e_0 v & \text{if } i = 0, \\ z^k \otimes e_i v & \text{if } i \neq 0, \end{cases}$$

$$f_i(z^k \otimes v) = \begin{cases} z^{k-1} \otimes f_0 v & \text{if } i = 0, \\ z^k \otimes f_i v & \text{if } i \neq 0, \end{cases}$$

$$K_i^{\pm 1}(z^k \otimes v) = q_i^{\pm \langle h_i, \text{wt}(v) \rangle}(z^k \otimes v) \ \ (i \in I).$$

We define a $U_q'(\mathfrak{g})$-module automorphism $z_M : M_{\text{aff}} \to M_{\text{aff}}$ of weight δ by

$$z^k \otimes v \mapsto z^{k+1} \otimes v \ \ (v \in M,\ k \in \mathbf{Z}).$$

Let M_1, M_2 be good $U_q'(\mathfrak{g})$-modules and let u_1, u_2 be dominant extremal weight vectors of M_1 and M_2, respectively. Set $z_1 = z_{M_1}$ and $z_2 = z_{M_2}$. Then there exists a unique $U_q'(\mathfrak{g})$-module homomorphism

$$R^{\text{norm}}_{M_1,M_2}(z_1, z_2) : (M_1)_{\text{aff}} \otimes (M_2)_{\text{aff}} \longrightarrow \mathbf{k}(z_1, z_2) \otimes_{\mathbf{k}[z_1^{\pm 1}, z_2^{\pm 1}]} (M_2)_{\text{aff}} \otimes (M_1)_{\text{aff}}$$

satisfying

$$R^{\text{norm}}_{M_1,M_2}(u_1 \otimes u_2) = u_2 \otimes u_1,$$
$$R^{\text{norm}}_{M_1,M_2} \circ z_1 = z_1 \circ R^{\text{norm}}_{M_1,M_2},$$
$$R^{\text{norm}}_{M_1,M_2} \circ z_2 = z_2 \circ R^{\text{norm}}_{M_1,M_2}.$$

The homomorphism $R^{\text{norm}}_{M_1,M_2}$ is called the *normalized R-matrix of M_1 and M_2*.

Note that $\operatorname{Im} R^{\text{norm}}_{M_1,M_2} \subset \mathbf{k}(z_2/z_1) \otimes_{\mathbf{k}[(z_2/z_1)^{\pm 1}]} (M_2)_{\text{aff}} \otimes (M_1)_{\text{aff}}$. We denote by $d_{M_1,M_2}(u) \in \mathbf{k}[u]$ the monic polynomial of the smallest degree such that

$$\operatorname{Im}\left(d_{M_1,M_2}(z_2/z_1)\, R^{\text{norm}}_{M_1,M_2}\right) \subset (M_2)_{\text{aff}} \otimes (M_1)_{\text{aff}}.$$

The polynomial $d_{M_1,M_2}(u)$ is called the *denominator* of $R^{\text{norm}}_{M_1,M_2}$.

The normalized R-matrix satisfies the Yang-Baxter equation. That is, for $U_q'(\mathfrak{g})$-modules M_1, M_2, M_3, we have

$$\begin{aligned}(R^{\text{norm}}_{M_2,M_3} \otimes 1) \circ (1 \otimes R^{\text{norm}}_{M_1,M_3}) \circ (R^{\text{norm}}_{M_1,M_2} \otimes 1) \\ = (1 \otimes R^{\text{norm}}_{M_1,M_2}) \circ (R^{\text{norm}}_{M_1,M_3} \otimes 1) \circ (1 \otimes R^{\text{norm}}_{M_2,M_3}).\end{aligned}$$

6. Quantum affine Schur-Weyl duality functor

Let $\{V_s \mid s \in \mathcal{S}\}$ be a family of good modules and let v_s be a dominant extremal weight vector in V_s with weight λ_s $(s \in \mathcal{S})$. Take an index set J endowed with the maps $X : J \to \mathbf{k}^\times$ and $s : J \to \mathcal{S}$. For each $i, j \in J$, let

$$\begin{aligned}R^{\text{norm}}_{V_{s(i)},V_{s(j)}}(z_i, z_j) : (V_{s(i)})_{\text{aff}} \otimes (V_{s(j)})_{\text{aff}} \\ \longrightarrow \mathbf{k}(z_i, z_j) \otimes_{\mathbf{k}[z_i^{\pm 1}, z_j^{\pm 1}]} (V_{s(j)})_{\text{aff}} \otimes (V_{s(i)})_{\text{aff}}\end{aligned}$$

be the normalized R-matrix sending $v_{s(i)} \otimes v_{s(j)}$ to $v_{s(j)} \otimes v_{s(i)}$.

Let $d_{V_{s(i)},V_{s(j)}}(z_j/z_i)$ be the denominator of $R^{\mathrm{norm}}_{V_{s(i)},V_{s(j)}}(z_i, z_j)$. We define a quiver Γ^J as follows.

(i) We take J to be the set of vertices.

(ii) We put d_{ij} many arrows from i to j, where d_{ij} the order of zero of $d_{V_{s(i)},V_{s(j)}}(z_j/z_i)$ at $z_j/z_i = X(j)/X(i)$.

Define the Cartan matrix $A^J = (a^J_{ij})_{i,j\in J}$ by

$$a^J_{ij} = \begin{cases} 2 & \text{if } i = j, \\ -d_{ij} - d_{ji} & \text{if } i \neq j. \end{cases} \tag{6.1}$$

Thus we obtain a symmetric Cartan datum $(A^J, P, P^\vee, \Pi, \Pi^\vee)$ associated with Γ^J.

Set

$$\mathsf{Q}^J_{ij}(u, v) := \begin{cases} 0 & \text{if } i = j, \\ (u - v)^{d_{ij}}(v - u)^{d_{ji}} & \text{if } i \neq j. \end{cases} \tag{6.2}$$

We will denote by $R^J(\beta)$ $(\beta \in Q_+)$ the Khovanov-Lauda-Rouquier algebra associated with (A^J, Q^J).

For each $\nu = (\nu_1, \ldots, \nu_n) \in J^\beta$, let $\widehat{\mathcal{O}}_{\mathbf{T}^n,X(\nu)} = \mathbf{k}[[X_1 - X(\nu_1), \ldots, X_n - X(\nu_n)]]$ be the completion of $\mathcal{O}_{\mathbf{T}^n,X(\nu)}$ at $X(\nu) := (X(\nu_1), \ldots, X(\nu_n))$ and set

$$V_\nu := (V_{s(\nu_1)})_{\mathrm{aff}} \otimes \cdots \otimes (V_{s(\nu_n)})_{\mathrm{aff}},$$

where $X_k = z_{V_{s(\nu_k)}}$ $(k = 1, \ldots, n)$.

We define

$$\widehat{V}_\nu := \widehat{\mathcal{O}}_{\mathbf{T}^n,X(\nu)} \otimes_{\mathbf{k}[X_1^{\pm 1},\ldots,X_n^{\pm 1}]} V_\nu \quad \text{and} \quad \widehat{V}^{\otimes\beta} := \bigoplus_{\nu\in J^\beta} \widehat{V}_\nu\, e(\nu).$$

The following proposition is one of the main results of [17].

Proposition 6.1 ([17]). *The space $\widehat{V}^{\otimes\beta}$ is a $(U'_q(\mathfrak{g}), R^J(\beta))$-bimodule.*

Hence we obtain a functor

$$\mathcal{F}_\beta : \mathrm{mod}\,(R^J(\beta)) \longrightarrow \mathrm{mod}\, U'_q(\mathfrak{g})$$

defined by

$$M \longmapsto \widehat{V}^{\otimes\beta} \otimes_{R^J(\beta)} M \qquad \text{for } M \in \mathrm{mod}\,(R^J(\beta)).$$

Write $\mathrm{mod}\,(R^J) := \bigoplus_{\beta\in Q_+} \mathrm{mod}\,(R^J(\beta))$ and set

$$\mathcal{F} = \bigoplus_{\beta\in Q_+} \mathcal{F}_\beta : \mathrm{mod}\,(R^J) \longrightarrow \mathrm{mod}\, U'_q(\mathfrak{g}).$$

The functor $\mathcal{F}$ is called the *quantum affine Schur-Weyl duality functor*. The basic properties of $\mathcal{F}$ are summarized in the following theorem.

Theorem 6.2 ([17]).

(a) *The functor $\mathcal{F}$ restricts to*

$$\mathcal{F} : \mathit{rep}\,(R^J) \longrightarrow \mathcal{C}_{\mathit{int}},$$

where $\mathit{rep}\,(R^J) := \bigoplus_{\beta\in Q_+} \mathit{rep}\,(R^J(\beta))$ and $\mathcal{C}_{\mathit{int}}$ denotes the category of finite dimensional integrable $U_q'(\mathfrak{g})$-modules.

(b) *For each $i \in J$, let $S(\alpha_i) := \mathbf{k}\,u(i)$ be the 1-dimensional graded simple $R^J(1)$-module defined by*

$$e(j)\,u(i) = \delta_{ij}\,u(i), \quad x_1\,u(i) = 0.$$

Then we have

$$\mathcal{F}(S(\alpha_i)) \cong (V_{s(i)})_{X(i)},$$

where $(V_{s(i)})_{X(i)}$ is the evaluation module of $V_{s(i)}$ at $z_i = X(i)$.

(c) *$\mathcal{F}$ is a tensor functor; i.e., there exists a canonical $U_q'(\mathfrak{g})$-module isomorphisms*

$$\mathcal{F}(R^J(0)) \cong \mathbf{k}, \quad \mathcal{F}(M \circ N) \cong \mathcal{F}(M) \otimes \mathcal{F}(N)$$

for $M \in \mathit{rep}\,(R^J(m))$, $N \in \mathit{rep}\,(R^J(n))$.

(d) *If the quiver Γ^J is of type A_n $(n \geq 1)$, D_n $(n \geq 4)$, E_6, E_7, E_8, then $\mathcal{F}$ is exact.*

7. The Categories $\mathcal{T}_N$ and $\mathcal{C}_N$

Take $\mathbf{k} = \mathbf{C}(q)$. Let $\mathfrak{g} = A_{N-1}^{(1)}$ be the affine Kac-Moody algebra of type $A_{N-1}^{(1)}$ and let $V = V(\varpi_1)$ be the fundamental representation of $U_q'(A_{N-1}^{(1)})$ of weight ϖ_1.

Set $\mathcal{S} = \{V\}$, $J = \mathbf{Z}$ and let $X : \mathbf{Z} \to \mathbf{k}^\times$ be the map given by $j \mapsto q^{2j}$ $(j \in \mathbf{Z})$. Then the normalized R-matrix $R_{V,V}^{\mathrm{norm}} : V_{z_1} \otimes V_{z_2} \longrightarrow V_{z_2} \otimes V_{z_1}$ has the denominator $d_{V,V}(z_2/z_1) = z_2/z_1 - q^2$. Hence we have

$$d_{ij} = \begin{cases} 1 & \text{if } j = i+1, \\ 0 & \text{otehrwise,} \end{cases}$$

which yields the quiver Γ^J of type A_∞. Take $P_J = \bigoplus_{k\in\mathbf{Z}} \mathbf{Z}\,\varepsilon_k$ to be the weight lattice and let $Q^J = \bigoplus_{k\in\mathbf{Z}} \mathbf{Z}\,(\varepsilon_k - \varepsilon_{k+1})$ be the root lattice. There is a bilinear form on P_J given by $(\varepsilon_a, \varepsilon_b) = \delta_{ab}$.

For $a \leq b$, let $l = b - a + 1$ and let $L(a,b) := \mathbf{k}\,u(a,b)$ be the 1-dimensional graded simple $R^J(\varepsilon_a - \varepsilon_{b+1})$-module defined by

$$x_s\,u(a,b) = 0, \quad \tau_t\,u(a,b) = 0 \ \ (1 \leq s \leq l,\ 1 \leq t \leq l-1),$$
$$e(\nu)\,u(a,b) = \begin{cases} u(a,b) & \text{if } \nu = (a, a+1, \ldots, b), \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$\mathcal{F}(L(a,b)) \cong \begin{cases} V(\varpi_l)_{(-q)^{a+b}} & \text{if } 0 \leq l \leq N, \\ 0 & \text{if } l > N, \end{cases}$$

where $\mathcal{F} : \mathrm{mod}\,(R^J(l)) \longrightarrow \mathrm{mod}\, U_q'(\mathfrak{g})$ is the quantum affine Schur-Weyl duality functor.

Recall that $\mathrm{Rep}\,(R^J(l))$ is the category of finite dimensional graded $R^J(l)$-modules. Set $\mathcal{R} := \bigoplus_{l\geq 0} \mathrm{Rep}\,(R^J(l))$ and let $\mathcal{S}$ be the Smallest Serre subcategory of $\mathcal{R}$ such that

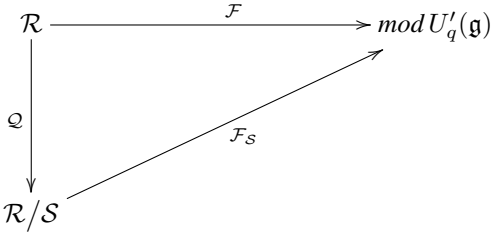
(i) $\mathcal{S}$ contains $L(a, a+N)$ for all $a \in \mathbf{Z}$,

(ii) $X \circ Y, Y \circ X \in \mathcal{S}$ for all $X \in \mathcal{R}, Y \in \mathcal{S}$.

Take the quotient category $\mathcal{R}/\mathcal{S}$ and let $\mathcal{Q} : \mathcal{R} \to \mathcal{R}/\mathcal{S}$ be the canonical projection functor. Then we have:

Proposition 7.1 ([17]).

(a) *The functor $\mathcal{F}$ factors through $\mathcal{R}/\mathcal{S}$. That is, there is a canonical functor $\mathcal{F}_{\mathcal{S}} : \mathcal{R}/\mathcal{S} \longrightarrow mod\, U_q'(\mathfrak{g})$ such that the following diagram is commutative.*



(b) *The functor $\mathcal{F}_{\mathcal{S}}$ sends a simple object in $\mathcal{R}/\mathcal{S}$ to a simple object in mod $U_q'(\mathfrak{g})$.*

Let $L_a := L(a, a+N-1)$ and $u_a := u(a, a+N-1) \in L_a$ $(a \in \mathbf{Z})$. Then $\mathcal{F}(L_a)$ is isomorphic to the trivial representation of $U_q'(\mathfrak{g})$. Let $S : P_J \to P_J$ $(\varepsilon_a \mapsto \varepsilon_{a+N-1})$ be an automorphism on P_J and let B be the bilinear form on P_J given by

$$B(x,y) := -\sum_{k>0}(S^k x, y) \ \text{ for all } x, y \in P_J.$$

We define a new tensor product $\star$ on $\mathcal{R}/\mathcal{S}$ by

$$X \star Y := q^{B(\alpha,\beta)}\, X \circ Y \quad \text{for } X \in (\mathcal{R}/\mathcal{S})_\alpha,\ Y \in (\mathcal{R}/\mathcal{S})_\beta.$$

Then there exists an isomorphism $R(a)(X) : L_a \star X \xrightarrow{\sim} X \star L_a$ which is functorial in $X \in \mathcal{R}/\mathcal{S}$. Moreover, the isomorphisms

$$R_a(L_b) : L_a \star L_b \xrightarrow{\sim} L_b \star L_a \ \text{ and } \ R_b(L_a) : L_b \star L_a \xrightarrow{\sim} L_a \star L_b$$

are inverses to each other. One can verify that $\{L_a, R_a(L_b) \mid a, b \in \mathbf{Z}\}$ forms a commuting family of central objects in $(\mathcal{R}/\mathcal{S}, \star)$ (see [17, Appendix A.6]).

Let $\mathcal{T}_N' := (\mathcal{R}/\mathcal{S})[L_a^{\star -1} \mid a \in \mathbf{Z}]$ be the localization of $\mathcal{R}/\mathcal{S}$ by this commuting family and define

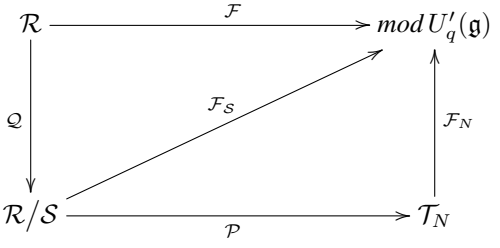
$$\mathcal{T}_N := (\mathcal{R}/\mathcal{S})[L_a \cong \mathbf{1} \mid a \in \mathbf{Z}].$$

We denote by $\mathcal{P} : \mathcal{R}/\mathcal{S} \to \mathcal{T}_N$ the canonical functor.

Theorem 7.2 ([17]).

(a) *The category $\mathcal{T}_N$ is a rigid tensor category ; i.e., every object in $\mathcal{T}_N$ has a right dual and a left dual.*

(b) *The functor $\mathcal{F}_{\mathcal{S}}$ factors through $\mathcal{T}_N$. That is, there exists a canonical functor $\mathcal{F}_N : \mathcal{T}_N \longrightarrow mod\, U_q'(\mathfrak{g})$ such that the following diagram is commutative.*



(c) *The functor $\mathcal{F}_N$ is exact and sends a simple object in $\mathcal{T}_N$ to a simple object in $mod\, U_q'(\mathfrak{g})$.*

Let $\mathcal{C}_N$ be the smallest full subcategory of $\mathcal{C}_{\text{int}}$ consisting of $U_q'(\mathfrak{g})$-modules M such that every composition factor of M appears as a composition factor of a tensor product of modules of the form $V(\varpi_1)_{q^{2j}}$ $(j \in \mathbf{Z})$. Thus $\mathcal{C}_N$ is an abelian category containing all $U_q'(\mathfrak{g})$-modules $V(\varpi_i)_{(-q)^{i+2a-1}}$ for $1 \le i \le N-1$ and $a \in \mathbf{Z}$. Moreover, $\mathcal{C}_N$ is stable under taking submodules, quotients, extensions and tensor products. Hence $\mathcal{F}_N$ restricts to an exact functor

$$\mathcal{F}_N : \mathcal{T}_N \longrightarrow \mathcal{C}_N.$$

Let $Irr\,(\mathcal{T}_N)$ (respectively, $Irr\,(\mathcal{C}_N)$) denote the set of isomorphism classes of simple objects in $\mathcal{T}_N$ (respectively, in $\mathcal{C}_N$). Define an equivalence relation on $Irr\,(\mathcal{T}_N)$ by setting $X \sim Y$ if and only if $X \xrightarrow{\sim} q^m Y$ for some $m \in \mathbf{Z}$. Set

$$Irr\,(\mathcal{T}_N)|_{q=1} := Irr\,(\mathcal{T}_N)\big/ \sim .$$

Theorem 7.3 ([17]).

(a) *The functor $\mathcal{F}_N$ induces a bijection between $Irr\,(\mathcal{T}_N)|_{q=1}$ and $Irr\,(\mathcal{C}_N)$.*

(b) *The exact functor $\mathcal{F}_N$ induces a ring isomorphism*

$$\phi_N : K(\mathcal{T}_N)|_{q=1} \xrightarrow{\sim} K(\mathcal{C}_N).$$

Therefore, the category $\mathcal{T}_N$ provides a graded lifting of $\mathcal{C}_N$ as a rigid tensor category.

8. The category $\mathcal{C}_Q$

In this section, we deal with affine Kac-Moody algebras $\mathfrak{g}$ of type $A_n^{(1)}$ $(n \ge 1)$, $D_n^{(1)}$ $(n \ge 4)$, $E_6^{(1)}$, $E_7^{(1)}$, $E_8^{(1)}$. Let $I = \{0, 1, \ldots, n\}$ be the index set for the simple roots of $\mathfrak{g}$ and set $I_0 = I \setminus \{0\}$. We denote by $\mathfrak{g}_0$ the finite dimensional simple Lie subalgebra of $\mathfrak{g}$

generated by e_i, f_i, h_i $(i \in I_0)$. Thus $\mathfrak{g}_0$ is of type A_n $(n \geq 1)$, D_n $(n \geq 4)$, E_6, E_7, E_8, respectively.

Let Q be the Dynkin quiver associated with $\mathfrak{g}_0$. A function $\xi : I_0 \to \mathbf{Z}$ is called a *height function* if $\xi_j = \xi_i - 1$ whenever we have an arrow $i \to j$.

Set

$$\widehat{I_0} := \{(i,p) \in I_0 \times \mathbf{Z} \mid p - \xi_i \in 2\mathbf{Z}\}.$$

The *repetition quiver* $\widehat{Q}$ is defined as follows.

(i) We take $\widehat{I_0}$ to be the set of vertices.

(ii) The arrows are given by

$$(i,p) \to (j,p+1), \quad (j,q) \to (i,q+1)$$

for all arrows $i \to j$ and $p, q \in Z$ such that $p - \xi_i \in \mathbf{Z}$, $q - \xi_j \in \mathbf{Z}$.

For all $i \in I_0$, let $s_i(Q)$ be the quiver obtained from Q by reversing the arrows that touch i. A reduced expression $w = s_{i_1} \cdots s_{i_l} \in W_0$ is said to be *adapted to* Q if i_k is the source of $s_{i_{k-1}} \cdots s_{i_1}(Q)$ for all $1 \leq k \leq l$. It is known that there is a unique Coxeter element $\tau \in W_0$ which is adaped to Q.

Set $\widehat{\Delta} := \Delta_+ \times \mathbf{Z}$, where Δ_+ is the set of positive roots of $\mathfrak{g}_0$. For each $i \in I_0$, let $B(i) := \{j \in I_0 \mid \text{ there is a path from } j \text{ to } i\}$ and define $\gamma_i := \sum_{j \in B(i)} \alpha_j$. We define a bijection $\phi : \widehat{I_0} \to \widehat{\Delta}$ inductively as follows.

(1) We begin with $\phi(i, \xi_i) := (\gamma_i, 0)$.

(2) If $\phi(i,p) = (\beta, j)$ is given, then we define

- $\phi(i, p-2) := (\tau(\beta), j)$ if $\tau(\beta) \in \Delta_+$,
- $\phi(i, p-2) := (-\tau(\beta), j-1)$ if $\tau(\beta) \in \Delta_-$,
- $\phi(i, p+2) := (\tau^{-1}(\beta), j)$ if $\tau^{-1}(\beta) \in \Delta_+$,
- $\phi(i, p+2) := (-\tau^{-1}(\beta), j+1)$ if $\tau^{-1}(\beta) \in \Delta_-$.

Let w_0 be the longest element of W_0 and fix a reduced expression $w_0 = s_{i_1} \cdots s_{i_l}$ which is adapted to Q. Set

$$J := \{(i,p) \in \widehat{I_0} \mid \phi(i,p) \in \Pi_0 \times \{0\}\},$$

where Π_0 denotes the set of simple roots of $\mathfrak{g}_0$. Take the maps $X : J \to \mathbf{k}^\times$ and $s : J \to \{V(\varpi_i) \mid i \in I_0\}$ defined by

$$X(i,p) = (-q)^{p+h}, \quad s(i,p) = V(\varpi_i) \quad \text{for } (i,p) \in J,$$

where h is the Coxeter number of $\mathfrak{g}_0$.

Theorem 8.1 ([18]). *For any $(i,p), (j,r) \in J$, assume that the normalized R-matrix*

$$R^{norm}_{V(\varpi_i), V(\varpi_j)}(z)$$

has a pole at $z = (-q)^{r-p}$ of order at most 1. Then the following statements hold.

(a) *The Cartan matrix A^J associated with (J, X, s) is of type $\mathfrak{g}_0$.*

(b) *There exists a quiver isomorphism*

$$Q^{rev} \xrightarrow{\sim} \Gamma^J, \quad k \mapsto \phi^{-1}(\alpha_k, 0) \ \ (k \in I_0),$$

where Q^{rev} is the reverse quiver of Q.

(c) *The functor $\mathcal{F} : \mathrm{rep}\,(R^J) \to \mathcal{C}_{int}$ is exact and*

$$\mathcal{F}(S(\alpha_k)) \cong V(\varpi_i)_{(-q)^{p+h}},$$

where $\phi(i,p) = (\alpha_k, 0)$.

Remark 8.2. When $\mathfrak{g}$ is of type $A_n^{(1)}$ $(n \geq 1)$ or $D_n^{(1)}$ $(n \geq 4)$, then the condition in Theorem 8.1 is satisfied. We conjecture that the same is true of $\mathfrak{g} = E_6^{(1)}$, $E_7^{(1)}$, $E_8^{(1)}$.

We now bring out the main subject of our interest in this section. Let $\mathcal{C}_Q$ be the smallest abelian full subcategory of $\mathcal{C}_{\text{int}}$ such that

(i) $\mathcal{C}_Q$ is stable under taking submodules, subquotients, direct sums and tensor products,

(ii) $\mathcal{C}_Q$ contains all $U_q'(\mathfrak{g})$-modules of the form $V(\beta)_z \big/ (z-1)^l\, V(\beta)_z$ $(\beta \in \Delta_+,\, l \geq 1)$. Here, $V(\beta) = V(\varpi_i)_{(-q)^{p+h}}$ such that $\phi(i,p) = (\beta, 0)$.

Let $\mathrm{Nilrep}\,(R^J(\beta))$ be the category of finite dimensional ungraded $R^J(\beta)$-modules such that all x_k's act nilpotently and set

$$\mathrm{Nilrep}\,(R^J) := \bigoplus_{\beta \in Q_+} \mathrm{Nilrep}\,(R^J(\beta)).$$

Note that every module in $\mathrm{Nilrep}\,(R^J)$ can be obtained by taking submodules, subquotients, direct sums and convolution products of $P(\alpha_k) \big/ (x_1^l)$ $(k \in I_0, l \geq 0)$, where $P(\alpha_k)$ is the projective cover of $S(\alpha_k)$. Thus we obtain a well-defined functor

$$\mathcal{F} : \mathrm{Nilrep}\,(R^J) \longrightarrow \mathcal{C}_Q,$$

which satisfies the following properties.

Theorem 8.3 ([17, 18]).

(a) *$\mathcal{F}$ is an exact tensor functor.*

(b) *$\mathcal{F}$ sends a simple object in $\mathrm{Nilrep}\,(R^J)$ to a simple object in $\mathcal{C}_Q$.*

It is straightforward to verify that $\mathcal{F}$ is a faithful functor. Since $\mathcal{C}_Q$ is the smallest abelian full subcategory of $\mathcal{C}_{\text{int}}$ satisfying the conditions (i) and (ii) given above, we conjecture that $\mathcal{F}$ is full and defines an equivalence of categories.

Remark 8.4. Note that our general approach to quantum affine Schur-Weyl duality applies to *all* quantum affine algebras and *any* choice of good modules. Thus we expect there are a lot more exciting developments to come. It is an interesting question whether our general construction can shed a new light on the hidden connection between quantum affine algebras and cluster algebras (cf. [12]).

Acknowledgements. This work was supported by NRF Grant #2013-035155 and NRF Grant #2013-055408. The author is very grateful to Masaki Kashiwara, Myungho Kim and Se-jin Oh for many valuable discussions and suggestions on this paper.

References

[1] T. Akasaka and M. Kashiwara, *Finite-dimensional representations of quantum affine algebras*, Publ. RIMS. Kyoto Univ. **33** (1997), 839-867.

[2] S. Ariki, *On the decomposition numbers of the Hecke algebra of* $G(M, 1, n)$, J. Math. Kyoto Univ. **36** (1996), 789-808.

[3] V. Chari and A. Pressley, *A guide to Quantum Groups*, Cambridge University Press, Cambridge, 1994.

[4] _____, *Quantum affine algebras and affine Hecke algebras*, Pacific J. Math. **174** (1996), 295-326.

[5] I. V. Cherednik, *A new interpretation of Gelfand-Tzetlin bases*, Duke Math. J. **54** (1987), 563-577.

[6] J. Chuang and R. Rouquier, *Derived equivalences for symmetric groups and* $\mathfrak{sl}_2$*-categorification*, Ann. Math. **167** (2008), 245–298.

[7] L. Crane and I. B. Frenkel, *Four-dimensional topological quantum field theory, Hopf categories, and the canonical bases*, J. Math. Phys. **35** (1994), 5136-5154.

[8] A. Ellis, M. Khovanov, and A. Lauda, *The odd nilHecke algebra and its diagrammatics*, Int. Math. Res. Notices **2014-4** (2014), 991-1062.

[9] A. Ellis and A. Lauda, *An odd categorification of quantum* $sl(2)$, arXiv:1307.7816.

[10] V. Ginzburg , N. Reshetikhin., and E. Vasserot, *Quantum groups and flag varieties*, Contemp. Math. **175** (1994), 101-130.

[11] J. Hong, S.-J. Kang, *Introduction to Quantum Groups and Crystal Bases*, Graduate Studies in Mathematics **42**, American Mathematical Society, Provodence, 2002.

[12] D. Hernandez and B. Leclerc, *Cluster algebras and quantum affine algebras*, Duke Math. J. **154** (2010), 265-341.

[13] D. Hill and W. Wang, *Categorification of quantum Kac-Moody superalgebras*, to appear in Trans. Amer. Math. Soc., arXiv:1202.2769.

[14] M. Jimbo, *A* q*-analogue of* $U(\mathfrak{gl}_{N+1})$*, Hecke algebra, and the Yang-Baxter equation*, Lett. Math.Phys. **11** (1986), 247-252.

[15] V. Kac, *Infinite Dimensional Lie algebras*, 3rd ed., Cambridge University Press, Cambridge, 1990.

[16] S.-J. Kang and M. Kashiwara, *Categorification of highest weight modules via Khovanov- Lauda-Rouquier algebras*, Invent. Math. **190** (2012), 699–742.

[17] S.-J. Kang, M. Kashiwara, and M. Kim, *Symmetric quiver Hecke algebras and R-matrices of quantum affine algebras*, arXiv:1304.0323.

[18] _____, *Symmetric quiver Hecke algebras and R-matrices of quantum affine algebras II*, arXiv:1308.0651.

[19] S.-J. Kang, M. Kashiwara, and S.-j. Oh, *Categorification of highest weight modules over quantum generalized Kac-Moody algebras*, Moscow Math. J. **13** (2103), 315–343.

[20] _____, *Supercategorification of quantum Kac-Moody algebras*, Adv. Math. **242** (2013), 116–162.

[21] _____, *Supercategorification of quantum Kac-Moody algebras II*, arXiv:1303.1916.

[22] S.-J. Kang, M. Kashiwra, and E. Park, *Geomtric realization of Khovanov-Lauda-Rouquier algebras associated with Borcherds-Cartan data*, Proc. London Math. Soc **(3) 107** (2013), 907-931.

[23] S.-J. Kang, M. Kashiwara, and S. Tsuchioka, *Quiver Hecke superalgebras*, to appear in J. reine angew. Math.

[24] S.-J. Kang, S.-j. Oh, and E. Park, *Categorification of quantum generalized Kac-Moody algebras and crystal bases*, Int. J. Math. **23** (2012), 1250116.

[25] M. Kashiwara, *On crystal bases of the q-analogue of universal enveloping algebras*, Duke Math. J. **63** (1991), 465–516.

[26] _____, *On level zero representations of quantized affine algebras*, Duke. Math. J. **112** (2002), 117–175.

[27] M. Khovanov and A. Lauda, *A diagrammatic approach to categorification of quantum groups I*, Represent. Theory **13** (2009), 309–347.

[28] _____, *A diagrammatic approach to categorification of quantum groups II*, Trans. Amer. Math. Soc. **363** (2011), 2685–2700.

[29] A. Lascoux, B. Leclerc, and J-Y. Thibon, *Hecke algebras at roots of unity and crystal bases of quantum affine algebras*, Commun. Math. Phys. **181** (1996), 205–263.

[30] G. Lusztig, *Canonical bases arising from quantized enveloping algebras*, J. Amer. Math. Soc. **3** (1990), 447–498.

[31] _____, *Introduction to Quantum Groups*, Birkhäuser, Boston, 1993.

[32] R. Rouquier, *2-Kac-Moody algebras*, arXiv:0812.5023.

[33] _____, *Quiver Hecke algebras and 2-Lie algebras*, arXiv:1112.3619.

[34] I. Schur, *Über Eine Klasse Von Matrizen, die Sich Einer Gegeben Matrix Zurodenen Lassen*, Ph.D. thesis (1901). Reprinted in Gesamelte Abhandlungen **1**, 1–70.

[35] _____, *Über die rationalen Darstellungen der allgemeinen linearen Gruppe*, Sitzungsberichte der Königlich Preussischen Adademie der Wissenschaften zu Berlin (1927), 58–75. Reprinted in Gessamelte Abhandlungen **3**, 68–85.

[36] M. Varagnolo and E. Vasserot, *Canonical bases and KLR algebras*, J. reine angew. Math. **659** (2011), 67–100.

[37] W. Wang, *Spin Hecke algebras of finite and affine types*, Adv. Math. **212** (2007), 723–748.

[38] Ben Webster, *Knot invariants and higher representation theory*, arXiv:1309.3796.

Department of Mathematical Sciences, Seoul National University, 599 Gwanak-Ro, Seoul 151-747, Korea

E-mail: sjkang@snu.ac.kr

Finitely Generated Groups with Controlled Pro-algebraic Completions

Martin Kassabov

Abstract. We construct finitely generated groups whose pro-algebraic completions are isomorphic to the product of the pro-algebraic completions of groups like $\mathrm{SL}_n(\mathbb{Z})$ and $\mathrm{SL}_n(\mathbb{Z}[x])$ for different n. This leads to examples of groups where the dimensions of the character varieties grow as any function with growth between linear and quadratic.

Mathematics Subject Classification (2010). Primary 20F69; Secondary 19C99, 20B30, 20E18, 20G05, 20G30, 20K25

Keywords. Pro-algebraic completion of groups, representation varieties, dimensions of character varieties.

1. Introduction

Let Γ be a finitely generated group, and for every positive integer d we consider the representations of Γ of dimension d over a fixed algebraically closed field $\mathbb{K}$ of characteristic 0. The set of these representations $R_\Gamma(d) = \mathrm{Hom}(\Gamma, \mathrm{GL}_d(\mathbb{K}))$ has a natural structure of an affine algebraic variety defined over $\mathbb{Q}$. There is a natural adjoint action of $\mathrm{GL}_d(\mathbb{K})$ on $R_\Gamma(d)$ and one can form the (categorical) quotient $X_\Gamma(d) = R_\Gamma(d)/\mathrm{GL}_d(\mathbb{K})$, which is also an algebraic variety defined over $\mathbb{Q}$. It is well-known that the points of this variety parameterize the isomorphism classes of completely reducible representations of Γ [14, 19].

We are mainly interested in the dimensions of these varieties $\varkappa_\Gamma(n) = \dim X_\Gamma(d)$. The group Γ is called rigid if $\varkappa_\Gamma(d) = 0$ for all d. If Γ is finitely generated and not rigid then $\varkappa_\Gamma(d)$ is bounded below by a linear function [18] and above by a quadratic function. One of our main results shows that any function satisfying a mild additional condition is the growth function of the dimension of the character varieties for some finitely generated group.

Theorem 1.1. *Let $f : \mathbb{N} \to \mathbb{N}$ be a function such that $f(d)/d$ is non-decreasing and $f(d) \leq d(d-1)/2$. Then there exists a finitely generated group Γ_f such that $\dim X_{\Gamma_f}(d) = f(d)$ for all $d \geq 3$, where $X_{\Gamma_f}(d)$ denotes the character variety of Γ_f.*

It can be shown that the groups Γ_f constructed in Theorem 1.1 are not linear. It is not known if there exists a linear group with growth of the dimensions of the character varieties strictly between linear and quadratic. The above theorem is actually an application of the following:

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Theorem 1.2. *There exists a finitely generated group* Γ *inside the infinite product* $\prod_{n\geq 3} \mathrm{SL}_n(\mathbb{Z})$ *such that* Γ *contains the direct sum* $\bigoplus_{n\geq 3} \mathrm{SL}_n(\mathbb{Z})$ *and any finite dimensional representation of* Γ *factors through one of finite products* $\prod_{n\geq 3}^{d} \mathrm{SL}_n(\mathbb{Z})$ *for some* d.

The group constructed in Theorem 1.2 is rigid[1] [2], i.e., its character variety is discrete and has only finitely many isomorphism types of representations of each degree. This is almost an immediate consequence of the theorem and the rigidity of the groups $\mathrm{SL}_n(\mathbb{Z})$ for $n \geq 3$. However this group is not super-rigid.

The idea behind this result is to construct groups with controlled pro-algebraic completion. The *pro-algebraic completion*[2] $\overline{\Gamma}$ of an affine pro-algebraic group is an inverse limit of algebraic groups which captures all finite dimensional representations of the group.

The condition that all finite dimensional representations factor through finite products is equivalent to saying that the pro-algebraic completion is isomorphic to a certain infinite product, which can be captured using the language of "congruence" kernels.

The inclusion $\Gamma \hookrightarrow \prod_{n\geq 3} \mathrm{SL}_n(\mathbb{Z})$ induces a map between the pro-algebraic completions

$$\iota : \overline{\Gamma} \to \prod_{n\geq 3} \overline{\mathrm{SL}_n(\mathbb{Z})}$$

and the property that Γ surjects onto the finite products implies that ι is surjective. In general such map is not injective because Γ might have finite dimensional representations which do not extend to the infinite product (equivalently are not continuous with respect to the topology on Γ induced from the product). In general the kernel of ι is called the "congruence" kernel.[3] The condition that all finite dimensional representations factor through finite products is exactly the same as the triviality of $\ker \iota$, i.e., Theorem 1.2 gives that ι is an isomorphism.

Theorem 1.2 also provides constructions of finitely generated rigid groups such that the connected component of the pro-algebraic completion is any infinite product $\prod_i S_i$ of simple algebraic groups S_i over $\mathbb{K}$ such that the $\mathbb{Q}$-ranks of S_i are strictly increasing.

The idea behind the proofs of these theorems is first to produce rings with controlled pro-algebraic completions. It is not clear if there is an analog of Theorem 1.2 for rings, however it is relatively easy to construct a ring $\boldsymbol{R}$ with relatively small congruence kernel. Then, one can form the Steinberg groups $\mathrm{St}_n(\boldsymbol{R})$ and it can be shown that they also have relatively small completion. The final step of the construction is to use two copies of the group $\mathrm{St}_n(\boldsymbol{R})$ satisfying addition relations which can be used to kill the congruence kernels. The step significantly uses that the completion turns out to be not only a normal subgroup but also a quotient of the pro-algebraic completion of $\mathrm{St}_n(\boldsymbol{R})$ (see Lemma 3.6).

The groups in Theorem 1.1 are subgroups inside $\prod_{n\geq 3} \mathrm{SL}_n(\mathbb{Z}[x_1, \ldots, x_{f(n)}])$ which are generated by two copies of the group Γ from Theorem 1.2. The triviality of the congruence kernel is an easy consequence of the same property for the group Γ, which reduces Theorem 1.1 to computing the dimensions of the character varieties for groups like $\mathrm{SL}_n(\mathbb{Z}[x_1, \ldots, x_k])$ (see Lemma 2.15).

[1] Also called representation rigid or SS-rigid.

[2] Sometimes the pro-algebriac completion is called Hochschild-Mostow group.

[3] The terminology of congruence kernel originated in the study of finite images of arithmetic groups.

2. Pro-algebraic completions of groups and rings

2.1. Pro-algebraic completions of groups. The pro-algebraic completion is an affine pro-algebric group:

Definition 2.1. An affine pro-algebraic group $\mathcal{G}$ is an inverse limit $\varprojlim G_i$ of affine algebraic groups G_i defined over the field $\mathbb{K}$. A subgroup $\Gamma \subset \mathcal{G}$ is called Zariski dense if the image of Γ in any G_i is Zariski dense.

There is a corresponding notion of morphism between pro-algebraic groups. Since the category of pro-algebraic groups is closed under inverse limits one can define the pro-algebraic completion as an universal object:

Definition 2.2. A pro-algebraic completion for Γ relative to the field $\mathbb{K}$ is a pair $(\overline{\Gamma}, i)$ consisting of a pro-algebraic group $\overline{\Gamma}$ and a homomorphism $i : \Gamma \to \overline{\Gamma}$ such that for any pro-algebraic group $\mathcal{G}$ and any homomorphism $i' : \Gamma \to \mathcal{G}$ there is a unique morphism of pro-algebraic group $\pi : \overline{\Gamma} \to \mathcal{G}$ such that $i' = \pi \circ i$.

It is immediate from the definition that a pro-algebraic completion is unique up to a unique isomorphism. Moreover the image of Γ under i is Zariski dense in $\overline{\Gamma}$. It is also easy to see that it is enough for a pro-algebraic completion for Γ only to satisfy the definition for the case that $\mathcal{G}$ is an affine algebraic group.

A more constructive way of describing the pro-algebraic completion is as an inverse limit

$$\overline{\Gamma} = \varprojlim_{\rho} \overline{\rho(\Gamma)}^{\mathrm{Zar}},$$

taken over all finite dimensional representations $\rho : \Gamma \to \mathrm{GL}_N(\mathbb{K})$ for some N, where $\overline{H}^{\mathrm{Zar}}$ denotes the Zariski closure of the subgroup H. Equivalently $\overline{\Gamma}$ is the Zariski closure of the image of Γ in the product

$$P = \prod_{\rho} \{\mathrm{GL}_N(\mathbb{K}) \mid \rho : \Gamma \to \mathrm{GL}_N(\mathbb{K})\}$$

under the diagonal homomorphism $\Gamma \to P$ induced by the ρ's.

Since any finite quotient of Γ can be considered as a finite dimensional representation with a finite image there is a canonical surjective map $\overline{\Gamma} \to \overline{\Gamma}^{\mathrm{f}}$ from the pro-algebraic completion to the pro-finite completion. The kernel of this map is the connected component $\overline{\Gamma}^0$ of the pro-algebraic group $\overline{\Gamma}$. The following result is well known:

Lemma 2.3 ([2]). *The pro-algebraic completion splits as semi-direct product*

$$\overline{\Gamma} = \overline{\Gamma}^0 \rtimes \overline{\Gamma}^{\mathrm{f}},$$

i.e., there exists a section $\overline{\Gamma}^{\mathrm{f}} \hookrightarrow \overline{\Gamma}$ *for the projection* $\overline{\Gamma} \to \overline{\Gamma}^{\mathrm{f}}$.

Remark 2.4. The pro-algebraic completion significantly depends on the characteristic of the field $\mathbb{K}$, for example if $\mathrm{char}\,\mathbb{K} > 0$ it is known that any map from $\mathrm{SL}_n(\mathbb{Z})$ to $\mathrm{GL}_N(\mathbb{K})$ has finite image, therefore the pro-algebraic completion coincides with the pro-finite completion. However, this is not the case in characteristic 0, see Example 2.14.

2.2. Algebraic rings. In this section, we present a quick survey of some basic properties of algebraic rings. The notion of algebraic rings was introduced by M. Greenberg [6, 7][4] and was used by Kassabov-Sapir [11] and I. Rapinchuk [17] to study representations of split Chevalley groups with coefficients in finitely generated rings and the Borel-Tits conjecture about abstract homomorphism between algebraic groups [4].

Definition 2.5. Let $\mathbb{K}$ be a fixed algebraically closed field. An algebraic ring over $\mathbb{K}$ is an affine algebraic variety $\mathcal{A}$ defined over $\mathbb{K}$ with two regular maps $\boldsymbol{\alpha}, \boldsymbol{\mu} : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ called 'addition' and 'multiplication' which define a ring structure (with 1) on the set $\mathcal{A}$. These maps are required to satisfy the standard axioms for associative ring.

The following result classifies algebraic rings over algebraically closed fields of characteristic 0.

Theorem 2.6 ([6, 11, 17]). *Let $\mathcal{A}$ be an associative algebraic ring with 1 over an algebraically closed field $\mathbb{K}$ of characteristic 0. Then the connected component $\mathcal{A}^0$ containing the element 0 is an ideal of finite index in $\mathcal{A}$, which is also a subring with 1 and $\mathcal{A}$ is a split extension*

$$0 \to \mathcal{A}^0 \to \mathcal{A} \to C \to 0$$

for some finite ring C. The additive group of $\mathcal{A}$ decomposes as direct sum $\mathcal{A} \simeq \mathcal{A}^0 \oplus C$, however this need not be a direct sum of rings since the finite ring C can act nontrivially on $\mathcal{A}^0$. Moreover the subring $\mathcal{A}^0$ is isomorphic as an algebraic ring to some finite dimensional algebra over $\mathbb{K}$. In particular the subring $\mathcal{A}^0$ satisfies the identity

$$s_M(x_1, \dots x_M) := \sum_{\sigma \in \mathrm{Sym}(M)} (-1)^\sigma x_{\sigma(1)} x_{\sigma(2)} \dots x_{\sigma(M)} = 0.$$

for any integer $M > \dim \mathcal{A}$.

Outline of the proof. The connected component $\mathcal{A}^0$ is a two-sided ideal in the ring $\mathcal{A}$. Under addition $\mathcal{A}^0$ is a connected commutative algebraic group and therefore breaks up into a product of semi-simple part and unipotent part. The multiplication of $\mathcal{A}$ gives that this group has many algebraic endomorphisms, which can be used to deduce that the semi-simple part is trivial. In the case of characteristic 0 any commutative unipotent group is isomorphic to the affine space $\mathbb{K}^n$ and it can be shown that the multiplication map $\boldsymbol{\mu}$ becomes bilinear, which turns $\mathcal{A}^0$ into an algebra over $\mathbb{K}$. The expression $s_M(x_1, \dots x_M)$ is an identity in the ring $\mathcal{A}^0$ because it holds in any $\mathbb{K}$ algebra of dimension less than M.

The splitting of the exact sequence is also a consequence of the splitting of the commutative algebraic group $\mathcal{A}$ into semi-simple part C and unipotent part $\mathcal{A}^0$. As an one-sided ideal $\mathcal{A}^0$ is finitely generated,[5] which can be used to deduce that the restriction of the multiplication map $\boldsymbol{\mu} : \mathcal{A}^0 \times \mathcal{A}^0 \to \mathcal{A}^0$ is surjective. The surjectivity of this map implies that $\mathcal{A}^0$ is a subring with an identity.

For complete proofs see [6] (in the connected case), [17] (in the commutative case), and [11] (in the case $\mathbb{K} = \mathbb{C}$). □

Remark 2.7. Theorem 2.6 does not hold over fields of positive characteristic, for example the ring of truncated Witt vectors in characteristic $p > 0$ have the structure of a connected

[4] Greenberg only considered connected algebraic rings.

[5] This statement might not be true if the field $\mathbb{K}$ has positive characteristic.

algebraic ring over $\mathbb{K}$ which is not isomorphic to an algebra over $\mathbb{K}$. It seems plausible that connected algebraic rings are isomorphic to quotients of algebras over the truncated Witt vectors (this is proven in the case of connected commutative rings in [6]).

Using the notion of algebraic ring one can define pro-algebraic rings and pro-algebraic completions.

Definition 2.8. Let R be a finitely generated ring. The pro-algebraic completion $\overline{R}$ of the ring R is the inverse limit of the images of all Zariski dense homomorphisms $\rho : R \to \mathcal{A}$ where $\mathcal{A}$ is algebraic ring.

Any finite ring can be given the structure of algebraic ring, i.e, the pro-algebraic completion captures all finite quotations of the ring R, therefore there exists a canonical projection from the pro-algebraic completion $\overline{R}$ to the pro-finite completion $\overline{R}^{\mathrm{f}}$. In fact the kernel of this projection is the connected component of 0 in the ring $\overline{R}$. There is also a splitting $\overline{R} = \overline{R}^0 \oplus \overline{R}^{\mathrm{f}}$ as additive groups (in general $\overline{R}^{\mathrm{f}}$ is not an ideal in $\overline{R}$).

Example 2.9. The pro-algebraic completion of the ring $\mathrm{Mat}_n(\mathbb{Z})$ is isomorphic to

$$\overline{\mathrm{Mat}_n(\mathbb{Z})} \simeq \begin{cases} \mathrm{Mat}_n(\overline{\mathbb{Z}}^{\mathrm{f}}) \oplus \mathrm{Mat}_n(\mathbb{K}) & \text{if } \operatorname{char}\mathbb{K} = 0 \\ \mathrm{Mat}_n(\overline{\mathbb{Z}}^{\mathrm{f}}) & \text{if } \operatorname{char}\mathbb{K} > 0, \end{cases}$$

where $\overline{\mathbb{Z}}^{\mathrm{f}}$ is the pro-finite completion of the ring $\mathbb{Z}$.

Example 2.10. The pro-algebraic completion of $\mathbb{Z}[x_1, \ldots, x_k]$ is a very large ring: its connected component is product of the formal completions of the stalks $\mathcal{O}_{\hat{x}}$ over the points in the affine space $\mathbb{A}^k$ and its semi-simple part is just the product of copies of the field $\mathbb{K}$ indexed by the points of $\mathbb{A}^k$, provided that the characteristic of $\mathbb{K}$ is 0.

2.3. The Steinberg group and pro-algebraic completions. By $\mathrm{Mat}_n(R)$ we denote the ring of $n \times n$ matrices over a ring R. The elementary matrix in $\mathrm{Mat}_n(R)$ with entry $r \in R$ in position (i, j) and 0 elsewhere is $e_{i,j}(r)$, for simplicity we will write $e_{i,j} := e_{i,j}(1)$. If $i \neq j$ then the matrix $E_{i,j}(r) := \mathrm{Id} + e_{i,j}(r)$ is invertible.

Definition 2.11. The group $\mathrm{EL}_n(R)$ is defined as the subgroup of the multiplicative group $\mathrm{Mat}_n(R)^*$ generated by elementary matrices $E_{i,j}(r)$. Similarly the Steinberg group $\mathrm{St}_n(R)$ is the group generated by elements $x_{i,j}(r)$ subject to the relations:

$$\begin{aligned} x_{i,j}(r_1)x_{i,j}(r_2) &= x_{i,j}(r_1 + r_2), \\ [x_{i,j}(r_1), x_{j,k}(r_2)] &= x_{i,k}(r_1 r_2), \\ [x_{i,j}(r_1), x_{p,q}(r_2)] &= 1. \end{aligned}$$

There is a surjective homomorphism $\mathrm{St}_n(R) \to \mathrm{EL}_n(R)$ sending $x_{i,j}(r) \to E_{i,j}(r)$, because $E_{ij}(r)$ satisfy the above relations. The kernel of this homomorphism is denoted by $K_{2,n}(R)$ and often is a central subgroup of $\mathrm{St}_n(R)$ (see [9, 16]).

It is well-known and easy to show that if the ring R is finitely generated and $n \geq 3$ then the groups $\mathrm{St}_n(R)$ and $\mathrm{EL}_n(R)$ are finitely generated. Similarly, if R is finitely presented and $n \geq 4$ then the group $\mathrm{St}_n(R)$ is finitely presented, however the proof is significantly more complicated [12].

The connection between the pro-algebraic completions of these groups and the pro-algebraic completions of the ring R is given by the following result which is essentially proven in [11] (see also [17]):

Theorem 2.12. *The pro-algebraic completions of the groups* $\mathrm{St}_n(R)$ *and* $\mathrm{EL}_n(R)$ *are quotients of* $\mathrm{St}_n(\overline{R})$ *and surject onto* $\mathrm{EL}_n(\overline{R})$*, where* $\overline{R}$ *is the pro-algebraic completion of the ring* R*.*

Sketch of the proof. Let $\pi : \mathrm{St}_n(R) \to \mathrm{GL}_N(\mathbb{K})$ be a finite dimensional representation of the group $\mathrm{St}_n(R)$. The Zariski closure $\mathcal{A} = \overline{\{\pi(x_{1,2}(r)) \,|\, r \in R\}}^{\mathrm{Zar}}$ has a structure of an algebraic ring – the addition is just the multiplication in the group $\mathrm{GL}_N(\mathbb{K})$ and the multiplication can be defined using the relation

$$[x_{1,2}(r_1), x_{2,3}(r_2)] = x_{1,3}(r_1 r_2),$$

by expressing $x_{i,j}(r)$ as conjugate of $x_{1,2}(r)$. There is an induced ring homomorphism $\bar{\pi} : R \to \mathcal{A}$ with Zariski dense image. This implies that the representation π factors through the Steinberg group $\mathrm{St}_n(\mathcal{A})$ and the representation $\mathrm{St}_n(\mathcal{A}) \to \mathrm{GL}_N(\mathbb{K})$ restricted to the root subgroups $X_{i,j}(*)$ is an injective rational map.

The existence of surjections from $\overline{\mathrm{EL}_n(R)}$ and $\overline{\mathrm{St}_n(R)}$ to $\mathrm{EL}_n(\overline{R})$ relies on the observation that for any algebraic ring $\mathcal{A}$ the group $\mathrm{EL}_n(\mathcal{A})$ is linear. □

Remark 2.13. The kernel of the map $\mathrm{St}_n(\overline{R}) \to \overline{\mathrm{EL}_n(R)}$ is related to the pro-algebraic completion of the quotient $K_{2,n}(\overline{R})$ by the image of $K_{2,n}(R)$ induced from the map $R \to \overline{R}$. In many cases of rings having R sufficiently many units it can be shown that $\overline{\mathrm{EL}_n(R)}$ is the same as $\mathrm{EL}_n(\overline{R})$.

Theorem 2.12 almost immediately implies that

Example 2.14. The pro-algebraic completion of the group $\mathrm{SL}_n(\mathbb{Z})$ for $n \geq 3$ is isomorphic to

$$\overline{\mathrm{SL}_n(\mathbb{Z})} \simeq \begin{cases} \mathrm{SL}_n(\overline{\mathbb{Z}}^{\mathrm{f}}) \oplus \mathrm{SL}_n(\mathbb{K}) & \text{if } \operatorname{char} \mathbb{K} = 0 \\ \mathrm{SL}_n(\overline{\mathbb{Z}}^{\mathrm{f}}) & \text{if } \operatorname{char} \mathbb{K} > 0, \end{cases}$$

where $\overline{\mathbb{Z}}^{\mathrm{f}}$ is the pro-finite completion of the ring $\mathbb{Z}$. This result can be deduced (and is equivalent to) the congruence subgroup property [3] and Margulis' super rigidity [15].

2.4. Growth of character varieties for $\mathbf{SL}_n(\mathbb{Z}[x_1, \dots, x_k])$. The groups Γ_f of Theorem 1.1 have pro-algebriac completions, which are infinite products of groups like $\mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_k])$ and the calculation of the growth of the character variety for Γ_f uses the dimensions of the character varieties for the groups $\mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_k])$. The following lemma is relatively easy to prove, however obtaining further details, for example the number of components of the character variety is significantly more complicated and only partial answers are known [1, 13, 18].

Lemma 2.15. *The degree* d *character variety* $X_G(d)$ *for the group* $G = \mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_k])$ *is a point for* $d < n$ *and has dimension* $k\lfloor d/n \rfloor$ *for* $d \geq n$,[6] *provided that* $n \geq 3$*.*

[6] Here $\lfloor \cdot \rfloor$ is the floor function – $\lfloor \alpha \rfloor$ denotes the largest integer less or equal to the real number α.

Proof. Let π be a representation of $\mathrm{SL}_n(\mathbb{Z}[x_1,\dots,x_k])$ in $\mathrm{GL}_d(\mathbb{K})$ corresponding to a point on the character variety and let $i : \mathbb{Z}[x_1,\dots,x_k] \to \mathcal{A} = \overline{\{\pi(x_{1,2}(r)) \,|\, r \in R\}}^{\mathrm{Zar}}$ be the corresponding map into an algebraic ring constructed in Theorem 2.12. The representation π lifts to a representation of the group $\mathrm{St}_n(\mathcal{A})$ which is injective on the root subgroups.

Our first step is to limit the possibilities for the ring $\mathcal{A}$. First we limit the index of the connected component of algebraic ring using that $\mathcal{A}$ contains a finite subring C of size $|\mathcal{A}/\mathcal{A}^0|$ and that $\mathrm{St}_n(C)$ does not have any representation of degree d where the root subgroups acts faithfully if $|C|$ is large enough.

The representation π is semi-simple (since it comes from a point on the character variety) and therefore the ring $\mathcal{A}^0$ is a semi-simple connected commutative algebraic ring. By Theorem 2.6 it is isomorphic to $\mathbb{K}^\alpha$ where $\alpha = \dim \mathcal{A}$. Thus, π yields a representation of the Steinberg group $\mathrm{St}_n(\mathbb{K}^\alpha)$ which is injective on the root subgroups, which is only possible if $\alpha n \leq d$, i.e., $\alpha \leq \lfloor d/n \rfloor$.

Therefore, there are only finitely many possibilities for the ring $\mathcal{A}$ and for each ring there are finitely many representations of $\mathrm{St}_n(\mathcal{A})$ into $\mathrm{GL}_d(\mathbb{K})$, since the group $\mathrm{St}_n(\mathbb{K})$ is rigid.[7] Thus, up to finite index the points on the character variety are parameterized by ring homomorphism from $\mathbb{Z}[x_1,\dots,x_k]$ to $\mathcal{A}$, i.e.,

$$\begin{aligned}\varkappa_{\mathrm{SL}_n(\mathbb{Z}[x_1,\dots,x_k])}(d) &= \dim X_d(\mathrm{SL}_n(\mathbb{Z}[x_1,\dots,x_k]) \leq \\ &= \max_{\mathcal{A}}\{\dim \mathrm{Hom}(\mathbb{Z}[x_1,\dots,x_k],\mathcal{A}\} = \max_{\mathcal{A}}\{k \dim \mathcal{A}\} \leq k\lfloor d/n \rfloor.\end{aligned}$$

The lower bound for the dimension is even easier to explain: the space of ring homomorphisms $\mathbb{Z}[x_1,\dots,x_k] \to \mathbb{K}$ is k dimensional, which implies that

$$\dim X_n(\mathrm{SL}_n(\mathbb{Z}[x_1,\dots,x_k]) \geq k.$$

This together with the observation that $\dim X_{m+n}(\Gamma) \geq \dim X_m(\Gamma) + \dim X_n(\Gamma)$ implies $\dim X_d(\mathrm{SL}_n(\mathbb{Z}[x_1,\dots,x_k]) \geq k\lfloor d/n \rfloor$. □

3. Main construction

3.1. Frame subgroups. The main idea of the proof of Theorem 1.2 is to construct finitely generated subgroups inside infinite products of linear groups which are dense and have the additional property that all their linear representations extend to the infinite product. These are subgroups inside infinite products such that the induced map between the pro-algebraic completions is an isomorphism. However this property is not enough for one of the main technical tools in Lemma 3.2, so we need an additional condition to the definition of a frame subgroup. This definition is a slight modification of the one in [10].

Let $\{S_n\}_{n=1}^\infty$ be an infinite family of discrete linear groups, and Γ be a dense subgroup of the infinite product $\mathfrak{S} = \prod_{n=1}^{\infty} S_n$. The inclusion $\Gamma \hookrightarrow \mathfrak{S}$ induces map between the pro-algebraic completions $\iota : \overline{\Gamma} \to \overline{\mathfrak{S}}$. The pro-algebraic completion of $\mathfrak{S}$ is equal to $\overline{\mathfrak{S}} = \prod \overline{S_n}$, since we only consider continuous representations. The density of Γ implies that the map ι is

[7] Here we are only considering representations of $\mathrm{St}_n(\mathbb{K})$ restricted to the root subgroups are given by rational functions.

surjective, however often this map is not injective. The map ι is injective if and only if every finite dimensional representation π of Γ factors through some finite product $\prod_{n=1}^{N} S_n$.

The following notion of frame subgroup is slightly stronger than the triviality of the congruence kernel

Definition 3.1. A finitely generated subgroup $\Gamma < \mathfrak{S}$ is *a frame* for $\mathfrak{S}$ if the following hold:

(a) Γ contains $\bigoplus_{n=1}^{\infty} S_n$;

(b) The natural surjection $\iota : \overline{\Gamma} \to \prod \overline{S_n}$ is an isomorphism.

One can think of condition (a) as saying that Γ is a good approximation of $\mathfrak{S}$ from 'within' while condition (b) says that G approximates very well $\mathfrak{S}$ from 'above'.

More precisely, for a finite subset $V \subset \mathbb{N}$ of integers define the *V-principal congruence subgroup* Γ_V to be the kernel of the projection of Γ onto $\prod_{n\in V} S_n$. Let $\Gamma(V)$ be the projection of Γ onto $\mathfrak{S}(V) := \prod_{n\notin V} S_n$. The m-th principal congruence subgroup Γ_m is just $\Gamma_{\{1,\dots,m\}}$ and $\Gamma(m) := \Gamma(\{1,\dots,m\})$.

Part (a) of the above definition is now equivalent to

$$\Gamma = \left(\prod_{n\in V} S_n\right) \times \Gamma_V, \quad \Gamma_V = \Gamma \cap \mathfrak{S}(V).$$

Therefore the congruence subgroup Γ_V can be identified with the projection $\Gamma(V)$.

On the other hand as mentioned above, part (b) of Definition 3.1 says that any finite dimensional representation π of Γ factors through a finite product $\prod_{n=1}^{N} S_n$.

We stress that the existence of even a single example of a frame is far from obvious at this stage — it is relatively easy to construct infinitely generated frame subgroups, e.g., $\bigoplus S_i$. The following Lemma allows us to find many frame subgroups provided we already know at least one:

Lemma 3.2. *Let $A_n, B_n < C_n$, ($n \in \mathbb{N}$) be linear groups with $C_n = \langle A_n, B_n\rangle$. Suppose that X (resp. Y) is a frame subgroup of the product $\mathfrak{A} = \prod_{n=1}^{\infty} A_n$ (resp. $\mathfrak{B} = \prod_{n=1}^{\infty} B_n$). Each of X and Y can be considered as a subgroup of $\mathfrak{C} := \prod_{n=1}^{\infty} C_n$ in the natural way. Then the group*

$$Z = \langle X, Y\rangle < \mathfrak{C}$$

is a frame in $\mathfrak{C}$.

Proof. It is clear that Z contains the direct product of C_n. Suppose that N is a kernel of some finite dimensional representation of the group Z. By hypothesis N contains the m-th principal congruence subgroups X_m, Y_m for some m. Identifying them with $X(m), Y(m)$ in $\mathfrak{C}(m)$ we see that N contains $\langle X(m), Y(m)\rangle = Z(m)$, which under our identification is Z_m. □

Remark 3.3. The above lemma does not hold under the weaker assumption that the maps $\iota_X : \overline{X} \to \prod \overline{A_n}$ and $\iota_Y : \overline{Y} \to \prod \overline{B_n}$ are isomorphisms.

3.2. The ring $\boldsymbol{R}$. Our first step is to construct a finitely generated ring which is almost a "frame" in some infinite product of rings. Recall that $\mathrm{Mat}_n(R)$ denotes the ring of $n \times n$ matrices over an associative ring R and $e_{i,j}(r)$ denotes the elementary matrix in $\mathrm{Mat}_n(R)$ with entry $r \in R$ in position (i,j). The ring $\boldsymbol{R}$ described below is similar to the ring used in [10].

Let $\boldsymbol{R}$ be the subring of $\prod_{n=5}^{\infty} \mathrm{Mat}_n(\mathbb{Z})$ generated by the five elements $1 = (\mathrm{Id}_n)$, $\mathbf{a} = (a_n)$, $\mathbf{a}^{-1}$, $\mathbf{b} = (b_n)$ and $\mathbf{c} = (c_n)$, defined as follows:

$$a_n = e_{1,2} + e_{2,3} + \ldots + e_{n,1}, \quad b_n = e_{1,2}, \quad c_n = e_{2,1}.$$

Note that $\boldsymbol{R}$ contains the elements

$$\mathbf{b}_k := \mathbf{b}\mathbf{a}^{k-1} = (e_{1,k+1})_n, \quad \mathbf{c}_k := \mathbf{a}^{1-k}\mathbf{c} = (e_{k+1,1})_n.$$

It is easy to see that the elements $\mathbf{b}_i, \mathbf{c}_i$ for $i = 1, 2$ generate a subring (without unit) of $\boldsymbol{R}$ isomorphic to $\mathrm{Mat}_3(\mathbb{Z})$ sitting diagonally in the top left corner of each factor.

The main technical result is that the ring $\boldsymbol{R}$ is almost a frame:

Theorem 3.4.

(i) *The ring $\boldsymbol{R}$ contains the direct sum* $\bigoplus_{n=5}^{\infty} \mathrm{Mat}_n(\mathbb{Z})$.

(ii) *The pro-algebraic completion $\overline{\boldsymbol{R}}$ of $\boldsymbol{R}$ is*

$$\overline{\boldsymbol{R}} = \overline{\mathbb{Z}[t, t^{-1}]} \bigoplus \prod_{n=5}^{\infty} \overline{\mathrm{Mat}_n(\mathbb{Z})}.$$

Proof. Let us observe that the commutator $[b_n, a_n^{-k} b_n a_n^k]$ is non-zero iff n divides $k-1$ or $k+1$, moreover if this commutator is non-zero then it is a matrix unit of the form $e_{i,j}$ and the ideal generated by it is the whole $\mathrm{Mat}_n(\mathbb{Z})$. This, together with the observation that $\mathrm{Mat}_n(\mathbb{Z})$ is generated by a_n, b_n and c_n implies that the ring generated by $\mathbf{a}$, $\mathbf{a}^{-1}$ and $\mathbf{b}$ contains the infinite direct sum.

Let $\pi : R \to \mathcal{A}$ be a map from R into an algebraic ring $\mathcal{A}$ with kernel J. By Theorem 2.6 we have that there exists an ideal of finite index $I \lhd R$ such that I/J can be embedded into a finite dimensional algebra A. Thus, there exists $N \geq 5$ such that $\mathbf{a}^N - 1 \in I$ and $\dim A \leq N$. Let $\tilde{\boldsymbol{R}} = \boldsymbol{R}/\tilde{J}$, where

$$\tilde{J} = J + \bigoplus_{n=5}^{2N} M_n(\mathbb{Z}) \quad \text{and} \quad \tilde{I} = I + \bigoplus_{n=5}^{2N} \mathrm{Mat}_n(\mathbb{Z}).$$

Then $\mathbf{b}_2 = \mathbf{b}\mathbf{a}^{-1}\mathbf{b}\mathbf{a} \in \tilde{I}$ because

$$\mathbf{b}\mathbf{a}^{-N-1}\mathbf{b}\mathbf{a} \in \bigoplus_{n=5}^{2N} \mathrm{Mat}_n(\mathbb{Z})$$

and

$$\mathbf{b}\mathbf{a}^{-1}\mathbf{b}\mathbf{a} - \mathbf{b}\mathbf{a}^{-N-1}\mathbf{b}\mathbf{a} \in \langle \mathbf{a}^N - 1 \rangle \subset I.$$

Similar argument gives also that $\mathbf{c}_2 \in \tilde{I}$. The relations

$$\mathbf{b} = \mathbf{b}_2\mathbf{a}^{-1}, \quad \mathbf{c} = \mathbf{a}\mathbf{c}_2$$

imply that $\mathbf{b}, \mathbf{c} \in \tilde{I}$ and $\boldsymbol{R}/\tilde{I}$ is generated by $\mathbf{a}$ and $\mathbf{a}^{-1}$.

The second part of Theorem 2.6 gives that the quotient I/J can be embedded into a finite dimensional algebra A of dimension less than N. It follows that I/J and also $\tilde{I}/\tilde{J}$ satisfy the identity

$$s_N(x_1, \dots x_N) := \sum_{\sigma \in \mathrm{Sym}(N)} (-1)^\sigma x_{\sigma(1)} x_{\sigma(2)} \dots x_{\sigma(N)} = 0.$$

Thus

$$\mathbf{p}_N := s_N(\mathbf{b}, \mathbf{a}^{-1}\mathbf{b}\mathbf{a}, \mathbf{a}^{-2}\mathbf{b}\mathbf{a}^2, \dots) \in \tilde{J}.$$

For each $n > N + 2$ one has

$$s_N(b_n, a_n^{-1} b_n a_n, a_n^{-2} b_n a_n^2, \dots) = e_{1,N+1}$$

thus $\mathbf{p}_N - \mathbf{b}_N \in \bigoplus_{n=5}^{N+2} M_n(\mathbb{Z})$, i.e., $\mathbf{b}_N \in \tilde{J}$. The equality $\mathbf{b} = \mathbf{b}_N \mathbf{a}^{-1} \mathbf{c}_{N-1} \mathbf{a}$ implies that the element $\mathbf{b}$ lies in $\tilde{J}$. A similar computation also shows that $\mathbf{c} \in \tilde{J}$.

This argument implies that $\boldsymbol{R}/\tilde{J}$ is generated by $\mathbf{a}$ and $\mathbf{a}^{-1}$, i.e., is an image of $\mathbb{Z}[t, t^{-1}]$. This together with the inclusion $\bigoplus \mathrm{Mat}_n(\mathbb{Z}) \subset \boldsymbol{R}$ gives that $\boldsymbol{R}/J$ is a quotient of

$$\mathbb{Z}[t, t^{-1}] \oplus \bigoplus_{n=5}^{2N} \mathrm{Mat}_n(\mathbb{Z}).$$

The second part of the proposition follows because $\overline{\boldsymbol{R}}$ is an inverse limit of the quotients of the form $R/\ker\pi$ and

$$\overline{\boldsymbol{R}} = \varprojlim_N \overline{\mathbb{Z}[t, t^{-1}] \oplus \bigoplus_{n=5}^{N} \mathrm{Mat}_n(\mathbb{Z})} = \overline{\mathbb{Z}[t, t^{-1}]} \oplus \prod_{n=5}^{\infty} \overline{\mathrm{Mat}_n(\mathbb{Z})}.$$

□

This result gives that the "congruence" kernel $U = \ker\{\iota : \overline{\boldsymbol{R}} \to \prod \overline{\mathrm{Mat}_n(\mathbb{Z})}\}$ is isomorphic to $\overline{\mathbb{Z}[t, t^{-1}]}$ and it is not only a subgroup of the pro-algebraic completion $\overline{\boldsymbol{R}}$, but it is in fact a direct summand. This fact is crucial for the proof of Theorem 3.9.

3.3. The first approximation of a frame subgroup. The first approximation of a frame subgroup will be the group Γ_0 generated by the elementary 4×4 matrices over the ring $\boldsymbol{R}$ – since the ring $\boldsymbol{R}$ is sitting in the direct product of rings and contains their direct sum, the group $\mathrm{EL}_4(\boldsymbol{R})$ has similar properties. Unfortunately $\mathrm{EL}_4(\boldsymbol{R})$ is not a frame subgroup since it has algebraic images which can not be extended to the infinite product, because the ring $\boldsymbol{R}$ has algebraic images coming from the $\overline{\mathbb{Z}[t, t^{-1}]}$ factor in the pro-algebraic completion.

Lemma 3.5. *The group Γ_0 is generated by the following set of sixteen elements:*

$$\{e_{i,j} \mid 1 \le i \ne j \le 4\} \bigcup \{e_{1,2}(x) \mid x \in \{\mathbf{a}, \mathbf{a}^{-1}, \mathbf{b}, \mathbf{c}\}\}.$$

In fact it can be shown that Γ_0 can be generated by only 3 elements.

Since $\boldsymbol{R}$ is a subring of $\prod_{n=5}^{\infty} \mathrm{Mat}_n(\mathbb{Z})$ we can consider Γ_0 as a subgroup of

$$\prod_{n=5}^{\infty} \mathrm{EL}_4(\mathrm{Mat}_n(\mathbb{Z})) = \prod_{n=5}^{\infty} \mathrm{SL}_{4n}(\mathbb{Z}).$$

Lemma 3.6. *The group Γ_0 contains $\bigoplus_{n=5}^{\infty} \mathrm{SL}_{4n}(\mathbb{Z})$ and the pro-algebraic completion of Γ_0 is*

$$\overline{\mathrm{EL}_4(\mathbb{Z}[t,t^{-1}])} \oplus \prod_{n=5}^{\infty} \overline{\mathrm{SL}_{4n}(\mathbb{Z})}.$$

Here $U = \overline{\mathrm{EL}_4(\mathbb{Z}[t,t^{-1}])}$ is the "congruence" kernel $U = \ker \iota$. More precisely, suppose $\pi : \Gamma_0 \to \mathrm{GL}_M(\mathbb{K})$ a finite dimensional representation of the group Γ_0 with kernel $H = \ker \pi$

(i) *There exists an N such that we have the following diagram:*

$$\begin{array}{ccc} & \Gamma_0 & \\ \swarrow & & \searrow \\ \mathrm{EL}_4(\mathbb{Z}[t,t^{-1}]) & & \Gamma_0/H \\ \searrow & & \swarrow \\ & \Gamma_0/\widetilde{H} = \widetilde{\Gamma}_0 & \end{array}$$

where $\widetilde{H} = H \cdot \bigoplus_{n=5}^{N} \mathrm{SL}_{4n}(\mathbb{Z})$. The map $\Gamma_0 \to \mathrm{EL}_4(\mathbb{Z}[t,t^{-1}])$ comes from the projection $R \to \mathbb{Z}[t,t^{-1}]$, sending $\mathbf{b}$ and $\mathbf{c}$ to 0 and $\mathbf{a}$ to t;

(ii) *Let $\tilde{\pi}$ denote the map $\Gamma_0 \to \widetilde{\Gamma}_0$. Then for each pair of indices we have $\tilde{\pi}(E_{i,j}(\mathbf{b})) = \tilde{\pi}(E_{i,j}(\mathbf{c})) = 1$.*

Proof. Theorem 2.12 implies that $\overline{\Gamma_0}$ is a quotient of $\mathrm{St}_4(\overline{\boldsymbol{R}})$ therefore it is a quotient of

$$\mathrm{St}_4(\overline{\mathbb{Z}[t,t^{-1}]}) \times \prod \mathrm{St}_4(\overline{\mathrm{Mat}_n(\mathbb{Z})}).$$

One can show that K_2 of these pro-algebraic rings is trivial (although this is not essential) which allow one to replace all St_4 with EL_4. Parts (i) – (iii) easily follow using the description of the congruence kernel of the ring $\boldsymbol{R}$. □

3.4. The frame sugroup. Lemma 3.6 states that Γ_0 is close to being a frame subgroup, but it is not quite one. The way to fix that and remove this "congruence" kernel U is similar to the method in [10]. We can take two copies of Γ_0 and use elements from one copy to kill the congruence kernel of the other.

The group Γ_0 contains two interesting subgroups: Let D_1 be the subgroup generated by $E_{1,4}(\mathbf{b}_i), E_{4,1}(\mathbf{c}_i)$ for $1 \le i \le 4$. It is easy to see that D_1 is isomorphic to $\mathrm{SL}_5(\mathbb{Z})$ embedded "diagonally" in $\prod \mathrm{SL}_{4n}(\mathbb{Z})$ where each copy of $\mathrm{SL}_5(\mathbb{Z})$ is embedded in $\mathrm{SL}_{4n}(\mathbb{Z})$ at positions $1, 3n+2, 3n+3, 3n+4, 3n+5$.

The subgroup D_1 contains an element q which corresponds to the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

As an element in $\prod \mathrm{SL}_{4n}(\mathbb{Z})$ this element is equal to $q = (q_{4n})$, where q_{4n} is a permutation matrix corresponding to the involution $(3n+2, 3n+3)(3n+4, 3n+5)$.

We remark that from Chevalley commutator relations it follows that $E_{1,4}(\mathbf{b}_i)$ and $E_{4,1}(\mathbf{c}_i)$ lie in the normal subgroup generated by $E_{12}(\mathbf{b})$ and $E_{12}(\mathbf{c})$, therefore the image of the whole group D_1 in the "congruence" kernel $\overline{\mathrm{EL}_4(\mathbb{Z}[t, t^{-1}])}$ is trivial.

Similarly, let D_2 be the subgroup generated by $E_{i,j}(\mathbf{1})$ for $1 \leq i \neq j \leq 3$. This group contains the element v corresponding to

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

As an element in $\prod \mathrm{SL}_{4n}(\mathbb{Z})$ this element is equal to $v = (v_{4n})$ where v_{4n} is a permutation matrix corresponding to $(1, n+1, 2n+1)(2, n+2, 2n+2)\dots(n, 2n, 3n)$. An easy computation shows that this element normally generates $\mathrm{SL}_3(\mathbb{Z})$, it even normally generates the whole group Γ_0.

Let θ_n denote the embedding of $\mathrm{SL}_n(\mathbb{Z})$ into $\mathrm{SL}_{2^n}(\mathbb{Z})$ coming from the action of $\mathrm{SL}_n(\mathbb{Z})$ on the exterior algebra $\bigwedge^* \mathbb{Z}^n$. We will need two lemmas about this embedding whose proofs are postponed until section 3.6:

Lemma 3.7.

(i) *For every $n \geq 20$ there exists an element $\varpi_n \in \mathrm{SL}_{2^n}(\mathbb{Z})$ such that $\mathrm{SL}_{2^n}(\mathbb{Z})$ is generated by the image of $\mathrm{SL}_n(\mathbb{Z})$ under θ_n and its conjugate by ϖ_n.*

(ii) *For every $n \geq 10$ there exists an element $\omega_n \in \mathrm{SL}_{2^n}(\mathbb{Z})$ such that*

$$i''_n(q_n)^{-1} i'_n(v_n) i''_n(q_n) = i'_n(v_n)^{-1}, \quad i'_n(q_n)^{-1} i''_n(v_n) i'(q_n) = i''_n(v_n)^{-1}. \quad (3.1)$$

where i'_n is the embedding θ_n and i''_n is obtained from θ_n by conjugation with ω_n. Here q_n and v_n are two commuting elements in $\mathrm{SL}_n(\mathbb{Z})$ similar to the ones described above, i.e., q_n is a permutation matrix corresponding to the product of two transpositions and v_n is a permutation matrix of order 3 corresponding to a permutation with support about $3n/4$.

Remark 3.8. It seems plausible that the elements ϖ_n and ω_n can be chosen to be the same. This will allow us to decrease the number of generators of the group constructed in Theorem 1.2 since the enlargement step in section 3.5 will not be necessary. Also, it is likely that elements with similar properties also exist for the embedding $\mathrm{SL}_n(\mathbb{Z}) \to \mathrm{SL}_{\binom{n}{n/2}}(\mathbb{Z})$ coming from the action of $\mathrm{SL}_n(\mathbb{Z})$ on $\bigwedge^{n/2} \mathbb{Z}^n$.

Let u_n denote the number 2^{4n} and let S_n be the subgroup of $\mathrm{SL}_{u_n}(\mathbb{Z})$ generated by the images of the homomorphism i'_{4n} and i''_{4n} from Lemma 3.7(ii). Consider the two homomorphisms

$$i', i'' : \prod_{n=5}^{\infty} \mathrm{SL}_{4n}(\mathbb{Z}) \to \prod_{n=5}^{\infty} S_n \subset \prod_{n=5}^{\infty} \mathrm{SL}_{u_n}(\mathbb{Z}),$$

given by $(g_n)_n \mapsto (i'_{4n}(g_n))_n$ and $(g_n)_n \mapsto (i''_{4n}(g_n))_n$.

Set $\Gamma'_0 = i'(\Gamma_0), \Gamma''_0 = i''(\Gamma_0)$ and let Γ_1 denote the subgroup of $\prod S_n$ generated by Γ'_0 and Γ''_0.

Theorem 3.9. *The group Γ_1 is a frame subgroup of $\prod S_n$.*

Proof. The group Γ_1 contains $\bigoplus_{n=5}^{\infty} S_n$, because each copy of Γ_0 contains the infinite direct sum of the images of $\mathrm{SL}_{4n}(\mathbb{Z})$ in S_n.

Suppose now that $\pi : \Gamma_1 \to \mathrm{GL}_N(\mathbb{K})$ is a finite dimensional representation of Γ_1 with kernel $H = \ker \pi$. By Proposition 3.6 (i) applied to Γ'_0 and Γ''_0, there exists an integer M such that if

$$\widetilde{H} = H.\bigoplus_{n=5}^{M} S_n \quad \widetilde{H'} = \widetilde{H} \cap \Gamma'_0, \quad \widetilde{H''} = \widetilde{H} \cap \Gamma''_0.$$

then $\widetilde{\Gamma'_0} = \Gamma'_0/\widetilde{H'}$ and $\widetilde{\Gamma''_0} = \Gamma''_0/\widetilde{H''}$ are images of $\mathrm{EL}_4(\mathbb{Z}[t, t^{-1}])$.

By Proposition 3.6 (ii) the images of $E_{i,j}(\mathbf{b})$ and $E_{i,j}(\mathbf{c})$ in $\widetilde{\Gamma'_0}$ and $\widetilde{\Gamma''_0}$ are trivial. In particular the group D_1 is in the kernel of both $\rho' = \rho \circ i'$ and $\rho'' = \rho \circ i''$, where ρ is the projection $\Gamma_1 \to \widetilde{\Gamma}$. This implies that $\rho'(q) = \rho''(q) = 1$.

By (3.1) the following relations hold in Γ_1:

$$i'(q)i''(v)i'(q) = i''(v)^{-1} \qquad i''(q)i'(v)i''(q) = i'(v)^{-1},$$

since these relations hold in every S_n. Therefore we have

$$\rho'(q)\rho''(v)\rho'(q) = \rho''(v)^{-1} \qquad \rho''(q)^{-1}\rho'(v)\rho''(q) = \rho'(v)^{-1}.$$

These, together with $\rho'(q) = \rho''(q) = 1$ and $v^3 = 1$ imply that $\rho'(v) = \rho''(v) = 1$. The elements $i'(v)$ and $i''(v)$ generate Γ'_0 and Γ''_0 as normal subgroups, which implies that ρ' and ρ'' have trivial images, i.e., $\widetilde{\Gamma_1} = \{1\}$. Therefore $\Gamma_1 = H.\bigoplus_{n=5}^{M} S_n$, hence H contains the congruence subgroup $(\Gamma_1)_M$ and Γ_1 is a frame subgroup. □

3.5. Proof of Theorem 1.2. The group Γ_1 is a frame in some 'unknown' infinite product. As mentioned before it might be possible to choose the elements ω_{4n} such that the embeddings i'_n and i''_n generate the full groups $\mathrm{SL}_{u_n}(\mathbb{Z})$, which will implies that Γ_1 is a frame in $\mathrm{SL}_{u_n}(\mathbb{Z})$. One way to bypass this problem is the following:

Lemma 3.10. *Let ϖ denote the element in $\prod \mathrm{SL}_{u_n}(\mathbb{Z})$ obtained by putting together all ϖ_{4n}. Then the group Γ_2 generated by Γ_1 and Γ_1^{ϖ} is a frame subgroup in $\prod \mathrm{SL}_{u_n}(\mathbb{Z})$.*

Proof. Part (i) of Lemma 3.7 implies that Γ_1 and Γ_1^{ϖ} satisfy the conditions of Lemma 3.2, which gives that Γ_2 is a frame subgroup. □

Before proving Theorem 1.2 we need another technical claim

Claim 3.11. *Let $\{n_i\}_{i=1}^K$ be a finite sequence of integers such that $n_i > m \geq 5$. Then the group $\prod \mathrm{SL}_{n_i}(\mathbb{Z})$ can be generated by two subgroups isomorphic to $\mathrm{SL}_m(\mathbb{Z})$.*

Proof. This relies on the well known fact that for each $n_i > m$ the alternating group Alt_{n_i} can be generated by two copies of Alt_m acting diagonally. This diagonal embedding can be extended to embedding of $\mathrm{SL}_m(\mathbb{Z})$ into $\mathrm{SL}_{n_i}(\mathbb{Z})$ which can be easily seen to generate the whole group. Combining these embeddings leads to two subgroups in $\prod \mathrm{SL}_{n_i}(\mathbb{Z})$ isomorphic to $\mathrm{SL}_m(\mathbb{Z})$ which generate the whole Cartesian product. □

Proof of Theorem 1.2. The product $\prod_{k\geq 3} \mathrm{SL}_n(\mathbb{Z})$ can be written as

$$\prod_{k\geq 3} \mathrm{SL}_n(\mathbb{Z}) = \prod_{k\geq 3}^{u_5-1} \mathrm{SL}_n(\mathbb{Z}) \times \prod_{n\geq 5} \left(\prod_{k=u_n}^{u_{n+1}-1} \mathrm{SL}_k(\mathbb{Z}) \right).$$

Let T_0 denote the product $\prod_{k\geq 3}^{u_5-1} \mathrm{SL}_n(\mathbb{Z})$ and T_n be $\prod_{k=u_n}^{u_{n+1}-1} \mathrm{SL}_k(\mathbb{Z})$. Using Claim 3.11 each of the groups T_n for $n \geq 5$ can be generated by two subgroups isomorphic to $\mathrm{SL}_{u_n}(\mathbb{Z})$, this together with Lemma 3.2 proves the existence of a frame subgroup Γ_3 in $\prod_{n\geq 5} T_n$. Since T_0 is finitely generated then $T_0 \times \Gamma_3$ is a frame subgroup in $\prod \mathrm{SL}_n(\mathbb{Z})$. □

3.6. Proof of Lemma 3.7. Consider the action of $\mathrm{SL}_n(\mathbb{Z})$ on $V_n := \bigwedge^* \mathbb{Z}^n$ which induces an embedding $\theta_n : \mathrm{SL}_n(\mathbb{Z}) \hookrightarrow \mathrm{SL}_{2^n}(\mathbb{Z})$. A key point is that θ_n sends any signed permutation matrix λ in $\mathrm{SL}_n(\mathbb{Z})$ to a signed permutation matrix in SL_{2^n} with large support, even if λ has tiny support. Let υ_n denote the embedding of $\mathrm{Sym}^{\pm}(n)$ into $\mathrm{Sym}^{\pm}(2^n)$ induced by θ_n where $\mathrm{Sym}^{\pm}(n)$ denotes the group of even signed permutations.

The image H_n of the embedding υ_n is a signed permutation group which acts on the set P of 2^n basis vectors of V_n with only $n+1$ orbits $P_{n,i}$ for $0 \leq i \leq n$ corresponding to the decomposition $V_n = \bigoplus_i \bigwedge^i \mathbb{Z}^n$. The group H contains a permutation matrix c of order 3 with very large support. Using the fact that with probability tending to 1 any two permutations of order 3 with large support generate the whole alternating group $\mathrm{Alt}(n)$, we can find a element ϖ_n such that the group generated by H and H^{ϖ_n} contains the alternating group. Part (i) of Lemma 3.7 follows immediately from the following

Claim 3.12. *The group $\mathrm{SL}_N(\mathbb{Z})$ is generated by an unipotent matrix U in Jordan normal form which has at least one Jordan block of size 1 and another one of size 2 and the alternating group $\mathrm{Alt}(N)$ acting by permutation matrices.*

Proof. The structure of U implies that there exists a 3-cycle $g \in \mathrm{Alt}(N)$ such that $[U, U^g]$ is an elementary matrix E. The orbit of E under conjugation by $\mathrm{Alt}(N)$ consists of all elementary matrices, which generate the group $\mathrm{SL}_N(\mathbb{Z})$. □

In order to prove the second part of Lemma 3.7 we need to understand the action of $\upsilon(q_n)$ and $\upsilon(\upsilon_n)$: We can split $\mathbb{Z}^n$ as a direct sum of $\mathbb{Z}^{n-4}$ and $\mathbb{Z}^4$ with υ_n acting on the first component and q_n acting on the second. This leads to decomposition of $\bigwedge^* \mathbb{Z}^n$ as tensor product

$$\bigwedge^* \mathbb{Z}^n = \left(\bigwedge^* \mathbb{Z}^{n-4} \right) \otimes \left(\bigwedge^* \mathbb{Z}^4 \right).$$

The second factor is 16 dimensional and $\upsilon(q_n)$ acts on it as a signed permutation matrix with 6 two cycles, 4 fixed points, moreover 2 of the points have positive signs and 2 have negative signs. This implies the following lemma

Lemma 3.13. *The image of the group G generated by $\upsilon(q_n)$ and $\upsilon(v_n)$ acts on the set P_n with $3M$ orbits of size 6, $2M$ orbits of size 3 fixed by q_n - on M of these orbits q_n fixes the corresponding basis vectors and on the other M orbits q_n negates the basis vectors; there are several additional orbits where v_n acts trivially, for some integer M.*

This lemma allow us to express the set P as union of G invariant sets of the following types

(a) an orbit of size 6 and an orbit of size 3 where both $\upsilon(q_n)$ and $\upsilon(v_n)$ are permutations;

(b) an orbit of size 6 and an orbit of size 3 where $\upsilon(q_n)$ is apermutation and $\upsilon(v_n)$ is a permutation times a central element;

(c) 4 orbits of size 6 and an orbit of size 3 where both $\upsilon(q_n)$ and $\upsilon(v_n)$ are permutations;

(d) an orbit of size 3 and an orbit of size 2 where both $\upsilon(q_n)$ and $\upsilon(v_n)$ are permutations;

(e) set where $\upsilon(q_n)$ acts trivially.

In order to show the existence of the required element ω_n it is enough to find a signed permutation matrix ς_n which satisfies the following relations

$$\upsilon(q_n)^{\varsigma}_n\upsilon(v_n)\upsilon(q_n)^{\varsigma}_n = \upsilon(v_n)^{-1} \quad \text{and} \quad \upsilon(q_n)\upsilon(v_n)^{\varsigma}_n\upsilon(q_n) = (\upsilon(v_n)^{\varsigma}_n)^{-1}. \tag{3.2}$$

The partition of the set P outlined above allows us to construct the element ς_n on each of the small pieces, which reduces the computation to finding suitable elements in signed permutation groups of relatively small sizes. Here is a brief construction of the elements ς_n in each of the cases:

(a) We can arrange the 9 points in a square of size 3, i.e., label the points (x, y) with $x, y \in \mathbb{F}_3$, it is possible to choose the labels such that $\upsilon(q) : (x, y) \to (x, -y)$ and $\upsilon(v) : (x, y) \to (x+1, y)$ where all computations are taken in the field with 3 elements. In this model the element $\varsigma : (x, y) \to (y, x)$ satisfies relation (3.2).

(b) This is the same as case (a) since the actions of q differ by an element in the center of groups.

(c) This is similar to case (a) however one uses 3 dimensional cube instead of a square: the points are labeled (x, y, z) with $x, y, z \in \mathbb{F}_3$ and $\upsilon(q) : (x, y, z) \to (x, -y, -z)$ and $\upsilon(v) : (x, y, z) \to (x+1, y, z)$. In this model the element $\varsigma : (x, y, z) \to (y, z, x)$ with satisfy relation (3.2).

(d) If $\upsilon(q)$ is the 3-cycle (123) and $\upsilon(v)$ is the transposition (45) we can take ς_n to be the involution $(24)(35)$.

(e) Since $\upsilon(q)$ is trivial, any element ς will satisfy the above relations.

Finally observe that the combination ς_n of the permutations described above on each piece will satisfy the relation (3.2). The element ω_n is just ς_n considered as a signed permutation matrix in $\mathrm{SL}_{u_n}(\mathbb{Z})$.

4. Application – proof of Theorem 1.1

Using Theorem 1.2 it is relatively quick to prove Theorem 1.1: a function $f(n) : \mathbb{N} \to \mathbb{N}$ will be called admissible if $f(n) \leq n(n-1)/2$ and $f(n)/n$ non-decreasing.

The next lemma follows from Theorem 1.2

Lemma 4.1. *For any admissible function f there exists a finitely generated frame subgroup inside*

$$\prod_n \mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_{f(n)}]).$$

Proof. Observe that the group $\mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_k])$ can be generated by two conjugated copies of $\mathrm{SL}_n(\mathbb{Z})$ for $k \leq n(n-1)/2$, one is the standard one and the other one is conjugated by any upper triangular matrix with entries the intermediates x_i. This allow us to apply Lemma 3.2 to the two embeddings of the frame subgroup constructed in the proof of Theorem 1.2. □

Let $I_\Gamma(n)$ denote the open subset of $X_\Gamma(n)$ which corresponds to the irreducible representations of Γ.

The dimension of $X_\Gamma(n)$ can be computed: a semi-simple representation can be written as direct sum of simple ones, i.e.,

$$\dim X_\Gamma(n) = \max\left\{\sum \dim I_\Gamma(n_i) \mid \sum n_i = n\right\}.$$

This formula reduces the computation of the dimensions of the character varieties of Γ to the dimensions of the irreducible components $\dim I_\Gamma(n)$.

Proof of Theorem 1.1. Let f be an admissible function and let Γ_f be a frame subgroup in $\prod_n \mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_{f(n)}])$. Since the group $\mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_{f(n)}])$ does not have any non-trivial representations of degree less than n, the representation and character varieties for Γ_f of degree d are the same as the ones for the finite product $G_d := \prod_{n \leq d} \mathrm{SL}_n(\mathbb{Z}[x_1, \dots, x_{f(n)}])$.

Using the above observation we need to compute the dimensions

$$\dim I_{G_d}(k), \quad \text{for } k \leq d.$$

However, the irreducible representations of product of groups are tensor products of irreducible representations of each factor therefore

$$\begin{aligned}\dim I_{G_d}(k) &= \max\left\{\sum \dim I_{\mathrm{SL}_i(\mathbb{Z}[x_1,\dots,x_{f(i)}])}(k_i) \mid \prod k_i = k,\ k_i > 2,\ k_i \neq k_j\right\} = \\ &= \max\left\{\sum f(k_i) \mid \prod k_i \leq k\right\},\end{aligned}$$

where we used that $\dim I_{\mathrm{SL}_i(\mathbb{Z}[x_1,\dots,x_{f(i)}]}(k_i)$ is 0 for $k_i < i$ and either $f(i)$ or 0 for $k_i \geq i$. If the function f is admissible the above maximum is equal to $f(k)$. Therefore for every $k \leq d$ we have

$$\dim I_{G_d}(k) = f(k).$$

This allows is to compute the dimensions of the character varieties

$$\begin{aligned}\dim X_{G_d}(k) &= \max\left\{\sum \dim I_{G_d}(k_i) \mid \sum k_i = k\right\} = \\ &= \max\left\{\sum f(k_i) \mid \sum k_i = k\right\} =\end{aligned}$$

$$= \max\left\{\sum k_i \frac{f(k_i)}{k_i} \mid \sum k_i = k\right\} \leq$$
$$\leq k \max\left\{\frac{f(k_i)}{k_i} \mid \sum k_i = k\right\} = f(k),$$

i.e., the dimensions of the character variety of degree k for G_d are bounded above by $f(k)$ for any $k \leq d$. There is a lower bound coming from

$$\dim X_{G_d}(k) \geq \dim X_{\mathrm{SL}_k(\mathbb{Z}[x_1,\ldots,x_{f(k)}])}(k) = f(k).$$

Therefore $\dim X_{\Gamma_f}(d) = f(d)$ for any d which completes the proof of Theorem 1.1. □

Remark 4.2. A modification of this construction allows us to construct a group $\tilde{\Gamma}_f$ with Kazhdan property T almost satisfying the conditions of Theorem 1.1, provided that the admissible function f satisfies the extra condition $\lim f(n)/n^2 \to 0$. The idea is to modify the construction of the ring $\boldsymbol{R}$ to obtain a finitely generated subring $\boldsymbol{R}_f$ inside

$$\prod \mathrm{Mat}_n(\mathbb{Z}[x_1 \ldots x_f(n)]).$$

The generators of the ring $\boldsymbol{R}_f$ are $\mathbf{a}^{\pm 1}$, $\mathbf{b}_f = (b_{n,f})$, $\mathbf{c}_f = (c_{n,f})$ and $\mathbf{y}_f = (y_{n,f})$, where the matrices

$$b_{n,f} = \sum_{i=0}^{n/d_n} e_{1+d_n i, 2+d_n i} \quad c_{n,f} = \sum_{i=0}^{n/d_n} e_{2+d_n i, 1+d_n i}$$

are modifications of b_n and c_n of rank about $f(n)/n$ and the new generators $y_{n,f}$ contain the variables x_i and

$$y_{n,f} = \sum_{i=0}^{n/d_n} \sum_{j=1}^{n} e_{1+d_n i, j} x_{in+j}$$

where $d_n = n^2/f(n)$.

A slight modification[8] of Theorem 3.4 shows that the pro-algebraic completion of this ring is equal to

$$\overline{\mathbb{Z}[t, t^{-1}]} \oplus \prod \overline{\mathrm{Mat}_n(\mathbb{Z}[x_1 \ldots x_f(n)])}.$$

Then, the group $\tilde{\Gamma}_f = \mathrm{EL}_3(\boldsymbol{R}_f)$ will have property T by the results of [5]. An argument similar to the one in the proof of Theorem 1.1 gives that

$$\varkappa_n(\tilde{\Gamma}_f) = f(\lfloor n/3 \rfloor).$$

Remark 4.3. Theorems 1.2 and 1.1 also hold over algebraically closed fields of positive characteristic. The construction and most of the proofs remains the same — the main difficulty is that there is no analog of Theorem 2.6 which significantly complicates the proof of Theorem 3.4.

Acknowledgements. The author was partially supported by NSF grant DMS 1303117 and Simons Foundation grant 305181. I would like to thank Nir Avni, Nikolay Nikolov, Mark Sapir and Igor Rapinchuk for the useful conversation which lead to this paper. I am grateful to everyone who reviewed the preliminary version of this paper for their suggestions which significantly improved the paper.

[8] In order to show that $\mathbf{y}$ is in the ideal J one need to use the relation $\mathbf{y} = \mathbf{bcy}$.

References

[1] N. Avni, *Arithmetic groups have rational representation growth*, Ann. of Math. (2) **174** (2011), no. 2, 1009–10

[2] H. Bass, A. Lubotzky, A. Magid, and S. Mozes, *The proalgebraic completion of rigid groups*, Geom. Dedicata **95** (2002), 19–58.

[3] H. Bass, J. Milnor, and J.-P. Serre, *Solution of the congruence subgroup problem for* $\mathrm{SL}_n\,(n \geq 3)$ *and* $\mathrm{Sp}_{2n}\,(n \geq 2)$, Inst. Hautes Études Sci. Publ. Math. No. 33 (1967), 59–137.

[4] A. Borel and J. Tits, *Homomorphismes "abstraits" de groupes algébriques simples*, Ann. of Math. (2) **97** (1973), 499–571.

[5] M. Ershov and A. Jaikin-Zapirain, *Property (T) for noncommutative universal lattices*, Invent. Math. **179** (2010), no. 2, 303–347.

[6] M.J. Greenberg, *Schemata over local rings*, Ann. of Math. (2) **73** (1961), 624–648.

[7] ———, *Algebraic rings*, Trans. Amer. Math. Soc. **111** (1964), 472-481.

[8] G. Hochschild and G. D. Mostow, Pro-affine algebraic groups, Amer. J. Math. **91** (1969), 1127–1140.

[9] A.J. Hahn and T. O. O'Meara, *The classical groups and* K*-theory*, Grundlehren der Mathematischen Wissenschaften **291**, Springer, Berlin (1989)

[10] M. Kassabov and N. Nikolov, *Cartesian products as profinite completions*, Int. Math. Res. Not. **2006**, Art. ID 72947, 17 pp.

[11] M. Kassabov and M. Sapir, *Nonlinearity of matrix groups*, J. Topol. Anal. **1** (2009), no. 3, 251-260.

[12] S. Krstić and J. McCool, *Presenting* $\mathrm{GL}_n(k\langle T\rangle)$, J. Pure Appl. Algebra **141** (1999), no. 2, 175–183

[13] M. Larsen and A. Lubotzky, *Representation growth of linear groups*, J. Eur. Math. Soc. (JEMS) **10** (2008), no. 2, 351–390.

[14] A. Lubotzky and A.R. Magid, *Varieties of representations of finitely generated groups*, Mem. Amer. Math. Soc. **58** # 336, Providence, RI (1985)

[15] G. A. Margulis, *Discrete subgroups of semisimple Lie groups*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 17, Springer, Berlin, 1991.

[16] J. Milnor, *Introduction to algebraic* K*-theory*, Princeton Univ. Press, Princeton, NJ, 1971.

[17] I. A. Rapinchuk, *On linear representations of Chevalley groups over commutative rings*, Proc. Lond. Math. Soc. (3) **102** (2011), no. 5, 951–983.

[18] ———, *On the character varieties of finitely generated groups*, preprint.

[19] A. Weil, *Remarks on the cohomology of groups*, Ann. of Math. (2) **80** (1964), 149–157.

Department of Mathematics, Cornell University, Ithaca, NY 14853, USA
E-mail: kassabov@math.cornell.edu

Model theory and algebraic geometry in groups, non-standard actions and algorithmic problems

Olga Kharlampovich and Alexei Myasnikov

Abstract. We discuss the modern theory of equations in groups, algebraic geometry and model theory in free and hyperbolic groups, as well as group actions on Λ-trees. One of our main tools is a combinatorial process that combines and generalizes a number of known results and algorithms, such as the Makanin-Razborov process for solving equations in groups, Rauzy-Veech induction in dynamical systems, classification of basic group actions in group theory and topology, and elimination and parametrization theorems in classical algebraic geometry. The development of algebraic geometry comes together with advances in the theory of fully residually free and fully residually hyperbolic groups, which are coordinate groups of irreducible algebraic varieties. We describe finitely generated groups elementarily equivalent to a free non-abelian group (another classification is given by Sela) and show that the first-order theory of a free or a torsion-free hyperbolic group is decidable (solution to Tarski's problems from 1940's). Furthermore, for such groups we give an algorithm for elimination of quantifiers to boolean combinations of $\forall\exists$-formulas. We also provide a description of definable sets in a torsion-free hyperbolic group (in particular, in a free group) and demonstrate that only cyclic subgroups and the whole group are definable in these groups (this solves Malcev's problem of 1965). In the group actions section we describe all finitely presented groups acting freely on Λ-trees (solution to Alperin's and Bass problem of 1990). At the end we outline some related open problems.

Mathematics Subject Classification (2010). Primary 20E05; Secondary 20A15, 20F67.

Keywords. Free group, model theory, group actions.

1. Fundamental questions in model-theoretic algebra

In this section we consider some fundamental model-theoretic questions that should be asked about a given algebraic structure (a group, a ring, etc.), or a class of structures, to understand its principal algebraic and logical properties. These questions include: elementary classification and decidability of the theory (Tarski's type questions), description of definable sets (Malcev's problems), quantifier elimination (to some set of formulas), elementary embeddings and model completeness (Robinson's questions), types of elements and stability, natural axioms of the theory, models of the universal theory and related algebraic geometry, Fraisse limits and existentially closed structures. Addressing these questions could be a hard task, but many interesting results appeared along this way.

The language L of group theory consists of multiplication $\cdot$, inversion $^{-1}$, and a constant symbol 1 for the identity in the group. For a given group G one may add all elements of G as constants to the language L, thus obtaining a language L_G. If G is generated by a finite set

Proceedings of the International Congress of Mathematicians, Seoul, 2014

A it suffices to include only elements of A into the language. By $\phi(p_1, \dots, p_n)$ we denote a first-order formula in the language L (or L_G) whose free variables are contained in the set $\{p_1, \dots, p_n\}$. We will also use tuple notation $\phi(P)$ where $P = (p_1, \dots, p_n)$ for variables. Each first-order formula can be represented as

$$\phi(P) = \exists x_1 \forall y_1 \dots \exists x_n \forall y_n \phi_0(P, X, Y),$$

where $\phi_0(P, X, Y)$ has no quantifiers. A formula without free variables is called a sentence. The first order theory $Th(G)$ of a group G is the set of all first-order sentences in L (or in L_G) that are true in G. $Th(G)$ is all the information about G describable in first-order logic. Groups G and H are elementarily equivalent if $Th(G) = Th(H)$. Similarly one defines the first order theory of a ring, or a field (the language in this case consists of addition, multiplication, zero and identity elements), or an arbitrary structure.

Let $\mathcal{K}$ be a class of groups (or rings). *Elementary classification* for K is a problem to describe (in algebraic terms) elementarily equivalent groups in $\mathcal{K}$, while *decidability problem* for $\mathcal{K}$ asks to describe groups G in $\mathcal{K}$ with decidable $Th(G)$. The class $\mathcal{K}$ could be the class of all groups, or all finitely generated groups, or some other interesting class of groups.

A subset $S \subseteq G^n$ is *definable* in a group (ring, structure) G if there exists a first-order formula $\phi(P)$ in L_G such that S is precisely the set of all tuples P in G^n for which $\phi(P)$ holds in G. We say that S is definable without parameters if it is defined by a formula $\phi \in L$. Malcev's type problems concern with description of all definable sets in a given group G, or some of them, say definable subgroups.

By quantifier elimination we understand *quantifier elimination* to some *set of formulas* E, called the *eliminating set*. In this case for every formula $\phi(P)$ of a language L (or L_G) there exists a formula $\phi^*(P)$, which is a boolean combination of formulas from E, such that ϕ is equivalent to ϕ^* in G. A quantifier elimination is computable (effective) if the function $\phi \to \phi^*$ is computable. If formulas from the set E are simple enough a quantifier elimination to E gives a powerful tool to study $Th(G)$. We do not discuss much of model completeness, or Fraisse limits, or existentially closed groups in a class $\mathcal{K}$, in this paper, so we refer for definitions to [28].

We discuss below several examples in algebra where the elementary theories were studied and the principle questions addressed.

Example 1.1. Tarski himself showed for the field $\mathbb{C}$ of complex numbers and its first-order theory $Th(\mathbb{C})$ that $Th(\mathbb{C}) = Th(F)$ if and only if F is an algebraically closed field of characteristic 0; $Th(\mathbb{C})$ is decidable; definable sets are precisely the constructible sets (boolean combinations of algebraic sets). The axioms of the theory state that the characteristic is 0, and every equation has a solution. Types of elements are described. This led to the development of the theory of algebraically closed fields. This theory admits elimination of quantifiers, namely, every formula is logically equivalent (in the theory of algebraically closed fields) to a boolean combination of quantifier-free formulas. $Th(\mathbb{C})$ is stable, $\aleph_1$-categorical and model complete. All fields are models of the $\forall$-theory of $\mathbb{C}$.

Example 1.2. Tarski showed, by the method of quantifier elimination, that the first-order theory of the real numbers $\mathbb{R}$ under addition and multiplication is decidable. While this result appeared only in 1948, it dates back to 1930. This is a very interesting result, because Church proved in 1936 that Peano arithmetic (the theory of natural numbers) is not decidable. The theory of real closed fields was developed by Artin and Schreier after Artin's solution to 17th Hilbert Problem. $Th(\mathbb{R})$ is decidable; $Th(\mathbb{R}) = Th(F)$ if and only if F is a real

closed field; $Th(\mathbb{R})$ is not stable (one can define an order in $\mathbb{R}$), intervals and their boolean combinations are definable sets.

A real closed field is an ordered field of characteristic 0, where every odd degree polynomial has a root and every element or its negative is a square (axioms of $\forall$-theory). There are many types (2^{ω_1}). Models of the $\forall$-theory are formal real fields. There is quantifier elimination to quantifier free formulas and formulas $\exists t(x = t^2)$.

Example 1.3. We briefly state some well-known results on elementary theories of abelian groups: elementary theory of an abelian group A has quantifier elimination to positive primitive formulas, i.e., formulas of the type $\exists x_1 \dots \exists x_n \phi$, where ϕ is a conjunction of positive atomic formulas (equations in the case of groups); this quantifier elimination is effective if and only if A has decidable index problem; the theory of A is stable; all abelian torsion-free groups are universally equivalent; definable subsets are boolean combinations of cosets; there is a precise elementary classification of all abelian groups in terms of some invariants; Fraisse limits in the class of finitely generated torsion-free abelian groups are $\mathbb{Q}$-vector spaces of countable dimension.

2. Results in free and hyperbolic groups

Around 1945 A. Tarski put forward two problems on elementary theories of free groups that served as a motivation for much of the research in group theory and logic for the last sixty years. A joint effort of mathematicians of several generations culminated in the following theorems, solving these Tarski's conjectures. Denote by F_n a free group of rank n.

Theorem 2.1 ([33]-[38], [71]-[73]). $Th(F_n) = Th(F_m)$, *for all* $m, n > 1$.

Theorem 2.2 ([33]-[39]). *The elementary theory* $Th(F)$ *of a free group* F *even with constants from* F *in the language is decidable.*

In the 60s Malcev was a leader of the very active Novosobirsk school studying similar questions. Malcev himself proved that the elementary theory of the class of all finite groups is undecidable, Ershov proved the undecidability of the theories of symmetric and finite simple groups, and also obtained axioms and the decidability of the first-order theory of the field of p-adic numbers $\mathbb{Q}_p$. The results of Ershov, Romanovskii and Noskov imply that the elementary theory of a finitely generated virtually solvable group is decidable if and only if the group is virtually abelian.

The following questions were asked by Malcev ([32], Problem 1.19) for a free non abelian group F: "Describe definable sets in F; describe definable subgroups in F; is the commutator subgroup $[F, F]$ of F definable in F?"

Before answering these questions we consider some examples.

Example 2.3. Let $W(P, A) = 1$ be an equation (with constants) in a group G. Then the algebraic set $V_G(W) = \{g \in G^n \mid W(g, A) = 1\}$ is definable in G.

Let $w(x_1, \dots, x_n) \in F(X)$ be a group word. Then the set

$$w[G] = \{g \in G \mid g = w(h_1, \dots, h_n) \text{ for some } h_1, \dots, h_n \in G\}$$

is called a *verbal set* and is defined in G by the formula

$$\phi(p) = \exists y_1 \dots \exists y_n (p = w(y_1, \dots, y_n)).$$

So $w[G] = \{g \in G \mid G \models \phi(g)\}$. In particular, the set of all commutators is definable in G.

Example 2.4. Another example of definable set is the set of all bases in $F_2 = F_2(a, b)$. This is based on Nielsen's theorem that states that elements $g, h \in F_2$ form a basis if and only if $[g, h]$ is conjugated either to $[a, b]$ or $[b, a]$. Hence the set of bases in F_2 is defined by the following formula

$$\phi(p_1, p_2) = \exists z([p_1, p_2] = z^{-1}[a, b]z \ \vee \ [p_1, p_2] = z^{-1}[b, a]z).$$

Therefore the set of primitive elements of F_2 is definable.

Example 2.5. The center and the centralizer of a finite subset are definable in any group G. In particular, maximal cyclic subgroups are definable in a free group or a torsion-free hyperbolic group (in the language with constants), and, therefore, all cyclic subgroups are definable.

For a word $w(x_1, \dots, x_n) \in F(X)$ the subgroup $w(G)$ in a group G generated by the verbal set $w[G]$ is called the *verbal subgroup* of G defined by w. The verbal subgroup $w(G)$ has *finite width* if there is a number k such that every element in $w(G)$ is a product of at most k values of the word w in G or their inverses. A verbal subgroup $w(G)$ of finite width is definable in G (without parameters).

The commutator subgroup $[G, G]$ is the verbal subgroup of G defined by the word $[x_1, x_2]$. Malcev asked about the commutator subgroup of F because if the commutator subgroup were definable the same formula in the free groups of different rank, this would imply that free groups of different ranks are not elementarily equivalent. In the case of abelian, nilpotent and solvable groups the situation is different than in a free group. Let A_m be a free abelian group of rank m. The verbal subgroup A_m^2 has width 1 in A_m, hence it is definable. A_m/A_m^2 is a vector space of dimension m over the field $\mathbb{Z}_2$ of two elements. Using definability of A_m^2 one can write a sentence D_m (without parameters) stating that the dimension of the space A_m/A_m^2 is precisely m. Therefore two free abelian groups of finite rank are elementarily equivalent if and only if they are isomorphic (therefore have the same rank).

Let G be a finitely generated free nilpotent group of rank m and class c. The commutator subgroup $[G, G]$ has finite width, hence it is definable in G. So the abelianization $G/[G, G] \simeq A_m$ is interpretable in G. Again, one can write down a sentence stating that the rank of the abelianization of G is precisely m. Two free nilpotent groups of finite rank are elementarily equivalent if and only if they are isomorphic. Similarly, two free solvable groups of finite rank are elementarily equivalent iff they are isomorphic.

Proper verbal subgroups in non-abelian free group F have infinite width [68]. The same is true for arbitrary non-elementary hyperbolic groups [59].

Theorem 2.6 ([38, 73]). *Every formula in the theory of F is equivalent to the boolean combination of $\forall\exists$-formulas.*

Every definable subset of F is defined by some boolean combination of formulas

$$\exists X \forall Y (\vee_{i=1}^{k} (U_i(P, X, Y) = 1 \wedge V_i(P, X, Y) \neq 1)), \tag{2.1}$$

where X, Y, P are tuples of variables.

Theorem 2.7 ([74]). *Every formula in the theory of a non-elementary torsion-free hyperbolic group G is equivalent to a boolean combination of $\forall\exists$-formulas. The theory is stable.*

Theorem 2.8 ([41]). *Let Γ be a torsion free hyperbolic group. There exists an algorithm, given a first-order formula ϕ to find a boolean combination of $\forall\exists$-formulas that define the same set as ϕ over Γ.*

The elementary theory of a torsion-free hyperbolic group is decidable.

Notice that in the language L_G every finite system of equations in a free group is equivalent to one equation (this is Malcev's result) and every finite disjunction of equations is equivalent to one equation (this is attributed to Gurevich). The same is true in a torsion-free hyperbolic group [40].

For a free and for a torsion-free hyperbolic group G a more precise result about quantifier elimination holds.

Theorem 2.9 ([40]). *Every definable set over G (in particular, over F) in the languige L_G is defined by some boolean combination of formulas*

$$\exists X \forall Y (U(P, X) = 1 \wedge V(P, X, Y) \neq 1), \tag{2.2}$$

where X, Y, P are tuples of variables.

The following theorem gives a solution to Malcev's problem, it describes definable subgroups in a torsion-free hyperbolic group.

Theorem 2.10 ([40]). *Proper non-cyclic subgroups of a torsion free hyperbolic group (in particular, of a free group F) are not definable.*

In the same paper [40] we prove that the set of primitive elements of F is not definable if $rank(F) > 2$.

The definition of a sub-multi pattern used in the theorem below is technical. It is Definition 6 in [40].

Theorem 2.11 ([40]). *For every definable set $P \subseteq F^m$ in a free group F, either P or its complement $\neg P$ is a sub-multipattern.*

This theorem implies Bestvina and Feighn's conjecture that every definable set in F is either negligible or co-negligible (Definition 15 in [40].)

Negligible subsets in that sense are also negligible in a sense of complexity theory. Recall that in complexity theory $T \subseteq F(X)$ is called generic if

$$\rho_n(T) = \frac{|T \cap B_n(X)|}{|B_n(X)|} \to 1, \ as \ n \to \infty,$$

where $B_n(X)$ is the ball of radius n in the Cayley graph of $F(X)$. A set is called negligible is its complement is generic.

3. Equations, algebraic geometry, and universal theory

The work of mathematicians on the Tarski conjectures was rather fruitful - several areas of group theory were developed along the way. It was clear from the beginning that one needs a precise description of solution sets of systems of equations over free groups and a robust

theory of finitely generated groups which satisfy the same universal (existential) formulas as a free non-abelian group. Basics of algebraic (or Diophantine) geometry over groups had been outlined by Baumslag, Miasnikov and Remeslennikov in [6], while the fundamentals of the elimination theory and the theory of fully residually free groups appeared in the works [33, 34]. Those two papers contain results that became fundamental for the proofs of the above theorems.

3.1. Milestones in the theory of equations in free groups. The first general results on equations in groups appeared in the 1960's [51]. Lyndon introduced an axiomatic theory of length functions and initiated the study of fully residually free groups. He proved [50] that the completion $F^{\mathbb{Z}[t]}$ of a free group F by the polynomial ring $\mathbb{Z}[t]$ is discriminated by F. This is, actually, a Fraisse model in the appropriate category. Notice, that the class of limit groups has joint embedding property and it is closed under taking subgroups, but it is not closed under amalgamation, so the standard Fraisse theorem does not apply. However, replacing arbitrary embeddings with $\exists_1$-embeddings (i.e., discriminating embeddings) one can reproduce the Fraisse construction and get similar results. If we further restrict the class of $\exists_1$-embeddings allowing only free products and centralizer extensions then the Fraisse models in this category will be all isomorphic to $F^{\mathbb{Z}[t]}$.

Malcev [56] described solutions of the equation $zxyx^{-1}y^{-1}z^{-1} = aba^{-1}b^{-1}$ in a free group. The description uses the group of automorphisms of the coordinate group of the equation, and the minimal solutions relative to these automorphisms - a very powerful idea, that nowadays is inseparable from the modern approach to equations.

Merzljakov proved [55] a remarkable theorem that any two nonabelian free groups of finite rank have the same positive theory, and also showed that positive formulas in free groups have definable Skolem functions, thus giving quantifier elimination of positive formulas in free groups to existential formulas. Recall that the positive theory of a group consists of all positive (without negations in their normal forms) sentences that are true in this group.

In the 1980's new crucial concepts were introduced. Makanin proved [52] the algorithmic decidability of the Diophantine problem over free groups, and showed that both the universal theory and the positive theory of a free group are algorithmically decidable. He created an extremely powerful technique (the Makanin elimination process) to deal with equations over free groups.

Shortly afterwards, Razborov described the solution set of an arbitrary system of equations over a free group in terms of what is known now as Makanin-Razborov diagrams [66, 67].

Solution sets of arbitrary quadratic equations over free groups were described in [15] and [23]. These equations came to group theory from topology and their role in group theory was not altogether clear then. Now they form one of the corner-stones of the theory of equations in groups, due to their relations to JSJ-decompositions of groups and NTQ systems.

3.2. Basic notions of algebraic geometry over groups. Following [6] and [35] we introduce here some basic notions of algebraic geometry over groups.

Let G be a group generated by a finite set A, $F(X)$ be a free group with basis $X = \{x_1, x_2, \dots x_n\}$, define $G[X] = G * F(X)$. If $S \subset G[X]$ then the expression $S = 1$ is called *a system of equations* over G. As an element of the free product, the left side of every equation in $S = 1$ can be written as a product of some elements from $X \cup X^{-1}$ (which are called *variables*) and some elements from A (*constants*). To emphasize this we sometimes

write $S(X, A) = 1$.

A *solution* of the system $S(X, A) = 1$ over a group G is a tuple of elements $g_1, \ldots, g_n \in G$ such that after replacement of each x_i by g_i the left hand side of every equation in $S = 1$ turns into the trivial element of G. To study equations over a given fixed group G it is convenient to consider the category of G-groups, i.e., groups which contain the group G as a distinguished subgroup. If H and K are G-groups then a homomorphism $\phi : H \to K$ is a G- homomorphism if $g^\phi = g$ for every $g \in G$, in this event we write $\phi : H \to_G K$. In this category morphisms are G-homomorphisms; subgroups are G-subgroups, etc. A solution of the system $S = 1$ over G can be described as a G-homomorphism $\phi : G[X] \longrightarrow G$ such that $\phi(S) = 1$. Denote by $ncl(S)$ the normal closure of S in $G[X]$, and by G_S the quotient group $G[X]/ncl(S)$. Then every solution of $S(X, A) = 1$ in G corresponds to a G-homomorphism $G_S \to G$, and vice versa. By $V_G(S)$ we denote the set of all solutions in G of the system $S = 1$, it is called the *algebraic set defined by* S. The algebraic set $V_G(S)$ uniquely corresponds to the normal subgroup

$$R(S) = \{T(x) \in G[X] \mid \forall A \in G^n (S(A) = 1 \to T(A) = 1\}$$

of the group $G[X]$. Notice that if $V_G(S) = \emptyset$, then $R(S) = G[X]$. The subgroup $R(S)$ contains S, and it is called the *radical of* S. The quotient group

$$G_{R(S)} = G[X]/R(S)$$

is the *coordinate group* of the system $S(X, A) = 1$. Again, every solution of $S(X) = 1$ in G can be described as a G-homomorphism $G_{R(S)} \to G$.

A group G is called a *CSA group* if every maximal abelian subgroup M of G is malnormal, i.e., $M^g \cap M = 1$ for any $g \in G - M$. The abbreviation CSA means conjugacy separability for maximal abelian subgroups. The class of CSA-groups is quite substantial. It includes all abelian groups, all torsion-free hyperbolic groups, all groups acting freely on Λ-trees and many one-relator groups (see, for example, [22].

We can define a *Zariski topology* on G^n by taking algebraic sets in G^n as a sub-basis for the closed sets of this topology.

A group G is called *equationally Noetherian* if every system $S(X) = 1$ with coefficients from G is equivalent over G to a finite subsystem $S_0 = 1$, where $S_0 \subset S$, i.e., $V_G(S) = V_G(S_0)$. It is known that linear groups (in particular, fully residually free groups) are equationally Noetherian (see [6, 9, 21]). Torsion-free hyperbolic groups are also equationally Noetherian [71]. If G is equationally Noetherian then the Zariski topology on G^n is *Noetherian* for every n, i.e., every proper descending chain of closed sets in G^n is finite. This implies that every algebraic set V in G^n is a finite union of irreducible subsets (called *irreducible components* of V), and such decomposition of V is unique. Recall that a closed subset V is *irreducible* if it is not a union of two proper closed subsets.

3.3. Fully residually free groups (limit groups) and Γ-limit groups. Finitely generated fully residually free groups (limit groups) play a crucial role in the theory of equations and first-order formulas over a free group. Recall that a group G is called *fully residually free* (or *freely discriminated*, or *ω-residually free*) if for any finite subset of non-trivial elements $g_1, \ldots, g_n \in G$ there exists a homomorphism ϕ of G into a free group F, such that $\phi(g_i) \neq 1$ for $i = 1, \ldots, n$. These groups are torsion-free, have the CSA property, each of their abelian subgroup is finitely generated, they are finitely presented [34] and linear.

Below, let Γ be a torsion-free hyperbolic group. A group G is fully residually Γ if for any finite set of non-trivial elements $g_1, \dots, g_n \in G$ there exists a Γ-homomorphism ϕ from G to Γ such that $\phi(g_i) \neq 1$ for $i = 1, \dots, n$. The following result appears for general algebras in [18]. We will formulate it for fully residually Γ groups.

Proposition 3.1. *Let G be a finitely generated Γ-group (containing a distinguished copy of Γ). Then the following conditions are equivalent:*

(1) *G is fully residually Γ;*

(2) *G is universally equivalent to Γ (in the language with constants);*

(3) *G is the coordinate group of an irreducible algebraic set over Γ;*

(4) *G is a Γ- limit group;*

(5) *G embeds into an ultrapower of Γ with Γ embeddded diagonally.*

3.4. Structure and embedding. Let Γ be a torsion-free hyperbolic group. An *iterated centralizer extension* of Γ can be obtained from Γ by a chain of HNN-extensions of a very specific type, so-called *extensions of centralizers*: $\Gamma = G_0 < G_1 < \dots < G_n$ where $G_{i+1} = \langle G_i, t_i \mid [C_{G_i}(u_i), t_i] = 1 \rangle$ (extension of the centralizer $C_{G_i}(u_i)$, where $u_i \in G_i$).

Theorem 3.2 ([34, 35]). *Given an irreducible system $S = 1$ over F one can effectively embed the coordinate group $F_{R(S)}$ into $F^{Z[t]}$ i.e., one can find $n \in \mathbb{N}$ and an embedding $F_{R(S)} \to G_n$ into an iterated centralizer extension G_n of F. The analogous result is true for a torsion-free hyperbolic group Γ in place of F [41, 42].*

Since every subgroup of a free group is free, this implies that every finitely generated fully residually free group is finitely presented (this is not true for fully residually Γ groups).

This allows one to study the coordinate groups of irreducible systems of equations via their splittings into graphs of groups. This also provides a complete description of limit groups (Γ-limit groups) and gives a lot of information about their algebraic structure. In particular, limit groups act freely on $\mathbb{Z}^n$-trees with lexicographic order, and all limit groups (strict Γ-limit groups), except for abelian and surface groups, have a non-trivial cyclic JSJ-decomposition.

Let K be an HNN-extension of a group G with associated subgroups A and B. K is called a *separated HNN-extension* if for any $g \in G$, $A^g \cap B = 1$.

Corollary 3.3 ([42]). *Let G be a Γ-limit group. There exists a group H isomorphic to G that can be obtained from a finite family $\Gamma_1, \dots, \Gamma_m$ of finitely generated subgroups of Γ and free abelian groups of finite rank by a finite sequence $\tau = (\tau_1, \dots, \tau_k)$ of operations of the following type: free products, free products with abelian amalgamated subgroups (at least one of which is a maximal abelian subgroup in its factor), free extensions of centralizers, and separated HNN-extensions with abelian associated subgroups (at least one of which is maximal).*

If G is given as the coordinate group of a finite system of equations over Γ or as a subgroup of an iterated centralizer extension of Γ (by a finite set of generators), then there is an algorithm to find a sequence τ as above, as well as the required family of subgroups $\Gamma_1, \dots, \Gamma_m$ of Γ (given by their finite generating sets), and an isomorphism between G and H.

Therefore if G is given as a subgroup of an iterated centralizer extension, we can find its representation as a coordinate group of a finite system of equations.

3.5. NTQ systems and groups. Elimination process (EP) is a symbolic rewriting process of a certain type that transforms formal systems of equations in groups or semigroups. Elimination processes proved to be crucial in solving various problems in groups: finding solutions of equations, finding non-trivial abelian splittings and JSJ decompositions, describing algebraic structure of the coordinate groups of irreducible systems of equations over a given group, classifying groups acting freely on Λ-trees, etc. They may differ a lot from one another, but they always have some common features. Most of these common features can be traced back to the original Makanin-Razborov process [66, 67], but there is a crucial one that appear first in 1996 in the paper [34], were it was used to obtain an effective description of solutions of equations in free (and fully residually free) groups in a particularly nice form. In general, this new feature can be described as an algorithm to reduce a given system of group equations over a group G (which is either free or has a free length function or is close in some sense to a free group) to a finite number of systems of equations in the *triangular quasi-quadratic* form (an analog of Gauss elimination process in the non-commutative setting). At the level of algebraic geometry this algorithm finds all irreducible components of a given algebraic set over G, in particular it embeds the coordinate group of a given system of equations in G into a finite direct product of NTQ groups (later Sela referred to NTQ groups as *ω-residually free towers* [71]). While at the level of groups the algorithm gives an embedding of a finitely generated residually G group into a finite direct product of fully residually G groups. Of course, this algorithm does not work for arbitrary groups G, but it does for quite a few of them (see, for example,[10, 11, 41, 44]).

Now we give formal definitions and describe some results.

Triangular quasi-quadratic (TQ) system is a finite system that has the following form

$$\begin{aligned} S_1\left(X_1, X_2, \ldots, X_n, A\right) &= 1, \\ S_2\left(X_2, \ldots, X_n, A\right) &= 1, \\ &\vdots \\ S_n\left(X_n, A\right) &= 1 \end{aligned}$$

where either $S_i = 1$ is quadratic in variables X_i, or $S_i = 1$ is a system $[x_j, x_k] = 1$ and, in addition, equations $[x, u] = 1$ for all $x, x_j, x_k \in X_i$ and some $u \in F_{R(S_{i+1},\ldots,S_n)}$ or S_i is empty.

A TQ system above is non-degenerate (NTQ) if for every i, $S_i(X_i, \ldots, X_n, A) = 1$ has a solution in the coordinate group $G_i = F_{R(S_{i+1},\ldots,S_n)}$, where $G_n = F$ (or $G_n = \Gamma$).

We proved in [33] (see also [35]) that *NTQ systems define irreducible algebraic sets* and, therefore, their coordinate groups, where the radical is computed in a non-abelian free group $F(A)$ (respectively, with the radical computed in a free product of $\Gamma(A)$ and a free group), that are called *NTQ groups*, are fully residually free (resp., fully residually Γ).

We represented a solution set of a system of equations canonically as a union of solutions of a finite family of NTQ groups.

Theorem 3.4 ([34, 35]). *One can effectively construct EP that starts on an arbitrary system*

$$S(X, A) = 1$$

over F and results in finitely many NTQ systems

$$U_1(Y_1, A) = 1, \ldots, U_m(Y_m, A) = 1$$

such that

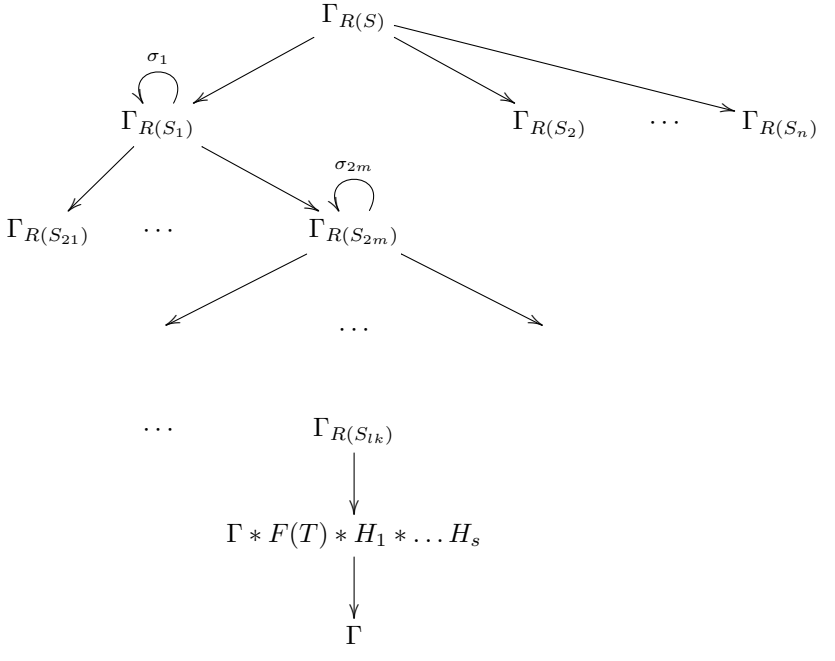
$$V_F(S) = P_1(V(U_1)) \cup \ldots \cup P_m(V(U_m))$$

for some word mappings $P_1, \ldots, P_m$. (P_i maps a tuple $\bar{Y}_i \in V(U_i)$ to a tuple $\bar{X} \in V_F(S)$. One can think about P_i as an A-homomorphism from $F_{R(S)}$ into $F_{R(U_i)}$, then any solution $\psi : F_{R(U_i)} \to F$ pre-composed with P_i gives a solution $\phi : F_{R(S)} \to F$).

Similarly one can effectively describe a solution set of a a system over a torsion-free hyperbolic group Γ [44].

Our elimination process can be viewed as a non-commutative analog of the classical elimination process in algebraic geometry.

A *Hom-diagram* is a finite rooted directed tree with groups assigned to its vertices and homomorphisms assigned to the edges. Here all groups, except, maybe, the one in the root, are fully residually Γ, (given by a finite presentation relatively to a finite number of finitely generated subgroups of Γ) arrows pointing down (except the last one) correspond to fixed epimorphisms (defined effectively in terms of generators) with non-trivial kernels, and loops correspond to automorphisms of the coordinate groups. A family of homomorphisms encoded in a path from the root to the leaf of this tree (each homomorphism in the family is a composition of a sequence of automorphisms and fixed epimorphisms assigned to edges) is called a *fundamental sequence* of homomorphisms.



where $H_1, \ldots, H_s$ are isomorphic to finitely generated subgroups of Γ, the arrows pointing to the leaf correspond to embeddings of $H_1, \ldots, H_s$ into Γ and arbitrary specializations of free variables from T.

Theorem 3.5 ([34, 35, 41]). *All solutions of the system of equations $S = 1$ in Γ can be effectively represented as homomorphisms from $\Gamma_{R(S)}$ into Γ encoded into a finite canonical Hom-diagram.*

Therefore the solution set of the system $S = 1$ consists of a finite number of fundamental sequences. And each fundamental sequence "factors through" one of the NTQ systems from Theorem 3.4 (or the analogous theorem for Γ). If $S = 1$ over Γ is irreducible, or, equivalently, $G = \Gamma_{R(S)}$ is fully residually Γ, then, obviously, one of the fundamental sequences discriminates G. This gives the following result.

Theorem 3.6 ([34, 35, 41]). *A finitely generated fully residually Γ group G is a subgroup of the coordinate group of an NTQ system. There is an algorithm to construct this embedding. There is an algorithm to construct an abelian JSJ decomposition of G.*

This corresponds to the extension theorems in the classical theory of elimination for polynomials.

Since NTQ groups are fully residually Γ, fundamental sequences corresponding to different NTQ groups discriminate fully residually Γ groups which are coordinate groups of irreducible components of system $S(X, A) = 1$. This implies

Theorem 3.7 ([34, 35, 41]). *There is an algorithm to find irreducible components for a system of equations $S(X, A) = 1$ over Γ. Equivalently, there is an algorithm to find maximal limit quotients of the coordinate group of $S(X, A) = 1$.*

3.6. Subgroups of limit groups.

Theorem 3.8. *Let G be a limit group.*

(1) *All finitely generated subgroups of G are quasi-isometrically embedded,*

(2) *Given a finite set of elements of G we can find a finite presentation of a subgroup generated by this set,*

(3) *All the classcal algorithmic problems for subgroups of G are decidable (see [38, §8]).*

4. Groups elementarily equivalent to non-abelian free groups

If an NTQ group, with the radical computed in a non-abelian free group, does not contain non-cyclic abelian groups we call it *regular NTQ group*. We have shown in [34] that regular NTQ groups are hyperbolic.

Theorem 4.1 ([38]). *Regular NTQ groups are exactly the finitely generated models of the elementary theory of a non-abelian free group.*

Sela also gave a description of finitely generated models of the elementary theory of a non-abelian free group in terms of *hyperbolic ω-residually free towers* ([73], Theorem 7). However, it was not quite correct, the definition of a hyperbolic ω-residually free tower given in [73] had to be changed. It was partially corrected in [63] and completely corrected in [64].

5. Group actions

We investigate non-Archimedean group actions, length functions and infinite words using the same elimination process that we use for solving equations in a free group. In [50] Lyndon

introduced real-valued length functions as a tool to carry over Nielsen cancelation theory from free groups to a more general setting. Some results in this direction were obtained in [3, 25–27, 65]. In [12] Chiswell described a crucial construction which shows that a group with a real-valued length function has an action on an $\mathbb{R}$-tree, and vice versa. Later, Morgan and Shalen realized that a similar construction and results hold for an arbitrary group with a Lyndon length function which takes values in an arbitrary ordered abelian group Λ (see [61]). In particular, they introduced Λ-trees as a natural generalization of $\mathbb{R}$-trees which they studied in relation with Thurston's Geometrization Program. Thus, actions on Λ-trees and Lyndon length functions with values in Λ are two equivalent languages describing the same class of groups. In the case when the action is free (the stabilizer of every point is trivial) we call groups in this class Λ-free. We refer to the book [13] for a detailed discussion on the subject.

One of the major events in combinatorial group theory in 1970's was the development of Bass-Serre theory. We refer to the book [76], where Serre laid down fundamentals of the theory of groups acting freely on simplicial trees. In particular, Bass-Serre theory makes it possible to extract information about the structure of a group from its action on a simplicial tree. Alperin and Bass [2] developed the initial framework of the theory of group actions on Λ-trees and stated the fundamental research goals: find the group theoretic information carried by an action (by isometries) on a Λ-tree; generalize Bass-Serre theory to actions on arbitrary Λ-trees.

A joint effort of several researchers culminated in a description of finitely generated groups acting freely on $\mathbb{R}$-trees [8, 20], which is now known as Rips' theorem: a finitely generated group acts freely on an $\mathbb{R}$-tree if and only if it is a free product of free abelian groups and surface groups (with an exception of non-orientable surfaces of genus $1, 2$, and 3). The key ingredient of this theory is the so-called "Rips machine", the idea of which comes from Makanin's algorithm for solving equations in free groups (see [52]). The Rips machine appears in applications as a general tool that takes a sequence of isometric actions of a group G on some "negatively curved spaces" and produces an isometric action of G on an $\mathbb{R}$-tree as the Gromov-Hausdorff limit of the sequence of spaces. Free actions on $\mathbb{R}$-trees cover all Archimedean actions, since every group acting freely on a Λ-tree for an Archimedean ordered abelian group Λ also acts freely on an $\mathbb{R}$-tree.

In the non-Archimedean case there were only partial results for particular choices of Λ. First of all, in [4] Bass studied finitely generated groups acting freely on $\Lambda_0 \oplus \mathbb{Z}$-trees with respect to the right lexicographic order on $\Lambda_0 \oplus \mathbb{Z}$, where Λ_0 is any ordered abelian group. In this case it was shown that the group acting freely on a $\Lambda_0 \oplus \mathbb{Z}$-tree splits into a graph of groups with Λ_0-free vertex groups and maximal abelian edge groups. Next, Guirardel (see [24]) obtained the structure of finitely generated groups acting freely on $\mathbb{R}^n$-trees (with the lexicographic order). In [46] the authors described the class of finitely generated groups acting freely and regularly on $\mathbb{Z}^n$-trees in terms of HNN-extensions of a very particular type. The action is *regular* if all branch points are in the same orbit. The importance of regular actions becomes clear from the results of [48], where we proved that a finitely generated group acting freely on a $\mathbb{Z}^n$-tree is a subgroup of a finitely generated group acting freely and regularly on a $\mathbb{Z}^m$-tree for $m \geqslant n$, and the paper [14], where is was shown that a group acting freely on a Λ-tree (for arbitrary Λ) can always be embedded in a length-preserving way into a group acting freely and regularly on a Λ-tree (for the same Λ).

In [45] we gave a partial solution (for finitely presented groups) of the following main problem of the Alperin-Bass program.

Problem. Describe finitely presented (finitely generated) Λ-free groups for an arbitrary ordered abelian group Λ.

We proved the following results.

Theorem 5.1. *Any finitely presented regular Λ-free group G can be represented as a union of a finite series of groups*

$$G_1 < G_2 < \cdots < G_n = G,$$

where

(1) *G_1 is a free group,*

(2) *G_{i+1} is obtained from G_i by finitely many HNN-extensions in which associated subgroups are maximal abelian, finitely generated, and the associated isomorphisms preserve the length induced from G_i.*

Theorem 5.2. *Any finitely presented Λ-free group can be isometrically embedded into a finitely presented regular Λ-free group.*

Theorem 5.3. *Any finitely presented Λ-free group G is $\mathbb{R}^n$-free for an appropriate $n \in \mathbb{N}$, where $\mathbb{R}^n$ is ordered lexicographically.*

Theorem 5.4. *Let G be a finitely presented group with a free Lyndon length function $l : G \to \Lambda$. Then the subgroup Λ_0 generated by $l(G)$ in Λ is finitely generated.*

The following result automatically follows from Theorem 5.1 and Theorem 5.2 by simple application of Bass-Serre Theory.

Theorem 5.5. *Any finitely presented Λ-free group G can be obtained from free groups by a finite sequence of amalgamated free products and HNN extensions along maximal abelian subgroups, which are free abelian groups of finite rank.*

The following result concerns abelian subgroups of Λ-free groups. For $\Lambda = \mathbb{Z}^n$ it follows from the main structural result for $\mathbb{Z}^n$-free groups and [47], for $\Lambda = \mathbb{R}^n$ it was proved in [24]. The statement 1) below answers Question 2 (page 250) from [13] in the affirmative for finitely presented Λ-free groups.

Theorem 5.6. *Let G be a finitely presented Λ-free group. Then:*

(1) *every abelian subgroup of G is a free abelian group of finite rank, which is uniformly bounded from above by the rank of the abelianization of G.*

(2) *G has only finitely many conjugacy classes of maximal non-cyclic abelian subgroups,*

(3) *G has a finite classifying space and the cohomological dimension of G is at most $\max\{2, r\}$ where r is the maximal rank of an abelian subgroup of G.*

Theorem 5.7. *Every finitely presented Λ-free group is hyperbolic relative to its non-cyclic abelian subgroups.*

This follows from the structural Theorem 5.1 and the Combination Theorem for relatively hyperbolic groups [16].

The following results answers affirmatively the strongest form of the Problem (GO3) from the Magnus list of open problems [5], in the case of finitely presented groups.

Corollary 5.8. *Every finitely presented Λ-free group is biautomatic.*

Proof. This follows from Theorem 5.7 and Rebbechi's result [70]. □

Definition 5.9. A *hierarchy* for a group G is a way to repeatedly build it starting with trivial groups by repeatedly taking amalgamated products $A *_C B$ and HNN extensions $A*_{C^t=D}$ whose vertex groups have shorter length hierarchies. The hierarchy is *quasi convex* if the amalgamated subgroup C is a finitely generated subgroups that embeds by a quasi-isometric embedding, and if C is malnormal in $A *_c B$ or $A*_{C^t=D}$.

Theorem 5.10. *Every finitely presented Λ-free group G has a quasi-convex hierarchy with abelian edge groups.*

Theorem 5.11 ([78]). *Suppose G is toral relatively hyperbolic and has a malnormal quasi convex hierarchy. Then G is virtually special (therefore has a finite index subgroup that is a subgroup of a right angled Artin group (RAAG)).*

As a corollary one gets the following result.

Corollary 5.12. *Every finitely presented Λ-free group G is locally undistorted, that is, every finitely generated subgroup of G is quasi-isometrically embedded into G.*

Corollary 5.13. *Every finitely presented Λ-free group G is virtually special, that is, some subgroup of finite index in G embeds into a right-angled Artin group.*

The following result answers in the affirmative to Question 3 (page 250) from [13] in the case of finitely presented groups.

Theorem 5.14. *Every finitely presented Λ-free group is right orderable.*

The following addresses Chiswell's question whether Λ-free groups are orderable or not.

Theorem 5.15. *Every finitely presented Λ-free group is virtually orderable, that is, it contains an orderable subgroup of finite index.*

Since right-angled Artin groups are linear and the class of linear groups is closed under finite extension we get the following

Theorem 5.16. *Every finitely presented Λ-free group is linear.*

Since every linear group is residually finite we get the following.

Corollary 5.17. *Every finitely presented Λ-free group is residually finite.*

Corollary 5.18. *Let G be a finitely presented Λ-free group. Then the following algorithmic problems are decidable in G:*

- *the Word and Conjugacy Problems;,*
- *the Diophantine Problem (decidability of arbitrary equations in G).*

Indeed, decidability of equations follows from [16]. Results of Dahmani and Groves [17] imply the following two corollaries.

Corollary 5.19. *Let G be a finitely presented Λ-free group. Then:*

- *G has a non-trivial abelian splitting and one can find such a splitting effectively,*
- *G has a non-trivial abelian JSJ-decomposition and one can find such a decomposition effectively.*

Corollary 5.20. *The isomorphism problem is decidable in the class of finitely presented groups that act freely on Λ-trees.*

Theorem 5.21. *The subgroup membership problem is decidable in every finitely presented Λ-free group.*

6. Open problems

6.1. Free, torsion-free hyperbolic, and toral relatively hyperbolic groups. The Diophantine problem in free groups is decidable, though the time complexity of the original Makanin's algorithm [52] is not primitive recursive [31]. Recent improvements on the time complexity of this problem allows one to put forward the following.

Problem 6.1. Is it true that the Diophantine problem in a free group is in NP (hence NP-complete)?

Affirmative solution to this problem would put solving equations in free groups, as well as some other rather complex algorithms in groups, into the realm of reasonable computations.

As we have mentioned in the previous sections the Tarski and Malcev's problems are now solved for torsion-free hyperbolic groups, as well as effective quantifier elimination to boolean combinations of $\forall\exists$-formulas. Elementary embeddings in the elementary theories of free and torsion-free hyperbolic groups were studied and described in [63, 64]. However, some principal model-theoretic questions (see Section 1) for free groups are still open. In particular, the following one is of crucial interest.

Problem 6.2.

1) Describe a natural system of axioms for the elementary theory of a free non-abelian group.
2) Describe a natural system of axioms for the elementary theory of a torsion-free hyperbolic group.

Problem 6.3. Describe existentially closed groups in the elementary theory of a free non-abelian (non-elementary torsion-free hyperbolic) group.

Problem 6.4. Solve principal model-theoretic questions for toral relatively hyperbolic groups.

6.2. Free products. Decidability of equations and description of algebraic sets in a free product of group $A * B$ was solved in [10].

Recently it was announced in [75] that if H_i and G_i are groups such that $Th(G_i) = Th(H_i)$, $i = 1, 2$, then $Th(H_1 * H_2) = Th(G_1 * G_2)$.

It seems it is a good time now to study elementary theories of free products.

Problem 6.5. Address the *principle model-theoretic problems* for free products of groups "modulo the factors".

In particular, the following concrete problems are of prime interest.

Problem 6.6 (Decidability of the elementary theory). Prove that if $Th(A)$ and $Th(B)$ are decidable then $Th(A * B)$ is also decidable.

Problem 6.7. Let G be a group and A a free factor of G. Prove that if $Th(G)$ is decidable then $Th(A)$ is also decidable.

To attack this problem one may use the elimination process developed in [10].

Finally, we suggest to study model theory of toral relatively hyperbolic groups. This is much more general class then torsion-free hyperbolic or limit groups, nevertheless, we believe that the methods developed for the groups above should suffice to address the larger class as well.

Problem 6.8. Solve principal model-theoretic questions for toral relatively hyperbolic groups.

Perhaps, a good start would be to address the following (easier) problem.

Problem 6.9. Solve principal model-theoretic questions for limit groups.

6.3. Right angled Artin groups. Right angled Artin groups play an increasingly important part in modern geometric and combinatorial group theory due to results of Wise [78], Agol [1] etc. Recall that a RAAG G is a group given by a presentation of the form $< a_1, \ldots, a_r | R >$, where R is a subset of the set $\{[a_i, a_j] | i, j = 1, \ldots, r\}$.

Casals-Ruiz and Kazachkov [11] obtained an algorithmic description of the solution set of a system of equations in RAAGs.

It is time now to develop algebraic geometry and model theory over RAAGS. Algebraic geometry over RAAGs is different from algebraic geometry over relatively hyperbolic groups. New types of "splittings" occur, and new "universal" splittings should be defined. Notice that it follows from [19] that the universal theory of any RAAG is decidable. We formulate here the following Tarski type questions.

Problem 6.10. Princilal model-theoretic questions for RAAGs. Let A be a right angled Artin group.

(1) Describe when two RAAGs are elementary equivalent.

(2) Is the elementary theory $Th(A)$ decidable?

(3) Describe a natural system of axioms for $Th(A)$.

(4) Which finitely generated groups are elementary equivalent to A?

It is known that two RAAGs defined by graphs Γ and Δ are isomorphic iff the graphs Γ and Δ are isomorphic [30]. This gives an easy classification of RAAGs up to isomorphism. However, according to our philosophy, to study algebraic properties of RAAGs, in particular, algebraic geometry and model theory of RAAGs, one need to look at all groups which are fully residually RAAGs. This requires a good understanding of finitely generated subgroups of RAAGs. Notice that the class $\mathcal{S}$ of all finitely generated subgroups of RAAGs is very large, as Wise's theorem reveals. For example, limit groups are virtually subgroups

of RAAG, as well as one-relator groups with torsion. We know already that RAAGs contain a lot of interesting groups as their subgroups. But we do not know yet if the class is so large that it gets out of hand. Solution to the following problem will clarify the situation.

Problem 6.11. Is the isomorphism problem for finitely generated quasi-convex subgroups of RAAGs decidable?

6.4. Pro-finite groups. Solutions to Tarsky's problems for free groups inspires us to pose the following problem.

Problem 6.12. Study the principal model-theoretic questions for free pro-p-groups of finite rank.

Notice that two finitely generated pro-p-groups are elementarily equivalent if and only if they are isomorphic [58]. The same is true for two finitely generated pro-finite groups [29], though in this case the argument is based on hard results from [62]. Observe, that the terms $\gamma_m(\hat{F}_n)$ of the lower central series of a free pro-p-groups $\hat{F}_n$ of finite rank have finite width, so they are definable in $\hat{F}_n$, hence the free nilpotent pro-p-groups which are quotients $\hat{F}_n/\gamma(\hat{F}_n)$ are interpretable in $\hat{F}_n$. That has some impact on the elementary theory of $\hat{F}_n$. Finitely generated nilpotent pro-p-groups with decidable elementary theory were described in [58], in particular, the theories of the quotients $\hat{F}_n/\gamma(\hat{F}_n)$ are decidable. Now we formulate two concrete related questions in this area.

Problem 6.13.

1) Find a set of axioms of $Th(\hat{F}_n)$.
2) Prove that the theory $Th(\hat{F}_n)$ is decidable.

To address these questions we believe one has to develop first the theory of equations and algebraic geometry over $\hat{F}_n$. To this end we formulate several questions.

Problem 6.14. Prove that the group $\hat{F}_n$ is equationally Noetherian.

Of course, equations in this case are the "pro-p-equations" (see [54] for definitions and some results).

Problem 6.15. Prove that the Diophantine problem for $\hat{F}_n$ is decidable.

6.5. Problems for free associative and Lie algebras. It is very interesting to study elementary theories of free associative and Lie algebras. It was shown in [60] that elementary theory of a finite dimensional free associative or Lie algebra with coefficients in a field with undecidable theory is also undecidable. So we may assume from the outset the we consider coefficients from a nice field, say a field of two elements or an algebraically closed field of characteristic zero.

Problem 6.16. Prove that the Diophantine problem is decidable for free associative algebras.

Problem 6.17. Are free associative finitely generated algebras equationally Noetherian?

Let $\mathcal{A}$ be such an algebra. If the answer is positive then the Zariski topology over $\mathcal{A}^n$ is Noetherian for every natural number n, so one can try to develop the algebraic geometry for

$\mathcal{A}$ along the same lines as for free groups. Otherwise, the treatment of associative algebras should be quite different.

Notice, that if $\mathcal{A}$ is equationally Noetherian, then the free Lie algebra associated with $\mathcal{A}$ is also equationally Noetherian. But whether the converse is true or not is not known. So we pose the following question independently.

Problem 6.18. Are free Lie finitely generated algebras $\mathcal{L}$ equationally Noetherian?

If free associative (Lie) algebras $\mathcal{A}$ are equationally noetherian, then by the unification theorem [18] finitely generated limits of $\mathcal{A}$ are precisely the finitely generated algebras universally equivalent to $\mathcal{A}$. In this case a lot of machinery of universal algebraic geometry is going to work. In any case the following problem is of principle interest.

Problem 6.19. Describe limits of free associative (Lie) algebras.

Unlike free groups, two free associative algebras of finite rank are elementarily equivalent if and only if they are isomorphic (these, and some other relevant facts, can be found in [60]).

The following general Tarski-type problems are of special interest.

Problem 6.20. Let $\mathcal{A}$ be a free associative (Lie) algebra of finite rank.

(1) Prove that if the ground field has decidable theory then the elementary theory $Th(\mathcal{A})$ algebra is also decidable.

(2) Describe a natural system of axioms for $Th(\mathcal{A})$.

(3) Does the elementary theory $Th(\mathcal{A})$ admit elimination of quantifiers (to boolean combinations of $\forall\exists$-sentences and the theory of the ground field)?

(4) Which finitely generated algebras are elementarily equivalent to $\mathcal{A}$?

6.6. Λ-free groups.

Conjecture 6.21. *Every finitely generated Λ-free group is finitely presented.*

This would imply that all the result mentioned in Section 5 hold also in arbitrary finitely generated Λ-free groups.

Conjecture 6.22. *Any finitely presented Λ-free group G is $\mathbb{Z}^k$-free for an appropriate $k \in \mathbb{N}$ and lexicographically ordered $\mathbb{Z}^k$.*

References

[1] I. Agol, *The virtual Haken conjecture. With an appendix by I. Agol, D. Groves, and J. Manning*, Doc. Math. **18** (2013), 1045–1087.

[2] R. Alperin and H. Bass, *Length functions of group actions on Λ-trees.* Combinatorial group theory and topology, (Ed. S. M. Gersten and J. R. Stallings), Annals of Math. Studies **111**, 265–378. Princeton University Press, 1987.

[3] R. Alperin and K. Moss, *Complete trees for groups with a real length function.* J. London Math. Soc. (2) **31** (1985), 55–68.

[4] H. Bass, *Groups acting on non–archimedian trees*, Arboreal group theory, 1991, 69–130.

[5] G. Baumslag, A. G. Myasnikov, and V. Shpilrain, *Open problems in combinatorial group theory*, Second Edition. In Combinatorial and geometric group theory, volume **296** of Contemporary Mathematics, pages 1–38. American Mathematical Society, 2002.

[6] G. Baumslag, A. Myasnikov, and V. Remeslennikov, *Algebraic geometry over groups I, Algebraic sets and ideal theory*, Journal of Algebra, 1999, v. **219**, 16–79.

[7] M. Bestvina and M. Feighn, *Combination theorem for negatively curved groups*, J. Differential Geom. **35** (1992), no. 1, 85–101, Addendum and correction to: "*A combination theorem for negatively curved groups*", J. Differential Geom. **35** (1992), no. 1, 85–101, J. Differential Geom. **43** (1996), no. 4, 783–788.

[8] M. Bestvina and M. Feighn, *Stable actions of groups on real trees*, Invent. Math. **121** no. 2 (1995), 287–321.

[9] R. Bryant, *The verbal topology of a group,* Journal of Algebra **48** (1977), 340–346.

[10] M. Casals-Ruiz and I. Kazachkov, *On Systems of Equations over Free Products of Groups*, Algebra **333** (2011), 368–426.

[11] _____, *On Systems of Equations over Free Partially Commutative Groups*, Memoirs Amer. Math. Soc. **212** (2011), no. 999, viii+153pp.

[12] I. Chiswell, *Abstract length functions in groups*, Math. Proc. Cambridge Philos. Soc., **80** no. 3 (1976), 451–463.

[13] _____, *Introduction to Λ-trees*, World Scientific, 2001.

[14] I. Chiswell and T. Muller, *Embedding theorems for tree-free groups*, Under consideration for publication in Math. Proc. Camb. Phil. Soc.

[15] L. P. Comerford Jr. and C. C. Edmunds, *Solutions of equations in free groups*, Walter de Gruyter, Berlin, New York, 1989.

[16] F. Dahmani, *Combination of convergence groups*, Geom. Topol. **7** (2003), 933–963.

[17] F. Dahmani and D. Groves, *The Isomorphism Problem for Toral Relatively Hyperbolic Groups*, Publ. Math., Inst. Hautes Etudes Sci. **107** no. 1 (2008), 211–290.

[18] E. Daniyarova, A. Miasnikov, and V. Remeslennikov, *Unification theorems in algebraic geometry*, Algebra and Discrete Mathematics, **1** (2008), 80-112, arXiv:0808.2522v1.

[19] V. Diekert and A. Muscholl, *Solvability of equations in graph groups is decidable*, Internat. J. Algebra Comput. **16** (2006), no. 6, 1047–1069.

[20] D. Gaboriau, G. Levitt, and F. Paulin, *Pseudogroups of isometries of $\mathbb{R}$ and Rips' Theorem on free actions on $\mathbb{R}$-trees*, Israel. J. Math. **87** (1994), 403–428.

[21] V. Guba, *Equivalence of infinite systems of equations in free groups and semigroups to finite subsystems*, Mat. Zametki, **40** (1986), 321–324.

[22] D. Gildenhuys, O.Kharlampovich, and A.Myasnikov, *CSA groups and separated free constructions*, Bull. Austr. Math. Soc., **52**(1) (1995), 63–84.

[23] R.I. Grigorchuk and P.F. Kurchanov, *On quadratic equations in free groups*, Contemp. Math., **131**(1) (1992), 159–171.

[24] V. Guirardel, *Limit groups and groups acting freely on $\mathbb{R}^n$-trees*, Geom. Topol. **8** (2004), 1427–1470.

[25] N. Harrison, *Real length funtions in groups*, Trans. Amer. Math. Soc. **174** (1972), 77–106.

[26] A. H. M. Hoare, *On length functions and Nielsen methods in free groups*, J. London Math. Soc. (2) **14** (1976), 188–192.

[27] _____, *Nielsen method in groups with a length function*, Math. Scand. **48** (1981), 153–164.

[28] W. Hodges. *Model theory*, Cambridge University Press, 1993.

[29] M. Jarden and A. Lubotsky, *Elementary equivalence of profinite groups*, Bull. London Math. Soc. (2008) **40**(5), 887–896.

[30] S. Humphries, *On representations of Artin groups and the Tits conjecture*, J. Algebra **169** (1994), 847–862.

[31] A. Koscielski and L. Pacholski, *Makanin's algorithm is not primitive recursive*, Theoretical Comp.Sci. **191**, 1998.

[32] Kourovka Notebook: *Unsolved Problems in Group Theory* (American Mathematical Society Translations Series 2) by Kourovskaia Tetrad. English, L. Ia Leifman and D. J. Johnson (Aug 1983).

[33] O. Kharlampovich and A. Myasnikov, *Irreducible affine varieties over a free group. 1: irreducibility of quadratic equations and Nullstellensatz*, J. of Algebra **200**:472–516, 1998.

[34] _____, *Irreducible affine varieties over a free group. II: Systems in triangular quasi-quadratic form and description of residually free groups*, J. of Algebra v. **200**, no. 2 (1998), 517–570.

[35] _____, *Description of Fully Residually Free Groups and Irreducible Affine Varieties Over a Free Group*, Banff Summer School 1996, Center de Recherchers Matematiques, CRM Proceedings and Lecture Notes, v. **17**, 1999, p.71-80. 571–613.

[36] _____, *Implicit function theorems over free groups*, J. Algebra **290** (2005), 1–203.

[37] _____, *Effective JSJ decompositions,* Group Theory: Algorithms, Languages, Logic, Contemp. Math., AMS, 2004, 87–212 (Math GR/0407089).

[38] _____, *Elementary theory of free non-abelian groups,* J. Algebra **302**, Issue 2, 451–552, 2006.

[39] O. Kharlampovich, A. Myasnikov, V. Remeslennikov, and D. Serbin, *Subgroups of fully residually free groups: algorithmic problems, Group theory, Statistics and Cryptography*, Contemp. Math., Amer. Math. Soc. **360**, 2004, 61–103.

[40] O. Kharlampovich and A. Myasnikov, *Definable sets in a hyperbolic group*, Intern. J. of Algebra and Computation **23** (2013) no 1, 91–110.

[41] ———, *Decidability of the elementary theory of a torsion-free hyperbolic group*, arXiv:1303.0760.

[42] ———, *Limits of relatively hyperbolic groups and Lyndon's completions*, Journal of the European Math. Soc. Volume **14**, Issue 3, 2012, pp. 659–680.

[43] ———, *Equations and fully residually free groups*, Combinatorial and geometric group theory, 203–242, Trends Math., Birkhauser/Springer Basel AG, Basel, 2010.

[44] O. Kharlampovich and J. Macdonald, *Effective embedding of residually hyperbolic groups into direct products of extensions of centralizers*, J. Group Theory **16** (2013), no. 5, 619–650.

[45] O. Kharlampovich, A. Myasnikov, and D. Serbin, *Actions, length functions, and non-archemedian words*, IJAC, 23 (2013), 2, 325–455.

[46] O. Kharlampovich, A. Myasnikov, V. Remeslennikov, and D. Serbin, *Groups with free regular length functions in* $\mathbb{Z}^n$, arXiv:0907.2356, Trans. Amer. Math. Soc., **364**:2847–2882, 2012.

[47] O. Kharlampovich, A. G. Myasnikov, V. N. Remeslennikov, and D. Serbin, *Exponential extensions of groups*, J. Group Theory **11**(1) (2008), 119–140.

[48] O. Kharlampovich, A. Myasnikov, and D. Serbin, *Regular completions of* $\mathbb{Z}^n$*-free groups*, Preprint, 2011.

[49] R. C. Lyndon, *Length functions in groups*, Math. Scand. **12** (1963), 209–234.

[50] ———, *Groups with parametric exponents*, Trans. Amer. Math. Soc. **96** (1960), 518–533.

[51] ———, *Equations in free groups*, Trans. Amer. Math. Soc. **96** (1960), 445–457.

[52] G.S. Makanin, *Equations in a free group (Russian)*, Izv. Akad. Nauk SSSR, Ser. Mat., **46**:1199–1273, 1982. transl. in Math. USSR Izv., V. **21**, 1983; MR 84m:20040.

[53] ———, *Decidability of the universal and positive theories of a free group*, Izv. Akad. Nauk SSSR, Ser. Mat., **48**(1):735–749, 1985. transl. in Math. USSR Izv., V. **25**, 1985; MR 86c:03009.

[54] S.G. Melesheva, *Equations and algebraic geometry over profinite groups*, Algebra and Logika, **49**(5), 2010.

[55] Ju. I. Merzljakov, *Positive formulae on free groups*, Algebra i Logika, **5**(4):25–42, 1966.

[56] A.I. Malcev, *On equation* $zxyx^{-1}y^{-1}z^{-1} = aba^{-1}b^{-1}$ *in a free group*, Algebra and Logic, **1** (1962), 45–50.

[57] A. Myasnikov, V. Remeslennikov, and D. Serbin, *Fully residually free groups and graphs labeled by infinite words*, to appear in IJAC.

[58] A. Myasnikov and V. Remeslennikov, *Recursive p-adic numbers and elementary theories of finitely generated pro-p-groups*, Math USSR Izv, 1988, **30** (3), 577–597.

[59] A. Myasnikov and A. Nikolaev, *Hyperbolic groups with infinite verbal width*, arXiv: 1107.3719.

[60] A. Myasnikov, *The structure of models and a criterion for the decidability of complete theories of finite- dimensional algebras (Russian)*, Izv. Akad. Nauk SSSR Ser. Mat. **53** (1989), no. 2, 379–397; translation in Math. USSR-Izv. **34** (1990), no. 2, 389–407

[61] J. Morgan and P. Shalen, *Valuations, Trees, and Degenerations of Hyperbolic Structures, I*, Annals of Math, 2nd Ser., **120** no. 3. (1984), 401–476.

[62] N. Nikolov and D. Segal, *On finitely generated profinite groups*, I: strong completeness and uniform bounds, Annals of Mathematics, **1**65 (2007), 171–238.

[63] C. Perin, *Elementary embeddings in torsion-free hyperbolic groups*, Annales Scientifiques de Ecole Normale Superieure (4), vol. **44** (2011), 631–681.

[64] ———, *Erratum: Elementary embeddings in torsion-free hyperbolic groups*, preprint, 2012.

[65] D. Promislow, *Equivalence classes of length functions on groups*, Proc. London Math. Soc (3) **51** (1985), 449–477.

[66] A. Razborov, *On systems of equations in a free group*, Math. USSR, Izvestiya, **25**(1) (1985), 115–162.

[67] ———, *On systems of equations in a free group*, PhD thesis, Steklov Math. Institute, Moscow, 1987.

[68] A.H. Rhemtulla, *A problem of bounded expressability in free products*, Proc. Cambridge Philos. Soc., **64** (1968), 573–584.

[69] L. Ribes, P. Zalesskii, *Conjugacy separability of amalgamated free products of groups*, J. Algebra, **179**(3) (1996), 751–774.

[70] D. Y. Rebbechi, *Algorithmic properties of relatively hyperbolic groups*, PhD Thesis, Univ. California, Davis, 2001. http://front.math.ucdavis.edu/math.GR/0302245.

[71] Z. Sela. *Diophantine geometry over groups I: Makanin-Razborov diagrams,* Publications Mathematiques de l'IHES **93** (2001), 31–105,

[72] ———, *Diophantine geometry over groups II-V*, Israel Journal of Math., 2003–2006.

[73] ———, *Diophantine geometry over groups VI: The elementary theory of a free group*, GAFA, **16**(2006), 707–730.

[74] ———, *Diophantine geometry over groups VII, The elementary theory of a hyperbolic group*, Proc. Lond. Math. Soc. **99**(3) (2009), no. 1, 217–273.

[75] ———, *Diophantine Geometry over Groups X: The Elementary Theory of Free Products of Groups*, arXiv:1012.0044.

[76] J.-P. Serre, *Trees*, New York, Springer, 1980.

[77] J.R. Stallings. *Finiteness of matrix representation*, Ann. Math. **124**:337–346, 1986.

[78] D. Wise, *The structure of groups with a quasiconvex hierarchy*, preprint.

Dept. Math. and Stat., Hunter College CUNY, Room 919 East, 695 Park Ave, New York, NY 10065, United States

E-mail: okharlampovich@gmail.com

Dept. Math., Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, NJ 07030, United States

E-mail: amiasnikov@gmail.com

Towards the eigenvalue rigidity of Zariski-dense subgroups

Andrei S. Rapinchuk

Abstract. We discuss the notion of weak commensurability of Zariski-dense subgroups of semi-simple algebraic groups over fields of characteristic zero, which enables one to match in a convenient way the eigenvalues of semi-simple elements of these subgroups. The analysis of weakly commensurable arithmetic groups has led to a resolution of some long-standing problems about isospectral locally symmetric spaces. This work has also initiated a number of questions in the theory of algebraic groups dealing with the characterization of absolutely almost simple simply connected algebraic groups having the same isomorphism classes of maximal tori over the field of definition. The recent results in this direction provide evidence to support a new conjectural form of rigidity for arbitrary Zariski-dense subgroups in absolutely almost simple algebraic groups over fields of characteristic zero based on the eigenvalue information ("eigenvalue rigidity").

Mathematics Subject Classification (2010). Primary 20G15; Secondary 11E72, 53C35.

Keywords. Algebraic groups, Zariski-dense subgroups, locally symmetric spaces.

1. Introduction

The purpose of my talk is two-fold. First, I would like to report on the results obtained in a series of papers written in collaboration with G. Prasad and other co-authors. In these papers, we introduced the notion of *weak commensurability* of Zariski-dense subgroups of semi-simple algebraic groups, determined the consequences of the weak commensurability of two S-arithmetic subgroups of absolutely almost simple algebraic groups over a field of characteristic zero, and applied these results to the analysis of length-commensurable isospectral locally symmetric spaces. Second, I would like to outline a variety of problems and results in the theory of algebraic groups and related areas that this work has led to. These problems have to do with the understanding of finite-dimensional division algebras having the same maximal subfields, and more generally, with the characterization of absolutely almost simple algebraic groups having the same isomorphism classes of maximal tori over the field of definition. The results in this new direction obtained in the last several years point to a new version of the rigidity phenomenon, some aspects of which apply not only in the classical case of lattices but in fact to arbitrary Zariski-dense subgroups. Its distinctive feature is that it is formulated in terms of the eigenvalues of semi-simple elements of a given Zariski-dense subgroup, which led us to call it *eigenvalue rigidity*. Its investigation is very much a work in progress, so along with available results, we will discuss several conjectures. Overall, the possibility of having some form of rigidity for arbitrary Zariski-dense subgroups

Proceedings of the International Congress of Mathematicians, Seoul, 2014

(which may well be free) looks quite exciting, and I would like to begin with a discussion of what kinds of results one can or cannot expect in this generality.

In the theory of algebraic/Lie groups, the term "rigidity" in a very general sense is used to describe a situation where, given a semi-simple algebraic group G over a field F, the structure of a "large" subgroup Γ of $G(F)$ determines the group G as well as the "location" of Γ inside $G(F)$. More concretely, when F is a non-discrete locally compact field, then under appropriate assumptions, any abstract isomorphism $\Gamma_1 \to \Gamma_2$ between two lattices $\Gamma_1, \Gamma_2 \subset G(F)$ extends to a rational automorphism of G (strong rigidity), or even any abstract representation $\Gamma \to \mathrm{GL}_n(F)$ (virtually) extends to a rational representation $G \to \mathrm{GL}_n$ (superrigidity). This implies, for example, that the entire geometry of a compact hyperbolic manifold of dimension $\geqslant 3$ (including its volume, the Laplace spectrum, the lengths of closed geodesics, etc.) is determined by the structure of its fundamental group. Among the algebraic consequences of structural rigidity, the following is most relevant for our discussion.

Let $\Gamma = \mathrm{SL}_n(\mathbb{Z})$, where $n \geqslant 3$, and suppose we are given an absolutely almost simple simply connected algebraic group G over a number field K with ring of integers $\mathcal{O}$. If Γ is (virtually) isomorphic to $G(\mathcal{O})$ as an abstract group, then $K = \mathbb{Q}$ (and hence $\mathcal{O} = \mathbb{Z}$), and $G \simeq \mathrm{SL}_n$ as algebraic groups over $\mathbb{Q}$. Thus, the structure of a higher rank arithmetic group uniquely determines the *field of definition* and the *ambient group* as an algebraic group over this field. The results we will present suggest that one should be able to recover this data (in a somewhat weaker form) not just from a higher rank arithmetic group, but in fact from any finitely generated Zariski-dense subgroup if in place of structural information one uses information about the eigenvalues of elements, expressed in terms of *weak commensurability*. More precisely, we will see that in this set-up the field of definition can still be recovered uniquely (cf. Theorem 3.2), while the ambient algebraic group over this field is conjecturally determined up to finitely many possibilities (cf. Conjecture 6.1). The finiteness is known to hold when the field of definition is a number field, and is supported in the general case by, for example, results on division algebras having the same maximal subfields (cf. 6.5). Moreover, in many situations, S-arithmetic groups are unique (up to commensurability) in their weak commensurability class (cf. Theorem 6.3(1)), and thus are eigenvalue rigid in a strong sense. Just like structural rigidity, eigenvalue rigidity has geometric applications to isospectral locally symmetric spaces (see 2.2 and 4.4). There are other aspects of eigenvalue rigidity dealing with questions of whether various properties of Zariski-dense subgroups (such as discreteness, co-compactness, arithmeticity) can be characterized in terms of the eigenvalue information (see 4.3), but here we will focus almost exclusively on the question of to what extent the latter determines the ambient algebraic group. As we already mentioned, it is precisely shifting the focus from the structure to eigenvalues that makes results of this kind possible for arbitrary Zariski-dense subgroups.

Before discussing the results, we need to explain how we match the eigenvalues of elements of two Zariski-dense subgroups, and on the other hand, why we care about these eigenvalues.

2. Weak commensurability

The following definition, introduced in [40], provides a way of matching the eigenvalues of matrices of different sizes.

Definition 2.1. Let F be an infinite field.

(1) Let $\gamma_1 \in \mathrm{GL}_{n_1}(F)$ and $\gamma_2 \in \mathrm{GL}_{n_2}(F)$ be *semi-simple matrices*, and let

$$\lambda_1, \dots, \lambda_{n_1} \quad \text{and} \quad \mu_1, \dots, \mu_{n_2}$$

be their eigenvalues (in a fixed algebraic closure $\overline{F}$). We say that γ_1 and γ_2 are *weakly commensurable* if there exist $a_1, \dots, a_{n_1}$, $b_1, \dots, b_{n_2} \in \mathbb{Z}$ such that

$$\lambda_1^{a_1} \cdots \lambda_{n_1}^{a_{n_1}} \;=\; \mu_1^{b_1} \cdots \mu_{n_2}^{b_{n_2}} \;\neq\; 1.$$

(2) Let $G_1 \subset \mathrm{GL}_{n_1}$ and $G_2 \subset \mathrm{GL}_{n_2}$ be reductive algebraic groups defined over F. Two *Zariski-dense subgroups* $\Gamma_1 \subset G_1(F)$ and $\Gamma_2 \subset G_2(F)$ are called *weakly commensurable* if every semi-simple element $\gamma_1 \in \Gamma_1$ of infinite order is weakly commensurable to some semi-simple element $\gamma_2 \in \Gamma_2$ of infinite order, and vice versa.

It should be noted that the definition of weak commensurability can be stated in several different ways. First, in the above notations, semi-simple elements $\gamma_1 \in G_1(F)$ and $\gamma_2 \in G_2(F)$ are weakly commensurable if and only if there exist maximal F-tori T_i of G_i for $i = 1, 2$ such that $\gamma_i \in T_i(F)$ and for some characters $\chi_i \in X(T_i)$ (defined over $\overline{F}$) we have

$$\chi_1(\gamma_1) = \chi_2(\gamma_2) \neq 1.$$

This reformulation shows that the notion of weak commensurability (of γ_1 and γ_2) does not depend on the choice of matrix realizations of G_1 and G_2, and is also more convenient for technical arguments.

Second, semi-simple elements $\gamma_1 \in G_1(F)$ and $\gamma_2 \in G_2(F)$ are weakly commensurable if and only if there exist F-rational representations

$$\rho_1 \colon G_1 \longrightarrow \mathrm{GL}_{m_1} \text{ and } \rho_2 \colon G_2 \longrightarrow \mathrm{GL}_{m_2}$$

such that $\rho_1(\gamma_1)$ and $\rho_2(\gamma_2)$ have a *nontrivial common eigenvalue* (these representations can vary from one element to another).

Informally speaking, weak commensurability appears to be a rather natural way (and perhaps even the only natural way) of matching the eigenvalues of (semi-simple) elements of two algebraic groups that does not depend on the choice of their matrix realizations. On the other hand, it is easy to construct examples of very different (certainly non-conjugate) matrices that are weakly commensurable, so one needs to discuss the utility of this notion. As we will see later, while being inconsequential for individual matrices and "small" (e.g., cyclic) subgroups, weak commensurability has remarkably strong consequences for "large" subgroups (viz., Zariski-dense and particularly S-arithmetic subgroups). Now, however, we would like to point out that the main motivation for the notion of weak commensurability in our work came from the famous problem in differential geometry about isospectral Riemannian manifolds best known as M.Kac's [30] question *Can one hear the shape of a drum?*

2.2. Geometric motivation. Let M be a Riemannian manifold. In differential geometry one considers the following sets of data associated with M:

- $\mathcal{E}(M)$ - spectrum of the Beltrami-Laplace operator;

- $L(M)$ - (weak) length spectrum, i.e. the collection of lengths of all closed geodesics in M.

Then one asks whether two Riemannian manifolds M_1 and M_2 are necessarily isometric if

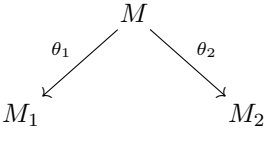
(1) $\mathcal{E}(M_1) = \mathcal{E}(M_2)$ (i.e., M_1 and M_2 are *isospectral*);

(2) $L(M_1) = L(M_2)$ (i.e., M_1 and M_2 are *iso-length spectral*)?

When asking a question of this kind, one of course needs to specialize the class of manifolds being considered, and in our work we focused on locally symmetric spaces of semi-simple groups having nonpositive curvature (recall that these are endowed with the standard Riemannian structure coming from the Killing form); this class includes such geometrically important spaces as hyperbolic manifolds and, in particular, Riemann surfaces. It is important to point out that for *compact* locally symmetric spaces, questions (1) and (2) are *related*, viz.

$$\mathcal{E}(M_1) \;=\; \mathcal{E}(M_2) \;\;\Rightarrow\;\; L(M_1) \;=\; L(M_2), \tag{S}$$

but *both* have a negative answer. Counter-examples for (arithmetically defined) Riemann surfaces were given by Vigneras [53], and then a more general group-theoretic construction was offered by Sunada [50]. Both constructions always produce pairs of *commensurable* locally symmetric spaces. We recall that Riemannian manifolds M_1 and M_2 are called *commensurable* if they admit a common finite-sheeted cover M, i.e. if there is a diagram:



in which M is a Riemannian manifold and θ_1, θ_2 are finite-sheeted locally isometric covering maps. This suggests that one should probably settle for a weaker version of the question, viz. whether M_1 and M_2 are necessarily *commensurable* given the fact that they are isospectral or iso-length-spectral. While this modified question still has a negative answer in the general case [35], our work, based on the analysis of weakly commensurable groups, shows that the answer is in the affirmative for many (arithmetically defined) locally symmetric spaces - cf. Theorem 4.5 (previously such results were available only for arithmetically defined Riemann surfaces [47] and hyperbolic 3-manifolds [12]). In fact, our results give the commensurability of pairs of locally symmetric spaces that satisfy the following condition:

(3) $\mathbb{Q} \cdot L(M_1) = \mathbb{Q} \cdot L(M_2)$.

This condition, called *length commensurability*, is conceivably much weaker than conditions (1) and (2), but surprisingly in most situations it has many of the same consequences. Its real advantage over (1) and (2) is that it is invariant under passing to commensurable manifolds.

The main point here is that the length-commensurability of finite volume locally symmetric spaces implies the weak commensurability of their fundamental groups. To give a precise statement, we need to fix some notations. Let G be an absolutely simple adjoint real algebraic group, let $\mathcal{G} = G(\mathbb{R})$ be the group of $\mathbb{R}$-points, considered as a real Lie group, and let $\mathfrak{X} = \mathcal{K}\backslash\mathcal{G}$, where $\mathcal{K}$ is a maximal compact subgroup of $\mathcal{G}$, be the associated symmetric space endowed with the Riemannian metric coming from the Killing form on the Lie algebra of $\mathcal{G}$. Furthermore, given a torsion-free discrete subgroup Γ of $\mathcal{G}$, we let $\mathfrak{X}_\Gamma = \mathfrak{X}/\Gamma$ denote

the corresponding locally symmetric space; we say that $\mathfrak{X}_\Gamma$ is *arithmetically defined* if the subgroup Γ is *arithmetic*[1]. Finally, given two simple real algebraic groups G_i $(i = 1, 2)$, we will denote the symmetric spaces of the groups $\mathcal{G}_i = G_i(\mathbb{R})$ by $\mathfrak{X}_i$, and the locally symmetric spaces obtained as quotients by torsion-free discrete subgroups Γ_i of $\mathcal{G}_i$ by $\mathfrak{X}_{\Gamma_i}$.

Theorem 2.3 ([43], Corollary 2.8). *Let $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ be two locally symmetric spaces having finite volume, of absolutely simple real algebraic groups G_1 and G_2. If $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ are length-commensurable, then Γ_1 and Γ_2 are weakly commensurable.*

While this result is straightforward for Riemann surfaces (see [43, 2.1]), its proof in the general case relies on the formula for the length of a closed geodesic c_γ in $\mathfrak{X}_\Gamma$ corresponding to a nontrivial semi-simple element $\gamma \in \Gamma$ as a function of the logarithms of eigenvalues of γ in the adjoint representation - see [40, Proposition 8.5(ii)] (note that this formula also explains why we care about the eigenvalues of semi-simple elements of discrete subgroups). So, to prove the weak commensurability of Γ_1 and Γ_2, we need to sort out the logarithms appearing in this formula, which requires transcendental number theory. More precisely, for rank one locally symmetric spaces of dimension > 2, we use the famous result of Gel'fond and Schneider that settled Hilbert's seventh problem - cf. [4]. In all other cases, our argument assumes the truth of Schanuel's conjecture (cf. [3]). This means that while all of our results on weak commensurability are, of course, unconditional, their geometric consequences are *conditional* (at least for locally symmetric spaces of rank > 1).

Since the locally symmetric spaces $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ are commensurable if and only if the subgroups Γ_1 and Γ_2 are commensurable as groups up to an isomorphism between G_1 and G_2 (see 3.4 below for the details of this notion), we see that in order to prove the commensurability of length-commensurable (in particular, isospectral or iso-length spectral) locally symmetric space, we need to answer the following question:

(C) *When does the weak commensurability of Γ_1 and Γ_2 imply their commensurability?*

3. First signs of eigenvalue rigidity

Before providing a rather definitive answer to Question (C) for S-arithmetic subgroups (see §4), we would like to present a few results demonstrating that weak commensurability captures some important characteristics in the case of arbitrary Zariski-dense subgroups. So, let G_1 and G_2 be absolutely almost simple algebraic groups over a field F of *characteristic zero*, and let $\Gamma_i \subset G_i(F)$ be a *finitely generated* Zariski-dense subgroup for $i = 1, 2$.

Theorem 3.1 ([40], Theorem 1). *If Γ_1 and Γ_2 are weakly commensurable, then either G_1 and G_2 are of the same Killing-Cartan type, or one of them is of type B_ℓ and the other of type C_ℓ for some $\ell \geqslant 3$.*

This result is already interesting because, in principle, Γ_1 and Γ_2 may very well be free groups, hence carry no structural information about the ambient algebraic groups. Note that what we really prove is that G_1 and G_2 have the same order of the Weyl group - it

[1]We recall that combining the celebrated results of Margulis [36] on the arithmeticity of higher rank irreducible lattices and of Corlette [15] and Gromov-Shoen [27], one obtains that a finite volume locally symmetric space $\mathfrak{X}_\Gamma$ of a simple real algebraic group is automatically arithmetically defined unless $\mathfrak{X}$ is either the real hyperbolic space $\mathbb{H}^n$ or the complex hyperbolic space $\mathbb{H}^n_{\mathbb{C}}$.

is known that this number uniquely determines the Killing-Cartan type of the group except for the ambiguity involving types B_ℓ and C_ℓ. As shown by Theorem 4.2 below, Zariski-dense, and even S-arthmetic, subgroups in groups of types B_ℓ and C_ℓ can indeed be weakly commensurable.

Now, given a Zariski-dense subgroup $\Gamma \subset G(F)$, where G is a semi-simple F-group, we let K_Γ denote the *trace field* of Γ, i.e. the subfield of F generated by the traces $\mathrm{tr}(\mathrm{Ad}\,\gamma)$ of all elements $\gamma \in \Gamma$ in the adjoint representation on the corresponding Lie algebra $\mathfrak{g} = L(G)$. By a result of Vinberg [51], the field $K = K_\Gamma$ is the minimal field of definition of $\mathrm{Ad}\ \Gamma$. This means that K is the minimal subfield of F such that all transformations in $\mathrm{Ad}\ \Gamma$ can be simultaneously represented by matrices over K in a suitable basis of $\mathfrak{g}$. If such a basis is chosen, then the Zariski closure of $\mathrm{Ad}\ \Gamma$ in $\mathrm{GL}(\mathfrak{g})$ is a semi-simple algebraic K-group $\mathcal{G}$. It is an F/K-form of the adjoint group $\overline{G}$, and we will call it the *algebraic hull* of $\mathrm{Ad}\ \Gamma$.

Theorem 3.2 ([40], Theorem 2). *Keep the notations and conventions introduced prior to Theorem 3.1. If Γ_1 and Γ_2 are weakly commensurable, then $K_{\Gamma_1} = K_{\Gamma_2}$.*

Now let K be the common trace field of two weakly commensurable Zariski-dense subgroups Γ_1 and Γ_2 as above, and let $\mathcal{G}_i$ be the algebraic hull of $\mathrm{Ad}\ \Gamma_i$ for $i = 1, 2$. We denote by L_i the minimal Galois extension of K over which $\mathcal{G}_i$ becomes an inner form of a split group.

Proposition 3.3 (cf. [44], Lemma 5.2). *In the above notations, $L_1 = L_2$.*

(We would like to mention the following useful consequence of this proposition: *Let G_1 and G_2 be absolutely almost simple groups over a field F of characteristic zero, and let E_i be the minimal Galois extension of F over which G_i becomes an inner form. If there exist finitely generated Zariski-dense subgroups $\Gamma_1 \subset G_1(F)$ and $\Gamma_2 \subset G_2(F)$ that are weakly commensurable, then $E_1 = E_2$.* Indeed, $E_i = FL_i$ in the above notations.)

3.4. *S*-arithmetic subgroups. We will now specialize to the case of S-arithmetic subgroups. We recall that if G is an absolutely almost simple algebraic group over a field F of characteristic zero, then Zariski-dense S-arithmetic subgroups of $G(F)$ can be described in terms of triples $(\mathcal{G}, K, S)$, where K is a number field contained in F, $\mathcal{G}$ is a F/K-form of the adjoint group $\overline{G}$, and S is a finite set of places of K containing all archimedean ones; the subgroups corresponding to such triples are called $(\mathcal{G}, K, S)$-*arithmetic*. We refer to [40, §1] and [43, 3.3] for the details of this description, and only indicate here that for a $(\mathcal{G}, K, S)$-arithmetic Zariski-dense subgroup Γ, the field K coincides with the trace field K_Γ, and the group $\mathcal{G}$ with the algebraic hull of $\mathrm{Ad}\ \Gamma$.

Furthermore, given two absolutely almost simple F-groups G_1 and G_2, we say that the subgroups $\Gamma_1 \subset G_1(F)$ and $\Gamma_2 \subset G_2(F)$ are *commensurable up to an F-isomorphism* between the adjoint groups $\overline{G}_1$ and $\overline{G}_2$ if there exists an F-isomorphism $\sigma\colon \overline{G}_1 \to \overline{G}_2$ such that the subgroups $\sigma(\pi_1(\Gamma_1))$ and $\pi_2(\Gamma_2)$ are commensurable as subgroups of $\overline{G}_2(F)$ in the usual sense (i.e., their intersection is of finite index in each of them), where $\pi_i\colon G_i \to \overline{G}_i$ is the canonical isogeny for $i = 1, 2$. (This notion of commensurability is precisely what we need for geometric applications, cf. §2.)

The following result shows that the description of S-arithmetic subgroups of absolutely almost simple groups in terms of triples is very convenient in the analysis of their commensurability.

Theorem 3.5. *Let G_1 and G_2 be absolutely almost simple algebraic groups defined over a field F of characteristic zero, and for $i = 1, 2$, let Γ_i be a Zariski-dense $(\mathcal{G}_i, K_i, S_i)$-arithmetic subgroup of $G_i(F)$. Then*

(1) *Γ_1 and Γ_2 are commensurable up to an F-isomorphism between $\overline{G}_1$ and $\overline{G}_2$ if and only if $K_1 = K_2$, $S_1 = S_2$, and $\mathcal{G}_1$ and $\mathcal{G}_2$ are K-isomorphic;*

(2) *if Γ_1 and Γ_2 are weakly commensurable, then $K_1 = K_2$ and $S_1 = S_2$.*

Thus, the study of the commensurability classes of weakly commensurable Zariski-dense S-arithmetic subgroups *is equivalent* to the study of K-forms $\mathcal{G}$ involved in their description. This leads to a complete resolution of question (C) for S-arithmetic subgroups that we will present in the next section.

4. Results for S-arithmetic groups and geometric consequences

The following two theorems summarize the main results dealing with the weak commensurability of S-arithmetic subgroups.

Theorem 4.1 (cf. Prasad-Rapinchuk [40, 43]). *Let G_1 and G_2 be absolutely almost simple algebraic groups over a field F of characteristic zero, and let $\Gamma_1 \subset G_1(F)$ and $\Gamma_2 \subset G_2(F)$ be Zarsiki-dense S-arithmetic subgroups.*

(1) *Assume that G_1 and G_2 are of the same Killing-Cartan type, which is different from A_n, D_{2n+1} $(n > 1)$, and E_6. If Γ_1 and Γ_2 are weakly commensurable, then they are commensurable.*

(2) *In all cases, S-arithmetic subgroups $\Gamma_2 \subset G_2(F)$ weakly commensurable to a given S-arithmetic subgroup $\Gamma_1 \subset G_1(F)$ form* finitely many *commensurability classes.*

(3) *If Γ_1 and Γ_2 as above are weakly commensurable, then Γ_1 contains unipotent elements if and only if Γ_2 does.*

(4) (arithmeticity theorem) *Let now F be a locally compact field of characteristic zero, and $\Gamma_1 \subset G_2(F)$ be an S-arithmetic lattice. If $\Gamma_2 \subset G_2(F)$ is a lattice weakly commensurable to Γ_1, then Γ_2 is also S-arithmetic.*

(In this theorem, "commensurability" means "commensurability up to an F-isomorphism between $\overline{G}_1$ and $\overline{G}_2$" as defined in 3.4.)

An interesting feature of this theorem is that for groups of type D_n, the answer to Question (C) is different depending on whether n is even or odd. Assertion (1) for type D_{2n} with $n > 2$ was originally proved in [41]. The case of type D_4 (including triality forms) was handled by Garibaldi [21] by a different method which applies to all groups of type D_{2n}. On the other hand, for each of the exceptional types A_n $(n > 1)$, D_{2n+1} $(n > 1)$, and E_6 one can construct weakly commensurable, but not commensurable, Zariski-dense S-arithmetic subgroups (see [40, §9])[2].

According to Theorem 3.1, to complete the investigation of weak commensurability for S-arithmetic subgroups, it remains to consider the case where one group is of type B_ℓ and the other of type C_ℓ for some $\ell \geqslant 3$.

[2]Note that these are precisely the types for which the multiplication by (-1) considered as an automorphism of the corresponding root system is *not* in the Weyl group.

Theorem 4.2 (Garibaldi-Rapinchuk [22]). *Let G_1 and G_2 be absolutely almost simple algebraic groups over a field F of characteristic zero having Killing-Cartan types B_ℓ and C_ℓ ($\ell \geqslant 3$), respectively, and let Γ_i be a Zariski-dense $(\mathcal{G}_i, K, S)$-arithmetic subgroup of $G_i(F)$. Then Γ_1 and Γ_2 are weakly commensurable if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ are* twins*, i.e.*

- *$\mathcal{G}_1$ and $\mathcal{G}_2$ are both split over all nonarchimedean places of K;*
- *$\mathcal{G}_1$ and $\mathcal{G}_2$ are* simultaneously *either split or anisotropic over all archimedean valuations of K.*

In §5, we will review some of the techniques involved in the proof of Theorems 4.1 and 4.2. But from a very general perspective, the essence of the argument is to obtain information about the algebraic hull $\mathcal{G}$ of an S-arithmetic group Γ that is weakly commensurable to a given S-arithmetic group – recall that according to Theorem 3.5, $\mathcal{G}$ uniquely determines Γ up to commensurability. So, to establish assertion (1) of Theorem 4.1, we prove that the algebraic hull $\mathcal{G}$ is itself unique when the type if different from A_n, D_{2n+1} $(n > 1)$, and E_6. Furthermore, for assertion (2), we prove that there are only finitely many possibilities for the $\mathcal{G}$'s. In §§6-7 we will indicate that the latter property is expected to hold not only for S-arithmetic, but in fact for arbitrary finitely generated Zariski-dense subgroups (see Conjecture 6.2). This phenomenon, if confirmed, would be a rather strong form of eigenvalue rigidity. We will now, however, briefly discuss a few other questions that one can ask in the context of weak commensurability.

4.3. Some other aspects of eigenvalue rigidity. First, it is easy to construct examples showing that a Zariski-dense subgroup weakly commensurable to a *rank-one* arithmetic subgroup need not be arithmetic (see [40, Remark 5.5]); in other words, assertion (4) of Theorem 4.1 fails if we drop the assumption that Γ_2 is a lattice. So, it would be extremely interesting to determine if a Zariski-dense subgroup weakly commensurable to a *higher rank* S-arithmetic subgroup is itself S-arithmetic (see Problem 10.1 in [43] and the subsequent discussion). This can potentially provide a new characterization of higher rank S-arithmetic subgroups involving weak commensurability (i.e., ultimately the eigenvalue information).

Second, one can ask if weak commensurability can be used to characterize the discreteness of Zariski-dense subgroups. More precisely, let G_1 and G_2 be connected absolutely almost simple algebraic groups over a nondiscrete locally compact field F, and let Γ_i be a finitely generated Zariski-dense subgroup of $G_i(F)$ for $i = 1, 2$. Assume that Γ_1 and Γ_2 are weakly commensurable. *Does the discreteness of Γ_1 imply the discreteness of Γ_2?* (Problem 10.2 in [43]). An affirmative answer to this question was given in [40, Proposition 5.6] for the case where F is a *nonarchimedean* local field, but the case $F = \mathbb{R}$ or $\mathbb{C}$ remains open.

Third, one can also ask if weak commensurability preserves cocompactness of lattices. Namely, let again G_1 and G_2 be connected absolutely almost simple algebraic groups over $F = \mathbb{R}$ or $\mathbb{C}$, and let $\Gamma_i \subset G_i(F)$ be a lattice for $i = 1, 2$. Assume that Γ_1 and Γ_2 are weakly commensurable. *Does the compactness of $G_1(F)/\Gamma_1$ imply the compactness of $G_2(F)/\Gamma_2$?* (Problem 10.3 in [43]). (We note that if G is a semi-simple algebraic group over a nonarchimedean local field F of characteristic zero, then any lattice $\Gamma \subset G(F)$ is automatically cocompact, and the problem in this case becomes vacuous.) We recall that the cocompactness of a lattice in a semi-simple real Lie group is equivalent to the absence of nontrivial unipotent elements in it, see [45, Corollary 11.13]. So, the above question is equivalent to the question of whether for two weakly commensurable *lattices*, the existence of nontrivial unipotent elements in one of them implies their existence in the other. The

combination of parts (3) and (4) of Theorem 4.1 provides an affirmative answer if one of the lattices is arithmetic. On the other hand, in this form the question itself is meaningful for arbitrary Zariski-dense subgroups (not necessarily discrete or of finite covolume), but no other cases have been considered so far.

4.4. Geometric applications. Combining Theorem 2.3, which reduced the length - commensurability of locally symmetric spaces to the weak commensurability of their fundamental groups, with the analysis of weak commensurability in Theorem 4.1, we obtain the following geometric result.

Theorem 4.5 ([40], Theorem 8.16). *Let G_1 and G_2 be connected absolutely simple real algebraic groups, and set $\mathcal{G}_i = G_i(\mathbb{R})$, for $i = 1, 2$. Then the set of arithmetically defined locally symmetric spaces $\mathfrak{X}_{\Gamma_2}$ of $\mathcal{G}_2$, which are length-commensurable to a given arithmetically defined locally symmetric space $\mathfrak{X}_{\Gamma_1}$ of $\mathcal{G}_1$, is a union of finitely many commensurability classes. It in fact consists of a single commensurability class if G_1 and G_2 have the same type different from A_n, D_{2n+1}, with $n > 1$, or E_6.*

This statement applies in various concrete geometric situations. For example, here is what it yields for hyperbolic manifolds.

Corollary 4.6. *Let M_1 and M_2 be arithmetically defined real hyperbolic d-manifolds where d is either even or is $\equiv 3 (\mathrm{mod}\, 4)$ and $d > 3$. If M_1 and M_2 are length-commensurable (in particular, compact and isospectral), then they are commensurable.*

Previously, this was known only for $d = 2$ [47] and $d = 3$ [12]. Length-commensurability implies commensurability also for all quaternionic hyperbolic manifolds. On the other hand, in the case of real hyperbolic manifolds of dimension $\equiv 1 (\mathrm{mod}\, 4)$ or of complex hyperbolic manifolds, one can construct examples of arithmetically defined length-commensurable, but not commensurable spaces. Furthermore, using Theorem 3.1 (and Proposition 3.3 to handle the isomorphism $\mathsf{A}_3 = \mathsf{D}_3$), one proves that an arithmetically defined complex hyperbolic space cannot be length-commensurable to either a real or a quaternionic arithmetically defined hyperbolic space. Employing Theorem 4.2, one also proves that arithmetically defined real and quaternionic hyperbolic spaces cannot be length-commensurable. (In fact, assuming Schanuel's conjecture in all cases, one can get rid of the arithmeticity assumption in these two statement, see [42], particularly Remark 8.5, and the discussion after Theorem 4.8 below.)

Next, parts (3) and (4) of Theorem 4.1, in conjunction with Theorem 2.3, imply the following rather surprising result which has so far defied all attempts to find a purely geometric proof.

Theorem 4.7 ([40], Theorem 8.19). *Let $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ be locally symmetric spaces of finite volume. Assume that one of the spaces is arithmetically defined. If the spaces are length-commensurable, then the other space is also arithmetically defined, and the compactness of one of the spaces implies the compactness of the other.*

In fact, if one of the spaces is compact and the other is not, the length spectra $L(\mathfrak{X}_{\Gamma_1})$ and $L(\mathfrak{X}_{\Gamma_2})$ are quite different - see [42, Theorem 5.9]. The question of whether the arithmeticity assumption in this theorem can be dropped boils down to one of the problems we discussed in 4.3.

Last but not least, implication (S) from 2.2 enables us to apply the above results to isospectral compact locally symmetric spaces. We then obtain the following.

Theorem 4.8 ([40], §10). *Let $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ be compact locally symmetric spaces, and assume that they are isospectral.*

(1) *If $\mathfrak{X}_{\Gamma_1}$ is arithmetically defined, then $\mathfrak{X}_{\Gamma_2}$ is also arithmetically defined.*

(2) *$G_1 = G_2 =: G$, hence $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ have the same universal cover.*

(3) *If at least one of the subgroups Γ_1 and Γ_2 is arithmetic, then unless G is of type A_n $(n > 1)$, D_{2n+1} $(n > 1)$ and E_6, the spaces $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ commensurable.*

We note that part (2) was proved in [40] (with the help of a result of Sai-Kee Yeung [55]) under the assumption that at least one of the groups Γ_1 or Γ_2 is arithmetic. Suppose now that both Γ_1 and Γ_2 are nonarithmetic. Then each space $\mathfrak{X}_i$ $(i = 1, 2)$ is either the real hyperbolic space $\mathbb{H}^{n_i}$ or the complex hyperbolic space $\mathbb{H}^{n_i}_{\mathbb{C}}$ for some $n_i \geqslant 2$, and the corresponding real adjoint algebraic group G_i is, respectively, either $\mathrm{PSO}(n_i, 1)$ or $\mathrm{PSU}(n_i, 1)$ in the standard notations. It follows from Theorem 3.1 that the isospectrality, hence length-commensurability, of $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ implies that either G_1 and G_2 must be of the same Cartan-Killing type, or one of them is of type B_ℓ and the other of type C_ℓ for some $\ell \geqslant 3$. In our situation, this can happen only if either $G_1 = G_2$ or (after a possible switch) $G_1 = \mathrm{PSO}(5, 1)$ and $G_2 = \mathrm{PSU}(3, 1)$ (of common type $\mathsf{D}_3 = \mathsf{A}_3$). In the latter case, $\mathfrak{X}_{\Gamma_1}$ is 5-dimensional, and $\mathfrak{X}_{\Gamma_2}$ is 6-dimensional. But according to Weyl's Law (see, for example, [24]) isospectral Riemannian manifolds are always of the same dimension. So, in this case $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ cannot be isospectral[3], leaving us with the only option $G_1 = G_2$, as required.

5. Generic elements. Isogeny Theorem

In this section, we would like to discuss two ingredients involved in the proofs of Theorems 4.1 and 4.2: the existence of generic elements in Zariski-dense subgroups and the Isogeny Theorem.

5.1. Generic elements. First, we need to recall the notion of a *generic K-torus*. Let G be a connected semi-simple algebraic group defined over an infinite field K. Fix a maximal K-torus T of G, and, as usual, let $\Phi = \Phi(G, T)$ denote the corresponding root system, and let $W(G, T)$ be its Weyl group. Furthermore, we let K_T denote the (minimal) splitting field of T in a fixed algebraic closure $\overline{K}$ of K. Then the natural action of the Galois group $\mathrm{Gal}(K_T/K)$ on the character group $X(T)$ of T induces an injective homomorphism

$$\theta_T \colon \mathrm{Gal}(K_T/K) \to \mathrm{Aut}(\Phi(G, T)).$$

We say that T is *generic* (over K) if

$$\theta_T(\mathrm{Gal}(K_T/K)) \supset W(G, T). \tag{5.1}$$

For example, any maximal K-torus of $G = \mathrm{SL}_{n,K}$ is of the form $T = \mathrm{R}^{(1)}_{E/K}(\mathbb{G}_{m,E})$ for some n-dimensional commutative étale K-algebra E. Then such a torus is generic over K if and only if E is a separable field extension of K and the Galois group of the normal closure L of E over K is isomorphic to the symmetric group S_n.

[3] In fact, it follows from the remark made after Proposition 3.3 that in this case $\mathfrak{X}_{\Gamma_1}$ and $\mathfrak{X}_{\Gamma_2}$ cannot even be length-commensurable as G_1 is an inner form of a split group over $\mathbb{R}$, and G_2 is an outer form.

Definition 5.2. Let G be a connected semi-simple algebraic group defined over a field K. A regular semi-simple element $g \in G(K)$ is called *generic* (over K) if the torus $T = Z_G(g)^\circ$ is generic (over K) as defined above (note that T is a K-torus, cf. [7, 9.1]).

Generic elements play a crucial role in our work, but they have also been used in a variety of other problems, including the study of the rigidity of actions (cf. [32, 37]) and the Auslander problem [1].

Theorem 5.3 (cf. [39], Theorem 3). *Let G be a connected absolutely almost simple algebraic group over a finitely generated field K of characteristic zero, and let $\Gamma \subset G(K)$ be a Zariski-dense subgroup. Then Γ contains a regular generic element (over K) of infinite order.*

Basically, our proof (which in fact applies to all connected semi-simple groups) shows that given a finitely generated Zariski-dense subgroup $\Gamma \subset G(K)$, one can produce a finite system of congruences (defined in terms of suitable valuations of K) such that the set of elements $\gamma \in \Gamma$ satisfying this system of congruences consists entirely of generic elements (and additionally this set is in fact a coset of a finite index subgroup in Γ, in particular, it is Zariski-dense in G). Recently, Gorodnik-Nevo [26], Jouve-Kowalski-Zywina [29], and Lubotzky-Rosenzweig [34] have developed different *quantitative* ways of showing that generic elements exist in abundance (in fact, these results demonstrate that "most" elements in Γ are generic). More precisely, the result of [26] gives the asymptotics of the number of generic elements of a given height in an arithmetic group, while the results of [34], generalizing earlier results of [29], are formulated in terms of random walks on groups and apply to arbitrary Zariski-dense subgroups in not necessarily connected semi-simple groups. These papers introduce several new ideas and techniques, but at the same time employ some elements of the argument from [39]. We also note that the proof of Theorems 4.1 and 4.2 uses not only Theorem 5.3 itself but also its different variants that provide generic elements with additional properties, e.g. having prescribed local behavior.

5.4. The Isogeny Theorem and its consequences. An important step in the proofs of Theorems 4.1 and 4.2 is the passage from the weak commensurability of two semi-simple elements to an isogeny, and in most cases even to an isomorphism, of the tori containing these elements. This is done with the help of the following technical statement which we called the *Isogeny Theorem*. After the theorem, we give a (less technical) corollary that, to a significant degree, reduces the analysis of weak commensurability to the investigation of absolutely almost simple algebraic groups having the same isomorphism/isogeny classes of maximal tori over the base field; this problem, together with some variations, will be discussed in the concluding §§6-7. We recall that a K-torus T is called (K-)*irreducible* if it does not contain any proper K-subtori; note that a maximal K-torus of an absolutely almost simple algebraic K-group which is generic over K, is automatically K-irreducible.

Theorem 5.5 ([40], Theorem 4.2). *Let G_1 and G_2 be two connected absolutely almost simple algebraic groups defined over an infinite field K, and let L_i be the minimal Galois extension of K over which G_i becomes an inner form of a split group. Suppose that for $i = 1, 2$, we are given a semi-simple element $\gamma_i \in G_i(K)$ contained in a maximal K-torus T_i of G_i. Assume that*

(i) *G_1 and G_2 are either of the same Killing-Cartan type, or one of them is of type B_n and the other is of type C_n;*

(ii) γ_1 *has infinite order;*

(iii) T_1 *is* K*-irreducible; and*

(iv) γ_1 *and* γ_2 *are weakly commensurable.*

Then:

(1) *there exists a* K*-isogeny* $\pi\colon T_2 \to T_1$ *which carries* $\gamma_2^{m_2}$ *to* $\gamma_1^{m_1}$ *for some integers* $m_1, m_2 \geqslant 1$*;*

(2) *if* $L_1 = L_2 =: L$ *and* $\theta_{T_1}(\mathrm{Gal}(L_{T_1}/L)) \supset W(G_1, T_1)$*, then*

$$\pi^*\colon X(T_1) \otimes_{\mathbb{Z}} \mathbb{Q} \to X(T_2) \otimes_{\mathbb{Z}} \mathbb{Q}$$

has the property that $\pi^*(\mathbb{Q} \cdot \Phi(G_1, T_1)) = \mathbb{Q} \cdot \Phi(G_2, T_2)$*. Moreover, if* G_1 *and* G_2 *are of the same Killing-Cartan type different from* $\mathsf{B}_2 = \mathsf{C}_2$*,* F_4 *or* G_2*, then a suitable rational multiple of* π^* *maps* $\Phi(G_1, T_1)$ *onto* $\Phi(G_2, T_2)$*, and if* G_1 *is of type* B_n *and* G_2 *is of type* C_n*, with* $n > 2$*, then a suitable rational multiple* λ *of* π^* *takes the long roots in* $\Phi(G_1, T_1)$ *to the short roots in* $\Phi(G_2, T_2)$*, while* 2λ *takes the short roots in* $\Phi(G_1, T_1)$ *to the long roots in* $\Phi(G_2, T_2)$*.*

It follows that in the situations where π^* can be, and has been, scaled so that $\pi^*(\Phi(G_1, T_1)) = \Phi(G_2, T_2)$, it induces K-isomorphisms $\tilde{\pi}\colon \widetilde{T}_2 \to \widetilde{T}_1$ and $\overline{\pi}\colon \overline{T}_2 \to \overline{T}_1$ between the corresponding tori in the simply connected and adjoint groups $\widetilde{G}_i$ and $\overline{G}_i$, respectively, that extend to $\overline{K}$-isomorphisms $\widetilde{G}_2 \to \widetilde{G}_1$ and $\overline{G}_2 \to \overline{G}_1$.

Furthermore, if G_1 and G_2 are absolutely almost simple algebraic groups over a field K of characteristic zero and $\Gamma_1 \subset G_1(K)$ and $\Gamma_2 \subset G_2(K)$ are weakly commensurable finitely generated Zariski-dense subgroups, then we already know that either G_1 and G_2 have the same type or one of them is of type B_ℓ and the other of type C_ℓ for some $\ell \geqslant 3$ (Theorem 3.1), and $L_1 = L_2$ (remark after Proposition 3.3). Thus, the important assumptions in Theorem 5.5 are satisfied automatically, and its application yields the following.

Corollary 5.6. *In the above situation, every generic maximal* K*-torus* T_1 *of* G_1 *whose intersection with* Γ_1 *contains an element of infinite order, is* K*-isogenous, and if both* G_1 *and* G_2 *are either simply connected or adjoint of the same type different from* $\mathsf{B}_2 = \mathsf{C}_2$*,* F_4 *and* G_2*, even* K*-isomorphic, to a generic maximal* K*-torus* T_2 *of* G_2 *whose intersection with* Γ_2 *contains an element of infinite order, and vice versa.*

If finitely generated Zariski-dense subgroups $\Gamma_1 \subset G_1(F)$ and $\Gamma_2 \subset G_2(F)$ are weakly commensurable, then by Theorem 3.2, they have a common trace field $K_{\Gamma_1} = K_{\Gamma_2} =: K$, which is finitely generated. Then Theorem 5.3 and its variants guarantee the existence in Γ_1 and Γ_2 of elements that are generic over K and its suitable finite extensions, and satisfy some additional conditions. Applying Theorem 5.5 and/or Corollary 5.6, we obtain that the algebraic hulls $\mathcal{G}_1$ and $\mathcal{G}_2$ of Γ_1 and Γ_2, respectively, share large families of maximal K-tori. In the case where Γ_1 and Γ_2 are S-arithmetic, this information about the common maximal tori turns out to be sufficient to prove Theorems 4.1 and 4.2. In the final two sections we will discuss the implementation of this approach for arbitrary Zariski-dense subgroups.

6. Arbitrary Zariski-dense subgroups

As we already explained, the results for arithmetic groups were obtained by analyzing the algebraic hulls of arithmetic groups which are weakly commensurable to a given one. While general Zariski-dense subgroups are not determined up to commensurability by their algebraic hull (even if they are lattices, cf. [52]), the latter remains an important invariant. At the same time, the results in the arithmetic case as well as some very recent results over general fields concerning simple algebraic groups with the same maximal tori and division algebras with the same maximal subfields, which we will discuss in the rest of this article, have led us to believe that the algebraic hull itself is *almost* determined by the presence of a Zariski-dense subgroup weakly commensurable to a given one in all situations. More precisely, we would like to propose the following *Finiteness Conjecture*.

Conjecture 6.1. *Let G_1 and G_2 be absolutely simple (adjoint) algebraic groups over a field F of characteristic zero, and let $\Gamma_1 \subset G_1(F)$ be a finitely generated Zariski-dense subgroup with trace field $K_{\Gamma_1} = K$. Then there exists a* finite *collection $\mathcal{G}_1^{(2)}, \dots, \mathcal{G}_r^{(2)}$ of F/K-forms of G_2 such that if $\Gamma_2 \subset G_2(F)$ is a finitely generated Zariski-dense subgroup that is weakly commensurable to Γ_1, then it is conjugate to a subgroup of one of the $\mathcal{G}_i^{(2)}(K)$'s ($\subset G_2(F)$).*

We already know that two weakly commensurable finitely generated Zariski-dense subgroups have the same trace field (Theorem 3.2). The above conjecture takes this result much farther by claiming that a finitely generated Zariski-dense subgroup weakly commensurable to a given one can exist only in finitely many simple algebraic groups over this field. (For example, if $G_0 = \mathrm{SO}_n(q_0)$, where q_0 is a nondegenerate quadratic form of dimension $n \geqslant 3$, $n \neq 4$, over a finitely generated field K of characteristic zero, and $\Gamma_0 \subset G_0(K)$ is a finitely generated Zariski-dense subgroup with trace field K, then according to the conjecture, there should exists a *finite* collection $q_1, \dots, q_r$ of nondegenerate n-dimensional quadratic forms over K such that if $G(K)$ for $G = \mathrm{SO}_n(q)$, with q a nondegenerate n-dimensional quadratic form over K, contains a finitely generated Zariski-dense subgroup that is weakly commensurable to Γ_0, then q *must* be similar to one of the q_i's, $i = 1, \dots, r$.)

Based on our results for S-arithmetic groups (cf., for example, Theorem 4.1(1)) and the results concerning division algebras algebras of exponent two having the same maximal subfields (see Corollary 6.8 and Theorem 7.10), one expects that in some situations one should be able to show that actually $r = 1$, which informally means that the ambient algebraic group is *uniquely* determined by the eigenvalue information of semi-simple elements in a finitely generated Zariski-dense subgroup.

Conjecture 6.1 is known to be true if K is a number field even when Γ_1 is not S-arithmetic (cf. [44, Theorem 5.1]) and also over general fields when G_1 is of type A_1. We recall that given a connected absolutely almost simple real algebraic subgroup of SL_n such that $\mathcal{G} = G(\mathbb{R})$ is noncompact and is not locally isomorphic to $\mathrm{SL}_2(\mathbb{R})$ and a lattice Γ in $\mathcal{G}$, then there exists a number field $K \subset \mathbb{R}$ such that Γ can be conjugated into $\mathrm{SL}_n(K)$, cf. [45, 7.67 and 7.68]. Combining these results, we conclude that Conjecture 6.1 is true when Γ_1 is a lattice in the group of real points of an absolutely almost simple real algebraic group. More evidence supporting this conjecture comes from the investigation of another natural problem in the theory of algebraic group — namely, characterizing absolutely almost simple algebraic K-groups having the same isomorphism/isogeny classes of maximal K-tori. The connection between the two is based on the Isogeny Theorem 5.5 and Corollary 5.6. While these two problems are not equivalent, their investigation usually involves many common

elements. To comment on these common aspects, we will temporarily shift the focus to the second problem. We will later see how the finiteness statements in the context of both problems fit into some more general conjectures about algebraic groups with reductive reduction - see Conjectures 7.5 and 7.8.

6.2. Simple algebraic groups over number fields with the same maximal tori. The tools used to prove Theorems 4.1 and 4.2 can be used to characterize absolutely almost simple algebraic groups over number fields having the same isomorphism/isogeny classes of maximal tori. We give the statements of these results below in order to demonstrate their complete similarity to the corresponding results concerning weak commensurability.

Theorem 6.3 (cf. [40], Theorem 7.5).

(1) *Let G_1 and G_2 be connected absolutely almost simple algebraic groups defined over a number field K, and let L_i be the smallest Galois extension of K over which G_i becomes an inner form of a split group. If G_1 and G_2 have the same K-isogeny classes of maximal K-tori, then either G_1 and G_2 are of the same Killing-Cartan type, or one of them is of type B_n and the other is of type C_n, and moreover, $L_1 = L_2$.*

(2) *Fix an absolutely almost simple K-group G. Then the set of isomorphism classes of all absolutely almost simple K-groups G' having the same K-isogeny classes of maximal K-tori is finite.*

(3) *Fix an absolutely almost simple simply connected K-group G whose Killing-Cartan type is different from A_n, D_{2n+1} $(n > 1)$ or E_6. Then any K-form G' of G (in other words, any absolutely almost simple simply connected K-group G' of the* same *type as G) that has the same K-isogeny classes of maximal K-tori as G, is isomorphic to G.*

The construction described in [40, §9] shows that the types excluded in (3) are honest exceptions, i.e., for each of those types one can construct non-isomorphic absolutely almost simple simply connected K-groups G_1 and G_2 of this type over a number field K that have the same isomorphism classes of maximal K-tori.

The case where G_1 and G_2 are of types B_ℓ and C_ℓ, respectively, is treated in the following theorem.

Theorem 6.4 ([22], Theorem 1.4). *Let G_1 and G_2 be absolutely almost simple algebraic groups over a number field K of types B_ℓ and C_ℓ, respectively, for some $\ell \geqslant 3$.*

(1) *The groups G_1 and G_2 have the same* isogeny *classes of maximal K-tori if and only if they are twins.*

(2) *The groups G_1 and G_2 have the same* isomorphism *classes of maximal K-tori if and only if they are twins, G_1 is adjoint, and G_2 is simply connected.*

We note that some aspects of the general problem of characterizing absolutely almost simple algebraic groups over local and global fields having the same isomorphism classes of maximal tori were considered in [20] and [31]. Another direction of research, which has already generated a number of results (cf. [5], [6] [21], [33], [41]) is the investigation of local-global principles for embedding tori into absolutely almost simple algebraic groups as maximal tori (in particular, for embedding of commutative étale algebras with involutive

automorphisms into simple algebras with involution); some number-theoretic applications of these results can be found, for example, in [17].

In order to get a better idea of what kind of results can be obtained over general fields, it is helpful to consider first the related problem of characterizing finite-dimensional division algebras having the same maximal subfields, which is somewhat reminiscent of Amitsur's famous theorem about central simple algebras having the same *generic* splitting fields (cf. [2], [25]).

6.5. Division algebras with the same maximal subfields. Let D_1 and D_2 be central division algebras of the same degree n over a field K. We say that D_1 and D_2 have the *same maximal subfields* if a degree n field extension L/K admits a K-embedding $L \hookrightarrow D_1$ if and only if it admits a K-embedding $L \hookrightarrow D_2$. We also let $\mathrm{Br}(K)$ denote the Brauer group of K, and for a (finite-dimensional) central simple K-algebra A, we let $[A] \in \mathrm{Br}(K)$ denote the corresponding Brauer class.

Definition 6.6. Let D be a central division K-algebra of degree n. The *genus* of D is defined to be

$$\mathbf{gen}(D) = \{\, [D'] \mid D' \text{is a central division} K\text{-algebra with the same maximal subfields as } D\}.$$

Two basic questions about the genus are:

(I) *When does* $\mathbf{gen}(D)$ *reduce to a single element?* (Then D is uniquely determined by its maximal subfields.)

(II) *What can one say about the size of* $\mathbf{gen}(D)$ *in the general case? In particular, when is* $\mathbf{gen}(D)$ *finite?*

We note that since the opposite algebra D^{op} has the same maximal subfields as D, the genus $\mathbf{gen}(D)$ can reduce to one element only if $D \simeq D^{\mathrm{op}}$, i.e. if $[D]$ has exponent 2 in $\mathrm{Br}(K)$. If K is a global field, then any central division algebra over K of exponent 2 is a quaternion algebra and furthermore it follows from the theorem of Albert-Hasse-Brauer-Noether (ABHN) that for any quaternion division K-algebra D we have $|\mathbf{gen}(D)| = 1$. Another consequence of (AHBN) is that $\mathbf{gen}(D)$ is finite for any central division algebra D over a global field K (see [9, 3.6]).

On the other hand, a construction proposed by M. Rost, M. Schacher, A. Wadsworth, and others (cf. [23, Example 2.1]), enables one to produce quaternion algebras over infinitely generated fields with nontrivial, and even infinite (see [38]), genus. So, both questions become nontrivial over fields more general than global fields, and the following two theorems, obtained jointly with V. Chernousov and I. Rapinchuk [8], [9], [46], contain some recent results in that direction.

The first theorem expands the variety of examples where the genus is trivial. We will say that a field F satisfies property $(*)$ if for any central division F-algebra D of exponent 2, the genus $\mathbf{gen}(D)$ reduces to a single element.

Theorem 6.7 (Stability theorem, [9, 46]). *If a field* k *of characteristic* $\neq 2$ *satisfies* $(*)$*, then so does the field of rational functions* $k(x)$.

(The stability property in characteristic 2 has not been investigated yet.)

Corollary 6.8. *Let k be either a finite field of characteristic $\neq 2$ or a number field, and let $K = k(x_1, \ldots, x_t)$ be a finitely generated purely transcendental extension of k. Then for any central division K-algebra of exponent 2, we have $|\mathbf{gen}(D)| = 1$.*

The second theorem establishes the finiteness of the genus over finitely generated fields.

Theorem 6.9. *Let D be a central division algebra of degree n over a finitely generated field K of characteristic not dividing n. Then the genus $\mathbf{gen}(D)$ is finite.*

Both theorems are based on an analysis of the ramification properties of division algebras in the genus. More precisely, given a discrete valuation v of K, we let $\mathcal{O}_{K,v}$ and $\overline{K}_v$ denote the corresponding valuation ring and residue field, respectively. Fix an integer $n > 1$ (which will later be either the degree or the exponent of D), and suppose that V is a set of discrete valuations of K that satisfy the following three conditions:

(A) *For any $a \in K^\times$, the set $V(a) := \{v \in V \mid v(a) \neq 0\}$ is finite;*

(B) *There exists a* finite *subset $V' \subset V$ such that the field of fractions of*

$$\mathcal{O} = \bigcap_{v \in V \setminus V'} \mathcal{O}_{K,v},$$

coincides with K;

(C) *For any $v \in V$, the characteristic of the residue field $\overline{K}_v$ is prime to n.*

(We note that if K is finitely generated, which will be the case in most of our applications, then (B) automatically follows from (A).) Due to (C), we can define for each $v \in V$ the corresponding *residue map*

$$\rho_v \colon {}_n\mathrm{Br}(K) \longrightarrow \mathrm{Hom}(\mathcal{G}^{(v)}, \mathbb{Z}/n\mathbb{Z}), \tag{R}$$

where ${}_n\mathrm{Br}(K)$ is the n-torsion in the Brauer group and $\mathcal{G}^{(v)}$ is the absolute Galois group of $\overline{K}_v$ (cf., for example, [48, §10] or [49, Ch.II, Appendix]). A class $[A] \in {}_n\mathrm{Br}(K)$ (or a finite-dimensional central simple K-algebra A representing this class) is said to be *unramified* at v if $\rho_v([A]) = 1$, and *ramified* otherwise. We let $\mathrm{Ram}_V(A)$ denote the set of all $v \in V$ where A is ramified; one shows that this set is always finite. We also define the unramified part of ${}_n\mathrm{Br}(K)$ with respect to V to be

$${}_n\mathrm{Br}(K)_V = \bigcap_{v \in V} \mathrm{Ker}\, \rho_v.$$

Then one shows [9, Theorem 2.2] that if ${}_n\mathrm{Br}(K)_V$ is finite, then for a central division algebra K-algebra D of degree n one has

$$|\mathbf{gen}(D)| \leqslant |{}_n\mathrm{Br}(K)_V| \cdot \varphi(n)^r, \text{ with } r = |\mathrm{Ram}_V(D)|. \tag{U}$$

Thus, to prove Theorem 6.9, one needs to show that for a finitely generated field K whose characteristic is prime to a given integer $n > 1$, there exists a set V of discrete valuations of K satisfying the above conditions (A)-(C) and such that the unramified Brauer group ${}_n\mathrm{Br}(K)_V$ is finite. This was first done by an explicit construction based on an analysis of the standard exact sequence for the Brauer group of a curve; this approach enables one to

give some explicit estimations on the size of the unramified Brauer group, hence of the genus, cf. [9, §4], [10]. Subsequently, a more general argument was pointed out to us J.-L. Colliot-Thélène (cf. [8]). More precisely, suppose our finitely generated field K is realized as the field of rational functions on a regular integral scheme X of finite type over $\mathrm{Spec}\ A$, where A is either a finite field or the ring of S-integers in a number field for some finite set S of its places, with n invertible in A, and let V be the set of discrete valuations of K associated with the divisors on X. Then the finiteness of ${}_n\mathrm{Br}(K)_V$ follows from Deligne's finiteness theorem for the étale cohomology of constructible sheaves [16] and Gabber's purity theorem [19]. The proof of Theorem 6.7 relies on the fact that if V is the set of all geometric places of the field of rational functions $k(x)$ (i.e., those that are trivial on k), where k is a field of characteristic $\neq 2$, then ${}_2\mathrm{Br}(k(x))_V$ reduces to ${}_2\mathrm{Br}(k)$ (cf. [25, Corollary 6.4.6]).

7. The genus of a simple algebraic group. Groups with reductive reduction.

In this concluding section, we will describe the ongoing project (cf. [11]) of obtaining the analogs of results from 6.5 for arbitrary absolutely almost simple algebraic groups, and connect this activity back to the Finiteness Conjecture 6.1. First, we need to extend Definition 6.6.

Definition 7.1. Let G be an absolutely almost simple simply connected algebraic group over a field K. The (K-)*genus* $\mathbf{gen}_K(G)$ (or simply $\mathbf{gen}(G)$ if this does not lead to any confusion) of G is the collection of K-isomorphism classes of K-forms G' of G that have the same K-isomorphism classes of maximal K-tori as G.

One can alternatively define the genus using "K-isogeny classes" of maximal tori in place of "K-isomorphism classes." While the exact relationship between these notions of genus has not been investigated, the Isogeny Theorem 5.5 and subsequent remarks strongly suggest that they should lead to the same qualitative results in most cases. On the other hand, A.S. Merkurjev proposed a different (in a way, more functorial) definition of the genus of an absolutely almost simple algebraic K-group G as the set of K-isomorphism classes of K-forms G' of G that have the same isomorphism/isogeny classes of maximal tori not only over K, but also over *any* field extension F/K. The results of Izboldin, Vishik and Karpenko indicate a connection between this genus for the spinor group $G = \mathrm{Spin}_n(q)$ of a quadratic form q and the motive of the projective quadric $q = 0$ in the category of Chow motives, so it makes sense to call this genus *motivic* (see [9, Remark 5.6] for more details). In this article, however, we will only use the notion of genus given in Definition 7.1.

Building on Theorem 6.9, it is natural to make the following conjecture.

Conjecture 7.2. *Let G be an absolutely almost simple simply connected algebraic group over a finitely generated field K of good characteristic*[4]*. Then $\mathbf{gen}_K(G)$ is finite.*

This conjecture is true over global fields (Theorem 6.3) and also for inner forms of type A_ℓ in the general case (see Theorem 7.6 below). While Conjecture 7.2 does not automatically imply our main Conjecture 6.1, we will now outline an approach that can potentially lead

[4]For each type, the following characteristics are defined to be *bad*: type A_ℓ - all primes dividing $(\ell+1)$, and also $p=2$ for outer forms; types B_ℓ, C_ℓ, D_ℓ - $p=2$, and also $p=3$ for ${}^{3,6}\mathsf{D}_4$; for type E_6 - $p=2,3,5$; for types E_7, E_8 - $p=2,3,5,7$; for types F_4, G_2 - $p=2,3$. All other characteristics for a given type are *good*.

to the proof of both conjectures, and also have some other implications. The considerations in 6.5 were based on an analysis of the ramification properties of central simple algebras at discrete valuations of the center. An adequate replacement of the notion of an unramified algebra for arbitrary absolutely almost simple groups is the notion of a group with *reductive reduction*. Let G be a connected absolutely almost simple simply connected algebraic group over a field K. Given a discrete valuation v of K, we let K_v denote the corresponding completion with valuation ring $\mathcal{O}_v$, valuation ideal $\mathfrak{p}_v$, and residue field $\overline{K}_v = \mathcal{O}_v/\mathfrak{p}_v$. One says that G has reductive reduction at v if there exists a reductive group scheme $\mathcal{G}$ over $\mathcal{O}_v$ with generic fiber $G \otimes_K K_v$. Then the reduction $\mathcal{G} \otimes_{\mathcal{O}_v} \overline{K}_v$ modulo $\mathfrak{p}_v$ will be denoted $\underline{G}^{(v)}$. A crucial point in the proof of the estimate (U) in 6.5, which reduces the finiteness of the genus $\mathbf{gen}(D)$ to the finiteness of the unramified Brauer group, was the fact that if $D' \in \mathbf{gen}(D)$, and $\chi = \rho_v([D])$, $\chi' = \rho_v([D'])$, where ρ_v is the residue map at v (cf. (R) in 6.5), then $\mathrm{Ker}\ \chi = \mathrm{Ker}\ \chi'$. In particular, if D is unramified at v then so is D' (thus, the property of being unramified is determined by maximal subfields). We have been able to prove the following analog of the latter fact for arbitrary absolutely almost simple groups.

Theorem 7.3 ([11]). *Assume that the residue field $\overline{K}_v$ is finitely generated and that G has reductive reduction at v. Then any $G' \in \mathbf{gen}_K(G)$ also has reductive reduction at v. Furthermore, the reduction $\underline{G'}^{(v)}$ lies in the genus $\mathbf{gen}_{\overline{K}_v}(\underline{G}^{(v)})$.*

Assume now that the field K is equipped with a set V of discrete valuations that satisfies the following two conditions

(A′) for any $a \in K^\times$, the set $V(a) := \{v \in V \mid v(a) \neq 0\}$ is finite;

(B′) for any $v \in V$, the residue field $\overline{K}^{(v)}$ is finitely generated.

Corollary 7.4. *If K and V satisfy conditions (A′) and (B′), then for any absolutely almost simple simply connected algebraic K-group G, there exists a finite subset $V_0 \subset V$ (depending on G) such that every $G' \in \mathbf{gen}_K(G)$ has reductive reduction at all $v \in V \setminus V_0$.*

The other ingredient of the proof of the finiteness of $\mathbf{gen}(D)$ in 6.5 was the finiteness of the unramified Brauer group ${}_n\,\mathrm{Br}(K)_V$ for a suitable set V of discrete valuations of K. One can expect the following general statement to be valid for the same sets V of valuations as in 6.5. Let again X be a regular integral scheme of finite type over $\mathrm{Spec}\ A$, where A is either a finite field or the ring of S-integers in a number field for some finite set S of its places, let K be the field of rational functions on X, and let V be the set of discrete valuations of K associated with the prime divisors on X (obviously, V satisfies conditions (A′) and (B′)).

Conjecture 7.5. *Let K and V be as above, and let G be an absolutely almost simple simply connected algebraic K-group such that* $\mathrm{char}\ K$ *is good for G. Then for any finite subset $V_0 \subset V$, the set of K-isomorphism classes of (inner) K-forms G' of G that have reductive reduction at all $v \in V \setminus V_0$, is finite.*

Over a number field K, the assertion of Conjecture 7.5 is an easy consequence of the finiteness results for Galois cohomology, cf. [49, Ch. III, 4.6]. (Interestingly, there are absolutely almost simple nonsplit algebraic groups over $\mathbb{Q}$ that have reductive reduction at all primes, see [28], [14], but there are no $\mathbb{Q}$-defined abelian varieties with smooth reduction at all primes [18].) At the time of this writing, our knowledge about the conjecture is limited to the following two theorems.

Theorem 7.6 (cf. [9], Theorem 5.3). *Conjectures 7.2 and 7.5 (for inner forms) are true for $G = \mathrm{SL}_{1,A}$ where A is a central simple K-algebra.*

Assume that $\mathrm{char}\, K \neq 2$ and let $\mu_2 = \{\pm 1\}$. Then for any discrete valuation v of K such that $\mathrm{char}\, \overline{K}_v \neq 2$ and any $i \geqslant 1$, one can define the residue map in Galois cohomology

$$\rho_v^i \colon H^i(K, \mu_2) \to H^{i-1}(\overline{K}_v, \mu_2)$$

extending (R) in 6.5 to all dimensions (see, for example, [13, 3.3] or [25, 6.8] for the details). Then for any set V of discrete valuations of K such that $\mathrm{char}\, \overline{K}_v \neq 2$ for all $v \in V$, one defines the unramified part $H^i(K, \mu_2)_V$ to be $\bigcap_{v \in V} \mathrm{Ker}\ \rho_v^i$ (of course, $H^2(K, \mu_2)_V = {}_2\,\mathrm{Br}(K)_V$).

Theorem 7.7 ([11]). *Let K be a finitely generated field of characteristic $\neq 2$, and let V be a set of discrete valuations of K as in Conjecture 7.5 such that $\mathrm{char}\, \overline{K}_v \neq 2$ for all $v \in V$. Assume that for any finite subset $V_0 \subset V$, the unramified cohomology groups $H^i(K, \mu_2)_{V \setminus V_0}$ are finite for all $i \geqslant 1$. Then for any $n \geqslant 5$, the set of K-isomorphism classes of the spinor groups $\mathrm{Spin}_n(q)$, where q is a nondegenerate n-dimensional quadratic form, that have reductive reduction at all $v \in V$, is finite.*

Now, our Finiteness Conjecture 6.1 would be a consequence of Conjecture 7.5 and the following.

Conjecture 7.8. *Let K and V be as in Conjecture 7.5, and assume that $\mathrm{char}\ K = 0$. Furthermore, let G_1 and G_2 be absolutely almost simple algebraic groups defined over a field $F \supset K$, and let $\Gamma_1 \subset G_1(F)$ be a Zariski-dense subgroup with trace field $K_\Gamma = K$. Then there exists a finite subset $V_0 \subset V$ (depending on Γ_1) such that if $\Gamma_2 \subset G_2(F)$ is weakly commensurable to Γ_1, then the (simply connected cover of the) algebraic hull $\mathcal{G}_2$ of Γ_2 has reductive reduction at all $v \in V \setminus V_0$.*

At this point, Conjecture 7.8 has been established for groups of type A_1 using the strong approximation theorem of Weisfeiler [54]. It seems that the same method should also be applicable in the general case.

The potential implications of Conjecture 7.5 reach beyond eigenvalue rigidity, e.g., it would also imply the finiteness of the Tate-Shafarevich set in some situations. More precisely, let K and V be as in Conjecture 7.5, and let G be an absolutely almost simple simply connected K-group. Consider the Tate-Shafarevich set

$$Ш(\overline{G}) := \mathrm{Ker}\left(H^1(K, \overline{G}) \longrightarrow \prod_{v \in V} H^1(K_v, \overline{G}) \right)$$

for the corresponding adjoint group $\overline{G}$. We can pick a finite subset $V_0 \subset V$ so that G has reductive reduction at all $v \in V \setminus V_0$. Now, let $\xi \in Ш(\overline{G})$, and let $G' = {}_\xi G$ be the corresponding twisted group. Then $G' \simeq G$ over K_v for all $v \in V$, and in particular, G' has reductive reduction at all $v \in V \setminus V_0$. Assuming Conjecture 7.5, we would have that the groups ${}_\xi G$ for $\xi \in Ш(\overline{G})$ form finitely many K-isomorphism classes; in other words, the image of $Ш(\overline{G})$ under the canonical map $H^1(K, \overline{G}) \xrightarrow{\lambda} H^1(K, \mathrm{Aut}\, G)$ is finite. But since $\overline{G} \simeq \mathrm{Int}\, G$ is of finite index in $\mathrm{Aut}\, G$, the map λ has finite fibers, so we obtain the finiteness of $Ш(\overline{G})$. In particular, Theorem 7.6 yields the following.

Corollary 7.9. *Let K and V be as in Conjecture 7.5, and let A be a central simple K-algebra of degree n not divisible by* $\mathrm{char}\ K$. *Then for $\overline{G} = \mathrm{PSL}_{1,A}$, the Tate-Shafarevich set $Ш(\overline{G})$ is finite.*

Finally, we would to point out that the techniques involved in Theorem 7.3 are instrumental not only for proving the finiteness of $\mathbf{gen}_K(G)$, but also for its quantitative analysis. For example, they give yet another instance where a K-form is *uniquely* determined by its maximal K-tori.

Theorem 7.10. *Let $K = k(x)$, where k is a global field of characteristic $\neq 2$. For any K-group G of type G_2, the genus $\mathbf{gen}_K(G)$ reduces to a single element.*

One expects a similar statement to hold over such a field K also for groups of types B_ℓ, C_ℓ ($\ell \geqslant 2$) and F_4 (maybe under some additional assumptions).

Acknowledgements. The author was partially supported by NSF grant DMS-1301800. The author is grateful to Vladimir Chernousov, Gopal Prasad and Igor Rapinchuk for comments and suggestions that helped to improve the exposition.

References

[1] Abels, H., Margulis, G.A., and Soifer, G.A., *The Auslander conjecture for groups leaving a form of signature $(n-2,2)$ invariant*, Probability in mathematics. Israel J. Math. **148** (2005), 11–21.

[2] Amitsur, S., *Generic splitting fields of central simple algebras*, Ann. math. **62** (1955), 8–43.

[3] Ax, J., *On Schanuel's conjecture*, Ann. math. **93** (1971), 252–268.

[4] Baker, A., *Transcendental Number Theory.* Cambridge Univ. Press, Cambridge, 1975.

[5] Bayer-Fluckiger, E., *Embeddings of maximal tori in orthogonal groups*, Ann. Inst. Fourier (Grenoble) (2014).

[6] ———, *Isometries of quadratic spaces*, J. Eur. Math. Soc. (JEMS) (2014).

[7] Borel, A., *Linear Algebraic Groups.* Second Enlarged Edition. Springer-Verlag, New York, 1991.

[8] Chernousov, V.I., Rapinchuk, A.S., Rapinchuk, I.A., *On the genus of a division algebra*, C. R. Acad. Sci. Paris, Ser. I **350** (2012), 807–812.

[9] ———, *The genus of a division algebra and the unramified Brauer group*, Bull. Math. Sci. **3** (2013), 211-240. I.A.

[10] ———, *Estimating the genus of a division algebra* (in preparation).

[11] ———, *On algebraic groups having the same maximal tori* (in preparation).

[12] Chinburg, T., Hamilton, E., Long, and D.D., Reid, A.W., *Geodesics and commensurability classes of arithmetic hyperbolic 3-manifolds*, Duke Math. J. **145** (2008), 25–44.

[13] Colliot-Thélène, J.-L., *Birational invariants, purity, and the Gersten conjecture, K-theory and Algebraic Geometry: Connections with Quadratic Forms and Division Algebras,* Proc. Symp. Pure Math. **58**, part 1, 1–64, AMS, 1995.

[14] Conrad, B., *Non-split reductive groups over $\mathbb{Z}$, Proceedings of the Summer School on Group Schemes*, Luminy, 2011.

[15] Corlette, K., *Archimedean superrigidity and hyperbolic geometry*, Ann. math. **135** (1992), no. 1, 165–182.

[16] Deligne, P., *Cohomologie étale*, SGA $4\frac{1}{2}$. Lect. Notes Math. **569**, Springer, 1977.

[17] Fiori, A., *Characterization of special points of orthogonal symmetric spaces*, J. Algebra **372** (2012), 397–419.

[18] Fontaine, J.-M., *Il n'y a pas de variété abélienne sur $\mathbb{Z}$*, Invent. math. **81** (1985), 515–538.

[19] Fujiwara, K., *A proof of the absolute purity conjecture (after Gabber)*, Algebraic Geometry 2000, Azumino (Hotaka), 153–183. Adv. Stud. Pure Math., **36**, Math. Soc. Japan, 2002.

[20] Garge, S., *Maximal tori determining the algebraic groups*, Pacif. J. Math. **220** (2005), no. 1, 69–85.

[21] Garibaldi, S., *Outer automorphisms of algebraic groups and determining groups by their maximal tori*, Michigan Math. J. **61** (2012), no. 2, 227–237.

[22] Garibaldi, S. and Rapinchuk, A.S., *Weakly commensurable S-arithmetic subgroups in almost simple algebraic groups of types B and C*, Algebra and Number Theory **7**(2013), no. 5, 1147–1178.

[23] Garibaldi, S. and Saltman, D., *Quaternion Algebras with the Same Subfields*, Quadratic forms, linear algebraic groups, and cohomology, 225-238. Dev. math. 18, Springer, New York, 2010.

[24] Gilkey, P., *Invariance Theory, the Heat Equation, and the Atiyah-Singer Index Theorem.* 2nd edition, CRC Press, Boca Raton, 1995.

[25] Gille, P. and Szamuely, T., *Central Simple Algebras and Galois Cohomology*. Cambridge Univ. Press, Cambridge, 2006.

[26] Gorodnik, A. and Nevo, A., *Spitting fields of elements in arithmetic groups*, Math. Res. Letters **18** (2011), 1281–1288.

[27] Gromov, M. and Shoen, R., *Harmonic maps into singular spaces and p-adic superrigidity for lattices in groups of rank one*, Publ. math. IHES **76** (1992), 165–246.

[28] Gross, B.H., *Groups over $\mathbb{Z}$*, Invent. math. **124** (1996), 263–279.

[29] Jouve, F., Kowalski, E., and Zywina, D., *Splitting fields of characteristic polynomials of random elements in arithmetic groups*, Israel J. Math. **193** (2013), no. 1, 263–307.

[30] Kac, M., *Can one hear the shape of a drum?*, Amer. Math. Monthly **73** (1966), no. 4, part 2, 1–23.

[31] Kariyama, K., *On conjugacy classes of maximal tori in classical groups*, J. Agebra **125** (1989), 133–149.

[32] Katok, A. and Spatzier, R.J., *Differential rigidity of Anosov actions of higher rank abelian groups and algebraic lattice actions*, Trudy Mat. Inst. Steklova **216** (1997), 292–319.

[33] Lee, T.-Y., *Embedding functors and their arithmetic properties*, arXiv:1211.3564.

[34] Lubotzky, A. and Rosenzweig, L., *The Galois group of random elements of linear groups*, Amer. J. Math. **136** (2014).

[35] Lubotzky, A., Samuels, B., and Vishne, U., *Division algebras and noncommensurable isospectral manifolds*, Duke Math. J. **135** (2006), 361–379.

[36] Margulis, G.A., *Discrete Subgroups of Semi-Simple Lie Groups.* Springer, Berlin, 1991.

[37] Margulis, G.A. and Qian, N., *Rigidity of weakly hyperbolic actions of higher real rank semisimple Lie groups and their lattices*, Ergodic Theory Dynam. Systems **21** (2001), 121–164.

[38] Meyer, J.S., *A division algebra with infinite genus*, Bull. London Math. Soc. (2014), doi:10.1112/blms/bdt104.

[39] Prasad, G. and Rapinchuk, A.S., *Existence of irreducible $\mathbb{R}$-regular elements in Zariski-dense subgroups*, Math. Res. Letters **10** (2003), 21–32.

[40] _____, *Weakly commensurable arithmetic groups and isospectral locally symmetric spaces*, Publ. math. IHES **109** (2009), 113–184.

[41] _____, *A local-global principle for embeddings of fields with involution into simple algebras with involution*, Comment. Math. Helv. **85** (2010), 583–645.

[42] _____, *On the fields generated by the lengths of closed geodesics in locally symmetric spaces*, Geom. Dedicata (2014).

[43] _____, *Generic elements in Zariski-dense subgroups and isospectral locally symmetric spaces*, Thin Groups and Superstrong Approximation, 211–252. MSRI Publications, **61** (2014).

[44] _____, *Weakly commensurable groups, with applications to differential geometry*, Handbook of Group Actions, ALM 31. Higher Education Press and Interational Press, Beijing-Boston, 2014.

[45] Raghunahan, M.S., *Discrete Subgroups of Lie groups.* Springer, Berlin, 1972.

[46] Rapinchuk, A.S. and Rapinchuk, I.A., *On division algebras having the same maximal subfields*, Manuscr. math. **132** (2010), 273–293.

[47] Reid, A., *Isospectrality and commensurability of arithmetic hyprebolic 2- and 3-manifolds*, Duke Math. J. **65** (1992), 215–228.

[48] Saltman, D., *Lectures on division algebras.* CBMS Regional Conference Series **94**, AMS, 1999.

[49] Serre, J.-P., *Galois Cohomology*. Springer, 1997.

[50] Sunada, T., *Riemannian coverings and isospectral manifolds*, Ann. math. **121** (1985), 169–186.

[51] Vinberg, E.B., *Rings of definition of dense subgroups of semisimple linear groups*, Math. USSR Izv. **5** (1971), 45–55.

[52] ———, *Some examples of Fuchsian groups sitting in* $SL_2(\mathbb{Q})$, preprint 12011 of the SFB-701. Universität Bielefeld (2012).

[53] Vignéras, M.-F., *Variétés Riemanniennes isospectrales et non isométriques*, Ann. math. **112** (1980), 21–32.

[54] Weisfeiler, B., *Strong approximation for Zariski-dense subgroups of semi-simple algebraic groups*, Ann. math. **120** (1984), no. 2, 271–315.

[55] Yeung, S.-K., *Isospectral Problem of Locally Symmetric Spaces*, Int. Math. Res. Not. IMRN, no. 12 (2011), 2810–2824.

Department of Mathematics, University of Virginia, Charlottesville VA 22904, USA
E-mail: asr3x@virginia.edu

Local and global Frobenius splitting

Karen E. Smith

Abstract. We survey recent progress in local and global Frobenius splitting, explaining the ideas that unify them, including a new way to look at test ideals.

Mathematics Subject Classification (2010). Primary 13A35; Secondary 14B15, 14B05, 14J45, 14E30.

Keywords. Frobenius splitting, F-purity, F-regularity, F-singularities, tight closure, test ideals, multiplier ideals, compatibly split ideals, log canonical, log terminal, rational singularity, Cohen-Macaulay.

1. Introduction

Frobenius splitting has inspired a vast arsenal of techniques in commutative algebra, algebraic geometry, and representation theory. A great many papers and books exist on the topic, with related techniques developed by different camps of researchers, often in language impenetrable to others. Only in the most recent years have many of the most elegant ideas coalesced into a simple and coherent story. I assume that my duty is to summarize some of the state of the art.

The story of Frobenius splitting begins in the 1970's with Hochster and Roberts' celebrated proof of the Cohen-Macaulayness of rings of invariants [41]. The term "Frobenius splitting" was coined a decade later in the exceptionally beautiful paper of Mehta and Ramanathan [65], which jump-started an entire industry devoted to better understanding (e.g., singularities and vanishing of cohomology for) naturally occurring varieties (such as Schubert varieties) in the representation theory of algebraic groups. On the commutative algebra side, Hochster and Roberts' legacy eventually spawned the theory of tight closure [44]. By the nineties, tight closure was being used to define "characteristic p analogs" of important notions in the minimal model program, a process still evolving today.

The last few years have witnessed the continued expansion of Frobenius splitting techniques into further reaches of mathematics; new applications have been found, for example, to Fomin and Zelevinsky's cluster algebras, and there is hope that some of these tools might be used to make progress towards the minimal model program in prime characteristic. Our understanding of *test ideals*—originally defined by Hochster and Huneke as a technical component of the theory of tight closure—has crystalized: now test ideals can be viewed as an instance of *compatible splitting* in the language of Mehta and Ramanathan, as a prime characteristic analog of multiplier ideals or within the broader context of Kawamata's centers of log canonicity, or as annihilators of certain important Artinian modules with Frobenius action. This triumvirate of perspectives on test ideals as major components of three quite

Proceedings of the International Congress of Mathematicians, Seoul, 2014

different areas—algebro-geometric representation theory, birational geometry, and commutative algebra, respectively—bespeaks the depth and importance of Frobenius splitting.

In this talk, I begin in the local setting, reviewing the definition of Frobenius splitting and the closely related but slightly stronger notion of F-regularity due to Hochster and Huneke. This notion ultimately to "characteristic p analogs" of rational, log canonical, log terminal, and DuBois singularities. In a recent application, a large class of cluster algebras (*locally acyclic cluster algebras*) are shown to have canonical singularities [6].

In the global setting, I'll summarize the recent progress on unraveling the geometry of Frobenius split and globally F-regular varieties. These varieties (which are also defined in characteristic zero by "reduction to characteristic p") are important because of the wealth of vanishing theorems about them and because they are characteristic p analogs of some central players in the birational classification of algebraic varieties.

We now understand that global splittings of Frobenius for projective varieties are the same as local splittings "at the vertex of the cone" with respect to any polarization of the variety. This means that the local commutative algebraic theory of Hochster-Roberts-Huneke is equivalent to the global projective theory of Mehta-Ramanathan, despite the fact that the theories developed almost independently during the end of the twentieth century (witness the remarkable disjointness of the monographs [48] and [20]).

A culminating idea is Karl Schwede's recent insight that the test ideal can be viewed as one ideal in a lattice of ideals (which he calls uniformly F-compatible) that are special with respect to the Frobenius map. Such a lattice of ideals had already been intensely studied in commutative algebra as annihilators of "modules with Frobenius action," but Schwede's dual insight gave rise to a simpler and more elegant definition of test ideals which immediately tied them to Mehta and Ramanathan's compatibly split ideals, as well as to Kawamata's theory of log canonical centers. In the last part of my talk, I will explain this new development, as well as an asymptotic theory of test ideals which leads to a simple proof of a theorem on the behavior of symbolic powers of ideals in any polynomial ring.

This paper should be viewed as an extended abstract for my ICM lecture; it is grossly inadequate as a serious survey of this ever expanding field. Having recently written a comprehensive exposition on this topic for MSRI based on a mini-course for the opening workshop for last year's Program in Commutative Algebra, I refer readers who prefer a more leisurely development of the ideas in the main cases there. There are also many other excellent surveys on recent developments, including [83] or [14], both of which contain a more technical and extensive survey of test ideals. Older surveys such as [91] explain more about reduction to characteristic p and the connections between singularities in the minimal model program and characteristic p techniques. Other possible surveys of interest include Huneke's lectures on Tight Closure [48] and Brion-Kumar's text on Mehta-Ramanathan's Frobenius splitting [20]. There are also many topics that I simply won't have time to touch, including connections with Hilbert-Kunz multiplicities (see [49]) or the difficult arithmetic issues arising in the work on F-pure thresholds (see [5] or [70]).

2. The Frobenius map for rings

Let R be any commutative ring of prime characteristic p. The Frobenius map (or p-th power map) is a ring homomorphism:

$$F : R \to R; \qquad r \mapsto r^p.$$

Viewing the second copy of R as an R-module via Frobenius, the Frobenius map is also an R-module map. A nice notation for this map is $F : R \to F_*R$, which is consistent with the notation for sheaves.

The ring R is said to be **F-finite** if the Frobenius map is finite. This is a mild assumption; for example, the class of F-finite rings is closed under localization, surjective image, completion at a maximal ideal, and finite extensions. In particular, finitely generated algebras over a perfect field are F-finite.

A basic fact due to Kunz [57] is that *an F-finite ring R is regular if and only if F_*R is a locally free R-module.*[1] Frobenius splitting can be viewed as a weakened version of this freeness:

Definition 2.1. A reduced ring R of characteristic p is **Frobenius split** (or **F-split**) if the Frobenius map $F : R \to F_*R$ splits (in the category of R-modules).

An F-finite regular ring is Frobenius split by Kunz's theorem. Any direct summand of a Frobenius split ring is Frobenius split itself. So Veronese subrings of polynomial rings are Frobenius split but not regular. So is the ring of invariants of a finite group acting on a polynomial ring over a field whose characteristic does not divide the order of the group.

The first systematic use of Frobenius splitting is in [41, 42], although Hochster and Roberts defined a different notion called F-purity which they proved equivalent to Frobenius splitting in the F-finite case (and possibly always).

2.1. Notation. For a reduced ring R, the Frobenius map $F : R \to F_*R$ is equivalent to the the inclusion of R^p-modules $R^p \hookrightarrow R$. This in turn is equivalent to the inclusion of R-modules $R \hookrightarrow R^{1/p}$, where $R^{1/p}$ is the subring of p-th roots of elements of R in an algebraic closure of the total field of fractions of R.

2.2. F-regularity. A refinement of Frobenius splitting is the notion of F-regularity, which is defined via iterates of Frobenius:

$$F^e : R \to R \qquad r \mapsto r^{p^e}.$$

Again, we denote this map by $R \to F^e_*R$ to emphasize that we view the second copy of R as an R-module via iterates of Frobenius.

Definition 2.2 ([43]). An F-finite ring R is **strongly F-regular**[2] if for every element $f \in R$ not in any minimal prime of R, there exists $e \in \mathbb{N}$ such that the R-module map $R \to F^e_*R$ sending 1 to $f \in F^e_*R$ splits. Put differently, this means that R is reduced, and for all f not in any minimal prime, there exists $e \in \mathbb{N}$ and $\phi \in \mathrm{Hom}_R(R^{1/p^e}, R)$ such that $\phi(f^{1/p^e}) = 1$.

An F-regular ring is not free over R^{p^e} (unless it is regular), but it will have *many* summands isomorphic to R^{p^e}. Indeed, every non-zero element of R will generate an R^{p^e}-module direct summand of R for sufficiently large e. It is easy to see that regular rings are F-regular, and that F-regular rings are Frobenius split. Furthermore, direct summands of F-regular rings are F-regular. F-regular rings have many nice properties:

[1] More generally, even if R is not F-finite, Kunz shows that an excellent ring of characteristic $p > 0$ is regular if and only if its Frobenius map is flat. F-finite rings are always excellent [58].

[2] Hochster and Huneke introduced three flavors of F-regularity, all conjectured to be equivalent. The equivalence is known, for example, for graded rings [61] and for rings with isolated non-$\mathbb{Q}$-Gorenstein points [63]. In recent years, the adjective "strongly" is sometimes dropped from this term, at least in some corners of the literature.

Theorem 2.3.

(1) *F-regular rings are Cohen-Macaulay and normal* [43]*;*

(2) *F-regular rings are pseudo-rational*[3] [90].

Corollary 2.4 (The Hochster-Roberts Theorem [41]). *Fix any ground field k. Let G be a linearly reductive algebraic group over k acting on a regular Noetherian k-algebra S. Then the ring of invariants*

$$S^G := \{f \in S \mid f \circ g = f \quad \text{for all } \ g \in G\}$$

is Cohen-Macaulay.

The proof in characteristic p is immediate: the linear reductivity implies that S^G is a *direct summand* of the regular ring S, hence it is F-regular, and therefore Cohen-Macaulay, by the theorem above. The characteristic zero statement is proved by reduction to characteristic p. Later, Boutot gave a different proof in the characteristic zero case, which does not use reduction to characteristic p, although it still exploits the philosophy of "splitting" [18].

2.3. Recent application to cluster algebras. Fomin and Zelevinsky introduced *cluster algebras* as a way to understand total positivity in a variety of contexts [26]. Fix a purely transcendental field $\mathcal{F} = k(x_1, \ldots, x_n)$ of dimension n over some ground field k. A cluster algebra is a k-subalgebra $\mathcal{A}$ of $\mathcal{F}$ generated by a distinguished collection of rational functions which come together in overlapping groups of algebraically independent generators for $\mathcal{F}$ over k called *clusters*. The clusters are not arbitrary, but are related to one another by a sequence of *mutations* (defined via specific combinatorial rules). Each cluster is indexed by a skew-symmetrizable matrix (or a quiver in a simplified setting) which are related to one another by a corresponding sequence of mutations. See [26, 27] and especially [7] for details.

Many of the motivating examples of cluster algebras are famously well-behaved rings, including for example, (localizations of) the Plucker homogeneous coordinate rings for Grassmannians and other rings from combinatorial representation theory. However, a number of pathological examples also satisfy the definition and there is a sense that the "right class" of cluster algebras is yet to be identified.

The acyclic cluster algebras form one well-understood and well-behaved class, but this class is far too restrictive to be of major importance in its own right [7]. Greg Muller recently introduced the **locally acyclic** cluster algebras, a *much wider* class which simultaneously rules out all of the pathological examples while being flexible enough to include nearly all of the interesting ones [68]. By definition, a cluster algebra is *locally acyclic* if its spectrum admits a "cluster cover" by acyclic cluster algebras—that is, a cover by affine open sets Spec $\mathcal{A}_y$, where y is a product of cluster variables such that $\mathcal{A}_y$ has a naturally induced cluster algebra structure which is acyclic. See [68].

Theorem 2.5 ([6]). *Every locally acyclic cluster algebra over a field of prime characteristic is strongly F-regular.*

It follows that every locally acyclic cluster algebra over a field of characteristic zero has rational singularities [90], and hence canonical singularities (since they are always Gorenstein) [23]. The theorem applies to a wide range of cluster algebras, including for example,

[3]Pseudorationality is a characteristic-free notion introduced in [60] which agrees with rational singularities in characteristic zero.

cluster algebras of Grassmannians, cluster algebras of *marked surfaces* with at least two marked points on the boundary [68, Theorem 10.6], as well as cluster algebras of double Bruhat cells and more generally, *positroid cells* [69].

2.3.1. The Laurent Phenomenon and upper cluster algebras. By the *Laurent Phenomenon*, each cluster variable is a Laurent polynomial in every other cluster [7]. This implies that a cluster algebra $\mathcal{A}$ is contained in its upper cluster algebra

$$\mathcal{U} := \bigcap_{\text{clusters } \mathbf{x} \subset \mathbb{A}} k[x_1^{\pm 1}, x_2^{\pm 1}, ..., x_n^{\pm 1}] \subset \mathcal{F}.$$

For locally acyclic cluster algebras, $\mathcal{A} = \mathcal{U}$, but the inclusion is strict in general [7, 68].

If $k[x_1^{\pm 1}, x_2^{\pm 1}, ..., x_n^{\pm 1}]$ is the Laurent ring in some cluster, there is a Frobenius splitting defined by

$$\phi\left(\lambda x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}\right) = \begin{cases} \phi(\lambda) x_1^{a_1/p} x_2^{a_2/p} \cdots x_n^{a_n/p} & \text{if } a_1, a_2, ..., a_n \in p\mathbb{Z} \\ 0 & \text{otherwise} \end{cases}$$

(where $\phi : F_* k \to k$ is any fixed splitting of the ground field). This induces a Frobenius splitting of the function field $\mathcal{F}$. It is shown in [6] that *all clusters induces the same splitting of* $\mathcal{F}$. Thus there is a canonical cluster splitting on any upper cluster algebra. This is closely related to the fact that the cluster log forms

$$\frac{dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n}{x_1 \cdots x_n}$$

do not depend on the choice of cluster (up to sign),[4] since a Frobenius splitting is a section of the sheaf $\mathcal{O}_X((1-p)K_X)$. See Section 4.2.

2.4. F-signature. A numerical refinement of F-regularity called the *F-signature* sharpens the classification of F-singularities by measuring the *growth rate* of the rank of a maximal free summand of the R-module $F_*^e R$ as e goes to infinity. This was first studied in [96].

Fix a local F-finite domain R. For each natural number e, we can decompose the R module $F_*^e R$ as a direct sum of indecomposable modules, and count the number of summands that are isomorphic to R. Let a_e denote the rank of a maximal free summand of the R-module $F_*^e R$. [Note that if R is Frobenius split then $a_e \geq 1$, and if R is regular then $a_e = p^{de}$ where d is the rank of the free module $F_* R$ over R.] For an F-regular ring, we expect many summands of $F_*^e R$ isomorphic to R, so we expect a_e to grow with e.

Definition 2.6. The **F-signature** of R is

$$s(R) = \lim_{e \to \infty} \frac{a_e}{p^{de}},$$

where $d = [K : K^p]$ for K the fraction field of R. This limit was recently shown to exist [101].

The F-signature is one if and only if R is regular, as proved by Huneke and Leuschke, who also coined the term "F-signature" [50]. The F-signature is positive if and only if R

[4] This observation also appears in the unpublished work of Allen Knutson and David Speyer.

is F-regular [3]. Thus each F-regular ring has an F-signature strictly between zero and one; we can think of F-signature as a measurement of how close an F-regular ring is to being regular. For example, the rational double points $xy = z^{n+1}$ have F-signature $1/(n+1)$ [50], reflecting the fact that the singularity is "worse" for larger n. Formulas for the F-signature of toric varieties are worked out in Von Korff's PhD thesis [102].

Tucker vastly generalizes and simplifies much of the literature on F-signature in [101], to which we refer for a nice exposition. The state of the art on F-signature can be found in the recent papers of Blickle, Schwede and Tucker [16] and [17], which include generalizations of F-signature to pairs (and triples). The F-signature is closely related to the Hilbert-Kunz multiplicity, a subject pioneered by Paul Monsky [67]; see Huneke's survey [49] or Brenner's paper [19]. No known examples of non-rational F-signatures are known (though some expect that they exist).

2.5. Characteristic zero definitions. Let k denote any field of characteristic 0, and let R be a finitely generated k-algebra. Fix a presentation

$$R \cong k[x_1, \ldots, x_n]/(f_1, \ldots, f_r).$$

Let A be the $\mathbb{Z}$-subalgebra of k generated by all coefficients of the polynomials $f_1, \ldots, f_r$, and set

$$R_A = A[x_1, \ldots, x_n]/(f_1, \ldots, f_r).$$

Since A is a finitely generated $\mathbb{Z}$-algebra, the residue field of A at each of its maximal ideals is finite. The map Spec $R_A \to$ Spec A can be viewed as a "family of models" for R. The closed fibers of this map are characteristic p schemes (of varying p) whereas the $k-$valued points reproduce R.

Definition 2.7. The ring R is said to have **dense Frobenius split type** (or **dense F-regular type**) if there is a dense set of maximal ideals μ in Spec A such that $A/\mu \otimes_A R_A$ is Frobenius split (or F-regular).[5]

Definition 2.7 does not depend on the presentation of R, nor on the choice of A. See [47].

Example 2.8.

(1) The ring $\mathbb{C}[x, y, z]/(y^2 - xz)$ has dense F-regular type. As a matter of fact, taking $A = \mathbb{Z}$, the closed fibers of the family are the rings $\mathbb{F}_p[x, y, z]/(y^2 - xz)$, which are F-regular for *every* prime number p.

(2) The ring $\mathbb{C}[x, y, z]/(x^3 + y^3 + z^3)$ has dense Frobenius split type, but *not* dense F-regular type, nor open Frobenius split type. Indeed, $\mathbb{F}_p[x, y, z]/(x^3 + y^3 + z^3)$ is Frobenius split if and only if $p \equiv 1 (\text{mod } 3)$, so that there is an infinite set of prime numbers p for which the "reduction mod p" is Frobenius split. On the other hand, for every $p \geq 5$ and every e, one can show that there is **no map** sending x^{1/p^e} to 1. So this ring is not dense F-regular type.

[5] We can also define **open** Frobenius split (F-regular) type to be when $A/\mu \otimes_A R_A$ is F-split (F-regular) for an *open* set of maximal ideals μ in Spec A. It is expected that dense and open F-regular type are equivalent; this is known in Gorenstein rings and related settings; see [90] and [36].

2.6. Connections with the singularities in the minimal model program. Frobenius splitting and F-regularity in characteristic zero are closely related to a number of important issues arising independently in algebraic geometry, including log canonical and log terminal singularities, positivity, and multiplier ideals.

Theorem 2.9. *Let R be a finitely generated domain over a field of characteristic zero. Then*

(1) *If R is Gorenstein, then R has dense F-regular type if and only if R has rational singularities [32, 66, 90].*

(2) *If R is $\mathbb{Q}$-Gorenstein, then R has dense F-regular type if and only if R has log terminal singularities [36].*

(3) *If R is $\mathbb{Q}$-Gorenstein and has dense Frobenius split type, then R has log canonical singularities [36].*

All three statements can also be made for "pairs;" see Section 3.2.

There are related notions which can be defined in terms of local cohomology modules: F-injectivity means that Frobenius acts injectively on the local cohomology modules of a local ring with support in the maximal ideal, while F-rationality means that these local cohomology modules have no nontrivial proper submodules stable under Frobenius. F-rational type is equivalent to rational singularities [90], [32], [66]. F-injective type implies DuBois singularities [76].

The converse of Theorem 2.6 (3) is conjectured to hold in general as well. This long-standing open question related to an important conjecture linking the *F-pure threshold* and the *log canonical threshold;* see Section 5.4.

3. The global theory

Let X denote a scheme of prime characteristic p. The Frobenius map $F : X \to X$ is the identity map on the underlying topological space of X, while the corresponding map of sheaves $\mathcal{O}_X \to F_*\mathcal{O}_X$ is the p-th power map locally on sections.

Consistent with the terminology for rings, we say that a scheme X is *F-finite* if the quasi-coherent sheaf $F_*\mathcal{O}_X$ is *coherent.* Our main interest is when X is a variety over a perfect field k of characteristic p; such a variety is always F-finite. Note that the Frobenius map is rarely a map of varieties, since it is not linear over the ground field k (unless $k = \mathbb{F}_p$).

Kunz's Theorem implies that an F-finite scheme X is regular if and only if the coherent $\mathcal{O}_X$-module $F_*\mathcal{O}_X$ is locally free. A scheme X is *locally Frobenius split* if the map $\mathcal{O}_X \to F_*\mathcal{O}_X$ splits locally in a neighborhood of each point, or equivalently, if all stalks $\mathcal{O}_{X,p}$ are Frobenius split. Likewise, we say X is *locally F-regular* if all stalks are F-regular. A locally F-regular scheme is normal, Cohen-Macaulay and pseudo-rational by Theorem 2.3.

It is much stronger, of course, to require a *global* splitting of the Frobenius map:

Definition 3.1. Let X be an F-finite scheme of prime characteristic.

(1) X is *Frobenius split* if the Frobenius map $\mathcal{O}_X \to F_*\mathcal{O}_X$ splits as a map of $\mathcal{O}_X$-modules [65];

(2) X is globally F-regular if for all effective Cartier divisors D, there is an e such that the composition

$$\mathcal{O}_X \to F_*^e\mathcal{O}_X \hookrightarrow F_*^e\mathcal{O}_X(D)$$

splits as a map of $\mathcal{O}_X$-modules [93].

Globally F-regular varieties are Frobenius split in a strong sense: there are typically *many* splittings of Frobenius. Indeed, consider any effective divisor D on a globally F-regular variety X. A splitting of (2) above is a map $F^e_*\mathcal{O}_X(D) \xrightarrow{\phi} \mathcal{O}_X$, which can be restricted to $F^e_*\mathcal{O}_X$ to induce a splitting of $\mathcal{O}_X \to F^e_*\mathcal{O}_X$ as well. So we have a splitting of the (iterated) Frobenius $\mathcal{O}_X \to F^e_*\mathcal{O}_X$ which factors through $F^e_*\mathcal{O}_X(D)$—that is, a *Frobenius splitting along D* [73, 74]. A strongly F-regular variety is Frobenius split (for some iterate) along every effective divisor.

Frobenius split varieties satisfy strong vanishing theorems:

Theorem 3.2. *Let X be a Frobenius split projective scheme over an F-finite field, and let $\mathcal{L}$ be an invertible sheaf.*

(1) *If $H^i(X, \mathcal{L}^n) = 0$ for $n \gg 0$, then $H^i(X, \mathcal{L}) = 0$ [65].*

(2) *In particular, if $\mathcal{L}$ is ample, then $H^i(X, \mathcal{L})$ vanishes for all $i > 0$.*

(3) *In particular, if X is Cohen-Macaulay and $\mathcal{L}$ is ample, then $H^i(X, \omega \otimes \mathcal{L}) = 0$ for all $i > 0$ by Serre duality.*

(4) *If X is globally F-regular and $\mathcal{L}$ is nef, then $H^i(X, \mathcal{L})$ vanishes for all $\mathcal{L}$ and all $i > 0$ [93].*

Thus it is worthwhile to have criteria for Frobenius splitting and globally F-regularity. One useful criterion is essentially due to Hochster and Roberts, although the interpretation here is from [91, 4.10.2]:

Proposition 3.3. *A projective variety X is Frobenius split if and only if the induced map $H^{\dim X}(X, \omega_X) \to H^{\dim X}(X, F^*\omega_X)$ is injective.*

Example 3.4. It follows that a variety with trivial canonical bundle (such as an elliptic curve, or in higher dimension, an abelian variety or a Calabi-Yau) is Frobenius split if and only if it is ordinary. The study of ordinarity for Abelian varieties is a difficult problem in number theory; see e.g. [64].

The only smooth projective curve that is globally F-regular is $\mathbb{P}^1$; this follows by applying Theorem 3.2 (4) to the canonical bundle. In higher dimension as well, global F-regularity is a type of positivity. Indeed, Mehta and Ramanathan pointed out that a Frobenius splitting can be viewed as a special kind of pluri-anticanonical form:

Lemma 3.5. *Let X be a normal projective variety over a perfect field. Then we have*

$$\mathcal{H}om_{\mathcal{O}_X}(F^e_*\mathcal{O}_X, \mathcal{O}_X) \cong F^e_*\omega_X^{1-p^e}.$$

So we expect that globally F-regular schemes will admit many effective divisors in the linear systems $|(1 - p^e)K_X|$. Indeed, this property essentially characterizes globally F-regular varieties:

Theorem 3.6 ([82, 93]). *If X is a globally F-regular projective variety of characteristic p, then X is log Fano.*

By log Fano we mean a normal projective variety which admits an effective $\mathbb{Q}$-divisor Δ such that $-K_X - \Delta$ is ample, and the pair (X, Δ) has (at worst) Kawamata log terminal singularities.[6]

The proof of Theorem 3.6 constructs the $\mathbb{Q}$-divisor Δ as follows: Fix an ample divisor D. A splitting of Frobenius along D is a map $F^e_*\mathcal{O}_X(D) \to \mathcal{O}_X$, which in turn can be viewed as a global section of $\mathrm{Hom}_{\mathcal{O}_X}(F^e_*\mathcal{O}_X(D), \mathcal{O}_X) \cong \omega_X^{1-p^e}(-D)$. This gives rise to an effective divisor D' in the linear system $|(1-p^e)K_X - D|$, and we set $\Delta = \frac{1}{p^e-1}D'$. Note that the construction very much depends on the characteristic, p.

The converse of Theorem 3.6 fails because of irregularities in small characteristic. For example, the cubic hypersurface defined by $x^3 + y^3 + z^3 + w^3$ in $\mathbb{P}^3$ is Fano in every characteristic $p \neq 3$, but not globally F-regular nor even Frobenius split in characteristic two. However, it *is* globally F-regular for all characteristics $p \geq 5$.

Theorem 3.7 ([82, 93]). *If X is a log Fano variety of characteristic zero, then X has globally F-regular type.*

The converse to Theorem 3.7 is open. If X has globally F-regular type, then in each characteristic p model, the proof of Theorem 3.6 constructs a "witness" divisor Δ_p establishing that the pair (X_p, Δ_p) is log Fano. But Δ_p depends on p and there is no *a priori* reason that there must be some divisor Δ on the characteristic zero variety which reduces mod p to Δ_p.

Conjecture 3.8. *A projective globally F-regular type variety (of characteristic zero) is log Fano.*

Conjecture 3.8 has been proved for surfaces [30] as well as for $\mathbb{Q}$-factorial Mori Dream spaces [29]. This raises the question: are globally F-regular type varieties (of characteristic zero) Mori Dream Spaces? Moreover, since log Fano spaces (of characteristic zero) are Mori Dream spaces by [9, Cor 1.3.2], the answer is necessarily *yes* if Conjecture 3.8 is true. What about in characteristic p?

Question 3.9. *Assume that X is globally F-regular. Is it true that the Picard group of X is finitely generated? Is it true that the Cox ring of X is always finitely generated?*

Similar issues arise regarding the geometry of Frobenius split varieties:

Theorem 3.10 ([82]). *If X is a normal Frobenius split projective variety of characteristic p, then X is log Calabi-Yau.*

By log Calabi-Yau we mean that X admits an effective $\mathbb{Q}$-divisor such that (X, Δ) is log canonical[7] and $K_X + \Delta$ is $\mathbb{Q}$-linearly equivalent to the trivial divisor.

Again, the converse fails because of irregularities in small characteristic. However, we do expect an analog of Theorem 3.7 to hold.

Conjecture 3.11 ([82]). *If X is a log Calabi-Yau variety of characteristic zero, then X has Frobenius split type.*

Conjecture 3.11 is known in dimension two [30] and for Mori Dream spaces [29].

[6] Kawamata log terminal singularity is usually defined in characteristic 0 using a resolution of singularities, but it can be defined in any characteristic as follows. A pair (X, Δ) consisting of a normal variety with an effective $\mathbb{Q}$-divisor is *Kawamata log terminal* if $K_X + \Delta$ is $\mathbb{Q}$-Cartier, and for *all* birational proper maps $\pi : Y \to X$ with Y normal, choosing K_Y so that $\pi_* K_Y = K_X$, each coefficient of $\pi^*(K_X + \Delta) - K_Y$ is strictly less than 1.

[7] We define log canonical in characteristic p similarly to how we defined klt singularities.

3.1. Local versus global splitting. Despite developing separately, the local and global theories are completely equivalent. Indeed, a projective variety is Frobenius split or globally F-regular if and only if "its affine cone" has that property:

Theorem 3.12 ([82, 93]). *Let X be any projective scheme over a perfect field. The following are equivalent:*

(1) *X is Frobenius split;*
(2) *the ring $S_{\mathcal{L}} = \bigoplus_{n\in\mathbb{N}} H^0(X, \mathcal{L}^n)$ is Frobenius split for all invertible $\mathcal{L}$;*
(3) *the section ring $S_{\mathcal{L}} = \bigoplus_{n\in\mathbb{N}} H^0(X, \mathcal{L}^n)$ is Frobenius split for some ample invertible $\mathcal{L}$.*

Likewise, a projective variety X is globally F-regular if and only if some (equivalently, every) section ring $S_{\mathcal{L}}$ with respect to an ample invertible sheaf $\mathcal{L}$ is F-regular.

Example 3.13. Grassmannians of any dimension and characteristic are globally F-regular. Indeed, the homogeneous coordinate ring for the Plücker embedding of any Grassmannian is F-regular [45]. More generally, all Schubert varieties are globally F-regular [59]. A normal projective toric variety (of any characteristic) is globally F-regular, since a section ring given by a torus invariant ample divisor will be generated by monomials, hence F-regular [95].

3.2. Pairs. During the last decade, a theory of "F-singularities of pairs" has flourished, inspired by the rich theory of pairs developed in birational geometry [54]. The idea to create tight closure theory for pairs was a major breakthrough, pioneered by Nobuo Hara and Keiichi Watanabe in [36]. Once defined, the theory of tight closure theory for pairs—including F-regularity, F-splitting and test ideals— rapidly developed in a long series of technical papers by the Japanese school of tight closure, including Hara, Watanabe, Takagi, Yoshida and others.

By *pair* in this context, we have in mind a normal irreducible scheme X of finite type over a perfect field, together with either a $\mathbb{Q}$-divisor Δ (or an ideal sheaf $\mathfrak{a}$ raised to some fractional exponent.[8]) In the geometric setting, an additional assumption—namely that $K_X + \Delta$ is $\mathbb{Q}$-Cartier—is usually imposed, because a standard technique involves pulling back to different birational models. One possible advantage to the algebra set-up is that it is not necessary to assume that $K_X + \Delta$ is $\mathbb{Q}$-Cartier for the definitions, although alternatives have also been proposed directly in the world of birational geometry as well; see [25]. See also [80].

Definition 3.14. Let X be a normal F-finite variety, and Δ an effective $\mathbb{Q}$-divisor on X.

(1) The pair (X, Δ) is sharply Frobenius split (respectively locally sharply Frobenius split) if there exists an $e \in \mathbb{N}$ such that the natural map

$$\mathcal{O}_X \to F^{p^e}_* \mathcal{O}_X(\lceil (p^e - 1)\Delta \rceil)$$

splits as an map of sheaves of $\mathcal{O}_X$-modules (respectively, splits locally at each stalk).

(2) The pair (X, Δ) is globally (respectively, locally) F-regular if for all effective divisors D, there exists an $e \in \mathbb{N}$ such that the natural map

$$\mathcal{O}_X \to F^{p^e}_* \mathcal{O}_X(\lceil (p^e - 1)\Delta \rceil + D)$$

[8] There are even triples $(X, \Delta, \mathfrak{a}^t)$ incorporating aspects of both variants.

splits as an map of sheaves of $\mathcal{O}_X$-modules (respectively, splits locally at each stalk).

Remark 3.15. A slightly different definition of Frobenius splitting for a pair (X, Δ) was first given by Hara and Watanabe [36]. The variant here, which fits better into our context, was introduced by Karl Schwede [77].

Theorem 3.16 ([36]). *Let (X, Δ) be a pair where X is a normal variety of prime characteristic and Δ is a $\mathbb{Q}$-divisor such that $K_X + \Delta$ is $\mathbb{Q}$-Cartier.*

(1) *If (X, Δ) is a locally F-regular pair, then (X, Δ) is Kawamata log terminal.*

(2) *If (X, Δ) is a locally sharply Frobenius split pair, then (X, Δ) is log canonical.*

Similarly, there are global versions: Theorem 3.7 and 3.10 also hold for "pairs." See [82]. In characteristic zero, the converse of (1) holds, as does its global analog. The local and global converses of (2) are conjectured; this appears to be a difficult problem with deep connections to arithmetic.

We can think of F-regularity as a "characteristic p analog" of Kawamata log terminal singularities, and (at least conjecturally) Frobenius splitting as a "characteristic p analog" of log canonical singularities. The analogy runs deep: F-pure thresholds become "characteristic p analogs" of log canonical thresholds [99], test ideals become "characteristic p analogs" of multiplier ideals [33, 94], centers of sharp F-purity become "characteristic p analogs" of log canonicity [78], F-injectivity becomes a "characteristic p analog" of Dubois singularities [76].

3.2.1. Possible applications to the minimal model program in characteristic p. Recently, attention has turned to solving the minimal model program in prime characteristic, where a big obstruction is the failure of vanishing theorems. There is hope that the Frobenius splitting and tight-closure inspired definitions will help overcome this difficulty. For example, it is not even known in characteristic p whether klt singularities are Cohen-Macaulay, even when resolution of singularities is assumed [55]. Perhaps F-regularity is the "right" class of singularities to consider in prime characteristic instead? As another example, the test ideal is better than the multiplier ideal at capturing some of the subtleties in prime characteristic, for example, under pull back under wildly ramified mappings [84]. The world of F-singularities is beginning to get implemented in the minimal model program (see e.g. [31]), but the final outcome of this endeavor is not yet clear. Another place where Frobenius techniques have been helpful is in effective generation of adjoint bundles; this goes back to [92] which is reproved in dual format in [81], and generalized recently in [72]. See also [53].

4. The test ideal

The *test ideal* is a distinguished ideal reflecting the Frobenius properties of a prime characteristic ring. Test ideals can be defined very generally for pairs on more or less arbitrary Noetherian schemes of characteristic p. However, the theory becomes most transparent in two special cases, the "classical commutative algebra case" and the "classical algebraic geometry case."

The test ideal is a "characteristic p analog" of the multiplier ideal in characteristic zero. This was first proved in the absolute case in [94] and (independently) [33]; the proofs were later adapted to the relative case when test ideals of pairs were introduced [35, 37].

In the classical commutative algebra case, the scheme is the spectrum of a local ring R and we are interested in the "absolute" test ideal. In this case, the test ideal $\tau(R)$ is essentially Hochster and Huneke's test ideal for tight closure.[9] The test ideal can be viewed as just one (the smallest) ideal in a lattice of ideals special with respect to the Frobenius map.

The idea of viewing the test ideal as one ideal in a special lattice given by Frobenius has been around for some time. For example, the test ideal of a Gorenstein ring (R, m) is the annihilator of the maximal proper submodule of the injective hull of the residue field of R stable under any action of Frobenius [62, 89, 90]. But Karl Schwede recently dualized this point of view, leading to a more accessible and elegant theory which ties the test ideal in with Mehta and Ramanathan's theory of compatibly split ideals.

4.1. Schwede's definition of the test ideal. Let R be an F-finite reduced ring of characteristic p.

Definition 4.1. Fix any R-linear map $\varphi : R^{1/p^e} \to R$. An ideal J of R is called φ-*compatible* if $\varphi(J^{1/p^e}) \subseteq J$. That is, J is φ-compatible if there is a commutative diagram

$$\begin{array}{ccc} R^{1/p^e} & \xrightarrow{\varphi} & R \\ \downarrow & & \downarrow \\ (R/J)^{1/p^e} & \dashrightarrow & R/J, \end{array}$$

where the vertical arrows are the natural surjections, showing that φ descends to a map $(R/J)^{1/p^e} \to R/J$.

Definition 4.2 ([78]). An ideal J in an F-finite ring is **uniformly F-compatible** if it is compatible with respect to *every* R-linear map $R^{1/p^e} \to R$, for all e. The **test ideal** is the smallest uniformly compatible ideal not contained in any minimal prime.

It is a non-obvious fact that there exists a smallest such ideal. This is essentially due to Hochster and Huneke in their proof the existence of "completely stable test elements" [43]. For a summary of the proof in this context, see [83].

The test ideal behaves well under localization. In addition, it is easy to see that a ring R is F-regular if and only if its test ideal is trivial. Thus the test ideal defines the locus of non-F-regular points of Spec R.

The set of uniformly F compatibly ideals forms a lattice closed under sum and intersection. If R is Frobenius split, all these ideals are radical as well. This lattice has been studied before: it is the precisely the lattice of F-ideals discussed in [89] and [90], as well as the lattice of annihilators of $\mathcal{F}(E)$-modules of [62], in their respective contexts. If we restrict to just one φ which happens to be a Frobenius splitting of R, this is the lattice of compatibly split ideals of Mehta and Ramanathan [65].

Remark 4.3 (For experts in tight closure theory). The test ideal defined here is the "big" test ideal in the tight closure terminology.[10] If R is complete local, for example, the test ideal

[9] Our terminology differs slightly from the tight closure literature, where our test ideal would be called the "big test ideal" for "non-finitistic tight closure."

[10] Of course, all versions of test ideals in the tight closure theory are conjectured to be equal, and are known to be equal in many cases, including $\mathbb{Q}$-Gorenstein [4] and graded [61] cases.

we define here is the same as the annihilator of the non-finitistic tight closure of zero in the injective hull of the residue field of R [62]. See [78] for a proof.

Experts in tight closure can easily see how the definition of the test ideal here relates to the one in the literature, and why there is a unique smallest uniformly F-compatible ideal, at least in the Gorenstein local case. Let (R, m) be a Gorenstein local domain of dimension d. As is well-known, the test ideal is the annihilator of the tight closure of zero in $H^d_m(R)$. In [89] and [90], the Frobenius stable submodules of $H^d_m(R)$ (including the tight closure of zero) are analyzed and their annihilators in R are dubbed "F-ideals;" there it is shown (also using Hochster and Huneke's test elements!) that there is a unique largest proper Frobenius stable submodule of $H^d_m(R)$, hence a unique smallest non-zero F-ideal, namely test ideal of R. The uniformly F-compatible ideals are precisely the F-ideals—that is, annihilators of submodules of the top local cohomology module $H^d_m(R)$ stable under Frobenius. This is not hard to check using Lemma 4.4; SeF-compatiblee [78] or [24, Thm 4.1]. In the non-Gorenstein case the uniformly F-compatible ideals are the annihilators of the $\mathcal{F}(E)$-modules of [62]; see [78].

The lattice of uniformly F-compatible ideals is especially nice in a Frobenius split ring. For a fixed splitting φ of Frobenius, the set of φ-compatibly split ideals is *finite.* See [56, 75, 85], or [24] for a related dual result.

It follows from the definitions that if R is Frobenius split and J is uniformly F-compatible, then the quotient R/J is Frobenius split. This is an analog of the fact that a log canonical center of a log canonical scheme is itself log canonical. Indeed, Schwede calls the prime uniformly F-compatible ideals *centers of F-purity* and shows that these can be viewed as "characteristic p analogs" of Kawamata's centers of log canonicity [78]. [All of these statements have corresponding versions for pairs; see the references for exact statements.] Schwede also shows that these centers of F-purity satisfy an analog of Kawamata's subadjunction [52]; see [75]. For a local ring, there is a maximal proper uniformly F-compatible ideal—this is the **splitting prime** of Aberbach and Enescu [1].

4.2. Trace of Frobenius. An elementary but powerful observation is that often uniform F-compatibility can be checked by checking compatibility with just one morphism. The following is a simple consequence of the fact that the canonical module of a local Gorenstein ring is cyclic:

Lemma 4.4. *If R is an F-finite Gorenstein local ring, then $\mathrm{Hom}_R(R^{1/p}, R)$ is a cyclic $R^{1/p}$-module.*

It is easy to check that composing the generator (e times) for $\mathrm{Hom}_R(F_*R,R)$ gives a generator for $\mathrm{Hom}_R(F^e_*R,R)$. So an ideal J in a local Gorenstein ring is uniformly F-compatible if and only if it is compatible with a $R^{1/p}$-module generator Ψ for $\mathrm{Hom}_R(R^{1/p},R)$.

For any F-finite[11] ring R, we can dualize the Frobenius map $R \to F_*R$ into ω_R:

$$\mathrm{Hom}_R(F_*R, \omega_R) \longrightarrow \mathrm{Hom}_R(R, \omega_R)$$

which produces an R-module map

$$F_*\omega_R \to \omega_R,$$

[11] F-finite rings always admit a canonical module [28, 13.6].

called the *trace* of Frobenius.[12] For smooth projective varieties, this is called the Cartier map. If R is local and Gorenstein, this trace map Ψ_e generates $\mathrm{Hom}_R(F^e_* R, R)$. See the surveys [14], [84] or [20], for more on the trace map.

Remark 4.5. Blickle's *Cartier algebras* give another point of view [10]. An R-module map $F^e_* R \to R$ is an additive map $R \xrightarrow{\phi} R$ satisfying $\phi(r^{p^e} x) = r\phi(x)$ for any $r, x \in R$ [11]. The **Cartier algebra**[13] $\mathcal{C}(R)$ is the subalgebra of $\mathrm{Hom}_{\mathbb{Z}}(R, R)$ generated by all p^{-e}-linear maps (as we range over all e). Clearly R is a module over $\mathcal{C}(R)$, and clearly its $\mathcal{C}(R)$-submodules are precisely the uniformly F-compatible ideals. The trace map can also be easily interpreted in this language: in the Gorenstein local case, the trace Ψ_e of Lemma 4.4 is literally the composition of Ψ with itself e-times, so that Ψ generates $\mathcal{C}(R)$ as an R-algebra. Generalizations of uniformly F-compatible ideals also come up in the work of Blickle (*e.g.* [10]) under the name of Cartier-submodules and crystals; see the survey [14].

5. Test ideals for pairs

Let R be an F-finite ring, and let $\mathfrak{a}$ be an ideal of R. For each non-negative real number t, we associate an ideal

$$t \in \mathbb{R}_{\geq 0} \ \rightsquigarrow \ \tau(R, \mathfrak{a}^t).$$

In the classical commutative algebra case, $\mathfrak{a} = R$, and all $\tau(R, \mathfrak{a}^t)$ produce the same ideal, the "absolute test ideal" $\tau(R)$. With the introduction of tight closure for pairs, test ideals for pairs naturally followed [37], but the original definition was quite technical. Schwede's definition is much nicer:

Definition 5.1 ([78]). Let R be a reduced F-finite ring and let $\mathfrak{a}$ be an ideal of R. The test ideal $\tau(R, \mathfrak{a}^t)$ is defined to be the smallest ideal J not contained in any minimal prime that satisfies

$$\varphi((\mathfrak{a}^{\lceil t(p^e-1)\rceil} J)^{1/p^e}) \subseteq J$$

for all $\varphi \in \mathrm{Hom}_R(R^{1/p^e}, R)$ (ranging over all $e \geq 1$).

The existence of such a smallest nonzero ideal is a non-trivial statement; again, the crucial point due to Hochster and Huneke in their proof of the existence of completely stable test elements. See [35, 83].

Just as multiplier ideals of pairs are particularly appealing to work with when the ambient scheme is regular, the same is true of test ideals. Although it is not obvious, the following characterization of test ideals below (from [12]) is equivalent to Schwede's in this more restrictive setting.

5.1. Test ideals in regular ambient rings. Let R be an F-finite regular domain, and $\mathfrak{a}$ any ideal of R. For each R-linear map $\phi : F^e_* R \to R$, we consider the image of $\mathfrak{a}$ under ϕ.

[12] This map is only as canonical as the choice of ω_R, so the "the" is slightly misleading. Of course in geometric situations where the canonical module is defined by differential forms, there is a canonical choice.

[13] Here we assume that R is reduced and of dimension greater than zero. In general, the definition of Cartier algebra is slightly more technical, but it reduces to this under very mild conditions. See [10].

Ranging over all $\phi \in \mathrm{Hom}_R(F^e_* R, R)$, we get an ideal

$$\mathfrak{a}^{[1/p^e]} := \sum_{\phi \in \mathrm{Hom}_R(F^e_* R, R)} \phi(\mathfrak{a}).$$

Lemma 5.2. *For any ideal $\mathfrak{a}$ in a Frobenius split ring R, we have*

$$\mathfrak{a}^{[1/p^e]} \subset (\mathfrak{a}^p)^{[1/p^{e+1}]}.$$

The lemma implies that given a rational number $t = n/p^e$ whose denominator is a power of p, there is an increasing sequence of ideals:

$$(\mathfrak{a}^n)^{[1/p^e]} \subset (\mathfrak{a}^{np})^{[1/p^{e+1}]} \subset (\mathfrak{a}^{np^2})^{[1/p^{e+2}]} \subset \cdots \tag{5.1}$$

which must eventually stabilize by the Noetherian property of the ring.

Definition 5.3 ([12]). Let R be an F-finite regular ring of characteristic p and let $\mathfrak{a}$ be an ideal of R. For each $t \in \mathbb{R}_{\geq 0}$, we define

$$\tau(R, \mathfrak{a}^t) := \bigcup_{e \in \mathbb{N}} (\mathfrak{a}^{\lceil tp^e \rceil})^{[\frac{1}{p^e}]}.$$

The sequence $\frac{\lceil tp^e \rceil}{p^e}$ could be replaced by any sequence of rational numbers (whose denominators are a power of p) converging to t from above: the lemma guarantees that all give an ascending chain of ideals stabilizing to the test ideal.

5.2. Properties of test ideals. All the properties of multiplier ideals on smooth ambient varieties carry over to test ideals with exceptionally simple proofs in this setting [12, 97]:

Theorem 5.4. *Let R be an F-finite regular ring of characteristic p, with ideals $\mathfrak{a}, \mathfrak{b}$. The following properties of the test ideal hold:*

(1) $\mathfrak{a} \subseteq \mathfrak{b} \ \Rightarrow\ \tau(R, \mathfrak{a}^t) \subseteq \tau(R, \mathfrak{b}^t)$ *for all* $t \in \mathbb{R}_{>0}$.

(2) $t \geq t' \ \Rightarrow\ \tau(R, \mathfrak{a}^t) \subseteq \tau(R, \mathfrak{a}^{t'})$.

(3) $\tau(R, (\mathfrak{a}^n)^t) = \tau(R, \mathfrak{a}^{nt})$ *for each positive integer n and each* $t \in \mathbb{R}_{>0}$.

(4) *Let W be a multiplicatively closed set in R, then*

$$W^{-1}\tau(R, \mathfrak{a}^t) = \tau(W^{-1}R, (W^{-1}\mathfrak{a})^t).$$

(5) *Let $\overline{\mathfrak{a}}$ denote the integral closure of $\mathfrak{a}$ in R. Then*

$$\tau(R, \overline{\mathfrak{a}}^t) = \tau(R, \mathfrak{a}^t) \text{ for all } t.$$

(6) $\mathfrak{a} \subseteq \tau(R, \mathfrak{a})$.

(7) *For each $t \in \mathbb{R}_{>0}$, there exists an $\varepsilon > 0$ such that $\tau(R, \mathfrak{a}^{t'}) = \tau(R, \mathfrak{a}^t)$ for all $t' \in [t, t + \varepsilon)$.*

(8) (Briançon-Skoda Theorem[14]) *If $\mathfrak{a}$ can be generated by r elements, then for each integer $\ell \geq r$ we have*

$$\tau(R, \mathfrak{a}^\ell) = \mathfrak{a}\tau(R, \mathfrak{a}^{\ell-1}).$$

[14] Also called Skoda's Theorem, or "Briançon-Skoda theorem with coefficients." See *e.g.* [2].

(9) (Restriction Theorem) *Let $x \in R$ be a regular parameter and $\mathfrak{a} \bmod x$ denote the image of $\mathfrak{a}$ in $R/(x)$, then*

$$\tau(R/(x), (\mathfrak{a} \bmod x)^t) \subseteq \tau(R, \mathfrak{a}^t) \bmod x.$$

(10) (Subadditivity Theorem) *If R is essentially of finite type over a perfect field, then $\tau(R, \mathfrak{a}^{tn}) \subseteq \tau(R, \mathfrak{a}^t)^n$ for all $t \in \mathbb{R}_{\geq 0}$ and all $n \in \mathbb{N}$.*[15]

Remark 5.5. Elementary proofs of all these properties are gathered together in [97]. The key fact used is the flatness of Frobenius. Most of these properties hold more generally, though most require some sort of restriction on the singularities of R and the proofs are considerably more technical. See *e.g.* [37, 98, 100].

5.3. Asymptotic test ideals and an application to symbolic powers. Asymptotic test ideal can be defined analogous to the asymptotic multiplier ideal first defined in [21]. This is quite simple and elegant in the case of a regular ambient scheme. See also [34, 100] for a more general setting.

Definition 5.6 ([21]). A sequence of ideals $\{\mathfrak{a}_n\}_{n\in\mathbb{N}}$ is called a *graded sequence* of ideals if

$$\mathfrak{a}_n \mathfrak{a}_m \subseteq \mathfrak{a}_{n+m}$$

for all n, m.

Graded sequences arise naturally in many contexts. For example, the sequence of base loci of the powers of a fixed line bundle form a graded sequence of ideals on a variety. The symbolic powers $\{\mathfrak{a}^{(n)}\}_{n\in\mathbb{N}}$ of any ideal $\mathfrak{a}$ in any ring form a graded sequence.

Given any graded sequence of ideals $\{\mathfrak{a}_n\}$, it follows from the definition and Property 5.4(1) that

$$\tau(R, \mathfrak{a}_n) = \tau(R, (\mathfrak{a}_n^m)^{1/m}) \subseteq \tau(R, \mathfrak{a}_{mn}^{1/m}).$$

In other words, the collection

$$\{\tau(R, \mathfrak{a}_m^{1/m})\}_{m\in\mathbb{N}}$$

has the property that any two ideals are dominated by a third in the collection. Since R is noetherian, this collection must have a maximal element; this stable ideal is called the **asymptotic test ideal:**

Definition 5.7. The **n-th asymptotic test ideal** of the graded sequence $\{\mathfrak{a}_n\}_{n\in\mathbb{N}}$ is the ideal

$$\tau_\infty(R, \mathfrak{a}_n) := \sum_{\ell\in\mathbb{N}} \tau(R, \mathfrak{a}_{\ell n}^{1/\ell}),$$

which is equal to

$$\tau(R, \mathfrak{a}_{mn}^{1/m})$$

for sufficiently large and divisible m.

By definition, it is clear that $\tau_\infty(R, \mathfrak{a}_n)$ satisfies appropriate analogs of all the properties listed in Theorem 5.4. In particular, we have the following consequence of the subadditivity theorem:

[15] More generally, the subadditivity property guarantees that for the *mixed test ideal* $\tau(\mathfrak{a}^t \mathfrak{b}^s)$ defined analogously as $\tau(\mathfrak{a}^{\lceil sp^e \rceil} \mathfrak{b}^{\lceil tp^e \rceil})^{[1/p^e]}$ for $e \gg 0$, we have $\tau(\mathfrak{a}^t \mathfrak{b}^s) \subseteq \tau(\mathfrak{a}^t)\tau(\mathfrak{b}^s)$ for all $t, s \in \mathbb{R}_{\geq 0}$.

Corollary 5.8. *For any graded sequence in an F-finite regular ring R, we have $\tau_\infty(R, \mathfrak{a}_{nm}) \subset (\tau_\infty(R, \mathfrak{a}_n))^m$ for all $n, m \in \mathbb{N}$.*

As an application, we prove

Theorem 5.9 (Ein-Lazasfeld-Smith; Hochster-Huneke). *Let I be an unmixed (e.g. prime) ideal in $k[x_1, \ldots, x_d]$. Then*

$$I^{(dn)} \subseteq I^n \text{ for all } n \in \mathbb{N}.$$

Proof of Theorem 5.9. The characteristic zero case follows from the prime characteristic case by a standard argument. We consider the graded sequence of ideals $\{I^{(n)}\}_{n\in\mathbb{N}}$. According to Theorem 5.4(3), we have $I^{(dN)} \subseteq \tau_\infty(R, I^{(dN)})$. By Corollary 5.8, we have

$$\tau_\infty(R, I^{(dN)}) \subseteq \tau_\infty(R, I^{(d)})^N$$

for all N. Hence it is enough to check that $\tau_\infty(I^{(d)}) \subseteq I$. For this, we can check at each associated prime $\mathfrak{p}$ of I, which means essentially that we can assume that R is local and that I is primary to the maximal ideal; that is, we need to show that

$$\tau_\infty(R_\mathfrak{p}, (I^d R_\mathfrak{p})) \subset I R_\mathfrak{p}.$$

In $R_\mathfrak{p}$, there is a reduction of I that can be generated by $\dim(R_\mathfrak{p}) \leq d$ elements, and hence according to Properties 5.4(5) we may assume that I itself can be generated by d elements. Then the Briançon-Skoda property 5.4(8) tells us

$$\tau_\infty(R_\mathfrak{p}, (I^d R_\mathfrak{p})) \subseteq I.$$

This finishes the proof. □

This theorem was first proved in [21] in characteristic zero, using asymptotic multiplier ideals. Later Hochster and Huneke gave a tight closure proof in the characteristic p case [46]. Takagi-Yoshida ([100]) generalized this result using test ideals. Our proof here is a straightforward adaptation of the original multiplier ideal proof in [21].

5.4. F-pure thresholds and F-jumping exponents. Having defined test ideals of a pair, analogs of the log canonical threshold and the jumping exponents (Cf. [22]) are the next obvious step. For simplicity, we restrict attention to the case an an ambient regular ring; the general case is much more technical.

Definition 5.10. Let R be an F-finite regular ring, and $\mathfrak{a}$ an ideal.

(1) The F-pure threshold $(R, \mathfrak{a})$ is the supremum, over all positive t, such that $\tau(R, \mathfrak{a}^t) = R$ [37];

(2) More generally, an F-jumping exponent is a real number ξ such that for all $\varepsilon > 0$, $\tau(R, \mathfrak{a}^{\xi-\varepsilon})$ is strictly larger than $\tau(R, \mathfrak{a}^\xi)$ [22].

Given a pair defined over a field of characteristic zero, one may "reduce mod p" and compare the F-pure threshold that arise to the log canonical threshold. The F-pure threshold of any characteristic p model is always less than or equal to the log canonical threshold; also as p goes to infinity, the F-pure thresholds converge to the log canonical threshold (this

follows from [36, 37]). In all known examples, there are in fact infinitely many p for which the F-pure threshold *equals* the log canonical threshold. An open question for over fifteen years is to show that this is always the case. If true, it follows that log canonical pairs (of characteristic zero) have dense Frobenius split type. Mustata and Srinivas have recently shown that this conjecture would follow from the following *weak ordinarily conjecture*: if X is a projective variety over a field of characteristic zero, then the Frobenius map acts injectively on $H^{\dim X}(X, \mathcal{O}_X)$ for infinitely many "mod p reductions" [71] (the extension to the singular setting is in [8]). The difficulty of these conjectures likely lies in some hard number theory. See the survey [70].

The F-pure threshold is very difficult to compute, with a fractal-like behavior in many cases, see [38, 39] and [40] for concrete computations of F-thresholds. See also [88]. See [5] for a beginners guide to the subject of F-pure threshold.

Discreteness and rationality of the F-jumping exponents has been another active research topic; it is vexing that the analogous properties for multiplier ideals are more-or-less obvious in characteristic zero. The F-jumping exponents are shown to be discrete and rational in the case of an ambient regular ring of finite type [12]. Since then, these results have been generalized to the $\mathbb{Q}$-Gorenstein case [15, 79]. The paper [86] gives an exceptionally well-written account of the state of the art. See also [13, 51, 87].

Acknowledgements. Deep thanks to my good friend Jeffrey Lagarias, who convinced be to write this paper when the demands of life threatened to cause me to punt, and to my dear fiancé Kai Rajala, who relieved me for enough hours to pull it off. Huge thanks also to my faithful post-doc Angelica Benito, who heroically took on massive editing tasks at the last minute, and my good friends Daniel Hernandez and Emily Witt for their careful proofreading.

References

[1] I. M. Aberbach and F. Enescu, *The structure of F-pure rings*, Math. Z. **250** (2005), 791–806.

[2] I. M. Aberbach and A. Hosry, *A less restrictive Briançon-Skoda theorem with coefficients*, J. Algebra **345** (2011), 72–80.

[3] I. M. Aberbach and G. J. Leuschke, *The F-signature and strong F-regularity*, Math. Res. Lett. **10** (2003), 51–56.

[4] I. M. Aberbach and B. MacCrimmon, *Some results on test ideals*, Proc. Edinburgh Math. Soc. (2) **42** (1999), 541–549.

[5] A. Benito, E. Faber, and K. E. Smith, *Measuring singularities with Frobenius: the basics*, Commutative Algebra: Expository Papers Dedicated to David Eisenbud on the Occasion of His 65th Birthday, Springer, New York, 57–97, 2013.

[6] A. Benito, G. Muller, J. Rajchgot, and K. E. Smith, *Singularities of locally acyclic cluster algebras*, Preprint 2014.

[7] A. Berenstein, S. Fomin, and A. Zelevinsky, *Cluster algebras. III*. Upper bounds and double Bruhat cells, Duke Math. J. **126** (2005), 1–52.

[8] B. Bhatt, K. Schwede, and S. Takagi, *The weak ordinarity conjecture and F-singularities*, arXiv:1307.3763.

[9] C. Birkar, P. Cascini, C. D. Hacon, and J. McKernan, *Existence of minimal models for varieties of log general type*, J. Amer. Math. Soc. **23** (2010), 405–468.

[10] M. Blickle, *Test ideals via algebras of p^{-e}-linear maps*, J. Algebraic Geom. **22** (2013), 49–83.

[11] M. Blickle and G. Böckle, *Cartier modules: finiteness results*, J. Reine Angew. Math. **661** (2011), 85–123.

[12] M. Blickle, M. Mustata, and K. E. Smith, *Discreteness and rationality of F-thresholds*, Michigan Math. J. **57** (2008), 43–61.

[13] ———, *F-thresholds of hypersurfaces*, Trans. Amer. Math. Soc. **361** (2009), 6549–6565.

[14] M. Blickle and K. Schwede, *p^{-1}-linear maps in algebra and geometry*, Commutative Algebra: Expository Papers Dedicated to David Eisenbud on the Occasion of His 65th Birthday, Springer, New York, 123–205 (2013).

[15] M. Blickle, K. Schwede, S. Takagi, and W. Zhang, *Discreteness and rationality of F-jumping numbers on singular varieties*, Math. Ann. **347** (2010), 917–949.

[16] M. Blickle, K. Schwede, and K. Tucker, *F-signature of pairs and the asymptotic behavior of Frobenius splittings*, Adv. Math. **231** (2012), 3232–3258.

[17] ———, *F-signature of pairs: continuity, p-fractals and minimal log discrepancies*, J. Lond. Math. Soc. (2) **87** (2013), 802–818.

[18] J.-F. Boutot, *Singularités rationnelles et quotients par les groupes réductifs*, Invent. Math. **88** (1987), 65–68.

[19] H. Brenner, *Irrational Hilbert-Kunz multiplicities*, arXiv:1305.5873.

[20] M. Brion and S. Kumar, *Frobenius splitting methods in geometry and representation theory*, Progress in Mathematics, Birkhäuser Boston Inc. **231** x+250, Boston, MA (2005).

[21] L. Ein, R. Lazarsfeld, and K. E. Smith, *Uniform bounds and symbolic powers on smooth varieties*, Invent. Math. **144** (2001), 241–252.

[22] L. Ein, R. Lazarsfeld, K. E. Smith, and D. Varolin, *Jumping coefficients of multiplier ideals*, Duke Math. J. **123** (2004), 469–506.

[23] R. Elkik, *Rationalité des singularités canoniques*, Invent. Math. **64** (1981), 1–6.

[24] F. Enescu and M. Hochster, *The Frobenius structure of local cohomology*, Algebra Number Theory **2** (2008), 721–754.

[25] T. de Fernex and C. D. Hacon, *Singularities on normal varieties*, Compos. Math. **145** (2009), 393–414.

[26] S. Fomin and A. Zelevinsky, *Cluster algebras. I. Foundations*, J. Amer. Math. Soc. **15** (2002), 497–529.

[27] ———, *Cluster algebras. II. Finite type classification*. Invent. Math. **154** (2003), no.1 63–121.

[28] O. Gabber, *Notes on some t-structures*, Geometric aspects of Dwork theory. Vol. I, II, Walter de Gruyter GmbH & Co. KG, Berlin, 711–734 (2004).

[29] Y. Gongyo, S. Okawa, A. Sannai, and S. Takagi, *Characterization of varieties of Fano type via singularities of Cox rings*, arXiv:1201.1133.

[30] Y. Gongyo and S. Takagi, *Surfaces of globally F-regular and F-split type*, arXiv: 1305.3056.

[31] C. Hacon and C. Xu, *On the three dimensional minimal model program in positive characteristic*, preprint arXiv:1302.0298.

[32] N. Hara, *A characterization of rational singularities in terms of injectivity of Frobenius maps*, Amer. J. Math. **120** (1998), 981–996.

[33] ———, *Geometric interpretation of tight closure and test ideals* Trans. Amer. Math. Soc. **353** (2001), 5 1885–1906.

[34] ———, *A* characteristic p analog of multiplier ideals and applications, Comm. Algebra **33** (2005), 3375–3388.

[35] N. Hara and S. Takagi, *On a generalization of test ideals*, Nagoya Math. J. **175** (2004), 59–74.

[36] N. Hara and K.-I. Watanabe, *F-regular and F-pure rings vs. log terminal and log canonical singularities*, J. Algebraic Geom. **11** (2002), 363–392.

[37] N. Hara and K.-I. Yoshida, *A generalization of tight closure and multiplier ideals*, Trans. Amer. Math. Soc. **355** (2003), 3143–3174.

[38] D. J. Hernández, *F-invariants of diagonal hypersurfaces*, to appear in Proceedings of the AMS, arXiv:1112.2425.

[39] ———, *F-pure thresholds of binomial hypersurfaces*, to appear in Proceedings of the AMS, arXiv:1112.2427.

[40] D. J. Hernández and P. Teixeira, *F-threshold functions: Syzygy gap fractals and the two variable homogeneous case*, Preprint 2014.

[41] M. Hochster and J. L. Roberts, *Rings of invariants of reductive groups acting on regular rings are Cohen-Macaulay*, Advances in Math. **13** (1974), 115–175.

[42] ———, *The purity of the Frobenius and local cohomology*, Advances in Math. **21** (1976), 117–172.

[43] M. Hochster and C. Huneke, *Tight closure and strong F-regularity*, Mém. Soc. Math. France (N.S.) (1989), 119–133.

[44] ———, *Tight closure, invariant theory, and the Briançon-Skoda theorem*, J. Amer. Math. Soc. **3** (1990), 31–116.

[45] ———, *Tight closure of parameter ideals and splitting in module-finite extensions*, J. Algebraic Geom. **3** (1994), 599–670.

[46] ———, *Comparison of symbolic and ordinary powers of ideals*, Invent. Math. **147** (2002), 349–369.

[47] ———, *Tight closure in equal characteristic zero*, A preprint of a manuscript (2006).

[48] C. Huneke, *Tight closure and its applications*, CBMS Regional Conference Series in Mathematics **88**, Published for the Conference Board of the Mathematical Sciences, Washington, DC x+137 (1996).

[49] ———, *Hilbert-Kunz multiplicity and the F-signature*, Commutative Algebra: Expository Papers Dedicated to David Eisenbud on the Occasion of His 65th Birthday, Springer, New York, 485–525 (2013).

[50] C. Huneke and G. J. Leuschke, *Two theorems about maximal Cohen-Macaulay modules*, Math. Ann. **324** (2002), 391–404.

[51] M. Katzman, G. Lyubeznik, and W. Zhang, *On the discreteness and rationality of F-jumping coefficients*, J. Algebra **322** (2009), 3238–3247.

[52] Y. Kawamata, *Subadjunction of log canonical divisors. II*, Amer. J. Math. **120** (1998), 893–899.

[53] D. S. Keeler, *Fujita's conjecture and Frobenius amplitude*, Amer. J. Math. **130**(5) (2008), 1327–1336.

[54] J. Kollár, *Singularities of pairs*, Algebraic geometry – Santa Cruz 1995, Proc. Sympos. Pure Math., Amer. Math. Soc. **62** (1997), 221–287.

[55] ———, *AimPL: The minimal model program in characteristic p*, available at http://aimpl.org/minimalmodcharp.

[56] S. Kumar and V. B. Mehta, *Finiteness of the number of compatibly split subvarieties*, Int. Math. Res. Not. IMRN **19** (2009), 3595–3597.

[57] E. Kunz, *Characterizations of regular local rings for characteristic p*, Amer. J. Math. **91** (1969), 772–784.

[58] ———, *On Noetherian rings of characteristic p*, Amer. J. Math. **98** (1976), 999–1013.

[59] N. Lauritzen, U. Raben-Pedersen, and J. F. Thomsen, *Global F-regularity of Schubert varieties with applications to D-modules*, J. Amer. Math. Soc. **19** (2006), 345–355.

[60] J. Lipman and B. Teissier, *Pseudorational local rings and a theorem of Briançon-Skoda about integral closures of ideals*, Michigan Math. J. **28** (1981), 97–116.

[61] G. Lyubeznik and K. E. Smith, *Strong and weak F-regularity are equivalent for graded rings*, Amer. J. Math. **121** (1999), 1279–1290.

[62] ———, *On the commutation of the test ideal with localization and completion*, Trans. Amer. Math. Soc. **353** (2001), 3149–3180.

[63] B. C. Maccrimmon, *Strong F-regularity and boundedness questions in tight closure*, ProQuest LLC, Ann Arbor, MI, 1996. Thesis (Ph.D.)–University of Michigan.

[64] B. Mazur, *Frobenius and the Hodge filtration*, Bull. Amer. Math. Soc. **78** (1972), 653–667.

[65] V. B. Mehta and A. Ramanathan, *Frobenius splitting and cohomology vanishing for Schubert varieties*, Ann. of Math. (2) **122** (1985), 27–40.

[66] V. B. Mehta and V. Srinivas, *A characterization of rational singularities*, Asian J. Math. **1** (1997), 249–271.

[67] P. Monsky, *The Hilbert-Kunz function*, Math. Ann. **263** (1983), 43–49.

[68] G. Muller, *Locally acyclic cluster algebras*, Adv. Math. **233** (2013), 207–247.

[69] G. Muller and D. E Speyer, *Cluster Algebras of Grassmannians are Locally Acyclic*, arXiv:1401.5137.

[70] M. Mustata, *IMPANGA lecture notes on log canonical thresholds*, In Contributions to algebraic geometry, EMS Ser. Congr. Rep. Eur. Math. Soc., Zürich, 2012, pp. 407–442. Notes by Tomasz Szemberg.

[71] M. Mustata and V. Srinivas, *Ordinary varieties and the comparison between multiplier ideals and test ideals*, Nagoya Math. J. **204** (2011), 125–157.

[72] Z. Patakfalvi, *Semi-positivity in positive characteristics*, arXiv:1208.5391.

[73] A. Ramanathan, *Frobenius splitting and Schubert varieties*, In Proceedings of the Hyderabad Conference on Algebraic Groups (Hyderabad, 1989) (1991), Manoj Prakashan, Madras, pp. 497–508.

[74] S. Ramanan and A. Ramanathan, *Projective normality of flag varieties and Schubert varieties*, Invent. Math. **79** (1985), 217–224.

[75] K. Schwede, *F-adjunction*, Algebra Number Theory **3** (2009), 907–950.

[76] ———, *F-injective singularities are Du Bois*, Amer. J. Math. **131** (2009), 445–473.

[77] ———, *A refinement of sharply F-pure and strongly F-regular pairs*, J. Commut. Algebra **2** (2010), 91–109.

[78] ———, *Centers of F-purity*, Math. Z. **265** (2010), 687–714.

[79] ———, *A note on discreteness of F-jumping numbers*, Proc. Amer. Math. Soc. **139** (2011), 3895–3901.

[80] ———, *Test ideals in non-$\mathbb{Q}$-Gorenstein rings*, Trans. Amer. Math. Soc. **363** (2011), 5925–5941.

[81] ———, *A canonical linear system associated to adjoint divisors in characteristic $p > 0$*, arXiv:1107.3833.

[82] K. Schwede and K. E. Smith, *Globally F-regular and log Fano varieties*, Adv. Math. **224** (2010), 863–894.

[83] K. Schwede and K. Tucker, *A survey of test ideals*, In Progress in Commutative Algebra 2. Closures, Finiteness and Factorization, C. Francisco, L. C. Klinger, S. M. Sather-Wagstaff, and J. C. Vassilev, Eds. Walter de Gruyter GmbH & Co. KG, Berlin, 2012, pp. 39–99.

[84] ———, *On the behavior of test ideals under finite morphisms*, arXiv:1003.4333, to appear in Journal of Algebraic Geometry.

[85] ———, *On the number of compatibly Frobenius split subvarieties, prime F-ideals, and log canonical centers*, Ann. Inst. Fourier (Grenoble) **60** (2010), 1515–1531.

[86] ———, *Test ideals of non-principal ideals: Computations, Jumping Numbers, Alterations and Division Theorems*, arXiv:1212.6956.

[87] K. Schwede, K. Tucker, and W. Zhang, *Test ideals via a single alteration and discreteness and rationality of F-jumping numbers*, Math. Res. Lett. **19** (2012), 191–197.

[88] T. Shibuta and S. Takagi, *Log canonical thresholds of binomial ideals*, Manuscripta Math. **130** (2009), no. 1, 45–61.

[89] K. E. Smith, *Tight closure of parameter ideals*, Invent. Math. **115** (1994), 41–60.

[90] ———, *F-rational rings have rational singularities*, Amer. J. Math. **119** (1997), 159–180.

[91] ———, *Vanishing, singularities and effective bounds via prime characteristic local algebra*, In Algebraic geometry – Santa Cruz 1995, vol. **62** of Proc. Sympos. Pure Math. Amer. Math. Soc., Providence, RI, 1997, pp. 289–325.

[92] ———, *Fujita's freeness conjecture in terms of local cohomology*, J. Algebraic Geom. **6**, 3 (1997), 417–429.

[93] ———, *Globally F-regular varieties: applications to vanishing theorems for quotients of Fano varieties*, Michigan Math. J. **48** (2000), 553–572.

[94] ———, *The multiplier ideal is a universal test ideal*, Comm. Algebra **28** (2000), no. 12, 5915–5929.

[95] ———, *Tight closure commutes with localization in binomial rings*, Proc. Amer. Math. Soc. **129** (2001), 667–669.

[96] K. E. Smith and M. Van den Bergh, *Simplicity of rings of differential operators in prime characteristic*, Proc. London Math. Soc. (3) **75** (1997), 32–62.

[97] K. E. Smith and W. Zhang, *Frobenius splitting in Commutative Algebra*, Preprint 2014.

[98] S. Takagi, *Formulas for multiplier ideals on singular varieties*, Amer. J. Math. **128** (2006), 1345–1362.

[99] S. Takagi and K.-i. Watanabe, *On F-pure thresholds*, J. Algebra **282** no. 1, (2004), 278–297.

[100] S. Takagi and K.-i. Yoshida, *Generalized test ideals and symbolic powers*, Michigan Math. J. **57** (2008), 711–724 .

[101] K. Tucker, *F-signature exists*, Invent. Math. **190** (2012), 743–765 .

[102] M. R. Von Korff, *The F-Signature of Toric Varieties*, ProQuest LLC, Ann Arbor, MI, 2012. Thesis (Ph.D.)–University of Michigan.

Department of Mathematics, University of Michigan, Ann Arbor, MI 48109
E-mail: kesmith@umich.edu

3. Number Theory

Motivic periods and $\mathbb{P}^1\backslash\{0,1,\infty\}$

Francis Brown

Abstract. This is a review of the theory of the motivic fundamental group of the projective line minus three points, and its relation to multiple zeta values.

Mathematics Subject Classification (2010). Primary 11M32; Secondary 14C15.

Keywords. Belyi's theorem, multiple zeta values, mixed Tate motives, modular forms.

1. Introduction

The role of the projective line minus three points $X = \mathbb{P}^1\backslash\{0,1,\infty\}$ in relation to Galois theory can be traced back to Belyi's theorem [4] (1979):

Theorem 1.1. *Every smooth projective algebraic curve defined over $\overline{\mathbb{Q}}$ can be realised as a ramified cover of $\mathbb{P}^1$, whose ramification locus is contained in $\{0,1,\infty\}$.*

Belyi deduced that the absolute Galois group of $\mathbb{Q}$ acts faithfully on the profinite completion of the fundamental group of X, i.e., the map

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Aut}(\widehat{\pi}_1(X(\mathbb{C}), b)) \tag{1.1}$$

where $b \in X(\mathbb{Q})$, is injective. In his famous proposal '*Esquisse d'un programme*' in 1984 [21], Grothendieck suggested studying the absolute Galois group of $\mathbb{Q}$ via its action on completions of fundamental groups of moduli spaces of curves $\mathcal{M}_{g,n}$ of genus g with n ordered marked points (X being isomorphic to $\mathcal{M}_{0,4}$) and their interrelations. A few years later, at approximately the same time, these ideas were developed in somewhat different directions in three enormously influential papers due to Drinfeld, Ihara, and Deligne [10, 15, 26]. Ihara's 1990 ICM talk gives a detailed account of the subject at that time [27]. However, the problem of determining the image of the map (1.1) remains completely open to this day.

In this talk I will mainly consider the pro-unipotent completion of the fundamental group of X, which seems to be a more tractable object than its profinite version, and closely follow the point of view of Deligne, and Ihara (see [27], §5).

1.1. Unipotent completion. Deligne showed [10] that the pro-unipotent completion of $\pi_1(X)$ carries many extra structures corresponding to the realisations of an (at the time) hypothetical category of mixed Tate motives over the integers. Since then, the motivic framework has now been completely established due to the work of a large number of different authors including Beilinson, Bloch, Borel, Levine, Hanamura, and Voevodsky. The definitive reference is [14], §§1-2.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

1. There exists an abstract Tannakian category $\mathcal{MT}(\mathbb{Z})$ of mixed Tate motives unramified over $\mathbb{Z}$. It is a $\mathbb{Q}$-linear subcategory of the category $\mathcal{MT}(\mathbb{Q})$ of mixed Tate motives over $\mathbb{Q}$ obtained by restricting certain Ext groups. It is equivalent to the category of representations of an affine group scheme G^{dR} which is defined over $\mathbb{Q}$ and is a semi-direct product
$$G^{dR} \cong U^{dR} \rtimes \mathbb{G}_m \ .$$
The subgroup U^{dR} is pro-unipotent, and its graded Lie algebra (for the action of $\mathbb{G}_m$) is isomorphic to the free graded Lie algebra with one generator
$$\sigma_3, \sigma_5, \sigma_7, \ldots$$
in every odd negative degree ≤ -3. The essential reason for this is that the algebraic K-theory of the integers $K_{2n-1}(\mathbb{Z})$ has rank 1 for $n = 3, 5, 7, \ldots$, and rank 0 otherwise, as shown by Borel [2, 3]. Note that the elements σ_{2n+1} are only well-defined modulo commutators.
2. The pro-unipotent completion $\pi_1^{un}(X, \vec{1}_0, -\vec{1}_1)$ is the Betti realisation of an object, called the motivic fundamental groupoid (denoted by π_1^{mot}), whose affine ring is a limit of objects in the category $\mathcal{MT}(\mathbb{Z})$.

The majority of these notes will go into explaining 2 and some of the ideas behind the following motivic analogue of Belyi's injectivity theorem (1.1):

Theorem 1.2. *G^{dR} acts faithfully on the de Rham realisation of $\pi_1^{mot}(X, \vec{1}_0, -\vec{1}_1)$.*

This theorem has an ℓ-adic version which can be translated into classical Galois theory ([27], §5.2), and relates to some questions in the literature cited above. Unlike Belyi's theorem, which is geometric, the proof of theorem 1.2 is arithmetic and combinatorial. The main ideas came from the theory of multiple zeta values.

1.2. Multiple zeta values. Let $n_1, \ldots, n_r$ be integers ≥ 1 such that $n_r \geq 2$. Multiple zeta values are defined by the convergent nested sums
$$\zeta(n_1, \ldots, n_r) = \sum_{1 \leq k_1 < \ldots < k_r} \frac{1}{k_1^{n_1} \ldots k_r^{n_r}} \quad \in \mathbb{R} \ .$$
The quantity $N = n_1 + \ldots + n_r$ is known as the weight, and r the depth. Multiple zeta values were first studied by Euler (at least in the case $r = 2$) and were rediscovered independently in mathematics by Zagier and Ecalle, and in perturbative quantum field theory by Broadhurst and Kreimer. They satisfy a vast array of algebraic relations which are not completely understood at the time of writing.

The relationship between these numbers and the fundamental group comes via the theory of iterated integrals, which are implicit in the work of Picard and were rediscovered by Chen and Dyson. In general, let M be a differentiable manifold and let $\omega_1, \ldots, \omega_n$ be smooth 1-forms on M. Consider a smooth path $\gamma : (0, 1) \to M$. The iterated integral of $\omega_1, \ldots, \omega_n$ along γ is defined (when it converges) by
$$\int_\gamma \omega_1 \ldots \omega_n = \int_{0 < t_1 < \ldots < t_n < 1} \gamma^*(\omega_1)(t_1) \ldots \gamma^*(\omega_n)(t_n) \ .$$

Kontsevich observed that when $M = X(\mathbb{C})$ and $\gamma(t) = t$ is simply the inclusion $(0,1) \subset X(\mathbb{R})$, one has the following integral representation

$$\zeta(n_1,\ldots,n_r) = \int_\gamma \omega_1 \underbrace{\omega_0\ldots\omega_0}_{n_1-1}\omega_1\underbrace{\omega_0\ldots\omega_0}_{n_2-1}\ldots\omega_1\underbrace{\omega_0\ldots\omega_0}_{n_r-1} \tag{1.2}$$

where $\omega_0 = \frac{dt}{t}$ and $\omega_1 = \frac{dt}{1-t}$. I will explain in §2.1 that this formula allows one to interpret multiple zeta values as periods of the pro-unipotent fundamental groupoid of X. The action of the motivic Galois group G^{dR} on the (de Rham version of) the latter should translate, via Grothendieck's period conjecture, into an action on multiple zeta values themselves. Thus one expects multiple zeta values to be a basic example in a Galois theory of transcendental numbers ([1], §23.5); the action of the Galois group should preserve all their algebraic relations.

Of course, Grothendieck's period conjecture is not currently known, so there is no well-defined group action on multiple zeta values. This can be circumvented using motivic multiple zeta values. The action of G^{dR} on the de Rham fundamental group of X can then be studied via its action on these objects.

1.3. Motivic periods. Let T be a neutral Tannakian category over $\mathbb{Q}$ with two fiber functors $\omega_B, \omega_{dR} : T \to \mathrm{Vec}_{\mathbb{Q}}$. Define the ring of motivic periods to be the affine ring of functions on the scheme of tensor isomorphisms from ω_{dR} to ω_B

$$\mathcal{P}_T^{\mathfrak{m}} = \mathcal{O}(\underline{\mathrm{Isom}}_T(\omega_{dR},\omega_B)) \ .$$

Every motivic period can be constructed from an object $M \in T$, and a pair of elements $w \in \omega_{dR}(M), \sigma \in \omega_B(M)^\vee$. Its matrix coefficient is the function

$$\phi \mapsto \langle \phi(w), \sigma\rangle : \underline{\mathrm{Isom}}_T(\omega_{dR},\omega_B) \to \mathbb{A}^1_{\mathbb{Q}}$$

where $\mathbb{A}^1_{\mathbb{Q}}$ is the affine line over $\mathbb{Q}$, and defines an element denoted $[M,w,\sigma]^{\mathfrak{m}} \in \mathcal{P}_T^{\mathfrak{m}}$. It is straightforward to write down linear relations between these symbols as well as a formula for the product of two such symbols. If, furthermore, there is an element $\mathrm{comp}_{B,dR} \in \underline{\mathrm{Isom}}_T(\omega_{dR},\omega_B)(\mathbb{C})$ we can pair with it to get a map

$$\mathrm{per} : \mathcal{P}_T^{\mathfrak{m}} \longrightarrow \mathbb{C} \tag{1.3}$$

called the period homomorphism. The ring $\mathcal{P}_T^{\mathfrak{m}}$ admits a left action of the group $G^{dR} = \underline{\mathrm{Isom}}_T(\omega_{dR},\omega_{dR})$, or equivalently, a left coaction

$$\mathcal{P}_T^{\mathfrak{m}} \longrightarrow \mathcal{O}(G^{dR}) \otimes_{\mathbb{Q}} \mathcal{P}_T^{\mathfrak{m}} \ . \tag{1.4}$$

Example 1.3. Let T be any category of mixed Tate motives over a number field. It contains the Lefschetz motive $\mathbb{L} = \mathbb{Q}(-1)$, which is the motive $H^1(\mathbb{P}^1\backslash\{0,\infty\})$. Its de Rham cohomology is the $\mathbb{Q}$-vector space spanned by the class $[\frac{dz}{z}]$ and its Betti homology is spanned by a small positive loop γ_0 around 0. The Lefschetz motivic period is

$$\mathbb{L}^{\mathfrak{m}} = [\mathbb{L}, [\tfrac{dz}{z}], [\gamma_0]] \in \mathcal{P}_T^{\mathfrak{m}} \ .$$

Its period is $\mathrm{per}(\mathbb{L}^{\mathfrak{m}}) = 2\pi i$. It transforms, under the rational points of the de Rham Galois group of T, by $\mathbb{L}^{\mathfrak{m}} \mapsto \lambda\mathbb{L}^{\mathfrak{m}}$, for any $\lambda \in \mathbb{Q}^\times$.

This construction can be applied to any pair of fiber functors to obtain different notions of motivic periods. Indeed, the elements of $\mathcal{O}(G^{dR})$ can be viewed as 'de Rham' motivic periods, or matrix coefficients of the form $[M, w, v]^{dR}$, where $w \in \omega_{dR}(M)$ and $v \in \omega_{dR}(M)^\vee$ (called framings). Whenever the fiber functors carry extra structures (such as 'complex conjugation' on ω_B or a 'weight' grading on ω_{dR}), then the ring of motivic periods inherits similar structures.

1.3.1. Motivic MZV's. Let $T = \mathcal{MT}(\mathbb{Z})$. The Betti and de Rham realisations provide functors ω_B, ω_{dR}, and integration defines a canonical element

$$\mathrm{comp}_{B,dR} \in \underline{\mathrm{Isom}}_T(\omega_{dR}, \omega_B)(\mathbb{C}).$$

Since the de Rham functor ω_{dR} is graded by the weight, the ring of motivic periods $\mathcal{P}^{\mathfrak{m}}_{\mathcal{MT}(\mathbb{Z})}$ is also graded.[1] It contains graded subrings

$$\mathcal{P}^{\mathfrak{m},+}_{\mathcal{MT}(\mathbb{Z}),\mathbb{R}} \subset \mathcal{P}^{\mathfrak{m},+}_{\mathcal{MT}(\mathbb{Z})} \subset \mathcal{P}^{\mathfrak{m}}_{\mathcal{MT}(\mathbb{Z})}$$

of geometric periods (periods of motives whose weights are ≥ 0), denoted by a superscript $+$, and those which are also invariant under complex conjugation (denoted by a subscript $\mathbb{R}$, since their periods lie in $\mathbb{R}$ as opposed to $\mathbb{C}$).

Next, one has to show that the integral (1.2) defines a period of an object M in $\mathcal{MT}(\mathbb{Z})$ (this can be done in several ways: [14, 20, 36]. This defines a matrix coefficient $[M, w, \sigma]^{\mathfrak{m}}$, where w encodes the integrand, and σ the domain of integration, which we call a motivic multiple zeta value (§2.2)

$$\zeta^{\mathfrak{m}}(n_1, \ldots, n_r) \in \mathcal{P}^{\mathfrak{m}}_{\mathcal{MT}(\mathbb{Z})} \,.$$

Its weight is $n_1 + \ldots + n_r$ and its period is (1.2). Most (but not all) of the known algebraic relations between multiple zeta values are also known to hold for their motivic versions. Motivic multiple zeta values generate a graded subalgebra

$$\mathcal{H} \subset \mathcal{P}^{\mathfrak{m},+}_{\mathcal{MT}(\mathbb{Z}),\mathbb{R}} \,. \tag{1.5}$$

The description §1.1, (1) of U^{dR} enables one to compute the dimensions of the motivic periods of $\mathcal{MT}(\mathbb{Z})$ in each degree by a simple counting argument:

$$\text{if} \quad d_N := \dim_{\mathbb{Q}} \left(\mathcal{P}^{\mathfrak{m},+}_{\mathcal{MT}(\mathbb{Z}),\mathbb{R}}\right)_N \quad \text{then} \quad \sum_{N\geq 0} d_N t^N = \frac{1}{1-t^2-t^3} \,. \tag{1.6}$$

This implies a theorem proved independently by Goncharov and Terasoma [14],[36].

Theorem 1.4. *The $\mathbb{Q}$-vector space spanned by multiple zeta values of weight N has dimension at most d_N, where the integers d_N are defined in* (1.6).

So far, this does not use the action of the motivic Galois group, only a bound on the size of the motivic periods of $\mathcal{MT}(\mathbb{Z})$. The role of $\mathbb{P}^1\backslash\{0, 1, \infty\}$ is that the automorphism group of its fundamental groupoid yields a formula for the coaction (1.4) on the motivic multiple zeta values (§2.5). The main theorem uses this coaction in an essential way, and is inspired by a conjecture of M. Hoffman [25].

[1] In the field of multiple zeta values, the 'weight' refers to one half of the Hodge-theoretic weight, so that $\mathbb{L}^{\mathfrak{m}}$ has degree 1 instead of 2. I shall adopt this terminology from here on.

Theorem 1.5. *The following set of motivic MZV's are linearly independent:*

$$\{\zeta^{\mathfrak{m}}(n_1,\ldots,n_r) \quad \textit{for} \quad n_i=\{2,3\}\}\ . \tag{1.7}$$

From the enumeration (1.6) of the dimensions, we deduce that $\mathcal{H}=\mathcal{P}^{\mathfrak{m},+}_{\mathcal{MT}(\mathbb{Z}),\mathbb{R}}$, and that (1.7) is a basis for $\mathcal{H}$. From this, one immediately sees that U^{dR} acts faithfully on $\mathcal{H}$, and theorem 1.2 follows easily. As a bonus we obtain that U^{dR} has canonical generators σ_{2n+1} (defined in §3.1), and, furthermore, by applying the period map we obtain the

Corollary 1.6. *Every multiple zeta value of weight N is a $\mathbb{Q}$-linear combination of $\zeta(n_1,\ldots,n_r)$, where $n_i\in\{2,3\}$ and $n_1+\ldots+n_r=N$.*

The point of motivic periods is that they give a mechanism for obtaining information on the action of G^{dR}, via the period map, from arithmetic relations between real numbers. For theorem 1.5, the required arithmetic information comes from a formula for $\zeta(2,\ldots,2,3,2,\ldots,2)$ proved by Zagier [38] using analytic techniques.

1.4. Transcendence of motivic periods. With hindsight, theorem 1.5 has less to do with mixed Tate motives, or indeed $\mathbb{P}^1\backslash\{0,1,\infty\}$, than one might think. Define a category H whose objects are given by the following data:

1. A finite-dimensional $\mathbb{Q}$-vector space M_B equipped with an increasing filtration called the weight, which is denoted by W.
2. A finite-dimensional $\mathbb{Q}$-vector space M_{dR} equipped with an increasing filtration W and a decreasing filtration F (the Hodge filtration).
3. An isomorphism $\mathrm{comp}_{B,dR}\ :\ M_{dR}\otimes\mathbb{C}\ \tilde{\to}\ M_B\otimes\mathbb{C}$ which respects the weight filtrations. The vector space M_B, equipped with W and the filtration F on $M_B\otimes\mathbb{C}$ induced by $\mathrm{comp}_{B,dR}$ is a $\mathbb{Q}$-mixed Hodge structure.

The category H is Tannakian ([10], 1.10), with two fiber functors, so it has a ring of motivic periods $\mathcal{P}^{\mathfrak{m}}_H$. Furthermore, the Betti and de Rham realisations define a functor $M\mapsto(M_B,M_{dR},\mathrm{comp}_{B,dR}):\mathcal{MT}(\mathbb{Z})\to H$, and hence a homomorphism

$$\mathcal{P}^{\mathfrak{m}}_{\mathcal{MT}(\mathbb{Z})}\longrightarrow\mathcal{P}^{\mathfrak{m}}_H\ . \tag{1.8}$$

This map is known to be injective, but we do not need this fact. The main theorem 1.5 is equivalent to saying that the images $\zeta^H(n_1,\ldots,n_r)\in\mathcal{P}^{\mathfrak{m}}_H$ of (1.7) for $n_i\in\{2,3\}$ are linearly independent. In this way, we could have dispensed with motives altogether and worked with objects in $\mathcal{P}^{\mathfrak{m}}_H$, which are elementary.[2] This leads to the following philosophy for a theory of transcendence of motivic periods in H (or another suitable category of mixed Hodge structures). It differs from standard approaches which emphasise finding relations between periods [28].

- Write down arithmetically interesting elements in, say $\mathcal{P}^{\mathfrak{m}}_H$, which come from geometry (i.e., which are periods in the sense of [28]).

[2] In fact, we should never need to compute relations explicitly using 'standard operations' such as those described in [28]; these are taken care of automatically by the Tannakian formalism, and the bound on the Ext groups of $\mathcal{MT}(\mathbb{Z})$ coming from Borel's theorems on algebraic K-theory.

- Compute the coaction (1.4) on these motivic periods, and use it to prove algebraic independence theorems.

Indeed, there is no reason to restrict oneself to mixed Tate objects, as the category H does not rely on any conjectural properties of mixed motives. The role of $\mathbb{P}^1\backslash\{0,1,\infty\}$ was to give an integral representation for the numbers (1.2) and provide a formula for the coaction.

1.4.1. Multiple modular values. Therefore, in the final part of this talk I want to propose changing the underlying geometry altogether, and replace a punctured projective line with (an orbifold) $M = \Gamma\backslash\backslash\mathbb{H}$, where $\mathbb{H}$ is the upper half plane, and $\Gamma \leq \mathrm{SL}_2(\mathbb{Z})$ is a subgroup of finite index. Because of Belyi's theorem 1.1, every smooth connected projective algebraic curve over a number field is isomorphic to an $\overline{\Gamma\backslash\mathbb{H}}$. Therefore the (pure) motivic periods obtained in this way are extremely rich[3]. It is reasonable to hope that the action of the Tannaka group on the *mixed* motivic periods of M should be correspondingly rich and should generate a large class of new periods suitable for applications in arithmetic and theoretical physics. Many of these periods can be obtained as regularised iterated integrals on $M = \Gamma\backslash\mathbb{H}$ (building on those considered by Manin in [32, 33]), and the philosophy of §1.4 concerning their Galois action can be carried out by computing a suitable automorphism group of non-abelian group cocyles. There still remains a considerable amount of work to put this general programme in its proper motivic context and extract all the arithmetic consequences.

1.5. Contents. In §2, I review the motivic fundamental group of X from its Betti and de Rham view points, define motivic multiple zeta values, and derive their Galois action from first principles. The only novelty is a direct derivation of the infinitesimal coaction from Ihara's formula. In §3, I state some consequences of theorem 1.5. In §4 I explain some results of Deligne concerning the motivic fundamental group of the projective line minus N^{th} roots of unity, and in §5 discuss the depth filtration on motivic multiple zeta values and its conjectural connection with modular forms. In §6 I mention some new results on multiple modular values for $\mathrm{SL}_2(\mathbb{Z})$, which forms a bridge between multiple zeta values and modular forms.

For reasons of space, it was unfortunately not possible to review the large recent body of work relating to associators, double shuffle equations ([1] §25, [16], [34]) and applications to knot theory, the Kashiwara-Vergne problem, and related topics such as deformation quantization; let alone the vast range of applications of multiple zeta values to high-energy physics and string theory. Furthermore, there has been recent progress in p-adic aspects of multiple zeta values, notably by H. Furusho and G. Yamashita, and work of M. Kim on integral points and the unit equation, which is also beyond the scope of these notes.

Many technical aspects of mixed Tate motives have also been omitted. See [14], §1-2 for the definitive reference.

2. The motivic fundamental group of $\mathbb{P}^1\backslash\{0,1,\infty\}$

Let $X = \mathbb{P}^1\backslash\{0,1,\infty\}$, and for now let $x, y \in X(\mathbb{C})$. The motivic fundamental groupoid of X (or rather, its Hodge realisation) consists of the following data:

[3]Grothendieck refers to $\mathrm{SL}_2(\mathbb{Z})$ as '*une machine à motifs*'

1. (Betti). A collection of schemes $\pi_1^B(X,x,y)$ which are defined over $\mathbb{Q}$, and which are equipped with the structure of a groupoid:

$$\pi_1^B(X,x,y)\times\pi_1^B(X,y,z)\longrightarrow\pi_1^B(X,x,z)$$

for any $x,y,z\in X(\mathbb{C})$. There is a natural homomorphism

$$\pi_1^{top}(X,x,y)\longrightarrow\pi_1^B(X,x,y)(\mathbb{Q}) \tag{2.1}$$

where the fundamental groupoid on the left is given by homotopy classes of paths relative to their endpoints. The previous map is Zariski dense.

2. (de Rham). An affine group scheme[4] over $\mathbb{Q}$ denoted by $\pi_1^{dR}(X)$.
3. (Comparison). A canonical isomorphism of schemes over $\mathbb{C}$

$$\mathrm{comp}_{B,dR}:\pi_1^B(X,x,y)\times_{\mathbb{Q}}\mathbb{C}\xrightarrow{\sim}\pi_1^{dR}(X)\times_{\mathbb{Q}}\mathbb{C}\,. \tag{2.2}$$

These structures are described below. Deligne has explained ([10], §15) how to replace ordinary base points with tangential base points in various settings. Denote such a tangent vector by

$$\overrightarrow{v}_x=\text{the tangent vector } v\in T_x(\mathbb{P}^1(\mathbb{C}))\text{ at the point } x\,.$$

Identifying $T_x(\mathbb{P}^1(\mathbb{C}))$ with $\mathbb{C}$, one obtains natural tangent vectors $\overrightarrow{1}_0$ and $-\overrightarrow{1}_1$ at the points 0 and 1 respectively, and a canonical path, or 'droit chemin'

$$\mathrm{dch}\in\pi_1^{top}(X,\overrightarrow{1}_0,-\overrightarrow{1}_1)$$

given by the straight line which travels from 0 to 1 in $\mathbb{R}$ with unit speed.

The reason for taking the above tangential base points is to ensure that the corresponding motive (theorem 2.1) has good reduction modulo all primes p: in the setting of $\mathbb{P}^1\backslash\{0,1,\infty\}$ there are no ordinary base points with this property.

The following theorem states that the structures $1-3$ are motivic.

Theorem 2.1. *There is an ind-object (direct limit of objects)*

$$\mathcal{O}(\pi_1^{mot}(X,\overrightarrow{1}_0,-\overrightarrow{1}_1))\in\mathrm{Ind}\,(\mathcal{MT}(\mathbb{Z})) \tag{2.3}$$

whose Betti and de Rham realisations are the affine rings $\mathcal{O}(\pi_1^B(X,\overrightarrow{1}_0,-\overrightarrow{1}_1))$*, and* $\mathcal{O}(\pi_1^{dR}(X))$*, respectively.*

Proof. (Sketch) The essential idea is due to Beilinson ([18], theorem 4.1) and Wojtkowiak [39]. Suppose, for simplicity, that M is a connected manifold and $x,y\in M$ are distinct points. Consider the submanifolds in $M\times\ldots\times M$ (n factors):

$$N_i=M^{i-1}\times\Delta\times M^{n-i-1}\qquad\text{for } i=1,\ldots,n-1$$

[4] It shall also be written $\pi_1^{dR}(X,x,y)$ but does not depend on the choice of base points. The fact that there is a canonical isomorphism $\pi_1^{dR}(X,x,y)=\pi_1^{dR}(X)$ is equivalent to saying that there is a 'canonical de Rham path' between the points x and y.

where Δ is the diagonal $M \subset M \times M$. Set $N_0 = \{x\} \times M^{n-1}$ and $N_n = M^{n-1} \times \{y\}$, and let $N \subset M^n$ be the union of the N_i, for $i = 0, \ldots, n$. Then

$$H_k(M^n, N) = \begin{cases} \mathbb{Q}[\pi_1^{top}(M, x, y)]/I^{n+1} & \text{if } k = n \\ 0 & \text{if } k < n \end{cases} \tag{2.4}$$

where the first line is the n^{th} unipotent truncation of the fundamental torsor of paths from x to y (I is the image of the augmentation ideal in $\mathbb{Q}[\pi_1^{top}(M, x)]$; see below). In the case when $M = \mathbb{P}^1 \backslash \{0, 1, \infty\}$, the left-hand side of (2.4) defines a mixed Tate motive. The case when $x = y$, or when x or y are tangential base points, is more delicate [14], §3. □

The Betti and de Rham realisations can be described concretely as follows.

1. (Betti). The Betti fundamental groupoid is defined to be the pro-unipotent completion of the ordinary topological fundamental groupoid. For simplicity, take $x = y \in X(\mathbb{C})$. Then there is an exact sequence

   $$0 \longrightarrow I \longrightarrow \mathbb{Q}[\pi_1^{top}(X(\mathbb{C}), x)] \longrightarrow \mathbb{Q} \longrightarrow 0$$

   where the third map sends the homotopy class of any path γ to 1 (thus I is the augmentation ideal). Then one has (Malčev, Quillen)

   $$\mathcal{O}(\pi_1^B(X, x)) = \lim_{N \to \infty} \left(\mathbb{Q}[\pi_1^{top}(X, x)]/I^{N+1} \right)^{\vee}$$

   The case when $x \neq y$ is defined in a similar way, since $\mathbb{Q}[\pi_1^{top}(X(\mathbb{C}), x, y)]$ is a rank one module over $\mathbb{Q}[\pi_1^{top}(X(\mathbb{C}), x)]$.
2. (de Rham). When $X = \mathbb{P}^1 \backslash \{0, 1, \infty\}$, one verifies that

   $$\mathcal{O}(\pi_1^{dR}(X)) \cong \bigoplus_{n \geq 0} H^1_{dR}(X)^{\otimes n}$$

   which is isomorphic to the tensor coalgebra on the two-dimensional graded $\mathbb{Q}$-vector space $H^1_{dR}(X) \cong \mathbb{Q}(-1) \oplus \mathbb{Q}(-1)$. We can take as basis the elements

   $$[\omega_{i_1} | \ldots | \omega_{i_n}] \quad \text{where } \omega_{i_k} \in \{\tfrac{dt}{t}, \tfrac{dt}{1-t}\}$$

   where the bar notation denotes a tensor product $\omega_{i_1} \otimes \ldots \otimes \omega_{i_n}$. It is a Hopf algebra for the shuffle product and deconcatenation coproduct and is graded in degrees ≥ 0 by the degree which assigns $\frac{dt}{t}$ and $\frac{dt}{1-t}$ degree 1.

Denoting $\overrightarrow{1}_0$ and $-\overrightarrow{1}_1$ by 0 and 1 respectively, one can write, for $x, y \in \{0, 1\}$

$${}_x\Pi_y^{\bullet} = \mathrm{Spec}(\mathcal{O}(\pi_1^{\bullet}(X, x, y)) \qquad \text{where} \quad \bullet \in \{B, dR, \mathrm{mot}\} .$$

It is convenient to write ${}_x\Pi_y$ instead of ${}_x\Pi_y^{dR}$. It does not depend on x or y, but admits an action of the motivic Galois group G^{dR} which is sensitive to x and y. If R is any commutative unitary $\mathbb{Q}$-algebra,

$${}_x\Pi_y(R) \cong \{S \in R\langle\langle x_0, x_1 \rangle\rangle^{\times} : \Delta S = S \otimes S\}$$

is isomorphic to the group of invertible formal power series in two non-commuting variables x_0, x_1, which are group-like for the completed coproduct Δ defined by $\Delta(x_i) = x_i \otimes 1 + 1 \otimes x_i$. The group law is given by concatenation of series.

2.1. Periods. The periods of the motivic fundamental groupoid of $\mathbb{P}^1\backslash\{0,1,\infty\}$ are the coefficients of the comparison isomorphism $\mathrm{comp}_{B,dR}$ (2.2) with respect to the $\mathbb{Q}$-structures on the Betti and de Rham sides. Let

$$ {}_0 1_1^B \quad \in \quad \pi_1^B(X, \vec{1}_0, -\vec{1}_1)(\mathbb{Q}) \quad \subset \quad \mathcal{O}(\pi_1^B(X, \vec{1}_0, -\vec{1}_1))^\vee $$

denote the image of dch under the natural map (2.1). It should be viewed as a linear form on the affine ring of the Betti π_1. For all $\omega_{i_k} \in \{\frac{dt}{t}, \frac{dt}{1-t}\}$,

$$ \langle \mathrm{comp}_{B,dR}([\omega_{i_1}|\dots|\omega_{i_n}]), {}_0 1_1^B \rangle = \int_{\mathrm{dch}} \omega_{i_1} \dots \omega_{i_n} \tag{2.5} $$

The right-hand side is the iterated integral from 0 to 1, *regularised* with respect to the tangent vectors 1 and -1 respectively, of the one-forms ω_{i_k}. No regularisation is necessary in the case when $\omega_{i_1} = \frac{dt}{1-t}$ and $\omega_{i_n} = \frac{dt}{t}$, and in this case the right-hand side reduces to the formula (1.2). In general, one can easily show:

Lemma 2.2. *The integrals* (2.5) *are $\mathbb{Z}$-linear combinations of MZV's of weight n.*

The *Drinfeld associator* is the de Rham image of dch

$$ \mathcal{Z} = \mathrm{comp}_{B,dR}({}_0 1_1^B) \in {}_0\Pi_1(\mathbb{C}) $$

Explicitly, it is the non-commutative generating series of the integrals (2.5)

$$ \mathcal{Z} = \sum_{i_k \in \{0,1\}} x_{i_1} \dots x_{i_n} \int_{\mathrm{dch}} \omega_{i_1} \dots \omega_{i_n} \tag{2.6} $$

$$ = 1 + \zeta(2)[x_1, x_0] + \zeta(3)([x_0, [x_0, x_1]] + [x_1, [x_1, x_0]]) + \cdots \tag{2.7} $$

It is an exponential of a Lie series.

2.2. Motivic multiple zeta values. By the previous paragraph, the affine ring of the de Rham fundamental group is the graded Hopf algebra

$$ \mathcal{O}({}_x\Pi_y) \cong \mathbb{Q}\langle e_0, e_1 \rangle $$

independently of $x, y \in \{0,1\}$. Its product is the shuffle product, and its coproduct is deconcatenation. Its basis elements can be indexed by words in $\{0,1\}$. By a general fact about shuffle algebras, the antipode is the map $w \mapsto w^*$ where

$$ (a_1 \dots a_n)^* = (-1)^n a_n \dots a_1 $$

is signed reversal of words. Thus any word w in $\{0,1\}$ defines a de Rham element in $\mathcal{O}({}_x\Pi_y)$. The augmentation map $\mathbb{Q}\langle e_0, e_1\rangle \to \mathbb{Q}$ corresponds to the unit element in the de Rham fundamental group and defines a linear form ${}_x 1_y^{dR} \in \mathcal{O}({}_x\Pi_y)^\vee$.

Define Betti linear forms ${}_x 1_y^B \in \mathcal{O}({}_x\Pi_y^B)^\vee$ to be the images of the paths

$$ \mathrm{dch} \text{ if } x = 0, y = 1 \quad ; \quad \mathrm{dch}^{-1} \text{ if } y = 1, x = 0 \quad ; \quad c_x \text{ if } x = y \,, $$

where dch is the straight path from 0 to 1, dch^{-1} is the reversed path from 1 to 0, and c_x is the constant (trivial) path based at x.

Out of this data we can construct the following motivic periods.

Definition 2.3. Let $x, y \in \{0,1\}$ and let w be any word in $\{0,1\}$. Let

$$I^{\mathfrak{m}}(x;w;y) = [\mathcal{O}({}_x\Pi_y^{\mathrm{mot}}), w, {}_x1_y^B]^{\mathfrak{m}} \qquad \in \quad \mathcal{P}^{\mathfrak{m},+}_{\mathcal{MT}(\mathbb{Z}),\mathbb{R}} \tag{2.8}$$

We call the elements $I^{\mathfrak{m}}$ motivic iterated integrals. The 'de Rham' motivic period is the matrix coefficient $[\mathcal{O}({}_x\Pi_y^{\mathrm{mot}}), w, {}_x1_y^{dR}]$ on $\mathcal{MT}(\mathbb{Z})$ with respect to the fiber functors ω_{dR}, ω_{dR}. It defines a function on G^{dR}. Its restriction to the prounipotent group U^{dR} defines an element $I^{\mathfrak{u}}(x;w;y) \in \mathcal{O}(U^{dR})$. The latter are equivalent to objects defined by Goncharov (which he also called motivic iterated integrals).

Definition 2.4. Define motivic (resp. unipotent) multiple zeta values by

$$\zeta^{\bullet}(n_1,\ldots,n_r) = I^{\bullet}(0;10^{n_1-1}\cdots 10^{n_r-1};1)\ , \quad \bullet = \mathfrak{m},\mathfrak{u}$$

It is important to note that $\zeta^{\mathfrak{m}}(2)$ is non-zero, whereas $\zeta^{\mathfrak{u}}(2) = 0$.[5] We immediately deduce from the definitions that

$$\begin{aligned} &\text{(i)} \quad I^{\mathfrak{m}}(x;w;x) = \delta_{w,\emptyset} \qquad \text{for } x \in \{0,1\} \\ &\text{(ii)} \quad I^{\mathfrak{m}}(x;w;y) = I^{\mathfrak{m}}(y;w^*;x) \end{aligned} \tag{2.9}$$

The first property holds because the constant path is trivial, the second follows from the antipode formula and because $\mathrm{dch}\circ\mathrm{dch}^{-1}$, or $\mathrm{dch}^{-1}\circ\mathrm{dch}$, is homotopic to a constant path. Finally, replacing multiple zeta values with their motivic versions, we can define a motivic version of the Drinfeld associator

$$\mathcal{Z}^{\mathfrak{m}} = \sum_{i_1,\ldots,i_n\in\{0,1\}} x_{i_1}\ldots x_{i_n} I^{\mathfrak{m}}(0;i_1,\ldots,i_n;1)\ . \tag{2.10}$$

It satisfies the associator equations defined by Drinfeld [15], on replacing $2\pi i$ by $\mathbb{L}^{\mathfrak{m}}$ (using the fact that $\zeta^{\mathfrak{m}}(2) = \frac{-(\mathbb{L}^{\mathfrak{m}})^2}{24}$), and the double shuffle equations of [34].

2.3. Action of the motivic Galois group. The category $\mathcal{MT}(\mathbb{Z})$ is a Tannakian category with respect to the de Rham fiber functor. Therefore the motivic Galois group acts on the affine ring $\mathcal{O}({}_0\Pi_1)$ of the de Rham realisation of the motivic fundamental torsor of path (2.3). A slight generalisation of theorem 2.1 shows that G^{dR} acts on the de Rham fundamental schemes

$${}_x\Pi_y \qquad \text{for all } x,y \in \{0,1\}$$

and furthermore, is compatible with the following structures:

- (Groupoid structure). The multiplication maps

$${}_x\Pi_y \times {}_y\Pi_z \longrightarrow {}_x\Pi_z$$

  for all $x,y,z \in \{0,1\}$.
- (Inertia). The action of U^{dR} fixes the elements

$$\exp(x_0) \text{ in } {}_0\Pi_0(\mathbb{Q}) \qquad \text{and} \qquad \exp(x_1) \text{ in } {}_1\Pi_1(\mathbb{Q})$$

[5] One can define a homomorphism $\mathcal{P}^{\mathfrak{m},+}_{\mathcal{MT}(\mathbb{Z}),\mathbb{R}} \to \mathcal{P}^{\mathfrak{u}}_{\mathcal{MT}(\mathbb{Z})}$ which sends $\zeta^{\mathfrak{m}}(n_1,\ldots,n_r)$ to $\zeta^{\mathfrak{u}}(n_1,\ldots,n_r)$ and prove that its kernel is the ideal generated by $\zeta^{\mathfrak{m}}(2)$.

The groupoid structure is depicted in figure 2.1.

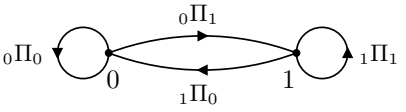


Figure 2.1. The groupoid ${}_x\Pi_y$ for $x, y \in \{0,1\}$. The diagram only represents the groupoid structure; the paths shown do not accurately depict the tangential base points.

The local monodromy map $\pi_1^{top}(\mathbb{G}_m, \vec{1}_0) \to \pi_1^{top}(X, \vec{1}_0)$ (where we write $\mathbb{G}_m$ for $\mathbb{P}^1\backslash\{0,\infty\}$), corresponding to monodromy around 0, has a motivic analogue which gives rise to the inertial condition. Its de Rham realisation is the map

$$\pi_1^{dR}(\mathbb{G}_m, \vec{1}_0) \to \pi_1^{dR}(X, \vec{1}_0) = {}_0\Pi_0$$

and is respected by G^{dR}. One shows that U^{dR} acts trivially on $\pi_1^{dR}(\mathbb{G}_m, \vec{1}_0)$, and furthermore that the element $\exp(x_0) \in {}_0\Pi_0(\mathbb{Q})$ is in the image of the previous map. This gives the first inertial condition.

Remark 2.5. It is astonishing that one obtains much useful information at all from such symmetry considerations. Nonetheless, it is enough to show the faithfulness of the action of G^{dR} (below). There are further structures respected by G^{dR}, such as compatibilities with automorphisms of $\mathbb{P}^1\backslash\{0,1,\infty\}$. They are not required.

2.4. Ihara action. Let $\mathcal{A}$ denote the group of automorphisms of the groupoid ${}_x\Pi_y$ for $x, y \in \{0,1\}$ which respects the structures $1, 2$ described in §2.3.

Proposition 2.6. *The scheme ${}_0\Pi_1$ is an $\mathcal{A}$-torsor. In particular, the action of $\mathcal{A}$ on $1 \in {}_0\Pi_1$ defines an isomorphism of schemes*

$$a \mapsto a(1) : \mathcal{A} \longrightarrow {}_0\Pi_1 \ . \tag{2.11}$$

The action of $\mathcal{A}$ on ${}_0\Pi_1$ defines, via this isomorphism, a new group law

$$\circ : {}_0\Pi_1 \times {}_0\Pi_1 \to {}_0\Pi_1 \ .$$

It is given explicitly on formal power series by Ihara's formula

$$A(x_0, x_1) \circ G(x_0, x_1) = G(x_0, Ax_1A^{-1})A \tag{2.12}$$

Proof. For the basic geometric idea, see [27], §2.3. Let $\mathfrak{a} \in \mathcal{A}$, and write $a_{xy}(\xi)$ for the action of $\mathfrak{a}$ on $\xi \in {}_x\Pi_y$. Write $a = a_{01}(1)$. Since ${}_0\Pi_0$ is a group, $\mathfrak{a}$ acts trivially on its identity element, and so $a_{00}(1) = 1$. Via the map ${}_0\Pi_1 \times {}_1\Pi_0 \to {}_0\Pi_0$ we have $a_{01}(1)a_{10}(1) = a_{00}(1)$ and hence $a_{10}(1) = a^{-1}$. The inertial conditions give

$$a_{00}(\exp(x_0)) = \exp(x_0) \quad \text{and} \quad a_{11}(\exp(x_1)) = \exp(x_1) \tag{2.13}$$

Now the composition of paths ${}_1\Pi_0 \times {}_0\Pi_0 \times {}_0\Pi_1 \to {}_1\Pi_1$ gives rise to an equation $1.\exp(x_1).1 = \exp(x_1)$. Applying $\mathfrak{a}$ to this gives by the second equation in (2.13)

$$a_{00}(\exp(x_1)) = a\exp(x_1)a^{-1} = \exp(ax_1a^{-1}) \tag{2.14}$$

which completely determines the action of $\mathcal{A}$ on ${}_0\Pi_0$. Via the map ${}_0\Pi_0 \times {}_0\Pi_1 \to {}_0\Pi_1$ we have the equation $g.1 = g$, and hence

$$a_{01}(g) = a_{00}(g).a\ . \tag{2.15}$$

Formula (2.12) follows from (2.13), (2.14), (2.15). One easily checks that a uniquely determines $\mathfrak{a}$, and so (2.11) is an isomorphism (see also [14], 5.9.) □

The groupoid and inertia structures are preserved by U^{dR}, giving a morphism

$$\rho : U^{dR} \longrightarrow \mathcal{A} \overset{(2.11)}{\cong} {}_0\Pi_1 \tag{2.16}$$

such that the following diagram commutes

$$\begin{array}{ccc} U^{dR} \times {}_0\Pi_1 & \longrightarrow & {}_0\Pi_1 \\ {\scriptstyle \rho\times \mathrm{id}}\downarrow & & \downarrow{\scriptstyle \mathrm{id}} \\ {}_0\Pi_1 \times {}_0\Pi_1 & \overset{\circ}{\longrightarrow} & {}_0\Pi_1 \end{array} \tag{2.17}$$

In principle this describes the action of the motivic Galois group on ${}_0\Pi_1$. Note, however, that the map (2.16) is mysterious and very little is known about it.

2.5. Dual formula. The coaction on motivic iterated integrals is dual to Ihara's formula. Dualising (2.17), we have

$$\Delta : \mathcal{O}({}_0\Pi_1) \longrightarrow \mathcal{O}(U^{dR}) \otimes \mathcal{O}({}_0\Pi_1)$$

It is equivalent, but more convenient, to consider the infinitesimal coaction

$$D\ :\ \mathcal{O}({}_0\Pi_1) \longrightarrow \mathcal{L} \otimes \mathcal{O}({}_0\Pi_1) \qquad \big(D(x) = \Delta(x) - 1 \otimes x \mod \mathcal{O}(U^{dR})^2_{>0}\big)$$

where $\mathcal{L} = \mathcal{O}(U^{dR})_{>0}/\big(\mathcal{O}(U^{dR})_{>0}\big)^2$ is the Lie coalgebra of indecomposables in $\mathcal{O}(U^{dR})$. The following formula is an infinitesimal variant of a formula due to Goncharov [19], relating to slightly different objects. In order to fill a gap in the literature, I will sketch how it follows almost immediately from Ihara's formula.

Proposition 2.7. *Let $a_0, \ldots, a_{n+1} \in \{0, 1\}$. The coaction D is given by*

$$\begin{aligned} D(I^{\mathfrak{m}}(a_0; a_1, \ldots, a_n; a_{n+1})) = \sum_{0\leq p<q\leq n} & \big[I^{\mathfrak{u}}(a_p; a_{p+1}, \ldots, a_q; a_{q+1})\big] \\ & \otimes I^{\mathfrak{m}}(a_0; a_1, \ldots, a_p, a_{q+1}, \ldots, a_n; a_{n+1})\ . \end{aligned} \tag{2.18}$$

where the square brackets on the left denote the map $[\ ] : \mathcal{O}(U^{dR})_{>0} \to \mathcal{L}$.

Proof. Denote the action of $\mathrm{Lie}\,\mathcal{A}$ on $\mathrm{Lie}\,{}_0\Pi_0$ by $\circ_0$. By (2.11), $\mathrm{Lie}\,\mathcal{A} \cong \mathrm{Lie}\,{}_0\Pi_1$ is the set of primitive elements in its (completed) universal enveloping algebra which we denote simply by $\mathcal{U}({}_0\Pi_1)$. By (2.13) and (2.14) we have $a \circ_0 x_0 = 0$ and $a \circ_0 x_1 = ax_1 - x_1a$. The antipode on $\mathcal{U}({}_0\Pi_1)$ is given by the signed reversal $*$. Since $a \in \mathcal{U}({}_0\Pi_1)$ is primitive, $a = -a^*$ and also

$$a \circ_0 x_0 = 0 \qquad \text{and} \qquad a \circ_0 x_1 = ax_1 + x_1a^*\ .$$

This extends to an action on $\mathcal{U}({}_0\Pi_0)$ via $a \circ_0 w_1w_2 = (a \circ_0 w_1)w_2 + w_1(a \circ_0 w_2)$. Now consider the action $a \circ_0 \cdot$ on the following words. All terms are omitted except those terms where a or a^* is inserted in-between the two bold letters:

$$\begin{aligned}
a \circ_0 w_1\mathbf{x_0x_0}w_2 &= \cdots + 0 + \cdots \\
a \circ_0 w_1\mathbf{x_0x_1}w_2 &= \cdots + w_1\mathbf{x_0ax_1}w_2 + \cdots \\
a \circ_0 w_1\mathbf{x_1x_0}w_2 &= \cdots + w_1\mathbf{x_1a^*x_0}w_2 + \cdots \\
a \circ_0 w_1\mathbf{x_1x_1}w_2 &= \cdots + \underbrace{w_1\mathbf{x_1ax_1}w_2 + w_1\mathbf{x_1a^*x_1}w_2}_{0} + \cdots
\end{aligned}$$

These four equations are dual to all but the first and last terms in (2.18), using the fact that $I^{\mathrm{u}}(x;w;x) = 0$ for $x = 0, 1$ (first and fourth lines), and the fact that $I^{\mathrm{u}}(1;w^*;0) = I^{\mathrm{u}}(0;w;1)$ (third line). A straightforward modification of the above argument taking into account the initial and final terms (using (2.15)) shows that the action $\circ_1$ of $\mathrm{Lie}\,\mathcal{A}$ on ${}_0\Pi_1$ is dual to the full expression (2.18). □

Armed with this formula, we immediately deduce that for all $n \geq 2$,

$$D\,\zeta^{\mathfrak{m}}(n) = [\zeta^{\mathfrak{u}}(n)] \otimes 1 \tag{2.19}$$

where we recall that $\zeta^{\mathfrak{u}}(2n) = 0$. One easily shows that $\zeta^u(2n+1) \neq 0$ for $n \geq 1$. See also [22]. Denote the map $w \mapsto [I^{\mathrm{u}}(0;w;1)] : \mathcal{O}({}_0\Pi_1)_{>0} \to \mathcal{L}$ simply by $\xi \mapsto [\xi^{\mathrm{u}}]$. From the structure §1.1, 1 of G^{dR} we have the following converse to (2.19) ([5], §3.2).

Theorem 2.8. *An element $\xi \in \mathcal{O}({}_0\Pi_1)$ of weight $n \geq 2$ satisfies $D\xi = [\xi^{\mathrm{u}}] \otimes 1$ if and only if $\xi \in \mathbb{Q}\,\zeta^{\mathfrak{m}}(n)$.*

This theorem, combined with (2.18), provides a powerful method for proving identities between motivic multiple zeta values. Applications are given in [6].

3. The main theorem and consequences

Theorem 1.5 is a result about linear independence. There is an analogous statement for algebraic independence of motivic multiple zeta values.

Definition 3.1. Let X be an alphabet (a set) and let $X^\times$ denote the free associative monoid generated by X. Suppose that X has a total ordering $<$, and extend it to $X^\times$ lexicographically. An element $w \in X^\times$ is said to be a Lyndon word if

$$w < u \quad \text{whenever} \quad w = uv \quad \text{and} \quad u, v \neq \emptyset\,.$$

For an ordered set X, let $\mathrm{Lyn(X)}$ denote the set of Lyndon words in X.

Theorem 3.2. *Let $X_{3,2} = \{2, 3\}$ with the ordering $3 < 2$. The set of elements*

$$\zeta^{\mathfrak{m}}(w) \quad \textit{where } w \in \mathrm{Lyn}(X_{3,2}^\times) \tag{3.1}$$

are algebraically independent over $\mathbb{Q}$, and generate the algebra $\mathcal{H}$ of motivic multiple zeta values.

Theorem 3.2 implies that every motivic multiple zeta value is equal to a unique polynomial with rational coefficients in the elements (3.1). It is often convenient to modify this generating family by replacing $\zeta^{\mathfrak{m}}(3,2,\ldots,2)$ (a three followed by $n-1$ two's) with $\zeta^{\mathfrak{m}}(2n+1)$ (by theorem 3.6). Taking the period yields the

Corollary. *Every multiple zeta value is a polynomial, with coefficients in $\mathbb{Q}$, in*

$$\zeta(w) \quad \text{where } w \in \mathrm{Lyn}(X_{3,2}^{\times}) \,. \tag{3.2}$$

Corollary. *The category $\mathcal{MT}(\mathbb{Z})$ is generated by $\pi_1^{mot}(\mathbb{P}^1\backslash\{0,1,\infty\}, \overrightarrow{1}_0, -\overrightarrow{1}_1)$ in the following sense. Every mixed Tate motive over $\mathbb{Z}$ is isomorphic, up to a Tate twist, to a direct sum of copies of sub-quotients of*

$$\mathcal{O}(\pi_1^{mot}(\mathbb{P}^1\backslash\{0,1,\infty\}, \overrightarrow{1}_0, -\overrightarrow{1}_1)) \,.$$

Corollary. *The periods of mixed Tate motives over $\mathbb{Z}$ are polynomials with rational coefficients of $(2\pi i)^{-1}$ and* (3.2).

More precisely [13], if $M \in \mathcal{MT}(\mathbb{Z})$ has non-negative weights (i.e. $W_{-1}M = 0$), then the periods of M are polynomials in (3.2) and $2\pi i$.

3.1. Canonical generators. Recall that the unipotent zeta values $\zeta^{\mathfrak{u}}$ are elements of $\mathcal{O}(U^{dR})$. As a consequence of theorem 3.2:

Corollary. *For every $n \geq 1$ there is a canonical element $\sigma_{2n+1} \in \mathrm{Lie}\, U^{dR}(\mathbb{Q})$ which is uniquely defined by $\langle \exp(\sigma_{2n+1}), \zeta^{\mathfrak{u}}(2m+1)\rangle = \delta_{m,n}$, and*

$$\langle \exp(\sigma_{2n+1}), \zeta^{\mathfrak{u}}(w)\rangle = 0 \qquad \text{for all } w \in \mathrm{Lyn}(X_{3,2}) \text{ such that } \deg_3 w > 1 \,.$$

The elements σ_{2n+1} can be taken as generators in §1.1 (1). It is perhaps surprising that one can define canonical elements of the motivic Galois group at all. These should perhaps be taken with a pinch of salt, since there may be other natural generators for the algebra of motivic multiple zeta values.

Corollary. *There is a unique homomorphism $\tau : \mathcal{H} \to \mathbb{Q}$ (see* (1.5)*) such that:*

$$\langle \tau, \zeta^{\mathfrak{m}}(2)\rangle = -\tfrac{1}{24}$$

and $\langle \tau, \zeta^{\mathfrak{m}}(w)\rangle = 0$ for all $w \in \mathrm{Lyn}(X_{3,2})$ such that $w \neq 2$.

Applying this map to the motivic Drinfeld associator defines a canonical (but not explicit!) rational associator:

$$\tau(\mathcal{Z}^{\mathfrak{m}}) \in {}_0\Pi_1(\mathbb{Q}) = \mathbb{Q}\langle\langle x_0, x_1\rangle\rangle$$

By acting on the canonical rational associator with elements σ_{2n+1}, one deduces that there exists a huge space of rational associators (which forms a torsor over $G^{dR}(\mathbb{Q})$). Such associators have several applications (see, for example [16]).

3.2. Transcendence conjectures.

Conjecture 3.3. *A variant of Grothendieck's period conjecture states that*

$$\mathrm{per} : \mathcal{P}^{\mathfrak{m}}_{\mathcal{MT}(\mathbb{Z})} \longrightarrow \mathbb{C}$$

is injective. In particular, its restriction to $\mathcal{H}$ is injective also.

The last statement, together with theorem 1.5, is equivalent to

Conjecture 3.4 (Hoffman). *The elements $\zeta(n_1,\ldots,n_r)$ for $n_i \in \{2,3\}$ are a basis for the $\mathbb{Q}$-vector space spanned by multiple zeta values.*

This in turn implies a conjecture due to Zagier, stating that the dimension of the $\mathbb{Q}$-vector space of multiple zeta values of weight N is equal to d_N (1.6), and furthermore that the ring of multiple zeta values is graded by the weight. Specialising further, we obtain the following folklore

Conjecture 3.5. *The numbers $\pi, \zeta(3), \zeta(5), \zeta(7), \ldots$ are algebraically independent.*

3.3. Idea of proof of theorem 1.5. The proof of linear independence is by induction on the number of 3's. In the case where there are no 3's, one can easily show by adapting an argument due to Euler that

$$\zeta(\underbrace{2,\ldots,2}_{n}) = \frac{\pi^{2n}}{(2n+1)!}\,.$$

The next interesting case is where there is one 3.

Theorem 3.6 (Zagier [38]). *Let $a, b \geq 0$. Then*

$$\zeta(\underbrace{2,\ldots 2}_{a},3,\underbrace{2,\ldots,2}_{b}) = 2\sum_{r=1}^{a+b+1}(-1)^r(A^r_{a,b}-B^r_{a,b})\,\zeta(2r+1)\,\zeta(\underbrace{2,\ldots,2}_{a+b+1-r})$$

where, for any $a,b,r \in \mathbb{N}$, $A^r_{a,b} = \binom{2r}{2a+2}$, and $B^r_{a,b} = \left(1-2^{-2r}\right)\binom{2r}{2b+1}$.

Zagier's proof of this theorem involves an ingenious mixture of analytic techniques. The next step in the proof of theorem 1.5 is to lift Zagier's theorem to the level of motivic multiple zeta values by checking its compatibility with the coaction (2.18) and using theorem 2.8. Since then, the proof of theorem 3.6 was simplified by Li [31], and Terasoma [37] has verified that it can be deduced from associator equations. Since the associator equations are known to hold between motivic multiple zeta values, it follows that, in principle, this part of the proof can now be deduced directly by elementary methods (i.e., without using theorem 2.8).

From the motivic version of theorem 3.6, one can compute the action of the abelianization of U^{dR} on the vector space built out of the elements $\zeta^{\mathfrak{m}}(n_1,\ldots,n_r)$, with $n_i = 2,3$, graded by the number of 3's. This action can be expressed by certain matrices constructed out of the combinatorial formula (2.18), whose entries are linear combinations of the coefficients $A^r_{a,b}$ and $B^r_{a,b}$ of theorem 3.6. The key point is that these matrices have non-zero determinant 2-adically, and are hence invertible. At its heart, this uses the fact that the $B^r_{a,b}$ terms in theorem 3.6 dominate with respect to the 2-adic norm due to the factor 2^{-2r}.

4. Roots of unity

There are a handful of exceptional cases when one knows how to generate certain categories of mixed Tate motives over cyclotomic fields and write down their periods. These results are due to Deligne [11], inspired by numerical computations due to Broadhurst in 1997 relating to computations of Feynman integrals.

Let $N \geq 2$ and let μ_N be the group of N^{th} roots of unity, and consider

$$\mathbb{P}^1 \backslash \{0, \mu_N, \infty\} \tag{4.1}$$

Fix a primitive N^{th} root ζ_N. One can consider the corresponding motivic fundamental groupoid (with respect to suitable tangential base points) and ask whether it generates the category $\mathcal{MT}(\mathcal{O}_N[\frac{1}{N}])$, where $\mathcal{O}_N$ is the ring of integers in the field $\mathbb{Q}(\zeta_N)$. Goncharov has shown that for many primes N, and in particular, for $N = 5$, this is false: already in weight two, there are motivic periods of this category which cannot be expressed as motivic iterated integrals on $\mathbb{P}^1 \backslash \{0, \mu_N, \infty\}$.

In certain exceptional cases, Deligne has proven a stronger statement:

Theorem 4.1. *For $N = 2, 3, 4, 6$ (resp. $N = 8$) the motivic fundamental group*

$$\pi_1^{mot}(\mathbb{P}^1 \backslash \{0, 1, \infty\}, \vec{1}_0, \zeta_N) \qquad \big(\textit{resp. } \pi_1^{mot}(\mathbb{P}^1 \backslash \{0, \pm 1, \infty\}, \vec{1}_0, \zeta_8)\big)$$

generates the categories $\mathcal{MT}(\mathcal{O}_N[\frac{1}{N}])$ for $N = 2, 3, 4, 8$, and $\mathcal{MT}(\mathcal{O}_N)$ for $N = 6$.

Iterated integrals on (4.1) can be expressed in terms of cyclotomic multiple zeta values[6] which are defined for $(n_r, \varepsilon_r) \neq (1, 1)$ by the sum

$$\zeta(n_1, \ldots, n_r; \varepsilon_1, \ldots, \varepsilon_r) = \sum_{0 < k_1 < k_2 < \ldots < k_r} \frac{\varepsilon_1^{k_1} \ldots \varepsilon_r^{k_r}}{k_1^{n_1} \ldots k_r^{n_r}}$$

where $\varepsilon_1, \ldots, \varepsilon_r$ are roots of unity. The weight is defined as the sum of the indices $n_1 + \ldots + n_r$ and the depth is the increasing filtration defined by the integer r. It is customary to use the notation

$$\zeta(n_1, \ldots, n_{r-1}, n_r \zeta_N) = \zeta(n_1, \ldots, n_r; \underbrace{1, \ldots, 1}_{r-1}, \zeta_N) \, .$$

One can define motivic versions relative to the canonical fiber functor ω ([14], §1.1) playing the role of what was previously the de Rham fiber functor (the two are related by $\omega_{dR} = \omega \otimes \mathbb{Q}(\zeta_N)$), and the Betti realisation functor which corresponds to the embedding $\mathbb{Q}(\zeta_N) \subset \mathbb{C}$. Denote these motivic periods by a superscript $\mathfrak{m}$. Recall that $\mathbb{L}^{\mathfrak{m}}$ is the motivic Lefschetz period of example 1.3, whose period is $2\pi i$. Let $X_{odd} = \{1, 3, 5, \ldots\}$ with the ordering $1 > 3 > 5 \ldots$. Rephrased in the language of motivic periods, Deligne's results for $N = 2, 3, 4$ yield:

1. ($N = 2$; algebra generators). The following set of motivic periods:

$$\{\mathbb{L}^{\mathfrak{m}}\} \cup \{\zeta^{\mathfrak{m}}(n_1, \ldots, n_{r-1}, -n_r) \text{ where } (n_r, \ldots, n_1) \in \mathrm{Lyn}(X_{odd})\}$$

[6] The conventions in [11] are opposite to the ones used here

are algebraically independent over $\mathbb{Q}$. The monomials in these quantities form a basis for the ring of geometric motivic periods[7] of $\mathcal{MT}(\mathbb{Z}[\frac{1}{2}])$.

2. ($N = 3, 4$; linear basis). The set of motivic periods

$$\zeta^{\mathfrak{m}}(n_1, \ldots, n_{r-1}, n_r\zeta_N)(\mathbb{L}^{\mathfrak{m}})^p \qquad \text{where } n_i \geq 1, p \geq 0$$

are linearly independent over $\mathbb{Q}$. They form a basis for the space of geometric motivic periods of $\mathcal{MT}(\mathcal{O}_N[\frac{1}{N}])$, for $N = 3, 4$ respectively.

By applying the period map, each case gives a statement about cyclotomic multiple zeta values. In the case $N = 2$, the underlying field is still $\mathbb{Q}$, and it follows from (i) that every multiple zeta value at 2^{nd} roots of unity (sometimes called an Euler sum) is a polynomial with rational coefficients in

$$(2\pi i)^2 \quad \text{and} \quad \zeta(n_1, \ldots, n_{r-1}, -n_r) \qquad \text{where} \quad (n_1, \ldots, n_r) \in \mathrm{Lyn}(X_{odd})\ .$$

This decomposition respects the weight and depth, where the depth of $(2\pi i)^n$ is 1. Thus an Euler sum of weight N and depth r can be expressed as a polynomial in the above elements, of total weight N and total depth $\leq r$.

5. Depth

The results of the previous section for $N = 2, 3, 4, 6, 8$ crucially use the fact that the depth filtration is dual to the lower central series of the corresponding motivic Galois group. A fundamental difference with the case $N = 1$ is that this fact is false for $\mathbb{P}^1\backslash\{0,1,\infty\}$, due to a defect closely related to modular forms.

Recall that ${}_0\Pi_1$ is a group for the Ihara action $\circ$. Let $\mathrm{Lie}({}_0\Pi_1)$ denote its Lie algebra. Its bracket is denoted by $\{\ ,\ \}$. Denote the images of the canonical generators §3.1 by $\sigma_{2n+1} \in \mathrm{Lie}({}_0\Pi_1)(\mathbb{Q})$, for $n \geq 1$. They are elements of the free graded Lie algebra on two generators x_0, x_1, and we have, for example

$$\sigma_3 = [x_0, [x_0, x_1]] + [x_1, [x_1, x_0]]$$

The higher σ_{2n+1} are of the form $\sigma_{2n+1} = ad(x_0)^{2n}(x_1)$ plus terms of degree ≥ 2 in x_1, but are not known explicitly except for small n. By theorem 1.2, the σ_{2n+1} freely generate a graded Lie subalgebra of $\mathrm{Lie}({}_0\Pi_1)(\mathbb{Q})$ which we denote by $\mathfrak{g}$. The depth filtration $\mathcal{D}$ on $\mathfrak{g}$ is the decreasing filtration given by the degree in the letter x_1. In 1993, Ihara and Takao observed that

$$\{\sigma_3, \sigma_9\} - 3\{\sigma_5, \sigma_7\} = \frac{691}{144}\, e_\Delta \tag{5.1}$$

where e_Δ is an element with integer coefficients of depth ≥ 4 (degree ≥ 4 in x_1), and the coefficient 691 on the right-hand side is the numerator of the Bernoulli number B_{12}. The element e_Δ is sparse:[8] indeed, computations in the early days gave the impression that the right-hand side is zero, although we now know that the σ_{2n+1} generate a free Lie algebra.

[7] recall that this is the subring of all motivic periods of the category $\mathcal{MT}(\mathbb{Z}[\frac{1}{2}])$ which is generated by motives M which have non-negative weights, i.e., $W_{-1}M = 0$.

[8] 'most' of its coefficients are zero, see [7], §8 for a closed formula for this element.

Relations such as (5.1) show that the structure of $\mathfrak{g}$ is related to arithmetic, but more importantly show that the associated depth-graded Lie algebra $\mathrm{gr}_{\mathcal{D}}\,\mathfrak{g}$ is not free, since the left-hand side of (5.1) vanishes in $\mathrm{gr}^2_{\mathcal{D}}\,\mathfrak{g}$. The depth filtration on $\mathfrak{g}$ corresponds, dually, to the depth filtration on motivic multiple zeta values, and (5.1) implies that motivic multiple zeta values of depth ≤ 2 are insufficient to span the space of all (real geometric) motivic periods of $\mathcal{MT}(\mathbb{Z})$ in weight 12 (one needs to include elements of depth ≥ 4 such as $\zeta^{\mathfrak{m}}(2,2,2,3,3)$ in a basis). By counting dimensions, this can be interpreted as a relation, viz:

$$28\,\zeta^{\mathfrak{m}}(3,9) + 150\,\zeta^{\mathfrak{m}}(5,7) + 168\,\zeta^{\mathfrak{m}}(7,5) = \frac{5197}{691}\zeta^{\mathfrak{m}}(12) \tag{5.2}$$

The corresponding relation for multiple zeta values was found in [17] and generalised to an infinite family corresponding to cuspidal cohomology classes of $\mathrm{SL}_2(\mathbb{Z})$. In particular, the family of motivic multiple zeta values

$$\zeta^{\mathfrak{m}}(2n_1+1,\ldots,2n_r+1)\zeta^{\mathfrak{m}}(2k)$$

cannot be a basis for $\mathcal{H}$, although it has the right dimensions in each weight (1.6). The Hoffman basis (1.7) gets around such pathologies, since, for example, its elements in weight 12 have depths between four and six.

In 1997, Broadhurst and Kreimer made exhaustive numerical computations on the depth filtration of multiple zeta values, which led them to the following conjecture, translated into the language of motivic multiple zeta values.

Conjecture 5.1 (Motivic version of the Broadhurst-Kreimer conjecture). *Let $\mathcal{D}$ denote the increasing filtration on $\mathcal{H}$ induced by the depth. Then*

$$\sum_{N,d\geq 0} \dim_{\mathbb{Q}}\left(gr^{\mathcal{D}}_d\mathcal{H}_N\right) s^d t^N = \frac{1+\mathbb{E}(t)s}{1-\mathbb{O}(t)s+\mathbb{S}(t)s^2-\mathbb{S}(t)s^4}\,, \tag{5.3}$$

where $\mathbb{E}(t) = \frac{t^2}{1-t^2}$, $\mathbb{O}(t) = \frac{t^3}{1-t^2}$, *and* $\mathbb{S}(t) = \frac{t^{12}}{(1-t^4)(1-t^6)}$.

Note that equation (5.3) specializes to (1.6) on setting s equal to 1. The series $\mathbb{E}(t)$ and $\mathbb{O}(t)$ are the generating series for the dimensions of the spaces of even and odd single motivic zeta values. The interpretation of $\mathbb{S}(t)$ as the generating series for cusp forms for $\mathrm{SL}_2(\mathbb{Z})$ suggests a deeper connection with modular forms which is well understood in depth two. By work of Zagier, and Goncharov, formula (5.3) has been confirmed in depths 2 and 3 (i.e., modulo s^4).

An interpretation for conjecture (5.3) in terms of the structure of $\mathrm{gr}_{\mathcal{D}}\mathfrak{g}$, as well as a complete conjectural description of generators and relations of $\mathrm{gr}_{\mathcal{D}}\mathfrak{g}$ in terms of modular forms for $\mathrm{SL}_2(\mathbb{Z})$ was given in [7]. A deeper geometric understanding of this conjecture would seem to require a framework which places multiple zeta values and modular forms on an equal footing, which is the topic of §6.

6. Multiple modular values

In this final paragraph, I want suggest applying the philosophy of §1.4 to iterated integrals on (orbifold) quotients of the upper half plane

$$\mathbb{H} = \{\tau \in \mathbb{C} : \mathrm{Im}\,(\tau) > 0\}$$

by finite index subgroups $\Gamma \leq \mathrm{SL}_2(\mathbb{Z})$. Iterated integrals of modular forms were first studied by Manin [32, 33]. Here, I shall only consider the case $\Gamma = \mathrm{SL}_2(\mathbb{Z})$.

6.1. Eichler-Shimura integrals. Denote the space of homogenous polynomials of degree $n \geq 0$ with rational coefficients by

$$V_n = \bigoplus_{i+j=n} \mathbb{Q} X^i Y^j$$

It admits a right action of Γ via the formula $(X,Y)|_\gamma = (aX + bY, cX + dY)$, where $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Let $f(\tau)$ be a modular form of weight k for Γ. Define

$$\underline{f}(\tau) = (2\pi i)^{k-1} f(\tau)(X - \tau Y)^{k-2} d\tau \qquad \in \qquad \Gamma(\mathbb{H}, \Omega^1_{\mathbb{H}} \otimes V_{k-2})$$

It satisfies the invariance property $\underline{f}(\gamma(\tau))\big|_\gamma = \underline{f}(\tau)$ for all $\gamma \in \Gamma$. For f a cusp form, the classical Eichler-Shimura integral (see, e.g., [29]) is

$$\int_0^\infty \underline{f}(\tau) = \sum_{n=1}^{k-1} c_n L(f,n) X^{k-n-1} Y^{n-1} \tag{6.1}$$

where c_n are certain explicit constants (rational multiples of a power of π) and $L(f,s)$ is the analytic continuation of the L-function $L(f,s) = \sum_{n\geq 1} \frac{a_n}{n^s}$ of f, where $f(\tau) = \sum_{n\geq 1} a_n q^n$ and $q = e^{2\pi i \tau}$. Manin showed that if f is a Hecke eigenform, there exist $\omega_f^+, \omega_f^- \in \mathbb{R}$ such that

$$\int_0^\infty \underline{f}(\tau) = \omega_f^+ P_f^+(X,Y) + \omega_f^- P_f^-(X,Y)$$

where $P_f^\pm(X,Y) \in V_{k-2} \otimes \overline{\mathbb{Q}}$ are polynomials with algebraic coefficients which are invariant (resp. anti-invariant) with respect to $(X,Y) \mapsto (-X,Y)$.

Recall that the Eisenstein series of weight $2k$, for $k \geq 2$, is defined by

$$e_{2k}(q) = -\frac{B_{2k}}{2k} + \sum_{n\geq 1} \sigma_{2k-1}(n) q^n \,, \qquad q = e^{2\pi i \tau}$$

where B_{2k} is the $2k^{\text{th}}$ Bernoulli number, and σ denotes the divisor function. The corresponding integrals for Eisenstein series diverges. Zagier showed how to extend the definition of the Eichler-Shimura integrals to the case e_{2k}, giving [29]

$$\frac{(2k-2)!}{2}\zeta(2k-1)(X^{2k-2} - Y^{2k-2}) - \frac{(2\pi i)^{2k-1}}{4k(2k-1)} \sum_{\substack{a+b=2k \\ a,b\geq 1}} \binom{2k}{a} B_a B_b X^{a-1} Y^{b-1} \tag{6.2}$$

Manipulating this formula leads to expressions for the odd Riemann zeta values in terms of Lambert series similar to the following formula due to Ramanujan:

$$\zeta(3) = \frac{7}{180}\pi^3 - 2\sum_{n\geq 1} \frac{1}{n^3(e^{2n\pi} - 1)} \,.$$

It converges very rapidy. One wants to think of (6.2) as pointing towards a modular construction of $\zeta^{\mathfrak{m}}(2k-1)$.

6.2. Regularisation. The theory of tangential base points ([10], §15) gives a general procedure for regularising iterated integrals on curves. If one applies this to the orbifold $\Gamma\backslash\backslash\mathbb{H}$, where $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, one can show that it yields the completely explicit formulae below, which generalise Zagier's formula for a single Eisenstein series. I shall only state the final answer. Via the map

$$\tau \mapsto q = \exp(2i\pi\tau) : \mathbb{H} \longrightarrow \{q \in \mathbb{C} : 0 < |q| < 1\} = D^\times$$

a natural choice of tangential base point (denoted $\vec{1}_\infty$) corresponds to the tangent vector 1 at $q = 0$. Since in this case we have explicit models $\mathbb{H} \subset \mathbb{C}$ for a universal covering space of $\Gamma\backslash\backslash\mathbb{H}$, and $\mathbb{C}$ for the universal covering of $D^\times$, one can compute all regularised iterated integrals by pulling them back to $\mathbb{C}$ as follows.

First, if $f = \sum_{n\geq 0} f_n q^n$ is the Fourier expansion of f, write

$$\underline{f}^\infty(\tau) = (2\pi i)^{k-1} f_0 (X - \tau Y)^{k-2} d\tau \qquad \in \qquad \Gamma(\mathbb{C}, \Omega^1_{\mathbb{C}} \otimes V_{k-2}) \tag{6.3}$$

Define a linear operator R on the tensor coalgebra on $\Gamma(\mathbb{C}, \Omega^1_{\mathbb{C}} \otimes V)$ by

$$\begin{aligned} R[\omega_1|\ldots|\omega_n] &= \sum_{i=0}^{n} (-1)^{n-i} [\omega_1|\ldots|\omega_i] \,\text{ш}\, [\omega_n^\infty|\ldots|\omega_{i+1}^\infty] \\ &= \sum_{i=1}^{n} (-1)^{n-i} \Big[[\omega_1|\ldots|\omega_{i-1}] \,\text{ш}\, [\omega_n^\infty|\ldots|\omega_{i+1}^\infty] \Big| \omega_i - \omega_i^\infty \Big] \,. \end{aligned}$$

where $V = \bigoplus_k V_k$ and ω^∞ is the 'residue at infinity' of ω defined by (6.3). The regularised iterated integral can be expressed as *finite* integrals

$$\int_\tau^{\vec{1}_\infty} [\underline{\omega}_1|\ldots|\underline{\omega}_n] = \sum_{i=0}^{n} \int_\tau^\infty R[\underline{\omega}_1|\ldots|\underline{\omega}_i] \int_\tau^0 [\underline{\omega}_{i+1}^\infty|\ldots|\underline{\omega}_n^\infty]$$

It takes values in $V_{k_1-2} \otimes \ldots \otimes V_{k_n-2} \otimes \mathbb{C}$ if $\omega_1, \ldots, \omega_n$ are of weights $k_1, \ldots k_n$, and hence admits a right action of Γ. The integrals in the right factor on the right-hand side are simply polynomials in τ and can be computed explicitly.

6.3. Cocycles. Choose a basis of Hecke normalised eigenforms f_i indexed by non-commuting symbols A_i, and form the generating series

$$I(\tau;\infty) = \sum_{i_k, n\geq 0} A_{i_1} \ldots A_{i_n} \int_\tau^{\vec{1}_\infty} [\underline{\omega}_{i_1}|\ldots|\underline{\omega}_{i_n}]$$

For every $\gamma \in \Gamma$, there exists a formal power series C_γ in the A_i such that

$$I(\tau;\infty) = I(\gamma(\tau);\infty)|_\gamma \, C_\gamma \tag{6.4}$$

which does not depend on τ. It satisfies the cocycle relation

$$C_{gh} = C_g\big|_h \, C_h \quad \text{ for all } g, h \in \Gamma \,.$$

The part of the cocycle C which involves iterated integrals of cusp forms was previously considered by Manin [32, 33]. Since the group Γ is generated by

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad , \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} ,$$

the cocycle C is determined by C_S and C_T. The series C_T can be computed explicitly and its coefficients lie in $\mathbb{Q}[2\pi i]$.

Definition 6.1. Define the ring of multiple modular values with respect to the group $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ to be the subring of $\mathbb{C}$ generated by the coefficients of C_S.

The series C_S is a kind of analogue of Drinfeld's associator $\mathcal{Z}$. Its terms of degree 1 in the A_i are precisely the Eichler-Shimura integrals (6.1) and (6.2). Setting $\tau = i$ in (6.4) gives integrals which converge extremely fast and are very well suited to numerical computation.

6.4. Galois action. One can mimic the Betti-de Rham aspects of the theory of the motivic fundamental group of $\mathbb{P}^1\backslash\{0,1,\infty\}$ as follows:

1. The coefficients of C_S can be interpreted as certain periods of the relative unipotent completion of Γ. This was defined by Deligne as follows. Let k be a field of characteristic 0 and S a reductive algebraic group over k. Suppose that Γ is a discrete group equipped with a Zariski dense homomorphism $\rho : \Gamma \to S(k)$. The completion of Γ relative to ρ is an affine algebraic group scheme $\mathcal{G}_\Gamma$, which sits in an exact sequence
$$1 \longrightarrow \mathcal{U}_\Gamma \longrightarrow \mathcal{G}_\Gamma \longrightarrow S \longrightarrow 1$$
where $\mathcal{U}_\Gamma$ is pro-unipotent. There is a natural map $\Gamma \to \mathcal{G}_\Gamma(k)$ which is Zariski dense, and whose projection onto $S(k)$ is the map ρ.
2. In 'geometric' situations, one expects the relative completion to be the Betti realisation of something which is motivic. Indeed, Hain has shown [23, 24] that $\mathcal{O}(\mathcal{G}_\Gamma)$ carries a mixed Hodge structure in this case. As a result, one can define Hodge-motivic periods and try to carry out §1.4.
3. The action of the unipotent radical of the Tannaka group of mixed Hodge structures acts via the automorphism group of a space of non-abelian cocyles of Γ with coefficients in $\mathcal{U}_\Gamma$. It is a certain semi-direct product of $\mathcal{U}_\Gamma$ with a group of non-commutative substitutions $\mathrm{Aut}(\mathcal{U}_\Gamma)^S$. An inertia condition corresponds, in the case $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, to the fact that C_T is fixed, and there are further constraints coming from the action of Hecke operators. The explicit expression for C_T yields precise information about the action of the Hodge-Galois group.

The following key example illustrates how multiple modular values for $\mathrm{SL}_2(\mathbb{Z})$ resolve the depth-defect for multiple zeta values as discussed in §5.

Example 6.2. On $\mathbb{P}^1\backslash\{0,1,\infty\}$ there are 2^{12} integrals of weight 12, namely

$$\int_{dch} \omega_{i_1} \ldots \omega_{i_{12}} \quad \text{where} \quad \omega_{i_j} \in \left\{ \frac{dt}{t}, \frac{dt}{1-t} \right\} .$$

However the space of multiple zeta values $\mathcal{Z}_{12}$ in weight 12 has dimension at most $d_{12} = 12$, so there are a huge number of relations. Indeed, modulo products of multiple zeta values of

lower weights, there are at most two elements of weight 12:

$$\zeta(3,3,2,2,2) \qquad \text{and} \qquad \zeta(3,2,3,2,2) \tag{6.5}$$

by the corollary to theorem 3.2. They are conjectured to be algebraically independent. Note that multiple zeta values of depths ≤ 2 (or ≤ 3 for that matter) will not suffice to span $\mathcal{Z}_{12}$ by equation (5.2).

On the other hand, we can consider the coefficients of C_S corresponding to regularised iterated integrals of Eisenstein series

$$\int_0^{\vec{1}_\infty} \underline{e}_{2a}(X,Y)\underline{e}_{2b}(X,Y) \in \mathbb{C}[X,Y] \tag{6.6}$$

If we are interested in periods modulo products, there are just two relevant cases: $(2a,2b) \in \{(4,10),(6,8)\}$. The description 3 above enables one to extract the relevant numbers from the coefficients of these polynomials. One finds experimentally that one obtains exactly the elements (6.5) modulo products, and that this is consistent with the coaction on the corresponding Hodge-motivic periods. Thus $\mathcal{Z}_{12}$ is spanned by exactly the right number of multiple modular values (which are linear combinations of the coefficients of (6.6).

The example shows that in weight 12, there are exactly two multiple modular values (modulo products) which are multiple zeta values, and they conjecturally satisfy no relations. By contrast, multiple zeta values in weight 12 are hugely over-determined, and satisfy a vast number of relations. Furthermore, the depth-defect described in §5 can be directly related to the appearance of special values of L-functions of cusp forms amongst certain coefficients of (6.6).

In conclusion, a rather optimistic hope is that a theory of motivic multiple modular values for congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$ might provide a more natural construction of the periods of mixed Tate motives over cyclotomic fields (and much more) than the motivic fundamental groupoid of the projective line minus Nth roots of unity, which suffers from the depth defect in the case $N = 1$ (§5), and from absent periods in non-exceptional cases such as $N = 5$ (§4).

Acknowledgements. The author's work is supported by ERC Grant 257638.

References

[1] Y. André, *Une introduction aux motifs*, Panoramas et Synthèses **17**, SMF (2004).

[2] A. Borel, *Stable real cohomology of arithmetic groups*, Annales Ecole Normale Sup. **7**, No. 4, (1974), 235–272.

[3] ———, *Cohomologie de SL_n et valeurs de fonctions zêta aux points entiers*, Annali della Scuola Norm. di Pisa, (1976), 613–635, + erratum.

[4] Belyi, *On Galois Extensions of a Maximal Cyclotomic Field*, Math. USSR-Izvestija 14:247–256 (1980)

[5] F. Brown, *Mixed Tate motives over* $\mathbb{Z}$, Annals of Math., volume 175, no. 1, 949–976 (2012).

[6] ———, *Decomposition of motivic multiple zeta values*, 'Galois-Teichmuller theory and Arithmetic Geometry', Adv. Stud. Pure Math. **63** (2012).

[7] ———, *Depth-graded motivic multiple zeta values*, http://arxiv.org/abs/1301.3053.

[8] P. Cartier, *Fonctions polylogarithmes, nombres polyztas et groupes pro-unipotents*, Sminaire Bourbaki, Astrisque No. **282** (2002), Exp. No. 885, 137–173.

[9] P. Deligne, *Catégories Tannakiennes*, Grothendieck Festschrift, vol. II, Birkhäuser Progress in Math. **87** (1990), 111–195.

[10] P. Deligne, *Le groupe fondamental de la droite projective moins trois points*, Galois groups over Q (Berkeley, CA, 1987), 79–297, Math. Sci. Res. Inst. Publ. **16** (1989)

[11] ———, *Le groupe fondamental unipotent motivique de* $\mathrm{G}_m - \mu_N$*, pour* $N = 2, 3, 4, 6$ *ou* 8, Publ. Math. Inst. Hautes Études Sci. **101** (2010).

[12] ———, *Multizêtas*, Séminaire Bourbaki, expos 1048, Astrisque 352 (2013).

[13] ———, *Letter to Brown and Zagier*, 28 April 2012.

[14] P. Deligne and A. B. Goncharov, *Groupes fondamentaux motiviques de Tate mixte*, Ann. Sci. École Norm. Sup. **38** (2005), 1–56.

[15] V. Drinfeld, *On quasi-triangular quasi-Hopf algebras and some group closely related with* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, Algebra i Analiz **2** (1990), no. 4, 149–181.

[16] H. Furusho, *Four groups related to associators*, arXiv:1108.3389 (2011).

[17] H. Gangl, M. Kaneko, and D. Zagier, *Double zeta values and modular forms*, Automorphic forms and zeta functions, 71–106, World Sci. Publ., Hackensack, NJ, 2006.

[18] A. Goncharov, *Multiple polylogarithms and Mixed Tate Motives*, arXiv:0103059 (2001).

[19] A. B. Goncharov, *Galois symmetries of fundamental groupoids and noncommutative geometry*, Duke Math. J. **128** (2005), 209–284.

[20] A. Goncharov and Y. Manin, *Multiple* ζ*-motives and moduli spaces* $\overline{\mathfrak{M}}_{0,n}$, Compos. Math. **140** (2004), no. 1, 1–14.

[21] A. Grothendieck, *Esquisse d'un programme*, http://www.math.jussieu.fr/ leila/grothendieckcircle/EsquisseFr.pdf.

[22] R. Hain and M. Matsumoto, *Weighted completion of Galois groups and Galois actions on the fundamental group of* $\mathbb{P}^1\backslash\{0,1,\infty\}$, Compositio Math. **139** (2003), no. 2, 119–167.

[23] R. Hain, *The Hodge de Rham theory of the relative Malcev completion*, Ann. Sci. École Norm. Sup. **31** (1998), 47–92.

[24] ———, *The Hodge-de Rham Theory of Modular Groups*, http://arxiv.org/abs/1403.6443.

[25] M. E. Hoffman, *The Algebra of Multiple Harmonic Series*, Journ. of Algebra **194** (1997), 477–495.

[26] Y. Ihara, *The Galois representation arising from* $\mathbb{P}^1 - \{0, 1, \infty\}$ *and Tate twists of even degree*, Galois groups over $\mathbb{Q}$, 299–313, Math. Sci. Res. Inst. Publ. **16** (1989).

[27] ———, *Braids, Galois Groups, and Some Arithmetic Functions*, Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990), 99–120.

[28] M. Kontsevich and D. Zagier, *Periods*, Mathematics unlimited-2001 and beyond, 771–808, Springer, Berlin, (2001).

[29] W. Kohnen and D. Zagier, *Modular forms with rational periods*, Modular forms (Durham, 1983), 197–249.

[30] M. Levine, *Tate motives and the vanishing conjectures for algebraic K-theory*, Algebraic K-theory and algebraic topology (Lake Louise, AB, 1991), 167–188.

[31] Z-H. Li, *Some identities in the harmonic algebra concerned with multiple zeta values*, Int. J. Number Theory **9** (2013), no. 3, 783–798.

[32] Y. Manin, *Iterated integrals of modular forms and non-commutative modular symbols*, Algebraic geometry and number theory, 565–597, Prog. Math. **253** (2006).

[33] ———, *Iterated Shimura integrals*, Moscow Math. J. **5** (2005), 869–881.

[34] G. Racinet, *Doubles mélanges des polylogarithmes multiples aux racines de l'unité*, Publ. Math. Inst. Hautes Études Sci. **95** (2002), 185–231.

[35] G. Shimura, *On the periods of modular forms*, Math. Annalen **229** (1977), 211–221.

[36] T. Terasoma, *Geometry of multiple zeta values*, International Congress of Mathematicians. Vol. II, (2006), 627–635.

[37] ———, *Brown-Zagier relation for associators*, arXiv:1301.7474 (2013).

[38] D. B. Zagier, *Evaluation of the multiple zeta values* $\zeta(2, \ldots, 2, 3, 2, \ldots, 2)$, Ann. of Math. (2) **175** (2012), no. 2, 977–1000.

[39] Z. Wojtkowiak, *Cosimplicial objects in algebraic geometry*, Algebraic K-theory and algebraic topology (Lake Louise, AB, 1991), 287–327, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **407** (1993).

IHES, 35 Route de Chartres, Bures-sur-Yvette, 91440 France
E-mail: brown@ihes.fr

Completed cohomology and the p-adic Langlands program

Matthew Emerton

Abstract. We discuss some known and conjectural properties of the completed cohomology of congruence quotients associated to reductive groups over $\mathbb{Q}$. We also discuss the conjectural relationships to local and global Galois representations, and a possible p-adic local Langlands correspondence.

Mathematics Subject Classification (2010). Primary 11F70; Secondary 22D12.

Keywords. p-adic Langlands program, p-adic Hodge theory, completed cohomology, Galois representations.

1. Introduction

The idea that there could be a p-adic version of the local Langlands correspondence was originally proposed by C. Breuil, in the case of the group $\mathrm{GL}_2(\mathbb{Q}_p)$, in his papers [11–13]. In [13], he also proposed a local-global compatibility between this (at the time conjectural) correspondence and p-adically completed cohomology of modular curves. Since then, the theory of p-adic Langlands has been extensively developed in the case of the group GL_2 over $\mathbb{Q}$; see e.g. [6, 29, 30, 32, 46, 50, 52]. Since excellent expositions of much of this material already exist [5, 14, 15], we have decided here to describe some aspects of the p-adic Langlands program that make sense for arbitrary groups. This has led us to focus on completed cohomology, since this is a construction that makes sense for arbitrary groups, and about which it is possible to establish some general results, and make some general conjectures.

Many of these conjectures are very much motivated by the conjectural relationship with Galois representations, and we have also tried to outline our expectations regarding this relationship, while trying not get bogged down in the myriad technical details that would necessarily attend a more careful discussion of this topic. Finally, we have tried to indicate how completed cohomology may be related, by the principle of local-global compatibility, to a still largely conjectural p-adic local Langlands correspondence for groups other than $\mathrm{GL}_2(\mathbb{Q}_p)$. Our focus is on drawing inferences about the possible structure of the local correspondence by interpolating from our expectations regarding completed cohomology. In this regard we mention also the paper [27], which somewhat literally interpolates p-adically completed cohomology (via the Taylor–Wiles–Kisin method) in order to construct a candidate for the p-adic local Langlands correspondence for GL_n of an arbitrary p-adic field.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

We close this introduction by simply listing a number of additional important papers [17–20] which also have the goal of extending the p-adic local Langlands correspondence, and local-global compatibility, beyond the case of $\mathrm{GL}_2(\mathbb{Q}_p)$. Unfortunately, lack of space prevents us from saying more about them here.

1.1. Notation. Throughout, we fix a prime p. We also fix a finite extension L of $\mathbb{Q}_p$, contained inside some given algebraic closure $\overline{\mathbb{Q}}_p$. We let $\mathcal{O}$ denote the ring of integers of L, and let ϖ denote a uniformizer of $\mathcal{O}$. The ring $\mathcal{O}$, and field L, will serve as our coefficients.

As usual, $\mathbb{A}$ denotes the ring of adèles over $\mathbb{Q}$, while $\mathbb{A}_f$ denotes the ring of finite adèles, and $\mathbb{A}_f^p$ denotes the ring of prime-to-p finite adèles.

2. Completed cohomology

Ideally, in the p-adic Langlands program, we would like to define spaces of p-adic automorphic forms, to serve as p-adic analogues of spaces of classical automorphic forms in the usual Langlands program. Unfortunately, for general reductive groups, no such definition is currently available. For groups that are compact at infinity, we can make such a definition (see (2.2.4) below), while for groups giving rise to Shimura varieties, we can use algebro-geometric methods to define spaces of p-adic automorphic forms, which however seem less representation-theoretic in nature than classical automorphic forms. Since one of our main goals is to employ representation-theoretic methods, we are thus led to find an alternative approach.

The approach we take here is to work with p-adically completed cohomology. This has the advantages of being definable for arbitrary groups, and of being of a representation-theoretic nature. For our purposes, it will thus serve as a suitable surrogate for a space of p-adic automorphic forms. For groups that are compact at infinity, it does recover the usual notion of p-adic automorphic forms that was already mentioned. (Its relationship to algebro-geometric notions of p-adic automorphic forms in the Shimura variety context is less clear; P. Scholze's paper [53] makes fundamental progress in this — and many other! — directions, but we won't attempt to discuss this here.) We therefore begin our discussion of global p-adic Langlands by recalling the basic definitions and facts related to completed cohomology, referring to [24] and [31] for more details.

2.1. Definitions and basic properties. Let us suppose that $\mathbb{G}$ is a reductive linear algebraic group over $\mathbb{Q}$. We write $G_\infty := \mathbb{G}(\mathbb{R})$ for the group of real-valued points of $\mathbb{R}$; this is a reductive Lie group. We let A_∞ denote the $\mathbb{R}$-points of the maximal $\mathbb{Q}$-split torus in the centre of $\mathbb{G}$, and let K_∞ denote a choice of maximal compact subgroup of G_∞. For any Lie group H, we let H° denote the subgroup consisting of the connected component of the identity.

The quotient $G_\infty/A_\infty^\circ K_\infty^\circ$ is a symmetric space on which G_∞ acts. We denote its dimension by d. We also write $l_0 :=$ rank of $G_\infty -$ rank of $A_\infty K_\infty$, and $q_0 := (d - l_0)/2$. If G_∞ is semisimple (so that in particular A_∞ is trivial) then these quantities coincide with the quantities denoted by the same symbols in [10] (which is why we notate them as we do). We note that q_0 is in fact an integer. (For the role played by these two quantities in our discussion, see (2.1.6), as well as Conjectures 3.1 and 3.2, below.)

If $K_f \subset \mathbb{G}(\mathbb{A}_f)$ is a compact open subgroup, then we write

$$Y(K_f) := \mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/A_\infty^\circ K_\infty^\circ K_f.$$

The double quotient $Y(K_f)$ is a finite union of quotients of the symmetric space $G_\infty^\circ/A_\infty^\circ K_\infty^\circ$ by cofinite volume discrete subgroups of G_∞.

2.1.1. Definitions. We fix a tame level, i.e. a compact open subgroup $K_f^p \subset \mathbb{G}(\mathbb{A}_f^p)$. If $K_p \subset G := \mathbb{G}(\mathbb{Q}_p)$ is a compact open subgroup, then of course $K_pK_f^p$ is a compact open subgroup of $\mathbb{G}(\mathbb{A}_f)$, and so we may form the space $Y(K_pK_f^p)$. We define the completed (co)homology at the tame level K_f^p as follows [24]:

$$\widetilde{H}^i := \varprojlim_s \varinjlim_{K_p} H^i\big(Y(K_pK_f^p), \mathcal{O}/\varpi^s\big) \quad \text{and} \quad \widetilde{H}_i := \varprojlim_{K_p} H_i\big(Y(K_pK_f^p), \mathcal{O}\big), \tag{2.1}$$

where in both limits K_p ranges over all compact open subgroups of $\mathbb{G}(\mathbb{Q}_p)$.

We equip each of $\widetilde{H}^i$ and $\widetilde{H}_i$ with its evident projective limit topology. In the case of $\widetilde{H}^i$, each of the $\mathcal{O}/\varpi^s$-modules $\varinjlim_{K_p} H^i\big(Y(K_pK_f^p), \mathcal{O}/\varpi^s\big)$ appearing in the projective limit is equipped with its discrete topology, and the projective limit topology on $\widetilde{H}^i$ coincides with its ϖ-adic topology; $\widetilde{H}^i$ is complete with respect to this topology. In the case of $\widetilde{H}_i$, each of the terms $H_i\big(Y(K_pK_f^p), \mathcal{O}\big)$ (which is a finitely generated $\mathcal{O}$-module) appearing in the projective limit is equipped with its ϖ-adic topology. The topology on $\widetilde{H}_i$ is then a pro-finite topology, and so in particular $\widetilde{H}_i$ is compact.

The completed cohomology and homology are related to one another in the usual way by duality over $\mathcal{O}$ [24]. In particular, if we ignore $\mathcal{O}$-torsion, then $\widetilde{H}_i$ is the $\mathcal{O}$-dual of $\widetilde{H}^i$ (equipped with its weak topology), and $\widetilde{H}^i$ is the continuous $\mathcal{O}$-dual of $\widetilde{H}_i$.

We can also form the limit

$$H^i := \varinjlim_{K_p} \varprojlim_s H^i\big(Y(K_pK_f^p), \mathcal{O}/\varpi^s\big) \cong \varinjlim_{K_p} H^i\big(Y(K_pK_f^p), \mathcal{O}\big). \tag{2.2}$$

There is a natural morphism

$$H^i \to \widetilde{H}^i, \tag{2.3}$$

which induces an embedding

$$\widehat{H}^i \hookrightarrow \widetilde{H}^i, \tag{2.4}$$

whose source is the ϖ-adic completion of H^i. Note that the morphism (2.3) need not be injective: although each of the terms $H^i\big(Y(K_pK_f^p), \mathcal{O}\big)$ in the direct limit defining H^i is a finitely generated $\mathcal{O}$-module, the limit H^i is merely countably generated as an $\mathcal{O}$-module, and hence may contain divisible elements. These then become zero after we pass to the ϖ-adic completion to obtain the embedding (2.4). (We give an example of this in (2.2.3) below.)

We furthermore remark that the transition maps in the direct limits of (2.1) and (2.2) need not be injective. (This is also illustrated by the example of (2.2.3).) However, if $K_p' \subset K_p$, then a trace argument shows that the restriction map

$$H^i\big(Y(K_pK_f^p), \mathcal{O}\big) \to H^i\big(Y(K_p'K_f^p), \mathcal{O}\big)$$

becomes injective after tensoring with L. Thus the kernel of the natural map $H^i\big(Y(K_pK_f^p), \mathcal{O}\big) \to H^i$ consists of torsion classes.

On the other hand, the kernel of the natural map

$$H^i\big(Y(K_pK_f^p),\mathcal{O}\big) \to \widetilde{H}^i \tag{2.5}$$

need not consist only of torsion classes; it is possible for non-torsion classes to have infinitely divisible image in H^i, and hence have vanishing image in $\widehat{H}^i$ (and so also have vanishing image in $\widetilde{H}^i$). (Again, we refer to (2.2.3) for an example of this.).

The morphism (2.4), although injective, need not be surjective. Its cokernel is naturally identified with $T_pH^{i+1} := \varprojlim_s H^{i+1}[\varpi^s]$; in other words we have a short exact sequence

$$0 \to \widehat{H}^i \to \widetilde{H}^i \to T_pH^{i+1} \to 0. \tag{2.6}$$

(Here we have written $H^{i+1}[\varpi^s]$ to denote the submodule of ϖ^s-torsion elements in H^{i+1}, and the projective limit is taken with respect to the multiplication-by-ϖ map from $H^{i+1}[\varpi^{s+1}]$ to $H^{i+1}[\varpi^s]$. The notation "T_p" is for *Tate module*.) In particular, if all the cohomology modules $H^{i+1}\big(Y(K_pK_f^p),\mathcal{O}\big)$ are torsion free, so that H^{i+1} is torsion free, then the morphism (2.4) *is* an isomorphism.

Although the restriction maps (2.5) can have non-trivial kernels, one can recover the cohomology at the various finite levels $K_pK_f^p$ via the Hochschild–Serre spectral sequence discussed in (2.1.3) below. The precise manner in which the cohomology at finite levels gets encoded in the completed cohomology is somewhat complicated in general. For instance, infinitely divisible torsion elements in H^i give rise to elements of T_pH^i, which by the discussion of the previous paragraph is naturally a quotient of $\widetilde{H}^{i-1}$. However, infinitely divisible non-torsion elements of H^i have no obvious incarnation as elements of completed cohomology, and the manner in which they are recovered by the Hochschild–Serre spectral sequence can be subtle.

2.1.2. Group actions. There is a natural continuous action of $G := \mathbb{G}(\mathbb{Q}_p)$ on each of $\widetilde{H}^i$ and $\widetilde{H}_i$, whose key property is encapsulated in the following result.

Theorem 2.1 ([24, 31]). *The G-action on $\widetilde{H}^i$ makes it a ϖ-adically admissible representation of G; i.e. it is ϖ-adically complete as an $\mathcal{O}$-module, and each quotient $\widetilde{H}^i/\varpi^s$ (which is a smooth representation of G over $\mathcal{O}/\varpi^s$) is admissible in the usual sense (for each compact open subgroup $K_p \subset G$, the submodule of K_p-invariants is finitely generated over $\mathcal{O}$).*

If $K_p \subset G$ is compact open, then we write $\mathcal{O}[[K_p]] := \varprojlim_{K_p'} \mathcal{O}[K_p/K_p']$, where the projective limit is taken over all open normal subgroups $K_p' \subset K_p$. Since G acts continuously on $\widetilde{H}^i$ and $\widetilde{H}_i$, it follows that for any such K_p, the K_p-action on each of these modules may be promoted to an action of $\mathcal{O}[[K_p]]$. One then has the following reformulation of the admissibility of the G-action on $\widetilde{H}^i$.

Theorem 2.2. *Each of the $\mathcal{O}[[K_p]]$-modules $\widetilde{H}_i$ is finitely generated.*

(We remark that if this finite generation statement holds for one choice of K_p then it holds for any such choice.)

2.1.3. Cohomology of local systems and the Hochschild–Serre spectral sequence. If W is a finitely generated $\mathcal{O}$-module equipped with a continuous representation of an open subgroup $K_p \subset G$, and K_p is sufficiently small, so that $\mathbb{G}(\mathbb{Q})$ acts with trivial stabilizers

on $\mathbb{G}(\mathbb{A})/A_\infty^\circ K_\infty^\circ K_p K_f^p$, then, for each open subgroup $K_p' \subset K_p$, the representation W determines a local system $\mathcal{W}$ of $\mathcal{O}$-modules on $Y(K_p K_f^p)$, defined via

$$\mathcal{W} := \mathbb{G}(\mathbb{Q})\backslash\Big(\big(\mathbb{G}(\mathbb{A}_f)/A_\infty^\circ K_\infty^\circ K_f^p\big) \times W\Big)/K_p'.$$

Suppose that W is furthermore torsion-free as an $\mathcal{O}$-module, and let $W^\vee$ denote the $\mathcal{O}$-dual of W, endowed with the contragredient K_p-action. There is then [24, 31] a Hochschild–Serre spectral sequence

$$E_2^{i,j} = \mathrm{Ext}^i_{\mathcal{O}[[K_p]]}(W^\vee, \widetilde{H}^j) \implies H^{i+j}\big(Y(K_p K_f^p), \mathcal{W}\big).$$

This gives a precise sense to the idea that p-adically completed cohomology captures *all* the cohomology (with arbitrary coefficients) at finite levels. It is the realization, in the context of completed cohomology, of the general philosophy (brought out especially in Hida's work, e.g. [45]) that when working with p-adic automorphic forms, by passing to infinite p-power level one automatically encompasses automorphic forms of all possible weights. (For the precise relationship with classical automorphic forms, see (2.1.6) below.)

In particular, if we take $W = \mathcal{O}$ (with the trivial K_p-action), we obtain a spectral sequence $E_2^{i,j} = H^i\big(K_p, \widetilde{H}^j\big) \implies H^{i+j}\big(Y(K_p K_f^p), \mathcal{O}\big)$ (where H^i denotes continuous cohomology), which recovers the cohomology at the finite level $K_p K_f^p$ from the completed cohomology. If we take a direct limit over all K_p, we obtain a spectral sequence $E_2^{i,j} = \varinjlim_{K_p} H^i\big(K_p, \widetilde{H}^j\big) \implies H^{i+j}$, which recovers the limits H^{i+j} of the cohomology at finite levels. The edge map $H^i \to \varinjlim_{K_p} (\widetilde{H}^i)^{K_p}$ is induced by the morphism (2.3). As already noted, the relationship between this spectral sequence, the morphism (2.3), and the exact sequence (2.6) is subtle. (See the example of (2.2.3) below.)

More generally, if we write $H^i(\mathcal{W}) := \varinjlim_{K_p'} H^i\big(Y(K_p' K_f^p), \mathcal{W}\big)$ (where K_p' runs over all open subgroups of the fixed K_p, of which W is a representation), then we have an edge map $H^i(\mathcal{W}) \to \varinjlim_{K_p'} \mathrm{Hom}_{K_p'}\big(W^\vee, \widetilde{H}^i\big)$ which relates the cohomology at finite levels with coefficients in $\mathcal{W}$ to the $W^\vee$-isotypic parts (for open subgroups K_p' of K_p) of $\widetilde{H}^i$.

2.1.4. Hecke actions. There is a finite set of primes Σ_0 (containing p) such that for $\ell \notin \Sigma_0$, we may factor $K_f^p = K_f^{p,\ell} K_\ell$ where $K_f^{p,\ell}$ is compact open in $\mathbb{G}(\mathbb{A}_f^{p,\ell})$, and K_ℓ is a hyperspecial maximal compact subgroup of $\mathbb{G}(\mathbb{Q}_\ell)$. (In particular, $\mathbb{G}$ is unramified at such a prime ℓ, so that it admits a hyperspecial maximal compact subgroup.) We may then consider the spherical Hecke algebra (i.e. the double coset algebra) $\mathcal{H}_\ell := \mathcal{H}\big(\mathbb{G}(\mathbb{Q}_\ell)//K_\ell, \mathcal{O}\big)$ with coefficients in $\mathcal{O}$. This algebra is commutative [41, 42], and acts naturally (by continuous operators) on any of the cohomology groups $H^i\big(Y(K_p K_f^p), \mathcal{W}\big)$ considered in (2.1.3).

If we let i range over all cohomological degrees, let K_p range over all compact open subgroups of G, and let W range over all representations of K_p on finitely generated *torsion* $\mathcal{O}$-modules, then $\prod_i \prod_{K_p} \prod_W \mathrm{End}\, H^i\big(Y(K_p K_f^p), \mathcal{W}\big)$ is a profinite ring. We fix a set of primes Σ containing Σ_0, and we define the global Hecke algebra $\mathbb{T}_\Sigma$ to be the closure in this profinite ring of the $\mathcal{O}$-subalgebra generated by the image of $\mathcal{H}_\ell$ for all $\ell \notin \Sigma$. (The reasons for allowing the possibility of Σ being larger than Σ_0 are essentially technical; it is not misleading to simply imagine that Σ is equal to Σ_0 and is fixed once and for all, e.g. by being taken to be as small as possible, given our fixed choice of tame level K_f^p.) By construction, $\mathbb{T}_\Sigma$ acts on each of the cohomology groups $H^i\big(Y(K_p K_f^p), \mathcal{W}\big)$ considered

in (2.1.3) (this is obvious if W is torsion, and follows in general by writing W as a projective limit $W \cong \varprojlim_s W/\varpi^s$), and also on $\widetilde{H}^i$ and $\widetilde{H}_i$. (In the case of $\widetilde{H}^i$ this follows from its definition in terms of limits of cohomology groups $H^i\big(Y(K_pK_f^p), \mathcal{O}/\varpi^s\big)$. It then follows for $\widetilde{H}_i$ by duality; more precisely, we can write $\widetilde{H}_i$ as the projective limit over s and K_p of $H_i\big(Y(K_pK_f^p), \mathcal{O}/\varpi^s\big)$, and this latter module is the $\mathcal{O}/\varpi^s$-dual of $H^i\big(Y(K_pK_f^p), \mathcal{O}/\varpi^s\big)$.)

These Hecke actions on $\widetilde{H}^i$ and $\widetilde{H}_i$ commute with the action of G. Thus there is an action of Hecke on the E_2-terms, as well the E_∞-terms, of the various Hochschild–Serre spectral sequences considered in (2.1.3), and these are in fact spectral sequences of $\mathbb{T}_\Sigma$-modules.

The $\mathcal{O}$-algebra $\mathbb{T}_\Sigma$ is a complete semi-local ring, i.e. it is a product of finitely many factors $\mathbb{T}_\mathfrak{m}$, each of which is a complete local ring. (This finiteness statement is proved via the methods of [3].) The topology on $\mathbb{T}_\Sigma$ is the product of the $\mathfrak{m}$-adic topologies on each of the complete local rings $\mathbb{T}_\mathfrak{m}$.

2.1.5. Variants in the non-compact case, and Poincaré duality. There are some variants of completed (co)homology which are useful in the case when the quotients $Y(K_f)$ are non-compact. (We recall that the quotients $Y(K_f)$ are compact precisely when the semisimple part of the group $\mathbb{G}$ is *anisotropic*, i.e. contains no torus that is split over $\mathbb{Q}$.)

Namely, replacing cohomology by cohomology with compact supports, we can define G-representations $\widetilde{H}^i_c$, H^i_c, $\widehat{H}^i_c$, and $T_pH^i_c$, and we have Hochschild–Serre spectral sequences for compactly supported cohomology. Similarly, we can define completed Borel–Moore homology $\widetilde{H}_i^{\mathrm{BM}}$, which is related to $\widetilde{H}^i_c$ by duality over $\mathcal{O}$.

There is a more subtle duality over $\mathcal{O}[[K_p]]$ (for a compact open subgroup $K_p \subset G$) which relates the usual and compactly supported variants of completed cohomology. It is most easily expressed on the homology side, where it takes the form of the Poincaré duality spectral sequence

$$E_2^{i,j} := \mathrm{Ext}^i_{\mathcal{O}[[K_p]]}(\widetilde{H}_j, \mathcal{O}[[K_p]]) \implies \widetilde{H}^{\mathrm{BM}}_{d-(i+j)}. \tag{2.7}$$

Of course, when the quotients $Y(K_f)$ are compact, compactly supported and usual cohomology coincide, as do usual and Borel–Moore homology. In general, we can relate them by considering the Borel–Serre compactifications $\overline{Y}(K_f)$. If we let $\partial(K_f)$ denote the boundary of $\overline{Y}(K_f)$, then we may compute the compactly supported cohomology of $Y(K_f)$ as the relative cohomology

$$H^i_c\big(Y(K_f), \mathcal{O}/\varpi^s\big) = H^i\big(\overline{Y}(K_f), \partial(K_f); \mathcal{O}/\varpi^s\big)$$

(and similarly we may compute Borel–Moore homology as relative homology), and so we obtain the long exact sequence of the pair relating the cohomology of $Y(K_f)$, the compactly supported cohomology of $Y(K_f)$, and the cohomology of $\partial(K_f)$. Passing to the various limits, we obtain a long exact sequence

$$\cdots \to \widetilde{H}^i_c \to \widetilde{H}^i \to \widetilde{H}^i(\partial) := \varprojlim_s \varinjlim_{K_p} H^i\big(\partial(K_pK_f^p), \mathcal{O}/\varpi^s\big) \to \widetilde{H}^{i+1}_c \to \cdots .$$

The completed cohomology of the boundary $\widetilde{H}^i(\partial)$ can be computed in terms of completed cohomology of the various Levi subgroups of $\mathbb{G}$; see [24] for more details.

We also mention that there is a $\mathbb{T}_\Sigma$-action on each of the objects introduced here, and that the various spectral sequences and long exact sequences considered here are all compatible with these actions.

2.1.6. The relationship to automorphic forms. We fix an isomorphism $\imath : \overline{\mathbb{Q}}_p \cong \mathbb{C}$. Since $L \subset \overline{\mathbb{Q}}_p$, we obtain induced embeddings $\mathcal{O} \subset L \hookrightarrow \mathbb{C}$.

Suppose that V is an algebraic representation of $\mathbb{G}$ defined over L, and suppose further that W is a K_p-invariant $\mathcal{O}$-lattice in V, for some compact open subgroup K_p of $G := \mathbb{G}(\mathbb{Q}_p) \subset \mathbb{G}(L)$. Write $V_{\mathbb{C}} := \mathbb{C} \otimes_L V = \mathbb{C} \otimes_{\mathcal{O}} W$ (the tensor products being taken with respect to the embeddings induced by $\imath$). If, for any compact open subgroup K'_p of K_p, we let $\mathcal{V}_{\mathbb{C}}$ denote the local system of $\mathbb{C}$-vector spaces on $Y(K'_pK^p_f)$ associated to $V_{\mathbb{C}}$, then we obtain a natural isomorphism

$$\mathbb{C} \otimes_{\mathcal{O}} H^i(\mathcal{W}) \cong \varinjlim_{K'_p} H^i\big(Y(K'_pK^p_f), \mathcal{V}_{\mathbb{C}}\big) =: H^i(\mathcal{V}_{\mathbb{C}}).$$

We may compute this cohomology using the space of automorphic forms on $\mathbb{G}$ [39].

Precisely, write $\mathcal{A}(K_f)$ for the space of automorphic forms on $\mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/K_f$, for any compact open subgroup $K_f \subset \mathbb{G}(\mathbb{A}_f)$, and write $\mathcal{A}(K^p_f) = \varinjlim_{K_p} \mathcal{A}(K_pK^p_f)$, where the direct limit is taken (as usual) over the compact open subgroups $K_p \subset G := \mathbb{G}(\mathbb{Q}_p)$. For simplicity, suppose that A°_∞ acts on $V_{\mathbb{C}}$ through some character χ (this will certainly hold if V is absolutely irreducible), and let $\mathcal{A}(K^p_f)_{\chi^{-1}}$ denote the subspace of $\mathcal{A}(K^p_f)$ on which A°_∞ acts through the character χ^{-1}. Let $\widetilde{G}_\infty$ denote the group of real points of the intersection of the kernels of all the rational characters of $\mathbb{G}$, and let $\widetilde{\mathfrak{g}}$ and $\mathfrak{k}$ denote the Lie algebras of $\widetilde{G}_\infty$ and K_∞ respectively. Then it is proved by J. Franke in [39] that there is a natural isomorphism

$$H^i(\mathcal{V}_{\mathbb{C}}) \cong H^i\big(\widetilde{\mathfrak{g}}, \mathfrak{k}; \mathcal{A}(K^p_f)_{\chi^{-1}} \otimes V_{\mathbb{C}}\big). \tag{2.8}$$

(In the case when the the quotients $Y(K_f)$ are compact, this result is known as *Matsushima's formula* [49]. In the case when $\mathbb{G} = \mathrm{GL}_2$ and $i = 1$, it is known as the *Eichler–Shimura isomorphism* [54].) The action of $\mathbb{T}_\Sigma$ on $H^i(\mathcal{W})$ induces an action of $\mathbb{T}_\Sigma$ on $H^i(\mathcal{V}_{\mathbb{C}})$, and the resulting systems of Hecke eigenvalues that appear in $H^i(\mathcal{V}_{\mathbb{C}})$ (which we may think of as being simultaneously $\mathbb{C}$-valued and $\overline{\mathbb{Q}}_p$-valued, by employing the isomorphism $\imath$) are automorphic, i.e. arise as systems of Hecke eigenvalues on the space of automorphic forms $\mathcal{A}(K^p_f)$.

Thus, to first approximation, we may regard the Hecke algebra $\mathbb{T}_\Sigma$ (or, more precisely, the $\overline{\mathbb{Q}}_p$-valued points of its Spec) as being obtained by ϖ-adically interpolating the automorphic systems of Hecke eigenvalues that occur in H^i. However, as we already noted, the map from H^i to $\widehat{H}^i$ can have a non-trivial kernel. More significantly (given that the Hochschild–Serre spectral sequence shows that, despite this possibility, the systems of Hecke eigenvalues appearing in H^i can be recovered from the completed cohomology), the inclusion $\widehat{H}^i \hookrightarrow \widetilde{H}^i$ can have a non-trivial cokernel T_pH^{i+1}, which arises from infinitely divisible torsion in H^{i+1}. Thus $\mathrm{Spec}\,\mathbb{T}_\Sigma$ sees not only the systems of eigenvalues arising from classical automorphic forms (and their ϖ-adic interpolations), but also systems of Hecke eigenvalues arising from torsion cohomology classes (and *their* ϖ-adic interpolations). We will discuss this point further in (3.1.3) below.

It will be helpful to say a little more about how one can use (2.8) to analyze cohomology. To this end, we decompose the space of automorphic forms $\mathcal{A}(K^p_f)_{\chi^{-1}}$ as the direct sum

$$\mathcal{A}(K^p_f)_{\chi^{-1}} = \mathcal{A}_{\mathrm{cusp}}(K^p_f)_{\chi^{-1}} \oplus \mathcal{A}_{\mathrm{Eis}}(K^p_f)_{\chi^{-1}} \tag{2.9}$$

of the cuspforms and the forms which are orthogonal to the cuspforms (in $\mathcal{A}(K^p_f)_\chi$) under the Petersson inner product (i.e. the L^2 inner product of functions on $\mathbb{G}(\mathbb{Q})\backslash\mathbb{G}(\mathbb{A})/A^\circ_\infty K^p_f$).

We label this complement to the space of cuspforms with the subscript Eis for *Eisenstein*, although its precise relationship with the space of Eisenstein series can be complicated; see e.g. [39, 40, 47]. Note that if the quotients $Y(K_f)$ are compact, then $\mathcal{A}(K_f^p)_{\chi^{-1}} = \mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$.

The decomposition (2.9) is a direct sum of $G_\infty \times G$-representations, and so the cohomology $H^i\big(\widetilde{\mathfrak{g}}, \mathfrak{k}; \mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}} \otimes V_{\mathbb{C}}\big)$ is a G-invariant direct summand of the cohomology $H^i\big(\widetilde{\mathfrak{g}}, \mathfrak{k}; \mathcal{A}(K_f^p)_{\chi^{-1}} \otimes V_{\mathbb{C}}\big)$, which via the isomorphism (2.8) we identify with a G-invariant direct summand $H^i_{\mathrm{cusp}}(\mathcal{V}_{\mathbb{C}})$ of $H^i(\mathcal{V}_{\mathbb{C}})$. It is also $\mathbb{T}_\Sigma$-invariant.

We may choose an everywhere positive element of $\mathcal{A}(K_f^p)$ on which $\mathbb{G}(\mathbb{A})$ acts via a character extending χ^2, multiplication by which induces an isomorphism between $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$ and $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi}$. The Petersson (i.e. the L^2) inner product between $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$ and $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi}$, taken together with this isomorphism, thus induces a pre-Hilbert space structure on $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$, with respect to which the $G_\infty \times G$-action is unitary up to a character. In particular, we find that $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$ is semisimple as a $G_\infty \times G$-representation, and so decomposes as a direct sum $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}} = \bigoplus_{\pi_\infty \otimes \pi_p} \pi_\infty \otimes \pi_p \otimes M(\pi_\infty \otimes \pi_p)$, where the direct sum is taken over (a set of isomorphism class representatives of) all the irreducible admissible representations of $G_\infty \otimes G$ over $\mathbb{C}$ on which A_∞° acts via χ^{-1} (which factor as the tensor product $\pi_\infty \otimes \pi_p$ of an irreducible admissible representation π_∞ of G_∞ on which A_∞° acts via χ^{-1}, and an irreducible admissible smooth representation π_p of G), and $M(\pi_\infty \otimes \pi_p) := \mathrm{Hom}_{G_\infty \times G}\big(\pi_\infty \otimes \pi_p, \mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}\big)$ is the (finite-dimensional) multiplicity space of $\pi_\infty \otimes \pi_p$ in $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$. Consequently, we obtain a direct sum decomposition

$$\begin{aligned} H^i_{\mathrm{cusp}}(\mathcal{V}_{\mathbb{C}}) &\cong H^i\big(\widetilde{\mathfrak{g}}, \mathfrak{k}; \mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}} \otimes V_{\mathbb{C}}\big) \\ &\cong \bigoplus_{\pi_\infty \otimes \pi_p} H^i(\widetilde{\mathfrak{g}}, \mathfrak{k}; \pi_\infty \otimes V_{\mathbb{C}}) \otimes \pi_p \otimes M(\pi_\infty \otimes \pi_p) \end{aligned} \quad (2.10)$$

(the point here being that the $(\widetilde{\mathfrak{g}}, \mathfrak{k})$-cohomology depends only on the structure of $\pi_\infty \otimes \pi_p$ as a G_∞-representation, which is to say, only on π_∞).

In this way, we can speak of the contribution of each of the irreducible summands $\pi_\infty \otimes \pi_p$ of $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$ to the cohomology $H^i_{\mathrm{cusp}}(\mathcal{V}_{\mathbb{C}})$. In the case when $H^i(\widetilde{\mathfrak{g}}, \mathfrak{k}; \pi_\infty \otimes V_{\mathbb{C}})$ is non-zero, so that $\pi_\infty \otimes \pi_p$ actually does contribute to cohomology, the multiplicity space $M(\pi_\infty \otimes \pi_p)$ is naturally a $\mathbb{T}_\Sigma$-module, and the direct sum decomposition (2.10) is compatible with the actions of G and $\mathbb{T}_\Sigma$ (where G-acts on π_p, and $\mathbb{T}_\Sigma$ acts on $M(\pi_\infty \otimes \pi_p)$).

We can now explain the significance of the quantities l_0 and q_0 introduced above. Namely, q_0 is the lowest degree in which *tempered* π_∞ can admit non-trivial $(\widetilde{\mathfrak{g}}, \mathfrak{k})$-cohomology. Furthermore, among the tempered representations of G_∞, it is precisely the *fundamental tempered representations* (i.e. those which are induced from discrete series representations of the Levi subgroup of a fundamental parabolic subgroup of G_∞) which can admit non-zero cohomology at all, and they do so precisely in degrees between q_0 and $q_0 + l_0$ [10, Ch. III, Thm. 5.1]. (This is a range of degrees of length $l_0 + 1$ symmetric about $d/2$ — which is one-half of the dimension of the quotients $Y(K_f)$.) A fundamental theorem of Harish-Chandra states that, when $\mathbb{G}$ is semisimple, the group G_∞ admits discrete series representations precisely if $l_0 = 0$; in this case q_0 is equal to $d/2$, and the fundamental tempered representations are precisely the discrete series representations. Key examples for which $l_0 = 0$ are given by groups $\mathbb{G}$ for which G_∞ is compact, and (the semisimple parts of) groups giving rise to Shimura varieties.

One representation of $G_\infty \times G$ that always appears in $\mathcal{A}(K_f^p)$ is the trivial representation $\mathbb{1}$. Thus there are induced morphisms

$$H^i(\widetilde{\mathfrak{g}}, \mathfrak{k}; \mathbb{1}) \to H^i\big(\widetilde{\mathfrak{g}}, \mathfrak{k}; \mathcal{A}(K_f^p)_{\mathbb{1}}\big) \cong \mathbb{C} \otimes_{\mathcal{O}} H^i, \tag{2.11}$$

whose sources are naturally identified with the cohomology spaces of the compact dual to the symmetric space $G_\infty^\circ / A_\infty^\circ K_\infty^\circ$. If the $Y(K_f)$ are not compact, then $\mathbb{1}$ lies in $\mathcal{A}_{\mathrm{Eis}}(K_f^p)_{\mathbb{1}}$, but it is typically not a direct summand, and so the morphisms (2.11) are typically not injective. Nevertheless, when i is small they will be injective, and indeed (if e.g. $\mathbb{G}$ is split, semisimple, and simply connected) an isomorphism, and these isomorphisms play a key role in the theory of homological stability [9]. We discuss the interaction of these isomorphisms with the theory of completed cohomology (in the case when $\mathbb{G} = \mathrm{SL}_N$) in (2.2.3) below.

2.1.7. Dimension theory. For any compact open subgroup $K_p \subset G$, the completed group ring $\mathcal{O}[[K_p]]$ is (left and right) Noetherian, and of finite injective dimension as a module over itself; more precisely, its injective dimension is equal to $\dim \mathbb{G} + 1$. This allows us to consider a (derived) duality theory for finitely generated $\mathcal{O}[[K_p]]$-modules: to any such module M we associate the Ext-modules $E^i(M) := \mathrm{Ext}^i_{\mathcal{O}[[K_p]]}(M, \mathcal{O}[[K_p]])$, for $i \geq 0$ (which necessarily vanish if $i > \dim \mathbb{G} + 1$); these are again naturally $\mathcal{O}[[K_p]]$-modules. (We use the left $\mathcal{O}[[K_p]]$-module structure on $\mathcal{O}[[K_p]]$ to compute the E^i, and then use the right $\mathcal{O}[[K_p]]$-module structure on $\mathcal{O}[[K_p]]$, converted to a left module structure by the usual device of applying the anti-involution $k \mapsto k^{-1}$ on K_p, to give the E^i an $\mathcal{O}[[K_p]]$-module structure.)

An important point is that $E^i(M)$ is canonically independent of the choice of K_p (in the sense that we may regard M as an $\mathcal{O}[[K_p']]$-module, for any open subgroup K_p' of K_p, but $E^i(M)$ is canonically independent of which choice of K_p' we make to define it). This has the consequence that if the K_p-representation on M extends to a G-representation, then the $E^i(M)$ are also naturally G-representations.

The Poincaré duality spectral sequence (2.7) may thus be rewritten as

$$E_2^{i,j} = E^i(\widetilde{H}_j) \implies \widetilde{H}^{\mathrm{BM}}_{d-(i+j)}.$$

All the objects appearing in it are G-representations, and it is G-equivariant.

If K_p is p-torsion free (which will be true provided K_p is sufficiently small) then the ring $\mathcal{O}[[K_p]]$ is in fact Auslander regular [55]. If K_p is furthermore pro-p, then the ring $\mathcal{O}[[K_p]]$ is an integral domain (in the sense that it contains no non-trivial left or right zero divisors). This has various implications for the theory of finitely generated $\mathcal{O}[[K_p]]$-modules. For example, it is reasonable to define such a module to be *torsion* if every element has a non-zero annihilator in $\mathcal{O}[[K_p]]$ (where K_p is chosen small enough to be pro-p and p-torsion free). More generally, we may make the following definition.

Definition 2.3. If M is a finitely generated $\mathcal{O}[[K_p]]$-module, then the *codimension* $\mathrm{cd}(M)$ of M is defined to be the least value of i for which $E^i(M) \neq 0$; the dimension $\dim M$ of M is defined to be $\dim \mathbb{G} + 1 - \mathrm{cd}(M)$.

If $\mathbb{G}$ is a torus, so that K_p is commutative, then $\dim M$ is precisely the dimension of the support of the localization of M over $\mathrm{Spec}\, \mathcal{O}[[K_p]]$. In general $\mathcal{O}[[K_p]]$ is non-commutative, and hence doesn't have a Spec over which we can localize a finitely generated module M. Nevertheless, the quantity $\dim M$ behaves as if it were the dimension of the support of M in

the (non-existent) Spec of $\mathcal{O}[[K_p]]$. (See e.g. [55, Prop. 3.5].) As one example, we note that M is torsion (in the sense defined above) if and only if $\mathrm{cd}(M) \geq 1$. As another, we note that $\mathrm{cd}\, E^i(M) \geq i$, with equality when $i = \mathrm{cd}\, M$. Also, we note that $\mathrm{cd}\, M = \infty$ (so, morally, the support of M is empty) precisely when $M = 0$.

By Theorem 2.2, completed homology is finitely generated over $\mathcal{O}[[K_p]]$, and so these dimension-theoretic notions apply to it. Information about the (co)dimension of $\widetilde{H}_i$ can be obtained by analyzing the Poincaré duality spectral sequence, since the functors E^i appear explicitly in it, and there is a tension in this spectral sequence between the top degree for the duality of $\mathcal{O}[[K_p]]$-modules (which is $\dim \mathbb{G} + 1$) and the top degree for usual Poincaré duality (which is d), which can sometimes be exploited. (See e.g. (2.2.2) below.) In making any such analysis, it helps to have some *a priori* information about the (co)dimensions of the $\widetilde{H}^i$. This is provided by the following result.

Theorem 2.4 ([23]). *Suppose that $\mathbb{G}$ is semisimple. Then $\widetilde{H}^i$ is torsion* (*i.e. its codimension is positive*) *unless $l_0 = 0$ and $i = q_0$, in which case* $\mathrm{cd}\, \widetilde{H}_i = 0$.

We make a much more precise conjecture about the codimensions of the $\widetilde{H}^i$ in Conjecture 3.2 below.

2.2. Examples. We illustrate the preceding discussion with some examples.

2.2.1. $\mathbf{GL_2}$ of $\mathbb{Q}$. If $\mathbb{G} = \mathrm{GL}_2$, then the quotients $Y(K_f)$ are classical modular curves. In particular, they have non-vanishing cohomology only in degrees 0 and 1, and so we consider completed cohomology in degrees 0 and 1. Since the cohomology of a curve is torsion free, we have $\widetilde{H}^0 = \widehat{H}^0 \cong \mathcal{C}\big(\Delta \times \mathbb{Z}_p^\times, \mathcal{O}\big)$, the space of continuous $\mathcal{O}$-valued functions on the product $\Delta \times \mathbb{Z}_p^\times$, where Δ is a finite group that depends on the choice of tame level, and so $\widetilde{H}_0$ (which is simply the $\mathcal{O}$-dual of $\widehat{H}^0$) is isomorphic to $\mathcal{O}[[\Delta \times \mathbb{Z}_p^\times]]$, which is of dimension two as a module over $\mathcal{O}[[K_p]]$ (for any compact open $K_p \subset \mathrm{GL}_2(\mathbb{Q}_p)$). We also have $\widetilde{H}^1 = \widehat{H}^1$. Again, $\widetilde{H}_1$ is the $\mathcal{O}$-dual of $\widetilde{H}^1$, and Theorem 2.4 shows that $\mathrm{cd}\, \widetilde{H}_1 = 0$. (Strictly speaking, we have to apply the theorem to SL_2 rather than GL_2, but it is then easy to deduce the corresponding result for GL_2, by considering the cup-product action of H^0 on H^1.)

2.2.2. $\mathbf{SL_2}$ of an imaginary quadratic field. Suppose that $\mathbb{G} = \mathrm{Res}_{F/\mathbb{Q}}\, \mathrm{SL}_2$, where F is an imaginary quadratic extension of $\mathbb{Q}$. The quotients $Y(K_f)$ are then connected, non-compact three-manifolds. The relevant (co)homological degrees are thus $i = 0, 1$, and 2. Since the $Y(K_f)$ are connected we see that $\widetilde{H}^0 = \widetilde{H}_0 = \mathcal{O}$. Theorem 2.4 implies that $\widetilde{H}_1$ has positive codimension. A consideration of the Poincaré duality spectral sequence then shows that $\widetilde{H}_2 = 0$, and that $\widetilde{H}_1$ is of codimension 1. This computation exploits both the fact that $\widetilde{H}_0 \neq 0$, and the gap between 3 (the dimension of the $Y(K_f)$) and 6 (the dimension of $\mathbb{G}$).

Since H^1 with coefficients in $\mathcal{O}$ of any space is ϖ-torsion free, we see that $\widetilde{H}^1$ is ϖ torsion-free, and coincides with the $\mathcal{O}$-dual of $\widetilde{H}_1$, while $\widetilde{H}^2$ is ϖ-torsion, and is naturally identified with the Pontrjagin dual of the $\mathcal{O}$-torsion submodule of $\widetilde{H}_1$ [24, Thm. 1.1]. We conjecture that in fact $\widetilde{H}_1$ is ϖ-torsion free (see Conjecture 3.2), and thus that $\widetilde{H}^2 = 0$.

2.2.3. $\mathbf{SL_N}$ of $\mathbb{Q}$ in low degrees. We first discuss $\widetilde{H}^0$ and $\widetilde{H}^1$, before turning to a discussion of higher degree cohomology from the point of view of homological stability.

The quotients $Y(K_f)$ are connected, and so $\widetilde{H}_0 = \mathcal{O}$. If $N \geq 3$, then SL_N satisfies the

congruence subgroup property. Furthermore, the groups $\mathrm{SL}_N(\mathbb{Z}_\ell)$ are perfect for all values of ℓ. Combining these two facts, we see that if $\Gamma(p^r)$ denotes the usual congruence subgroup of level p^r (i.e. the kernel of the surjection $\mathrm{SL}_N(\mathbb{Z}) \to \mathrm{SL}_N(\mathbb{Z}/p^r)$), then $H_1\big(\Gamma(p^r), \mathcal{O}\big) = \mathcal{O} \otimes_{\mathbb{Z}_p} \Gamma(p^r)^{\mathrm{ab}} \cong \mathcal{O} \otimes_{\mathbb{Z}_p} \Gamma(p^r)/\Gamma(p^{2r})$. Thus, if we take the tame level to be trivial (i.e. "level one"), then we see that the transition maps in the projective system defining $\widetilde{H}_1$ are eventually zero, implying that $\widetilde{H}_1 = 0$. Similarly, we see that $H^1 = \widetilde{H}^1 = 0$. (If we allowed a more general tame level, then $\widetilde{H}_1$ could be non-zero, but would be finite. Since H^1 and $\widetilde{H}^1$ are always torsion free, they would continue to vanish.)

We now consider cohomology in higher degree, but in the stable range, in the sense that we now explain. (The tame level can now be taken to be arbitrary.) Borel's result [9] on homological stability for SL_N shows that when i is sufficiently small (compared to N), the cohomology $L \otimes_{\mathcal{O}} H^i$ is independent of N; indeed, it consists precisely of the contributions to cohomology arising from the trivial automorphic representation, in the sense discussed in (2.1.6). If we define H^i_{stab} to be the stable value of H^i, then Borel shows that $\bigoplus H^i_{\mathrm{stab}}$ (as an algebra under cup product) is isomorphic to an exterior algebra generated by a single element in each degree $i = 0, 5, 9, 13, \ldots$.

In fact, stability also holds for completed cohomology.

Theorem 2.5 ([25]). *If i is sufficiently small compared to N, then $\widetilde{H}^i$ is independent of N, is finitely generated as an $\mathcal{O}$-module, and affords the trivial representation of G.*

We write $\widetilde{H}^i_{\mathrm{stab}}$ to denote the stable value of $\widetilde{H}^i$. F. Calegari has succeeded in computing $\widetilde{H}^i_{\mathrm{stab}}$ modulo torsion, contingent on a natural non-vanishing conjecture for certain special values of the p-adic ζ-function (a conjecture which holds automatically if p is a regular prime).

Theorem 2.6 ([22, Theorem 2.3]). *Suppose either that p is a regular prime, or that appropriate special values of the p-adic zeta function are non-zero. Then there is an isomorphism of graded vector spaces $\bigoplus_{i \geq 0} L \otimes_{\mathcal{O}} \widetilde{H}^i_{\mathrm{stab}} \cong L[x_2, x_6, x_{10}, \ldots]$ (where $L[x_2, x_6, x_{10}, \ldots]$ denotes the graded ring generated by the indicated sequence of polynomial variables x_i, $i \equiv 2 \bmod 4$, with x_i placed in degree i).*

Note in particular that $\bigoplus_i L \otimes_{\mathcal{O}} \widetilde{H}^i_{\mathrm{stab}}$ vanishes in all odd degrees, while $\bigoplus_i L \otimes_{\mathcal{O}} H^i_{\mathrm{stab}}$ is generated by classes in odd degrees. Thus, when $\mathbb{G} = \mathrm{SL}_N$, the map (2.3) is identically zero (modulo torsion) when i lies in the stable range!

Assuming the hypotheses of the theorem, we conclude that the Borel classes (i.e. the non-zero elements of $L \otimes_{\mathcal{O}} H^i_{\mathrm{stab}}$) become infinitely divisible when we pull them back up the p-power level tower. It is interesting to consider how they reappear in the Hochschild–Serre spectral sequence. If we again ignore torsion, then $H^i\big(\mathrm{SL}_N(\mathbb{Z}_p), L\big)$ coincides with the Lie algebra cohomology of $\mathfrak{sl}_N$ [48], which is (stably in N) an exterior algebra on generators in degrees $3, 5, 7, \ldots$. Since all of the non-zero $\widetilde{H}^i_{\mathrm{stab}}$ have trivial G-action, we see that we obtain non-trivial Ext terms in odd degrees in the Hochschild–Serre spectral sequence, and the Borel classes are recovered from these.

If we consider the short exact sequence (2.6) for non-zero even degrees, we see that $\widehat{H}^i$ vanishes (modulo torsion), while $\widetilde{H}^i$ is non-zero. Thus the term $T_p H^{i+1}$ must be non-zero; this provides a rather compelling illustration of the manner in which torsion classes can accumulate as we pass to the limit of the p-power level tower.

2.2.4. Groups that are compact at infinity. If G_∞ (or, more generally, if the quotient G_∞/A_∞) is compact, then $A^\circ_\infty K^\circ_\infty$ equals G°_∞, and so the quotients $Y(K_f)$ are simply finite

sets. Thus the only interesting degree of cohomology is $i = 0$. In this case the inverse limit $Y(K^p_f) := \varprojlim_{K_p} Y(K_pK^p_f)$ is a pro-finite set, which is in fact a compact p-adic analytic manifold, equipped with an analytic action of G. If K_p is taken sufficiently small (small enough that the $\mathbb{G}(\mathbb{Q})$-action on $\mathbb{G}(\mathbb{A})/G^\circ_\infty K_pK^p_f$ is fixed-point free), then K_p acts with trivial stabilizers on $Y(K_pK^p_f)$, and $Y(K_pK^p_f)$ is the disjoint union of finitely many (say n) principal homogeneous spaces over K_p.

One immediately verifies that $\widetilde{H}^0 \cong \mathcal{C}\big(Y(K^p_f), \mathcal{O}\big)$, the space of continuous $\mathcal{O}$-valued functions on $Y(K^p_f)$, while $\widetilde{H}_0 \cong \mathcal{O}[[Y(K^p_f)]]$, the space of $\mathcal{O}$-valued measures on $Y(K^p_f)$ (which is the $\mathcal{O}$-dual of $\mathcal{C}\big(Y(K^p_f), \mathcal{O}\big)$). In particular, if K_p is sufficiently small, then we see that $\widetilde{H}_0$ is free of rank n over $\mathcal{O}[[K_p]]$.

It is natural (e.g. in light of Gross's definition of algebraic modular forms in this context [42]) to define $\mathcal{C}\big(Y(K^p_f), \mathcal{O}\big)$ to be the space of p-adic automorphic forms on $\mathbb{G}(\mathbb{A})$ of tame level K^p_f, and so in this case we see that completed cohomology does indeed coincide with the natural notion of p-adic automorphic forms.

3. *p*-adic Langlands

We describe the manner in which we expect completed cohomology, and the Hecke algebra acting on it, to be related to deformation rings of global Galois representations. This conjectural relationship suggests various further conjectures, as we explain, as well as a relationship to a hypothetical p-adic local Langlands correspondence.

3.1. The connection between completed cohomology and Galois representations. If π is an automorphic representation of $\mathbb{G}(\mathbb{A})$ for which $H^i(\widetilde{\mathfrak{g}}, \mathfrak{k}; \pi_\infty \otimes V_{\mathbb{C}}) \neq 0$ for some algebraic representation V of $\mathbb{G}$ and some degree i, then π is C-algebraic [21, Lemma 7.2.2]. Thus, we expect that associated to π there should be a continuous representation of $G_{\mathbb{Q}}$ into the $\overline{\mathbb{Q}}_p$-valued points of the C-group of $\mathbb{G}$ [21, Conj. 5.3.4]. In light of the isomorphism (2.8), we thus expect that the systems of Hecke eigenvalues that occur in $H^i(\mathcal{V})$ should have associated representations of $G_{\mathbb{Q}}$ into the C-group of $\mathbb{G}$. For systems of Hecke eigenvalues occurring on torsion classes in cohomology, conjectures of Ash [2] again suggest that there should be associated Galois representations.

These expectations have been proved correct in many cases; for example, if $\mathbb{G}$ is the restriction of scalars to $\mathbb{Q}$ of GL_n over a totally real or a CM number field. (See [44] in the case of characteristic zero systems of eigenvalues, and [53] in the case of torsion systems of eigenvalues. See also [43] for an overview of these results.)

Returning for a moment to the general case, one further expects that the Galois representations obtained should be *odd*, in the sense that complex conjugation in $G_{\mathbb{Q}}$ maps to a certain prescribed conjugacy class in the C-group of $\mathbb{G}$. (See [7, Prop. 6.1] for a description of the analogous conjugacy class in the L-group of $\mathbb{G}$.)

Let $\mathfrak{m}$ denote a maximal ideal in $\mathbb{T}_\Sigma$, write $\mathbb{F} := \mathbb{T}_\Sigma/\mathfrak{m}$, and let $\overline{\mathbb{F}}$ denote a chosen algebraic closure of $\mathbb{F}$. Then, by the above discussion, we believe that associated to $\mathfrak{m}$ there should be a continuous representation of $G_{\mathbb{Q}}$ into the $\overline{\mathbb{F}}$-valued points of the C-group of $\mathbb{G}$. Assuming that this representation exists, we will denote it by $\overline{\rho}_{\mathfrak{m}}$.

To be a little more precise: the representation $\overline{\rho}_{\mathfrak{m}}$ should have the property that it is unramified at the primes outside Σ, and that for any $\ell \notin \Sigma$, the semisimple part of $\overline{\rho}_{\mathfrak{m}}(\mathrm{Frob}_\ell)$ should

be associated to the local-at-ℓ part of the system of Hecke eigenvalues by a suitable form of the Satake isomorphism (see [21, 42]). In general, this condition may not serve to determine $\overline{\rho}_{\mathfrak{m}}$ uniquely; even in the case $\mathbb{G} = \mathrm{GL}_n$, it determines $\overline{\rho}_{\mathfrak{m}}$ only up to semisimplification; for other choices of $\mathbb{G}$, as well as this issue, one can have the additional phenomenon that even a representation which doesn't factor through any parabolic subgroup of the C-group is not determined by its pointwise conjugacy classes ("local conjugacy does not determine global conjugacy"). We do not attempt to deal with this issue in general here; rather we simply ignore it, and continue our discussion as if $\overline{\rho}_{\mathfrak{m}}$ were unambiguously defined.

As we observed above, we may regard $\mathbb{T}_{\mathfrak{m}}$ (or its Spec) as interpolating systems of Hecke eigenvalues associated to classical C-algebraic automorphic forms, and/or to torsion classes in cohomology. Since the deformations of $\overline{\rho}_{\mathfrak{m}}$ can be interpolated into a formal deformation space, it is then reasonable to imagine that we may deform $\overline{\rho}_{\mathfrak{m}}$ to a representation of $G_{\mathbb{Q}}$ into the $\mathbb{T}_{\mathfrak{m}}$-valued points of the C-group of $\mathbb{G}$. Here again there are many caveats: firstly, if $\overline{\rho}_{\mathfrak{m}}$ is "reducible" (i.e. factors through a proper parabolic of the C-group), then we would have to work with some form of *pseudo-character* or *determinant*, as in [28, 53]; also, there is a question of rationality, or "Schur index" — it may be that if we want $\overline{\rho}_{\mathfrak{m}}$ to be defined over $\mathbb{F}$, and its deformation to be defined over $\mathbb{T}_{\mathfrak{m}}$, then we may have to extend our scalars, or else replace the C-group by an inner twist. Again, we don't attempt to address these issues here.

Rather, we begin with some general conjectures about dimension and vanishing that are motivated by the preceding discussion, and then continue by discussing the relationships between completed cohomology and p-adic Hodge theory and a possible p-adic local Langlands correspondence. Finally, we turn to a discussion of some specific examples, where we can make our generalities more precise.

3.1.1. Conjectures on dimension and vanishing. We begin by making a somewhat vague conjecture, which can be thought of as a rough expression of our hopes for reciprocity in the context of global p-adic Langlands: namely (continuing with the notation introduced in the preceding discussion), we conjecture that $\mathbb{T}_{\mathfrak{m}}$ is universal (in some suitable sense) for parameterizing *odd* formal deformations of $\overline{\rho}_{\mathfrak{m}}$ whose ramification away from p is compatible with the tame level structure K^p_f (via some appropriate form of ϖ-adic local Langlands for the group $\mathbb{G}$ at primes $\ell \nmid p$ which, again, we won't attempt to formulate here; but see [36] in the case when $\mathbb{G} = \mathrm{GL}_n$). The global Euler characteristic formula for Galois cohomology lets us compute the expected dimension of such a universal deformation ring, and this motivates in large part the following concrete conjecture [24, §8].

Conjecture 3.1. *Each local factor $\mathbb{T}_{\mathfrak{m}}$ of $\mathbb{T}_{\Sigma}$ is Noetherian, reduced, and ϖ-torsion free, of Krull dimension equal to $\dim \mathbb{B} + 1 - l_0$, where $\dim \mathbb{B}$ denotes the dimension of a Borel subgroup $\mathbb{B}$ of $\mathbb{G}$.*

This is known in some cases (which we will recall below). We know of no way to prove the Noetherianness of $\mathbb{T}_{\mathfrak{m}}$, or the statement about its Krull dimension, other than to relate the Hecke algebra to a deformation ring of Galois representations, and then use techniques from the theory of Galois representations to compute the dimension.

In some situations we *can* prove that $\mathbb{T}_{\mathfrak{m}}$ is torsion free and reduced (see e.g. the discussion of (3.1.2) below), and it seems reasonable to conjecture these properties in general. Indeed, these properties are closely related to the following conjecture [24, Conj. 1.5].

Conjecture 3.2. *$\widetilde{H}_i = 0$ if $i > q_0$, while $\mathrm{cd}\, \widetilde{H}_{q_0} = l_0$, and $\mathrm{cd}\, \widetilde{H}_i > l_0 + q_0 - i$ if $i < q_0$. Furthermore, $\widetilde{H}_{q_0}$ is ϖ-torsion free, and $\mathbb{T}_{\Sigma}$ acts faithfully on $\widetilde{H}_{q_0}$.*

As noted in [24, Thm. 1.6], the truth of this conjecture for $\mathbb{G}$ and all its Levi subgroups implies the analogous statement for $\widetilde{H}_i^{\mathrm{BM}}$. Also, the vanishing conjecture for $\mathbb{G} \times \mathbb{G}$ implies (via the Künneth formula) the torsion-freeness of $\widetilde{H}_{q_0}$.

One of the ideas behind this conjecture is that "all the interesting Hecke eigenvalues should appear in degree q_0". This is inspired by the fact (recalled in (2.1.6)) that tempered automorphic representations don't contribute to cohomology in degrees below q_0, so that the Galois representations associated to the systems of Hecke eigenvalues appearing in degrees below q_0 should be "reducible" (i.e. factor through a proper parabolic subgroup of the C-group). In a Galois deformation space with unrestricted ramification at p, the reducible representations should form a proper closed subset, and so should be approximable by "irreducible" Galois representations (i.e. representations which don't factor through a proper parabolic). Thus we don't expect to see any systems of Hecke eigenvalues in degrees below q_0 which can't already be observed in degree q_0, and hence we expect that $\mathbb{T}_\Sigma$ will act faithfully on $\widetilde{H}_{q_0}$.

We also expect that $\widetilde{H}_i$ should be "small" if $i < q_0$, since the possible systems of Hecke eigenvalues which it can carry are (or should be) constrained. However, if the $\widetilde{H}_i$ are small enough for $i < q_0$, then the Poincaré duality spectral sequence (combined with an analysis of the completed cohomology of the boundary, which can be treated inductively, by reducing to the case of lower rank groups) implies the vanishing of $\widetilde{H}_i$ in degrees $> q_0$.

If we believe that $\mathbb{T}_\Sigma$ acts faithfully on $\widetilde{H}_{q_0}$, and that $\mathbb{T}_\Sigma$ has dimension $\dim \mathbb{B} + 1 - l_0$ (as predicted by Conjecture 3.1), then conjecturing that $\widetilde{H}_{q_0}$ has codimension l_0 is morally equivalent to conjecturing that the fibres of $\widetilde{H}_{q_0}$ over the points of $\operatorname{Spec} \mathbb{T}_\Sigma$ are of dimension $\dim \mathbb{G}/\mathbb{B}$. This latter statement fits nicely with the fact that generic irreducible representations of $G := \mathbb{G}(\mathbb{Q}_p)$ have Gelfand–Kirillov dimension equal to $\dim \mathbb{G}/\mathbb{B}$, and the analogy between the dimension of $\mathcal{O}[[K_p]]$-modules and Gelfand–Kirillov dimension [24, Remark 1.19]. Unfortunately, we don't know how to make this idea precise, since we don't know how to prove (in any generality) this relationship between the Krull dimension of $\mathbb{T}_\Sigma$, the dimension of $\widetilde{H}_{q_0}$, and the dimension of the fibres of the latter over points of $\operatorname{Spec} \mathbb{T}_\Sigma$, even assuming that $\mathbb{T}_\Sigma$ acts faithfully on $\widetilde{H}_{q_0}$. Nevertheless, the idea that these dimensions should be related is an important motivation for the conjecture.

We note that if $\mathbb{T}_\Sigma$ acts faithfully on $\widetilde{H}_{q_0}$, and $\widetilde{H}_{q_0}$ is ϖ-torsion free, then $\mathbb{T}_\Sigma$ is ϖ-torsion free. This motivates the conjecture of ϖ-torsion freeness in Conjecture 3.1.

Conjecture 3.2 has essentially been proved by P. Scholze in many cases for which the group $\mathbb{G}$ gives rise to Shimura varieties [53, Cor. IV.2.3]. (He has non-strict inequalities rather than strict inequalities for the codimension of the cohomology in degrees $< q_0$.)

We mention one more example here, namely the case when $\mathbb{G}$ is the restriction of scalars of SL_2 from an imaginary quadratic field. In this case we have $l_0 = 1$, and the codimension statement of Conjecture 3.2 follows from the computation sketched in (2.2.2).

3.1.2. Localization at a non-Eisenstein system of Hecke eigenvalues. Suppose that $\mathfrak{m}$ is a maximal ideal in $\mathbb{T}_\Sigma$, and that the associated representation $\overline{\rho}_\mathfrak{m}$ (which we assume exists) is "irreducible", i.e. does not factor through any proper parabolic subgroup of the C-group of $\mathbb{G}$. We refer to such a $\mathfrak{m}$ as *non-Eisenstein*. For any $\mathbb{T}_\Sigma$-module M, we write $M_\mathfrak{m} := \mathbb{T}_\mathfrak{m} \otimes_{\mathbb{T}_\Sigma} M$ to denote the localization of M at $\mathfrak{m}$.

As already noted, we expect that all the systems of Hecke eigenvalues appearing in cohomological degrees $< i$ are "reducible", i.e. do factor through a proper parabolic subgroup ,

and so we conjecture that, when $\mathfrak{m}$ is a non-Eisenstein maximal ideal, $H^i\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} = 0$ for all $i < q_0$ and all local systems $\mathcal{W}$ associated to K_p-representations W on finitely generated $\mathcal{O}$-modules (where K_p is any sufficiently small compact open subgroup of G).

We remark that if we ignore torsion, then this follows (at least morally) from Arthur's conjectures [1]. (Namely, the result is true for the boundary cohomology, and hence it suffices to check it for the interior cohomology, i.e. the image of compactly supported cohomology in the usual cohomology. However, the interior cohomology is contained in the L^2-cohomology, and now Arthur's conjectures imply that any automorphic representation π contributing to L^2-cohomology that is non-tempered at ∞ — as must be the case for an automorphic representation that contributes to cohomology in degree $< q_0$ — is non-tempered at every prime, and hence should give rise to a reducible Galois representation.) We believe that this statement should be true for torsion cohomology as well.

Let us suppose, then, that $H^i\big(Y(K_pK^p),\mathcal{W}\big)_{\mathfrak{m}} = 0$ for all $i < q_0$. Just considering the case when $W = \mathcal{O}$ (the trivial representation), we find that $\widetilde{H}^i_{\mathfrak{m}} = 0$ for $i < q_0$. Certainly it should be the case that $H^i\big(\partial(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} = 0$ when $\mathfrak{m}$ is non-Eisenstein. We then find that $H^i_c\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} = 0$ for all $i < q_0$ as well. *Suppose now that* $l_0 = 0$, so that $q_0 = d/2$. Classical Poincaré duality then gives that $H^i\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} = 0$ for all $i > q_0$. Presuming that Conjecture 3.2 is true, so that $\widetilde{H}^i = 0$ for $i > q_0$ and $\widetilde{H}^{q_0}$ is ϖ-torsion free, a consideration of the Hochschild–Serre spectral sequence shows that $\mathrm{Ext}^i_{\mathcal{O}[[K_p]]}(W^\vee, \widetilde{H}^{q_0}_{\mathfrak{m}}) = 0$ for all $i > 0$ and all representations of K_p on finitely generated torsion free $\mathcal{O}$-modules. From this one easily deduces that $(\widetilde{H}_{q_0})_{\mathfrak{m}}$ is projective as an $\mathcal{O}[[K_p]]$-module.

In short, we have given a plausibility argument for the following conjecture, which refines Conjecture 3.2 in the context of localizing at a non-Eisenstein maximal ideal.

Conjecture 3.3. *If* $\mathfrak{m}$ *is a* non-Eisenstein *maximal ideal in* $\mathbb{T}_\Sigma$*, and if* $l_0 = 0$*, then* $(\widetilde{H}_i)_{\mathfrak{m}} = 0$ *for* $i \neq q_0$*, and* $(\widetilde{H}_{q_0})_{\mathfrak{m}}$ *is a projective* $\mathcal{O}[[K_p]]$*-module, for any sufficiently small subgroup* K_p *of* G*.* (*Here* sufficiently small *means that* $\mathbb{G}(\mathbb{Q})$ *acts with trivial stabilizers on* $\mathbb{G}(\mathbb{A})/A^\circ_\infty K^\circ_\infty K_pK_f^p$.)

If $l_0 = 0$, and if Conjecture 3.3 holds for some non-Eisenstein maximal ideal $\mathfrak{m}$, then we see (from the Hochschild–Serre spectral sequence) that for any sufficiently small K_p (as in the statement of the conjecture) and any representation of K_p on a finitely generated torsion free $\mathcal{O}$-module W, we have $H^i\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} = 0$ if $i \neq q_0$, while $H^{q_0}\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} \cong \mathrm{Hom}_{\mathcal{O}[[K_p]]}(W^\vee, \widetilde{H}^{q_0}_{\mathfrak{m}})$. In particular, we see that $H^{q_0}\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}}$ is ϖ-torsion free. We also see that $\widehat{H}^{q_0}_{\mathfrak{m}} = \widetilde{H}^{q_0}_{\mathfrak{m}}$ (since $H^{q_0+1}_{\mathfrak{m}}$ vanishes, and hence so does $T_pH^{q_0+1}_{\mathfrak{m}}$), and that $\mathbb{T}_{\mathfrak{m}}$ acts faithfully on $\widetilde{H}^{q_0}_{\mathfrak{m}}$.

Presuming that $H^i\big(\partial(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} = 0$ (which should certainly be true when $\mathfrak{m}$ is non-Eisenstein), we conclude that the natural map

$$H^{q_0}_c\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}} \to H^{q_0}\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}}$$

is an isomorphism, and thus that $H^{q_0}\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}}$ consists entirely of interior cohomology. In particular (being torsion free) it embeds into the L^2-cohomology. From this we deduce that the image of $\mathbb{T}_\Sigma$ acting on $H^{q_0}\big(Y(K_pK_f^p),\mathcal{W}\big)_{\mathfrak{m}}$ is reduced, and hence (considering all possible K_p and W) that $\mathbb{T}_{\mathfrak{m}}$ itself is reduced, as well as being ϖ-torsion free. This provides some evidence for the reducedness and torsion-freeness statements in Conjecture 3.1.

We now recall some known results in the direction of Conjecture 3.3. In the case when $\mathbb{G} = \mathrm{GL}_2$, the concept of a non-Eisenstein maximal ideal is well-defined, and the above conjecture holds [32, Cor. 5.3.19]. In the case when $\mathbb{G} = \mathrm{GL}_N$, this notion is again well-defined, and the vanishing of $(\widetilde{H}_i)_{\mathfrak{m}}$ is proved for i in the stable range in [25]. Again, if $\mathbb{G}$ is a form of $U(n-1,1)$ over $\mathbb{Q}$, then the notion of a non-Eisenstein maximal ideal is well-defined, and in [34] we prove vanishing of $(\widetilde{H}_i)_{\mathfrak{m}}$ for i in a range of low degrees, for certain maximal ideals $\mathfrak{m}$. In particular, in the case of $U(2,1)$, we are able to deduce Conjecture 3.3, provided not only that $\mathfrak{m}$ is non-Eisenstein, but that the associated Galois representation $\overline{\rho}_{\mathfrak{m}}$ has sufficiently large image, and is irreducible locally at p, satisfying a certain regularity condition.

As one more example, note that if G_∞/A_∞ is compact (in which case certainly $l_0 = 0$), then Conjecture 3.3 holds. Indeed, in this case we saw in (2.2.4), for sufficiently small K_p, that $\widetilde{H}_0$ is free over $\mathcal{O}[[K_p]]$, even without localizing at a non-Eisenstein maximal ideal.

If $l_0 \neq 0$, then we don't expect that $(\widetilde{H}_{q_0})_{\mathfrak{m}}$ should be projective over $\mathcal{O}[[K_p]]$; indeed, this would be incompatible with Conjecture 3.2. Rather, we expect that it should be pure of codimension l_0, in the sense of [55, Def. 3.1].

3.1.3. The relationship with p-adic Hodge theory. Let us continue to suppose that $l_0 = 0$ and that $\mathfrak{m}$ is a non-Eisenstein maximal ideal in $\mathbb{T}_\Sigma$, and let us suppose that Conjecture 3.3 holds. As we have observed, this implies that for any (sufficiently small) compact open subgroup K_p of G, and any representation W of K_p on a finitely generated torsion free $\mathcal{O}$-module, we have

$$H^i\big(Y(K_pK_f^p), \mathcal{W}\big)_{\mathfrak{m}} \cong \mathrm{Hom}_{\mathcal{O}[[K_p]]}(W^\vee, \widetilde{H}^{q_0}_{\mathfrak{m}}) \cong \mathrm{Hom}_{\mathcal{O}[[K_p]]}\big((\widetilde{H}_{q_0})_{\mathfrak{m}}, W\big)$$

(the first isomorphism being given by the Hochschild–Serre spectral sequence, and the second by duality).

We now suppose that W is an $\mathcal{O}$-lattice in an irreducible algebraic representation V of $\mathbb{G}$ over L. Since $H^i\big(Y(K_pK_f^p), \mathcal{W}\big)_{\mathfrak{m}}$ is ϖ-torsion free, it is a lattice in $H^i\big(Y(K_pK_f^p), \mathcal{V}\big)_{\mathfrak{m}}$. Considering the description of this cohomology in terms of automorphic forms (as in (2.1.6)) and the conjectures regarding Galois representations associated to automorphic forms [21, Conj. 5.3.4], we infer that the Galois representations associated to the systems of Hecke eigenvalues appearing in $H^i\big(Y(K_pK_f^p), \mathcal{V}\big)_{\mathfrak{m}}$ should be potentially semistable locally at p, with Hodge–Tate weights related to the highest weight of the algebraic representation V.

Since $\mathfrak{m}$ is non-Eisenstein, it should make sense to speak of the formal deformation ring $R_{\overline{\rho}_{\mathfrak{m}}}$ of $\overline{\rho}_{\mathfrak{m}}$, and we conjecture that there is a natural isomorphism $\mathbb{T}_{\mathfrak{m}} \cong R_{\overline{\rho}_{\mathfrak{m}}}$. We see that the $\mathbb{T}_{\mathfrak{m}}$-modules $H^i\big(Y(K_pK_f^p), \mathcal{W}\big)_{\mathfrak{m}}$ (where W ranges over the various lattices in the various algebraic representations V) should then be supported on the set of points of $\mathrm{Spec}\, R_{\overline{\rho}_{\mathfrak{m}}}$ corresponding to deformations of $\overline{\rho}_{\mathfrak{m}}$ that are potentially semistable at p.

The Fontaine–Mazur conjecture [38], when combined with Langlands reciprocity, should give a precise description of the support of $H^i\big(Y(K_pK_f^p), \mathcal{W}\big)_{\mathfrak{m}}$, in purely Galois-theoretic terms. Namely, any potentially semistable deformation of $\overline{\rho}_{\mathfrak{m}}$ of the appropriate Hodge–Tate weights should be motivic, and hence should be associated to an automorphic form on the quasi-split inner form of $\mathbb{G}$. Whether this automorphic form can then be transferred to $\mathbb{G}$, and can contribute to cohomology at level $K_pK_f^p$, should be answered by an analysis of the local Langlands correspondence for $\mathbb{G}$.

Conversely, if we can prove directly that the support of $H^i\big(Y(K_pK_f^p), \mathcal{W}\big)_{\mathfrak{m}}$ is as predicted by the Fontaine–Mazur–Langlands conjecture, then we can deduce this conjecture

for deformations of $\overline{\rho}_{\mathfrak{m}}$ that are classified by $\operatorname{Spec}\mathbb{T}_{\mathfrak{m}}$. We recall in (3.2.2) below how this strategy is compatible with an optimistic view-point on how p-adic local Langlands might behave. We briefly recall in (3.3.1) how this strategy was employed in [32] and [46] in the case when $\mathbb{G} = \mathrm{GL}_2$.

Let us fix a sufficiently small open subgroup K_p. Since $(\widetilde{H}_{q_0})_{\mathfrak{m}}$ is projective over $\mathcal{O}[[K_p]]$ by assumption, it is a direct summand of a finitely generated free $\mathcal{O}[[K_p]]$-module, and hence $\widetilde{H}^{q_0}_{\mathfrak{m}}$ is a direct summand of $\mathcal{C}(K_p,\mathcal{O})^{\oplus n}$, for some $n > 0$. Thus $L \otimes_{\mathcal{O}} \widetilde{H}^{q_0}_{\mathfrak{m}}$ is a direct summand of $\mathcal{C}(K_p, L)^{\oplus n}$, the space of continuous L-valued functions on K_p. Now the theory of Mahler expansions shows that the affine ring $L[\mathbb{G}]$ embeds with dense image in $\mathcal{C}(K_p, L)$ [51, Lemma A.1]. Recall that, as a $\mathbb{G}$-representation, we have $L[\mathbb{G}] \cong \bigoplus_V V \otimes_L \operatorname{Hom}_{\mathbb{G}}(V, L[\mathbb{G}])$, where V runs over (a set of isomorphism class representatives of) all irreducible representations of $\mathbb{G}$. (At this point, we assume for simplicity that L is chosen so that all the irreducible representations of $\mathbb{G}$ over $\overline{\mathbb{Q}}_p$ are in fact defined over L; thus these irreducible V are in fact absolutely irreducible.) Now the inclusion of $L[\mathbb{G}]$ into $\mathcal{C}(K_p, L)$ induces an isomorphism $\operatorname{Hom}_{\mathbb{G}}(V, L[\mathbb{G}]) \cong \operatorname{Hom}_{K_p}\big(V, \mathcal{C}(K_p, L)\big)$ (both are naturally identified with $V^\vee$, the contragredient of V). We conclude that the natural morphism $\bigoplus_V V \otimes_L \operatorname{Hom}_{K_p}\big(V, \mathcal{C}(K_p, L)\big) \to \mathcal{C}(K_p, L)$ is injective with dense image. This property is clearly preserved under passing to direct sums and direct summands, and hence the natural morphism $\bigoplus_V V \otimes_L \operatorname{Hom}_{K_p}(V, \widetilde{H}^{q_0}_{\mathfrak{m}} \otimes_{\mathcal{O}} L) \to \widetilde{H}^{q_0}_{\mathfrak{m}} \otimes_{\mathcal{O}} L$ is also injective with dense image. Replacing V by $V^\vee$ (which clearly changes nothing, since we are summing over all irreducible algebraic representations of $\mathbb{G}$), and recalling that (by Hochschild–Serre) $\operatorname{Hom}_{K_p}(V^\vee, \widetilde{H}^{q_0}_{\mathfrak{m}} \otimes_{\mathcal{O}} L) \cong H^{q_0}\big(Y(K_pK_f^p), \mathcal{V}\big)_{\mathfrak{m}}$, we find that $\mathbb{T}_{\mathfrak{m}}$ acts faithfully on $\bigoplus_V H^{q_0}\big(Y(K_pK_f^p), \mathcal{V}\big)_{\mathfrak{m}}$ (since, as we noted above, it follows from Conjecture 3.3 that $\mathbb{T}_{\mathfrak{m}}$ acts faithfully on $\widetilde{H}^{q_0}_{\mathfrak{m}}$).

Assuming that we can identify $\mathbb{T}_{\mathfrak{m}}$ with a deformation space of Galois representations, and that we know that $H^{q_0}\big(Y(K_pK_f^p), \mathcal{V}\big)_{\mathfrak{m}}$ is indeed supported on an appropriate locus of potentially semistable deformations, the preceding analysis shows that these loci (as V varies) are Zariski dense in $\operatorname{Spec}\mathbb{T}_{\mathfrak{m}}$.

In forthcoming work [37], the author and V. Paškūnas will apply a more sophisticated version of this argument to deduce additional Zariski density statements for various collections of potentially semistable loci in global Galois deformation spaces.

Note that these density results rely crucially on the projectivity statement of Conjecture 3.3 for their proof. As we already noted, we can't expect such a statement to be true when $l_0 > 0$, and (at least if $\mathbb{G}$ is semisimple, or, more generally, if the semisimple part of $\mathbb{G}$ satisfies $l_0 > 0$) we don't expect the potentially semistable points to be Zariski dense in global deformation spaces in this case. (See [26] for an elaboration on this point.) In particular, in such contexts, we don't expect that $\widehat{H}^{q_0}$ will equal $\widetilde{H}^{q_0}$, and so $\widetilde{H}^{q_0}$, and $\operatorname{Spec}\mathbb{T}_{\Sigma}$, should receive a non-trivial contribution from torsion classes (via the term $T_pH^{q_0+1}$ in the exact sequence (2.6)).

3.2. p-adic local Langlands. Until now, our discussion has been entirely global in nature. We now turn to describing how these global considerations might be related to a possible p-adic local Langlands correspondence.

3.2.1. The basic idea. Let us return for a moment to the direct sum decomposition (2.10). Local-global compatibility for classical Langlands reciprocity says that if $\pi_\infty \otimes \pi_p$ is a direct summand of $\mathcal{A}_{\mathrm{cusp}}(K_f^p)_{\chi^{-1}}$ which contributes to cohomology, lying in a Hecke eigenspace

that corresponds to some p-adic Galois representation ρ, then the local factor π_p and the Weil–Deligne representation attached to ρ (which should be defined, since ρ should be potentially semistable at p) should correspond via the local Langlands correspondence (or perhaps more generally via the local form of Arthur's conjectures [1], if π_∞ and π_p are not tempered). In particular, the local factor at p and the local Galois representation at p should be related in a purely local manner.

The basic idea of a p-adic local Langlands correspondence is that the same should be true when we take into account the structure of completed cohomology. To explain this, we continue the discussion of the preceding paragraph, and, in addition, we place ourselves in the context introduced in (3.1.3) (in particular, we continue to assume that the hypotheses and conclusions of Conjecture 3.3 hold). Thus, we assume that ρ is a deformation of $\overline{\rho}_{\mathfrak{m}}$, and that π_∞ and π_p are tempered, so that we have an embedding $\pi_p \hookrightarrow H^{q_0}(\mathcal{V})_{\mathfrak{m}} \cong \varinjlim_{K_p} \mathrm{Hom}_{K_p}(V^\vee, L \otimes_{\mathcal{O}} \widetilde{H}^{q_0}_{\mathfrak{m}})$. We may rewrite this as an embedding $V^\vee \otimes_L \pi_p \hookrightarrow L \otimes_{\mathcal{O}} \widetilde{H}^{q_0}_{\mathfrak{m}}$ and we can then take the closure of $V^\vee \otimes \pi_p$ in $H^{q_0}(\mathcal{V})_{\mathfrak{m}}$, to obtain a unitary Banach space representation $\widehat{V^\vee \otimes_L \pi_p}$ of G over L. A slightly more refined procedure is to form the intersection $(V^\vee \otimes_L \pi_p)^\circ := (V^\vee \otimes_L \pi_p) \cap \widetilde{H}^{q_0}_{\mathfrak{m}}$ (the intersection being taken in $L \otimes_{\mathcal{O}} \widetilde{H}^{q_0}_{\mathfrak{m}}$). This is a G-invariant $\mathcal{O}$-lattice contained in $V^\vee \otimes_L \pi_p$, and the Banach space $\widehat{V^\vee \otimes_L \pi_p}$ is then obtained by completing $V^\vee \otimes_L \pi_p$ with respect to this lattice. A natural question to ask, then, in the spirit of a local Langlands correspondence and local-global compatibility, is whether $\widehat{V^\vee \otimes_L \pi_p}$ depends only on the restriction to p of the associated Galois representation ρ. Since by assumption $\mathfrak{m}$ is non-Eisenstein, the Galois representation ρ should admit an essentially unique integral model ρ°, and we could further ask whether $(V^\vee \otimes_L \pi_p)^\circ$ depends only on the restriction to p of ρ°. This question goes back to C. Breuil's first work on the p-adic Langlands correspondence in the case when $\mathbb{G} = \mathrm{GL}_2$ (see especially the introduction of [13]). It has been largely resolved in that case, but remains open in general.

3.2.2. An optimistic scenario. The most optimistic conjecture that one might entertain regarding a p-adic local Langlands correspondence is that for any formal deformation ring $R_{\overline{r}}$ parameterizing the deformations of a representation $\overline{r}$ of the *local* Galois group $G_{\mathbb{Q}_p}$ into the $\overline{\mathbb{F}}$-valued points of the C-group of G, there is a profinite $R_{\overline{r}}$-module M, finitely generated over $R_{\overline{r}}[[K_p]]$ for some (and hence every) compact open subgroup $K_p \subset G$, equipped with a continuous G-action extending its K_p-module structure, which realizes the p-adic local Langlands correspondence for the group $\mathbb{G}$ in the following sense: in the context of (3.2.1) (and continuing with the notation of that discussion), the fibre over the restriction to p of ρ° (which is a point in $\mathrm{Spec}\, R_{\overline{r}}$, if $\overline{r}$ is the restriction of $\overline{\rho}_{\mathfrak{m}}$ to $G_{\mathbb{Q}_p}$) is isomorphic to the $\mathcal{O}$-dual of $(V^\vee \otimes_L \pi_p)^\circ$. (Here we suppress the issue of whether $R_{\overline{r}}$ should be understood to be a framed deformation ring, or a pseudo-deformation ring, or ....) In short, M should be a local analogue of the completed homology space $(\widetilde{H}_{q_0})_{\mathfrak{m}}$.

Since M is finitely generated over $R_{\overline{r}}[[K_p]]$, for any representation W of K_p on a finitely generated torsion free $\mathcal{O}$-module, the $R_{\overline{r}}$-module $\mathrm{Hom}^{\mathrm{cont}}_{K_p}(M, W)^d$ (where d denotes the continuous $\mathcal{O}$-dual) is a finitely generated $R_{\overline{r}}$-module. Another property one might require of M is that when W is a K_p-invariant $\mathcal{O}$-lattice in an irreducible algebraic $\mathbb{G}$-representation V, then $\mathrm{Hom}^{\mathrm{cont}}_{K_p}(M, W)^d$ is supported on an appropriate locus of potentially semistable

representations, corresponding to the fact that $\mathrm{Hom}_{\mathcal{O}[[K_p]]}\big((\widetilde{H}_{q_0})_{\mathfrak{m}}, W\big)$ (which is isomorphic to $H^{q_0}\big(Y(K_pK_f^p), \mathcal{W}\big)$) is supported on a locus of potentially semistable representations in $\mathrm{Spec}\,\mathbb{T}_{\mathfrak{m}}$.

Ideally, one might ask for the support of $\mathrm{Hom}_{K_p}^{\mathrm{cont}}(M, W)^d$ to be the full potentially semistable locus of appropriate Hodge–Tate weights (corresponding to the highest weight of V) and Weil–Deligne representations (corresponding to the particular choice of K_p and the nature of the local Langlands correspondence for $\mathbb{G}$). As we indicated in (3.1.3), such a result, combined with the local-global compatibility between M and $(\widetilde{H}_{q_0})_{\mathfrak{m}}$, would prove that any global Galois representation corresponding to a point of $\mathrm{Spec}\,\mathbb{T}_{\mathfrak{m}}$ which is potentially semistable of appropriate Hodge–Tate weights and with an appropriate Weil–Deligne representation, in fact arises from a system of Hecke eigenvalues occurring in $\widetilde{H}^{q_0}\big(Y(K_pK_f^p), \mathcal{V}\big)_{\mathfrak{m}}$, thus verifying the Fontaine–Mazur–Langlands conjecture for such points.

Note that a local Galois representation lies in the support of $\mathrm{Hom}_{K_p}^{\mathrm{cont}}(M, W)^d$ precisely if the dual to the fibre of M at this point receives a non-zero K_p-equivariant homomorphism from W. In particular, the dual to the fibre at such a point contains locally algebraic vectors. In the case of GL_2, the idea of describing p-adic local Langlands for potentially semistable representations in terms of locally algebraic vectors goes back to C. Breuil's first work on the subject [11, 12].

The optimistic scenario described here has been realized for the group $G = \mathrm{GL}_2(\mathbb{Q}_p)$; this is the theory of p-adic local Langlands for $\mathrm{GL}_2(\mathbb{Q}_p)$ [6, 29, 30, 50] (see also [5, 14]). Whether it can be realized for other groups, or is overly optimistic, remains to be seen.

3.3. Examples. We again illustrate our discussion with some examples.

3.3.1. $\mathbf{GL_2}$ of $\mathbb{Q}$. As already mentioned, the theory of p-adic local Langlands for $\mathrm{GL}_2(\mathbb{Q}_p)$ provides a structure satisfying all the desiderata of (3.2.2). The local-global compatibility between this structure and $\widetilde{H}^1_{\mathfrak{m}}$ (the localization of completed cohomology at a non-Eisenstein maximal ideal of the Hecke algebra) has been proved in [32] (see also [14, 15]), under a mild hypothesis on the local behaviour of $\overline{\rho}_{\mathfrak{m}}$. In particular, the strategy of (3.1.3) then applies to prove the Fontaine–Mazur–Langlands conjecture for points of $\mathrm{Spec}\,\mathbb{T}_{\mathfrak{m}}$.

Under the slightly stronger hypotheses that p is odd and $\overline{\rho}_{\mathfrak{m}}$ remains irreducible on restriction to $G_{\mathbb{Q}(\zeta_p)}$ (the *Taylor–Wiles condition*) one can prove that $\mathbb{T}_{\mathfrak{m}} \cong R_{\overline{\rho}_{\mathfrak{m}}}$ [8], [32, Thm. 1.2.3]. Since Conjecture 3.3 holds in this context [32, Cor. 5.3.19], the method of (3.1.3) (applied with $K_p = \mathrm{GL}_2(\mathbb{Z}_p)$) allows us to deduce the density of the crystalline loci in $\mathrm{Spec}\,R_{\overline{\rho}_{\mathfrak{m}}}$; the extension of this method to be described in [37] will allow us to deduce other density results about various potentially semistable loci. Here is one such simple variant: instead of considering the restriction to $\mathrm{GL}_2(\mathbb{Z}_p)$ of the family V of algebraic representations of GL_2, we instead consider the family of representations $\sigma \otimes V$, where σ is some fixed supercuspidal type and V is algebraic. This allows us to show that potentially crystalline points of (any fixed) supercuspidal type are Zariski dense in $\mathbb{T}_{\mathfrak{m}}$ (and hence in $R_{\overline{\rho}_{\mathfrak{m}}}$, if the Taylor–Wiles condition is satisfied).

3.3.2. A definite quaternion algebra ramified at p. Let D be the quaternion algebra over $\mathbb{Q}$ that is ramified at ∞ and p, and let $\mathbb{G} = D^\times$. Then G_∞/A_∞ is compact, and so, as noted above, Conjecture 3.3 holds. The classical Jacquet–Langlands correspondence allows us to attach odd two-dimensional Galois representations to systems of Hecke eigenvalues. In particular, if $\mathfrak{m}$ is non-Eisenstein, then $\mathbb{T}_{\mathfrak{m}}$ is a quotient of $R_{\overline{\rho}_{\mathfrak{m}}}$.

Assuming that p is odd and $\overline{\rho}_{\mathfrak{m}}$ satisfies the Taylor–Wiles condition, we conclude that in fact $\mathbb{T}_{\mathfrak{m}} \cong R_{\overline{\rho}_{\mathfrak{m}}}$. Indeed, we saw in the preceding example that the potentially crystalline points of supercuspidal type are Zariski dense in $R_{\overline{\rho}_{\mathfrak{m}}}$. Since Fontaine–Mazur–Langlands holds in this context, they all arise from classical modular forms, and hence (by Jacquet–Langlands) from classical automorphic forms on $D^\times$. Thus $\operatorname{Spec} \mathbb{T}_{\mathfrak{m}}$ contains a Zariski dense set of points in $\operatorname{Spec} R_{\overline{\rho}_{\mathfrak{m}}}$, and so the two Specs coincide. As one interesting consequence of this, we note that (assuming that the Taylor–Wiles condition holds) the Hecke algebras at $\mathfrak{m}$ for GL_2 and $D^\times$ are naturally isomorphic (both being isomorphic to $R_{\overline{\rho}_{\mathfrak{m}}}$). One can think of this as a p-adic interpolation of the classical Jacquet–Langlands correspondence. It is interesting to note that even though under the classical Jacquet–Langlands correspondence automorphic forms on GL_2 that are principal series at p don't match with a corresponding Hecke eigenform on $D^\times$, they are not excluded from this p-adic Jacquet–Langlands correspondence.

As another consequence, note that if we choose K_p to be the units in the maximal order of $(D \otimes \mathbb{Q}_p)^\times$, and apply the density argument of (3.1.3), we deduce that the representations which are *genuinely semistable* at p (i.e. semistable, and not crystalline) are Zariski dense in $\operatorname{Spec} \mathbb{T}_{\mathfrak{m}}$, and hence in $\operatorname{Spec} R_{\overline{\rho}_{\mathfrak{m}}}$, provided the Taylor–Wiles condition holds.

3.3.3. Definite quaternion algebras over totally real fields. Consider the case where G is the restriction of scalars to $\mathbb{Q}$ of the units $D^\times$ in a totally definite quaternion algebra D over a totally real field F; and suppose that F is unramified at p, and D is split at every prime in F above p. Put ourselves in the situation of (3.2.1), with the additional assumption that V is the trivial representation, and π_p is a tamely ramified principal series. In this case π_p contains, as a $\mathrm{GL}_2(\mathbb{Z}_p \otimes_{\mathbb{Z}} \mathcal{O}_F)$-subrepresentation, a principal series type σ, which is a representation of $\mathrm{GL}_2(\mathcal{O}_F/p)$ induced from a character of the Borel subgroup. In the paper [16], Breuil gave a conjectural description of the isomorphism class of the lattice $\sigma^\circ := \sigma \cap \widetilde{H}^0_{\mathfrak{m}}$, purely in terms of the restriction to p of ρ°. Under mild assumptions on $\overline{\rho}_{\mathfrak{m}}$, this was proved in [35]. This gives some evidence towards the possibility that $(V^\vee \otimes_L \pi_p)^\circ$ may be of a purely local nature.

3.3.4. Compact unitary groups. In [27], we apply Taylor–Wiles patching to pass from completed cohomology over a unitary group that is compact at infinity to a G-representation on a module M_∞ over the ring R_∞ obtained by adjoining a certain number of formal variables to a local deformation ring. More precisely: for any finite extension K of $\mathbb{Q}_p$, any n such that $p \nmid 2n$, and any representation $\overline{r} : G_K \to \mathrm{GL}_n(\overline{\mathbb{F}}_p)$ that admits a potentially crystalline lift which is *potentially diagonalizable* (in the sense of [4]), we can (by [33, Cor. A.7]) choose a unitary group $\mathbb{G}$ and Hecke maximal ideal $\mathfrak{m}$ so that $G := \mathbb{G}(\mathbb{Q}_p)$ is isomorphic to a product of copies of $\mathrm{GL}_n(K)$, and such that the restriction of $\overline{\rho}_{\mathfrak{m}}$ to G_K is equal to $\overline{r}$. The G-representation M_∞, which is a module over $R_{\overline{r}}[[x_1, \dots, x_n]]$ for some some $n \geq 0$, is then obtained by patching the completed cohomology for $\mathbb{G}$.

The module M_∞ satisfies several of the desiderata of (3.2.2): it is finitely generated over $R_\infty[[K_p]]$, and the modules $\operatorname{Hom}^{\mathrm{cont}}_{K_p}(M, W)^d$ (for K_p-invariant lattices W in algebraic representations) are supported on a union of components of the appropriate potentially semistable loci. Many questions about M_∞ remain open, however: whether the support of M_∞ equals the entirety of $\operatorname{Spec} R_\infty$; whether all potentially semistable points are contained in the support of $\operatorname{Hom}^{\mathrm{cont}}_{K_p}(M, W)^d$ (for an appropriate choice of W and K_p); and whether M_∞ is in fact of a purely local nature.

One further property of M_∞ is that it is projective in the category of profinite topo-

logical $\mathcal{O}[[K_p]]$-modules. In [37], we hope to show that M_∞ is in fact of full support on $\operatorname{Spec} R_{\overline{r}}[[x_1, \dots, x_n]]$ in many cases, and thus extend the method described in (3.1.3) to deduce density results for potentially semistable representations in local deformation spaces.

Acknowledgements. The author's work is partially supported by NSF grant DMS-1303450. The ideas expressed here owe much to the many discussions with my colleagues and collaborators that I've enjoyed over the years, and I would like to thank them, especially Ana Caraiani, Pierre Colmez, David Geraghty, Michael Harris, David Helm, Florian Herzig, Mark Kisin, David Savitt, and Sug Woo Shin. I am particularly indebted to Christophe Breuil, whose ideas regarding the p-adic Langlands program have been so influential and inspiring. Special thanks are owed to Frank Calegari, Toby Gee, and Vytas Paškūnas; many of the ideas described here were worked out in collaboration with them.

I am also grateful to Frank Calegari, Tianqi Fan, Toby Gee, Michael Harris, Florian Herzig, Daniel Le, David Savitt, and the Imperial College Study Group (Rebecca Bellovin, Kevin Buzzard, Toby Gee, David Helm, Judith Ludwig, James Newton, and Jack Shotton) for their careful reading of various preliminary versions of this note.

References

[1] Arthur, J., *Unipotent automorphic representations: conjectures*, in Orbites unipotentes et représentations II, Astérisque **171–172** (1989), 13–71.

[2] Ash, A., *Galois representations attached to mod p cohomology of* $\mathrm{GL}(n, \mathbb{Z})$, Duke Math. J. **65** (1992), 235–255.

[3] Ash, A. and Stevens, G., *Modular forms in characteristic ℓ and special values of their L-functions*, Duke Math. J. **53** (1986), 849–868.

[4] Barnet-Lamb, T., Gee, T., Geraghty, D., and Taylor, R., *Potential automorphy and change of weight*, Ann. Math. (to appear).

[5] Berger, L., *La correspondance de Langlands local p-adique pour* $\mathrm{GL}_2(\mathbf{Q}_p)$, Astérisque **339** (2011), 157–180.

[6] Berger, L. and Breuil, C., *Sur quelques représentations potentiellement cristallines de* $\mathrm{GL}_2(\mathbb{Q}_p)$, Astérisque **330** (2010), 265–281.

[7] Bergeron, N. and Venkatesh, A., *The asymptotic growth of torsion homology for arithmetic groups*, J. Inst. Math. Jussieu **12** (2013), 391–447.

[8] Böckle, G., *On the density of modular points in universal deformation spaces*, Amer. J. Math. **123** (2001), 985–1007.

[9] Borel, A., *Stable real cohomology of arithmetic groups*, Ann. Scient. Éc. Norm. Sup. (4) **7** (1974), 235–272.

[10] Borel, A. and Wallach, N., Continuous cohomology, discrete subgroups, and representations of reductive groups, Mathematical Surveys and Monographs 67, American Mathematical Society, Providence, RI, second edition, 2000.

[11] Breuil, C., *Sur quelques représentations modulaires et p-adiques de $\mathrm{GL}_2(\mathbb{Q}_2)$ II*, J. Inst. Math. Jussieu **2** (2003), 1–36.

[12] ———, *Invariant $\mathcal{L}$ et série spéciale p-adique*, Ann. Scient. Éc. Norm. Sup. (4) **37** (2004), 559–610.

[13] ———, *Série spéciale p-adique et cohomologie étale complété*, Astérisque **331** (2010), 65–115.

[14] ———, *The emerging p-adic Langlands programme*, Proceedings of the I.C.M. Vol. II, 203–230, Hindustan Book Agency, New Delhi, 2010.

[15] ———, *Correspondance de Langlands p-adique, compatibilité local-global et applications*, Astérisque **348** (2012), 119–147.

[16] ———, *Sur un problème de compatibilité local-global modulo p pour GL_2,* J. Reine Angew. Math. (to appear).

[17] ———, *Vers le socle localement analytique pour GL_n I*, preprint.

[18] ———, *Vers le socle localement analytique pour GL_n II*, preprint.

[19] Breuil, C. and Diamond, F., *Formes modulaires de Hilbert modulo p et valeurs d'extensions entre caractères galoisiens*, Ann. Scient. Éc. Norm. Sup. (to appear).

[20] Breuil, C. and Herzig, F., *Ordinary representation of $G(\mathbb{Q}_p)$ and fundamental algebraic representations*, preprint.

[21] Buzzard, K. and Gee, T., *The conjectural connections between automorphic representations and Galois representations*, Proceedings of the LMS Durham Symposium 2011 (to appear).

[22] Calegari, F., *The stable homology of congruence subgroups*, preprint.

[23] Calegari, F. and Emerton, M., *Bounds for multiplicities of unitary representations of cohomological type in spaces of cusp forms*, Ann. Math **170** (2009), 1437–1446.

[24] ———, *Completed cohomology — a survey*, Non-abelian fundamental groups and Iwasawa theory, 239–257, London Math. Soc. Lecture Note Ser., 393, Cambridge Univ. Press, Cambridge, 2012.

[25] ———, *Homological stability for completed cohomology*, preprint.

[26] Calegari, F. and Mazur, B., *Nearly ordinary Galois deformations over arbitrary number fields*, J. Inst. Math. Jussieu **8** (2009), 99–177.

[27] Caraiani, A., Emerton, M., and Gee, T., Geraghty, D., Paškūnas, V., Shin, S.-W., *Patching and the p-adic local Langlands correspondence*, preprint.

[28] Chenevier, G., *The p-adic analytic space of pseudocharacters of a profinite group, and pseudorepresentations over arbitrary rings*, Proceedings of the LMS Durham Symposium 2011 (to appear).

[29] Colmez, P., *Representations de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *et* (φ,Γ)*-modules*, Astérisque **330** (2010), 281–509.

[30] Colmez, P., Dospinescu, G., and Paškūnas, V., *The* p*-adic local Langlands correspondence for* $\mathrm{GL}_2(\mathbb{Q}_p)$, preprint.

[31] Emerton, M., *On the interpolation of systems of eigenvalues attached to automorphic Hecke eigenforms*, Invent. Math. **164** (2006), 1–84.

[32] ———, *Local-global compatibility in the* p*-adic Langlands program for* $\mathbb{G}L_{2/\mathbb{Q}}$, preprint.

[33] Emerton, M., Gee, T., *A* geometric perspective on the Breuil–Mézard conjecture, J. Inst. Math. Jussieu (to appear).

[34] ———, p*-adic Hodge theoretic properties of étale cohomology with mod* p *coefficients, and the cohomology of Shimura varieties*, preprint. http://arxiv.org/abs/1203.4963.

[35] Emerton, M., Gee, T., and Savitt, D., *Lattices in the cohomology of Shimura curves*, Invent. Math. (to appear).

[36] Emerton, M. and Helm, D., *The local Langlands correspondence for* GL_n *in families*, Ann. Scient. Éc. Norm. Sup. (to appear).

[37] Emerton, M., Paškūnas, V., in preparation.

[38] Fontaine, J.-M. and Mazur, B., *Geometric Galois representations*, Elliptic curves, modular forms, and Fermat's last theorem (J. Coates, S.T. Yau, eds.), 41–78, Int. Press, Cambridge, MA, 1995.

[39] Franke, J., *Harmonic analysis in weighted* L_2*-spaces*, Ann. Scient. Éc. Norm. Sup. (4) **31** (1998), 181–279.

[40] ———, A topological model for *some summand of the Eisenstein cohomology of congruence subgroups*, Eisenstein series and applications, 27–85, Progr. Math., 258, Birkhäuser Boston, Boston, MA, 2008.

[41] Gross, B., *On the Satake isomorphism*, in Galois representations in arithmetic algebraic geometry (Durham, 1996), 223–237, London Math. Soc. Lecture Note Ser., 254, Cambridge Univ. Press, Cambridge, 1998.

[42] ———, *Algebraic modular forms*, Israel J. Math. **113** (1999), 61–93.

[43] Harris, M., *Automorphic Galois representations and the cohomology of Shimura varieties*, this volume.

[44] Harris, M., Lan, K.-W., Taylor, R., and Thorne, J., *On the rigid cohomology of certain Shimura varieties*, preprint.

[45] Hida, H., *Galois representations into* $\mathrm{GL}_2(\mathbb{Z}_p[[X]])$ *attached to ordinary cusp forms*, Invent. Math. **85** (1986), 545–613.

[46] Kisin, M., *The Fontaine–Mazur conjecture for* GL_2, J. Amer. Math. Soc. **22** (2009), 641–690.

[47] Langlands, R.P., *On the notion of an automorphic representation*, Automorphic forms, representations and L-functions, 203–207, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979.

[48] Lazard, M., *Groupes analytic p-adiques*, Publ. Math. IHES **26** (1965).

[49] Matsushima, Y., *A formula for the Betti numbers of compact locally symmetric Riemannian manifolds*, J. Diff. Geom. **1** (1967), 99–109.

[50] Paškūnas, V., *The image of Colmez's Montreal functor*, Publ. Math. IHES **118** (2013), 1–191.

[51] ———, *Blocks for mod p representations of* $\mathrm{GL}_2(\mathbb{Q}_p)$, Proceedings of the LMS Durham Symposium 2011 (to appear).

[52] ———, *On the Breuil–Mézard conjecture*, to appear in Duke Math. J.

[53] Scholze, P., *On torsion in the cohomology of locally symmetric varieties*, preprint.

[54] Shimura, G., *Introduction to the arithmetic theory of automorphic functions*, Publications of the Mathematical Society of Japan, No. 11, Iwanami Shoten, Publishers, Tokyo; Princeton University Press, Princeton, N.J., 1971.

[55] Venjakob, O., *On the structure theory of the Iwasawa algebra of a p-adic Lie group*, J. Eur. Math. Soc. **4** (2002), 271–311.

Department of Mathematics, University of Chicago, 5734 S. University Ave., Chicago, IL 60637
E-mail: emerton@math.uchicago.edu

Theta correspondence: recent progress and applications

Wee Teck Gan

Abstract. We describe some recent progress in the theory of theta correspondence over both local and global fields. We also discuss applications of these recent developments to the local Langlands conjecture, the Gross-Prasad conjecture and the theory of automorphic forms for the metaplectic groups.

Mathematics Subject Classification (2010). Primary 11F67; Secondary 22E50.

Keywords. Theta correspondence, Siegel-Weil formula, local Langlands conjecture, Gross-Prasad conjecture, Shimura-Waldspurger correspondence.

1. History

In this paper, we shall report on some recent progress in the theory of theta correspondence, as well as some applications to number theory and representation theory. The use of theta correspondence has a long history, but its status as a theory was formally initiated by R. Howe in the influential paper [28] written in the 1970's but only published much later. This built upon the work of A. Weil [91] in the 1960's which provided a representation theoretic treatment of theta functions via his construction of the so-called Weil representations. In this introduction, we give a brief account of the historical development since the late 1970's; we apologise for omitting the contributions of many people.

As a theory, theta correspondence has its own share of internal problems which needed to be addressed, but from the onset, it was perceived mainly as a tool for constructing representations and automorphic forms. In particular, it gives natural constructions of certain instances of Langlands functorial lifting. In this vein, one of the first successes is Waldspurger's complete and elegant description [84, 85] of the Shimura correspondence between cuspidal representations of PGL_2 and the metaplectic double cover Mp_2. Another is Howe and Piatetski Shapiro's construction [29] of nontempered cuspidal representations on U_3 and Sp_4, contradicting the naive Ramanujan-Petersson conjecture. Such construction of nontempered cuspidal representations was later extended by J.S. Li [49–51] and C. Moeglin [61] to the general setting, resulting in the construction of many interesting examples of unitary representations and square-integrable automorphic forms.

The 1980's saw many key developments in the theory of theta correspondence. Firstly, motivated by Waldspurger's work, Rallis initiated a program [76–78] aimed at determining the cuspidality and nonvanishing of global theta liftings. This led to two important series of work. One is the work of Piatetski-Shapiro and Rallis [73] on the doubling zeta integral,

Proceedings of the International Congress of Mathematicians, Seoul, 2014

which is a Rankin-Selberg integral representing the standard L-function of classical groups. Another is the work of Kudla and Rallis [41, 42] on the Siegel-Weil formula, culminating in their paper [43]. Combining these two series of work resulted in the Rallis inner product formula in some instances. This characterises the nonvanishing of global theta liftings in terms of the analytic properties of the standard L-functions. In the course of their work, Kudla and Rallis were led to a local conjecture [44] about the nonvanishing of the local theta correspondence. They made significant progress towards this so-called conservation relation conjecture, proving it in many cases.

Secondly, starting with Kudla's paper [38], the local theta correspondence over p-adic fields was systematically investigated, culminating in Waldspurger's proof of the so-called Howe duality conjecture when $p \neq 2$ [86]. This Howe duality conjecture was shown by Howe himself [28] in the archimedean case, and for unramified groups in the nonarchimedean case [63]. Following this, significant understanding of the archimedean theta correspondence was obtained in the work of Adams-Barbasch [2–4] and Moeglin [60]. In particular, Adams-Barbasch determined the local theta correspondence over $\mathbb{C}$ completely [3], and also extended the local Shimura-Waldspurger correspondence from $\mathrm{Mp}_2(\mathbb{R})$ to $\mathrm{Mp}_{2n}(\mathbb{R})$ for general n [4]. In the p-adic case, the analogous results were conjectured but left open. In particular, Adams [1] formulated a conjecture on the functoriality of the theta correspondence in the language of A-packets, and D. Prasad [74, 75] formulated some precise conjectures describing the local theta correspondence in the (almost) equal rank case in terms of the local Langlands correspondence.

Since the mid-1990's, significant work continued to be done in classical theta correspondence, such as by Roberts [79], Moeglin [62], Muić [66–69], Muić-Savin [70], S.Y. Pan [71] and Ginzburg-Jiang-Soudry [21]. However, as many of the early pioneers turned their attention to other worthy endeavours, the field became relatively quiet compared with the flurry of activities in the 80's and early 90's and many of the problems highlighted some twenty years ago lie dormant and unresolved.

It was not until about 6 or 7 years ago that a new generation of researchers revisited these problems and it is a pleasure and privilege to report on the recent resolution of many of these problems here. This brings a certain degree of closure to the developments from 30 years ago, but we shall also highlight some exciting future directions.

2. Theta correspondence

In this section, we describe the basic setup and questions in the theory of theta correspondence.

2.1. Dual pairs. For simplicity, let F be a field of characteristic 0, and let $E = F$ or an étale quadratic F-algebra, with $\mathrm{Aut}(E/F) = \langle c \rangle$. With $\epsilon = \pm$, let V be a finite-dimensional ϵ-Hermitian space over E and W an $-\epsilon$-Hermitian space. Then $V \otimes_E W$ inherits a natural symplectic form over F and one has a natural map of isometry groups

$$\mathrm{U}(V) \times \mathrm{U}(W) \longrightarrow \mathrm{Sp}(V \otimes_E W).$$

The images of $\mathrm{U}(V)$ and $\mathrm{U}(W)$ are mutual commutants of each other, and such a pair of groups is called a *reductive dual pair*.

For ease of exposition, we shall henceforth focus on the case when E/F is a quadratic

field extension, V is Hermitian and W is skew-Hermitian, so that $\mathrm{U}(V)$ and $\mathrm{U}(W)$ are unitary groups.

2.2. Invariants of spaces. The spaces V and W have a natural invariant known as the discriminant:

$$\mathrm{disc}V \in F^\times/\mathrm{N}_{E/F}(E^\times) \quad \text{and} \quad \mathrm{discW} \in \delta^{\dim W} \cdot F^\times/\mathrm{N}_{E/F}(E^\times)$$

where δ is a fixed trace zero element in $E^\times$. When F is a local field, it is convenient to encode the discriminant in a sign $\pm$:

$$\epsilon(V) = \omega_{E/F}(\mathrm{disc}V) \quad \text{and} \quad \epsilon(W) = \omega_{E/F}(\delta^{-\dim W} \cdot \mathrm{disc}W)$$

where $\omega_{E/F}$ is the nontrivial quadratic character of $F^\times/N_{E/F}(E^\times)$. Note that $\epsilon(W)$ depends on the choice of δ. Moreover, if F is nonarchimedean, Hermitian spaces are classified by $\dim(V)$ and $\epsilon(V)$; likewise for skew-Hermitian spaces.

2.3. Weil representation. Assume that F is a local field. The symplectic group $\mathrm{Sp}(V \otimes_E W)$ has a nonlinear S^1-cover $\mathrm{Mp}(V \otimes_E W)$ known as the metaplectic group. This metaplectic group has a distinguished representation ω_ψ depending on a nontrivial additive character ψ of F. If the embedding i can be lifted to a homomorphism

$$\tilde{i} : \mathrm{U}(V) \times \mathrm{U}(W) \longrightarrow \mathrm{Mp}(V \otimes_E W),$$

then we obtain a representation $\omega_\psi \circ \tilde{i}$ of $\mathrm{U}(V) \times \mathrm{U}(W)$.

For the case of unitary groups considered here, a splitting can be specified [39] by picking two characters χ_V and χ_W of $E^\times$ such that

$$\chi_V|_{F^\times} = \omega_{E/F}^{\dim V} \quad \text{and} \quad \chi_W|_{F^\times} = \omega_{E/F}^{\dim W}.$$

Thus, $\mathrm{U}(V) \times \mathrm{U}(W)$ has a Weil representation $\omega_{V,W,\psi}$ that depends on ψ and the splitting data (χ_V, χ_W), which we shall suppress from the notation.

2.4. Local theta correspondonce. We will write $\mathrm{Irr}(\mathrm{U}(W))$ for the set of equivalence classes of irreducible smooth representations of $\mathrm{U}(W)$. For $\pi \in \mathrm{Irr}(\mathrm{U}(W))$, one considers the maximal π-isotypic quotient of $\omega_{V,W,\psi}$:

$$\omega_{V,W,\psi} \twoheadrightarrow \pi \boxtimes \Theta(\pi)$$

where $\Theta(\pi)$ is some smooth representation of $\mathrm{U}(V)$. We shall denote by $\theta(\pi)$ the maximal semisimple quotient of $\Theta(\pi)$. The goal of local theta correspondence is to determine the representations $\Theta(\pi)$ and $\theta(\pi)$ as much as possible. For example, the *Howe duality conjecture* states that if $\Theta(\pi)$ is nonzero, then it has a unique irreducible quotient, i.e. $\theta(\pi)$ is irreducible.

Note that the case when $E = F \times F$ is also necessary for global applications. In this case, the dual pair is $\mathrm{GL}_m \times \mathrm{GL}_n$. The study of this local theta correspondence is essentially completed in the paper [58] of A. Minguez. For example, the Howe duality conjecture was completely resolved. Hence we shall say no more about this case in this paper.

2.5. Theta functions. We turn now to the global setting. Thus, let k be a global field with ring of adeles $\mathbb{A}$, and let K/k be a quadratic field extension. Let V and W be a Hermitian and skew-Hermitian space over K, and fix a pair of Hecke characters χ_V and χ_W of

$\mathbb{A}_K^\times$ as before. For a nontrivial additive character $\psi = \otimes_v \psi_v$ of $k\backslash\mathbb{A}$, the adelic group $\mathrm{U}(V_\mathbb{A}) \times \mathrm{U}(W_\mathbb{A})$ possesses an abstract Weil representation $\omega_{V,W,\psi} = \otimes_v \omega_{V,W,\psi_v}$. It was shown by Weil that there is a natural equivariant map

$$\theta : \omega_{V,W,\psi} \longrightarrow \mathcal{A}(\mathrm{U}(V) \times \mathrm{U}(W)),$$

where the latter space denotes the space of automorphic forms on the dual pair. This map, called the "formation of theta functions", gives an automorphic realisation of $\omega_{V,W,\psi}$.

2.6. Global theta correspondence. One may use the functions $\theta(\phi)$ for $\phi \in \omega_{V,W,\psi}$ as kernel functions for the transfer of automorphic forms from $\mathrm{U}(W_\mathbb{A})$ to $\mathrm{U}(V_\mathbb{A})$. More precisely, if $f \in \mathcal{A}(\mathrm{U}(W))$, we set

$$\theta(\phi, f)(g) = \int_{\mathrm{U}(W_k)\backslash\mathrm{U}(W_\mathbb{A})} \theta(\phi)(gh) \cdot \overline{f(h)}\, dh,$$

where dh stands for the Tamagawa measure. This integral converges if f is a cusp form. Thus, if $\pi \subset \mathcal{A}(\mathrm{U}(W))$ is a cuspidal representation, then we obtain an equivariant map

$$\theta : \omega_{V,W,\psi} \otimes \overline{\pi} \longrightarrow \mathcal{A}(\mathrm{U}(V)).$$

The image of this map is denoted by $\Theta(\pi)$ and is called the global theta lift of π. Observe that one has, by definition, an equivariant map

$$\omega_{V,W,\psi} \twoheadrightarrow \pi \boxtimes \Theta(\pi).$$

The basic questions in global theta correspondence are whether the representation $\Theta(\pi)$ is cuspidal and whether it is nonzero. Note that if $\Theta(\pi)$ is nonzero and cuspidal, then it is semisimple, in which case $\Theta(\pi)$ is a quotient of the abstract representation $\otimes_v \Theta(\pi_v)$. Thus, if the Howe duality conjecture holds, then $\Theta(\pi) \cong \otimes_v \theta(\pi_v)$. In particular, the question of "what is $\Theta(\pi)$?" is essentially a local one.

3. Local developments

In this section, we discuss some recent developments concerning the local theta correspondence over a nonarchimedean local field F of residual characteristic $p > 0$. The following theorem, known since 1990, summarizes some basic results of Howe, Kudla, Mœglin-Vignéras-Waldspurger and Waldspurger (see [28, 38, 63, 86]).

Theorem 3.1.

(i) *For any $\pi \in \mathrm{Irr}(\mathrm{U}(W))$, the representation $\Theta(\pi)$ is either zero or of finite length.*

(ii) *If π is supercuspidal, then $\Theta(\pi)$ is either zero or irreducible (and thus is equal to $\theta(\pi)$). Moreover, for any irreducible supercuspidal π and π',*

$$\Theta(\pi) \cong \Theta(\pi') \neq 0 \Longrightarrow \pi \cong \pi'.$$

(iii) *If $p \neq 2$, the Howe duality conjecture holds.*

3.1. Local problems. The main remaining problems in local theta correspondence are thus:

(a) Establish the Howe duality conjecture when the residual characteristic of F is $p = 2$.

(b) Determine when $\Theta(\pi)$ is nonzero, in terms of some basic invariants of π.

(c) When $\Theta(\pi)$ is nonzero, understand the representation $\theta(\pi)$ as much as possible, either by computing some of its invariants or determining it precisely, such as in terms of the local Langlands correspondence.

For (a), we have the following two results. The first is due to Li-Sun-Tian [52] whereas the second is a recent result of the author and S. Takeda [19].

Theorem 3.2. *The representation $\theta(\pi)$ is multiplicity-free.*

Theorem 3.3.

(i) *If π is tempered, then $\theta(\pi)$ is either zero or irreducible. Moreover, for any irreducible tempered π and π',*
$$\theta(\pi) \cong \theta(\pi') \neq 0 \Longrightarrow \pi \cong \pi'.$$

(ii) *If $|\dim V - \dim W| \leq 1$, the Howe duality conjecture holds (for any residue characteristic p).*

In the rest of the section, we shall focus on problems (b) and (c).

3.2. Witt towers and first occurrence. Rallis observed that it is fruitful to consider theta correspondence in a family. Let V_0 be an anisotropic Hermitian space over E, and for $r \geq 0$, let
$$V_r = V_0 \oplus \mathbb{H}^r$$
where $\mathbb{H}$ is the hyperbolic plane. The collection $\{V_r \mid r \geq 0\}$ is called a Witt tower of spaces. We note that any given space V is a member of a unique Witt tower of spaces $\{V_r\}$, where V_0 is the anisotropic kernel of V.

One can then consider a family of theta correspondences associated to the tower of reductive dual pairs $(\mathrm{U}(W), \mathrm{U}(V_r))$. For $\pi \in \mathrm{Irr}(\mathrm{U}(W))$, the smallest non-negative integer r_0 such that $\Theta_{V_{r_0},W}(\pi) \neq 0$ is called the *first occurrence index* of π for the Witt tower $\{V_r\}$. By [63, p. 67], such an r_0 exists and $r_0 \leq \dim W$. Moreover, $\Theta_{V_r,W}(\pi) \neq 0$ for all $r \geq r_0$. Thus one way to rephrase problem (b) is to determine the first occurrence index of π in any given Witt tower.

3.3. Conservation relation. Harris-Kudla-Sweet [27] and Kudla-Rallis [44] discovered that the first occurrence indices of π for two different Witt towers $\{V_r\}$ and $\{V'_r\}$ are not independent of each other. More precisely, one considers two Witt towers $\{V_r\}$ and $\{V'_r\}$ such that
$$\dim V_0 \equiv \dim V'_0 \mod 2 \quad \text{but} \quad \epsilon(V_0) \neq \epsilon(V'_0).$$
Hence we may consider the first occurrence indices r_0 for the tower $\{V_r\}$ and r'_0 for the tower $\{V'_r\}$. The following is a basic theorem in the subject:

Theorem 3.4. *For any $\pi \in \mathrm{Irr}(\mathrm{U}(W))$, with first occurrence indices r_0 and r'_0 in two related Witt towers, we have*
$$\dim V_{r_0} + \dim V'_{r'_0} = 2 \dim W + 2.$$

This theorem was called the conservation relation conjecture of Kudla-Rallis. Kudla-Rallis [44] and Gong-Grenie [26] showed the inequality $\geq$ in the statement of the theorem, and also established the reverse inequality $\leq$ in many cases, for example for all supercuspidal representations. A simple and completely different proof of the theorem in the supercuspidal case was also discovered by A. Minguez [59]. Finally, a recent paper of Sun-Zhu [82] established the theorem in full. A corollary is the following dichotomy statement:

Corollary 3.5. *Let V and V' be two spaces in the Witt towers $\{V_r\}$ and $\{V'_r\}$ such that*

$$\dim V + \dim V' = 2 \cdot \dim W.$$

For any $\pi \in \mathrm{Irr}(\mathrm{U}(W))$, exactly one of the theta lifts $\Theta_{V,W}(\pi)$ and $\Theta_{V',W}(\pi)$ is nonzero.

These results place some constraints on the first occurrence indices but fall short of determining these indices. To go further, we need to introduce some basic invariants of π.

3.4. Local doubling zeta integral. The proof of Theorem 3.4 uses as a key tool the local doubling zeta integral, which was discovered by Piatetski-Shapiro and Rallis [73]. Analogous to the local zeta integral in Tate's thesis, the doubling zeta integral can be used to define the standard γ-factors for a pair (π, χ), with $\pi \in \mathrm{Irr}(\mathrm{U}(W))$ and χ a character of $E^\times$. Though this family of zeta integrals was discovered in the mid-1980's, the precise treatment and definition of the local factor $\gamma(s, \pi, \chi, \psi)$ was only carried out by Lapid-Rallis [48] in 2003. From the γ-factors, one can then define the local L-factor $L(s, \pi, \chi)$ and the local ϵ-factor $\epsilon(s, \pi, \chi, \psi)$ following a standard procedure of Shahidi.

However, there is another way to define $L(s, \pi, \chi)$ and $\epsilon(s, \pi, \chi, \psi)$ from a family of zeta integrals: one could define $L(s, \pi, \chi)$ as the GCD of the family of zeta integrals as the data varies. The two ways of defining these local L-factors have complementary strengths, and one would really like them to give the same L-factors and ϵ-factors. This is finally proved in a recent paper [95] of S. Yamana, thus bringing the theory of the doubling zeta integral to a definitive conclusion.

3.5. Epsilon dichotomy. The local factors defined by the doubling zeta integral are very useful for the study of theta correspondence. As an example, in the context of Corollary 3.5, the following result [14, 27] determines exactly which of $\Theta_{V,W}(\pi)$ and $\Theta_{V',W}(\pi)$ is nonzero in the equal rank case.

Theorem 3.6. *Assume that $\dim V = \dim W$. Let $\pi \in \mathrm{Irr}(\mathrm{U}(W))$ with central character ω_π. Then $\Theta_{V,W}(\pi) \neq 0$ if and only if*

$$\epsilon(\tfrac{1}{2}, \pi, \chi_V^{-1}, \psi) = \omega_\pi(-1) \cdot \chi_V(\delta)^{\dim W} \cdot \epsilon(V) \cdot \epsilon(W).$$

3.6. Poles of local γ-factors. As another example, the location of poles of the local γ-factors provides information on the first occurrence index.

Theorem 3.7. *Suppose that V and V' are two spaces in two related Witt towers such that $\dim V = \dim V'$. Assume that $l := \dim W - \dim V > 0$. Let π be an irreducible tempered representation of $\mathrm{U}(W)$.*

(i) *If one of $\Theta_{V,W}(\pi)$ and $\Theta_{V',W}(\pi)$ is nonzero, then $\gamma(s, \pi, \chi_V^{-1}, \psi)$ has a pole at $s = \frac{l+1}{2}$.*

(ii) *Suppose that either π is supercuspidal, or $l = 1$ and π is square integrable. Then the converse of* (i) *also holds.*

Corollary 3.8. *Let π be an irreducible tempered representation of* $\mathrm{U}(W)$. *Assume that $\gamma(s, \pi, \chi_V^{-1}, \psi)$ is holomorphic in* $\mathrm{Re}(s) \geq 1/2$. *Then we have: the first occurrence indices of π in the two Witt towers are given by:*

$$\begin{cases} \dim V_{r_0} = \dim V_{r_0'} = \dim W + 1 & \text{if } \dim W \neq \dim V_0 \mod 2, \\ \{\dim V_{r_0}, \dim V_{r_0'}\} = \{\dim W, \dim W + 2\} & \text{if } \dim W = \dim V_0 \mod 2. \end{cases}$$

Moreover, if $\dim V$ *is the smaller of the two elements in the second case, then $\epsilon(V)$ is determined by Theorem 3.6.*

Thus, one has a precise determination of the first occurrence indices in the tempered case when the relevant local γ-factor is entire in $\mathrm{Re}(s) \geq 1/2$. If π is supercuspidal, we shall see in a moment that one can determine the first occurrence indices precisely in general.

3.7. Prasad's conjecture. Consider the (almost) equal rank case when $\dim V - \dim W = 0$ or 1. For $\pi \in \mathrm{Irr}(\mathrm{U}(W))$, D. Prasad [74, 75] has given precise conjectures describing $\Theta_{V,W}(\pi)$ in terms of the local Langlands correspondence (LLC). We briefly recall the statement of the LLC.

The LLC for unitary groups postulates that each $\pi \in \mathrm{Irr}(\mathrm{U}(W))$ is classified by two invariants (ϕ, η):

(a) $\phi = WD_E \longrightarrow \mathrm{GL}_n(\mathbb{C})$ (where WD_E is the Weil-Deligne group of E and $n = \dim W$) is a conjugate-dual representation of WD_E of sign $(-1)^{n-1}$ (see [13]);

(b) η is a collection of signs. Namely, if we decompose $\phi = \oplus_i m_i \phi_i$, with ϕ_i irreducible and I_ϕ denotes the set of indices such that ϕ_i is also conjugate-dual of sign $(-1)^{n-1}$, then $\eta = (\eta_i)$ is a collection of signs indexed by I_ϕ, satisfying

$$\epsilon(W) = \prod_{i \in I_\phi} \eta_i^{m_i}. \tag{3.1}$$

The LLC for quasi-split unitary groups has been proved by C. P. Mok [65], following Arthur's book [6] for the symplectic and orthogonal groups. The case of non-quasi-split unitary groups is the ongoing work of several people.

If $\pi \in \mathrm{Irr}(\mathrm{U}(W))$ has L-parameter (ϕ, η), then Prasad's conjecture determines the L-parameter $(\theta(\phi), \theta(\eta))$ of $\theta(\pi) \in \mathrm{Irr}(\mathrm{U}(V))$ when it is nonzero.

Theorem 3.9. *Suppose that* $\dim V - \dim W = 0$ *or* 1. *Let $\pi \in \mathrm{Irr}(\mathrm{U}(W))$ and consider $\theta(\pi)$ on* $\mathrm{U}(V)$. *Then we have:*

(i) *If* $\dim V = \dim W$, *then $\theta(\pi)$ is nonzero when the condition in Theorem 3.6 holds. Moreover, $\theta(\phi) = \phi \otimes \chi_V^{-1}\chi_W$, so that $I_\phi = I_{\theta(\phi)}$ and*

$$\theta(\eta)_i / \eta_i = \epsilon(1/2, \phi_i \otimes \chi_V^{-1}, \psi(2 \cdot \mathrm{Tr}_{E/F} -)).$$

(ii) *If* $\dim V = \dim W + 1$, *then set $\theta(\phi) = (\phi \otimes \chi_V^{-1}\chi_W) \oplus \chi_W$, so that $\#I_{\theta(\phi)} = \#I_\phi$ or $\#I_\phi + 1$ depending on whether ϕ contains χ_V as a summand or not.*

(a) *if ϕ does not contain χ_V, then $\theta(\pi) \neq 0$. Its L-parameter is given by the $\theta(\phi)$ defined above, and*

$$\theta(\eta)_i = \eta_i \quad \textit{for all } i \in I_\phi.$$

The extra sign in $\theta(\eta)$ is associated to χ_V and is determined by the analog of the requirement (3.1) for the space V, so that

$$\theta(\eta)_{\chi_V} = \epsilon(W) \cdot \epsilon(V).$$

(b) *if ϕ contains χ_V (so that χ_V contributes to $I(\phi)$), then $\theta(\pi)$ is nonzero if and only if $\epsilon(V) = \epsilon(W) \cdot \eta_{\chi_V}$. Its L-parameter is given by $\theta(\phi)$ defined above and*

$$\theta(\eta)_i = \eta_i \quad \textit{for all } i \in I_\phi.$$

This resolves problem (c) completely in the (almost) equal rank case, and is shown in a recent preprint [15] of the author with Ichino.

3.8. First occurrence of supercuspidals. Putting the above results together, we can now determine the first occurrence of a supercuspidal representation $\pi \in \mathrm{Irr}(\mathrm{U}(W))$ in terms of some basic invariants of π. Set

$$\kappa = \begin{cases} 0, \text{ if } \dim W \neq \dim V_0 \mod 2; \\ 1/2, \text{ if } \dim W = \dim V_0 \mod 2, \end{cases}$$

and let l_0 be defined by:

$$\frac{l_0+1}{2} := \max\left(\{\kappa\} \cup \{s_0 : \gamma(s, \pi, \chi_V^{-1}, \psi) \text{ has a pole at } s = s_0\}\right).$$

Then it is known that l_0 is an integer of the same parity as $\dim W$–$\dim V_0$ and $-1 \leq l_0 \leq \dim W$. Moreover,

$$\{\dim V_{r_0}, \dim V'_{r_0}\} = \{\dim W - l_0, \dim W + 2 + l_0\}.$$

If $l_0 = 0$ or -1, then the first occurrence indices were already given in Corollary 3.8. If $l_0 > 0$, then $\dim W + 2 + l_0 > \dim W - l_0$, and so we need to determine which of $\dim V_{r_0}$ and $\dim V_{r'_0}$ is smaller. If $\dim V$ is the smaller of the two, we shall specify V by giving the sign $\epsilon(V)$:

- If $\dim W \equiv \dim V_0 \mod 2$, then $\epsilon(V)$ is determined by Theorem 3.6.
- If $\dim W \neq \dim V_0 \mod 2$, then $\epsilon(V)$ is determined by Theorem 3.9(ii)(b).

This resolves problem (b) for supercuspidal representations. The recent work of Moeglin [62] makes significant progress towards the general case.

4. Global developments

In this section, we survey some global developments. Hence, k is a number field with ring of adeles $\mathbb{A}$. Let K/k be a quadratic field extension, and consider a dual pair $\mathrm{U}(W) \times \mathrm{U}(V)$ of unitary groups for K/k. Write $[\mathrm{U}(W)]$ to denote the space $\mathrm{U}(W_k)\backslash\mathrm{U}(W_{\mathbb{A}})$. If π is a

cuspidal representation of $\mathrm{U}(W)$, we have its global theta lift $\Theta(\pi)$ on $\mathrm{U}(V)$. As in the local case, it is useful to consider a Witt tower $\{V_r\}$ and the associated global theta lifts $\Theta_{V_r}(\pi)$. It was shown by Rallis that there exists a minimal r_0 such that $\Theta_{V_{r_0}}(\pi) \neq 0$, in which case it is a cuspidal representation. The subsequent global theta lifts (for $r > r_0$) are noncuspidal and hence nonzero.

In view of this, the main question in global theta correspondence is to determine the nonvanishing of $\Theta_{V_r}(\pi)$ in terms of basic invariants of π. We may assume that $\Theta_{V_k}(\pi) = 0$ for $k < r$, so that $\Theta_{V_r}(\pi)$ is cuspidal. Write V for V_r henceforth.

4.1. Inner product. Rallis' approach [78] to answering this question is to compute the Petersson inner product $\langle \theta(\phi, f), \theta(\phi, f) \rangle$. The Rallis inner product formula relates this inner product to the special L-values of π. The mechanism for the Rallis inner product formula relies on the following see-saw diagram of dual pairs:

$$\begin{array}{ccc} \mathrm{U}(W \oplus W^-) & & \mathrm{U}(V) \times \mathrm{U}(V) \\ \Big| & \diagup\!\diagdown & \Big| \\ i & & \\ \mathrm{U}(W) \times \mathrm{U}(W^-) & & \mathrm{U}(V)^\Delta, \end{array}$$

where W^- denotes the skew–Hermitian space obtained from W by multiplying the form by -1, so that $\mathrm{U}(W^-) = \mathrm{U}(W)$. Then one has:

$$\begin{aligned} &\langle \theta(\phi_1, f_1), \theta(\phi_2, f_2) \rangle \\ &= \int_{[\mathrm{U}(V)]} \left(\int_{[\mathrm{U}(W)]} \theta(\phi_1)(g_1, h) \cdot \overline{f_1(g_1)}\, dg_1 \right) \cdot \left(\int_{[\mathrm{U}(W)]} \overline{\theta(\phi_2)(g_2, h)} \cdot f_2(g_2)\, dg_2 \right) dh \\ &= \int_{[\mathrm{U}(W) \times \mathrm{U}(W)]} \left(\int_{[\mathrm{U}(V)]} \theta(\phi_1)(g_1, h) \cdot \overline{\theta(\phi_2)(g_2, h)}\, dh \right) \cdot \overline{f_1(g_1)} \cdot f_2(g_2)\, dg_1\, dg_2 \quad (4.1) \end{aligned}$$

where in the last equality, we have formally exchanged the integrals. This inner integral (if it converges) can be interpreted as the global theta lift of the constant function 1 of $\mathrm{U}(V)^\Delta$ to $\mathrm{U}(W \oplus W^-)$. The inner integral converges absolutely, so that the above exchange is valid, if one is in the Weil's convergent range:

$$r = 0 \quad \text{or} \quad \dim V - r > \dim W.$$

To proceed further, one would like to give a different interpretation of the inner integral. This is the content of the so-called Siegel-Weil formula: it identifies the inner integral with an Eisenstein series. Even in Weil's convergent range, this Siegel-Weil formula was achieved in a series of papers by Weil [92], Kudla-Rallis [41, 42], Ichino [34] and Yamana [93, 94], stretching over 40 years.

4.2. The regularized theta integral. Henceforth, we shall consider life outside Weil's convergent range, so that $r > 0$ and $\dim V - r \leq \dim W$, in which case we have

$$0 < \dim V \leq 2 \dim W \quad \text{and} \quad r \leq \dim W.$$

Consider the Weil representation Ω of $\mathrm{U}(W \oplus W^-) \times \mathrm{U}(V)$. We are interested in the

theta integral

$$I(\phi)(g) = \frac{1}{\tau(\mathrm{U}(V))} \cdot \int_{[\mathrm{U}(V)]} \Theta(\phi)(g,h)\, dh.$$

for $\phi \in \Omega$ and where $\tau(\mathrm{U}(V))$ denotes the Tamagawa number. The integral diverges, but under the above conditions, Kudla-Rallis [43] discovered a regularization of this theta integral.

More precisely, one can find an element z of the Bernstein center of $\mathrm{U}(V_v)$ at some place v of k such that $\Theta(z \cdot \phi)$ is rapidly decreasing as a function on $[\mathrm{U}(V)]$ and hence the integral $I(z \cdot \phi)$ converges. One considers the (spherical) Eisenstein series $E(s)$ associated to the family of degenerate principal series representations induced from the maximal parabolic subgroup of $\mathrm{U}(V)$ stabilising a maximal isotropic subspace of V. At the point $s = \rho_V = \frac{\dim V - r}{2}$,

$$\mathrm{Res}_{s=\rho_V} E(s) = \kappa$$

is a constant function. Moreover, one has $z \cdot E(s) = P_z(s) \cdot E(s)$ for some function $P_z(s)$. Now one sets

$$B(s,\phi) = \frac{1}{\kappa \cdot P_z(s) \cdot \tau(\mathrm{U}(V))} \cdot \int_{[\mathrm{U}(V)]} \Theta(z \cdot \phi)(g,h) \cdot E(s,h)\, dh.$$

This meromorphic function is the regularised theta integral and one is interested in its analytic behaviour at the point $s = \rho_V$.

The Laurent expansion of $B(s,\phi)$ at $s = \rho_V$ has the form

$$B(s,\phi) = \frac{B_{-1}(\phi)}{s - \rho_V} + B_0(\phi) + \cdots \quad \text{when } \dim V \leq \dim W;$$

and

$$B(s,\phi) = \frac{B_{-2}(\phi)}{(s-\rho_V)^2} + \frac{B_{-1}(\phi)}{s-\rho_V} + \cdots \quad \text{when } \dim W < \dim V \leq 2 \dim W.$$

We shall refer to these two cases as the *first term range* and the *second term range* respectively. Each Laurent coefficient B_i gives a linear map

$$B_i : \omega \to \mathcal{A}(\mathrm{U}(W \oplus W^-))$$

and the one which is important for the inner product formula is the residue B_{-1}.

4.3. Siegel Eisenstein series. The purpose of the Siegel-Weil formula is to identify the automorphic forms $B_{-2}(\phi)$ and $B_{-1}(\phi)$ with the analogous Laurent coefficients of a Siegel-Eisenstein series $A(s,\phi)$ associated to ϕ.

More precisely, the diagonally embedded subspace $W^\Delta \subset W \oplus W^-$ is maximal isotropic, so that its stabiliser in $\mathrm{U}(W \oplus W^-)$ is a Siegel parabolic subgroup P, which has Levi factor $\mathrm{GL}(W^\Delta)$. Let

$$I_P(s) = \mathrm{Ind}_P^{\mathrm{U}(W \oplus W^-)} \chi_V |\det|^s.$$

be the associated Siegel principal series representation. Now the Weil representation Ω can be realised on $\mathcal{S}(W^\nabla \otimes V)$ (where W^∇ is an isotropic complement to W^Δ), and the map $\phi \mapsto f_\phi$ with

$$f_\phi(g) = (\Omega(g)\phi)(0) \quad \text{for } g \in \mathrm{U}(W_{\mathbb{A}} \oplus W_{\mathbb{A}}^-)$$

defines a $\mathrm{U}(W \oplus W^-)$-equivariant and $\mathrm{U}(V)$-invariant map

$$\Omega \longrightarrow I_P(s_{V,W}) \quad \text{with } s_{V,W} := (\dim V - \dim W)/2.$$

One then sets

$$A(s, \phi) = E(s, f_\phi).$$

Observe that in the first term range, $s_{V,W} \leq 0$, whereas in the second term range, $s_{V,W} > 0$. If $s = s_{V,W} > 0$, the Laurent expansion of the Siegel-Eisenstein series $A(s, \phi)$ there has the form

$$A(s, \phi) = \frac{A_{-1}(\phi)}{s - s_{V,W}} + A_0(\phi) + \cdots$$

As for B_i, each A_i is a linear map $A_i : \Omega \to \mathcal{A}(\mathrm{U}(W \oplus W^-))$.

4.4. First term identity. Assume that we are in the first term range, so that $s_{V,W} \leq 0$. Let V' be the space in the same Witt tower as V such that

$$\dim V + \dim V' = 2 \dim W \quad (\text{so } \dim V' \geq \dim V).$$

The space V' is called the complementary space to V with respect to W and is such that $s_{V',W} \geq 0$. We shall write A'_i and B'_i for the relevant Laurent coefficients in the context of V'. Ikeda has defined in [35] a natural $\mathrm{U}(W \oplus W^-) \times \mathrm{U}(V)$-equivariant map

$$\mathrm{Ik} : \Omega' \longrightarrow \Omega,$$

where Ω' is the Weil representation for $\mathrm{U}(V') \times \mathrm{U}(W \oplus W^-)$. Then the first term identity established in [32, 33, 35, 43, 93] is the following identity:

Theorem 4.1. *Assume that we are in the first term range. Then for all $\phi \in \Omega$,*

$$c \cdot A'_{-1}(\phi') = A_0(\phi) = 2 \cdot B_{-1}(\phi),$$

where c is an explicit constant, $\phi' \in \Omega'$ is such that $\mathrm{Ik}(\pi_K \phi') = \phi$ and π_K is the projection onto the K-fixed space (with K a maximal compact subgroup of $\mathrm{U}(V'_{\mathbb{A}})$).

4.5. Second term identity. In a recent paper [16], the regularised Siegel-Weil formula is extended to the second term range. More precisely, we have:

Theorem 4.2 (Siegel-Weil formula). *Suppose that $0 < r \leq \dim W$ and $\dim W < \dim V \leq \dim W + r$, so that we are in the second term range.*

(i) *(First term identity) For all $\phi \in \Omega$, one has*

$$A_{-1}(\phi) = B_{-2}(\phi).$$

(ii) *(Second term identity) For all $\phi \in \Omega$, one has*

$$A_0(\phi) = B_{-1}(\phi) - c \cdot \{B'_0(\mathrm{Ik}(\pi_K \phi))\} \mod \mathrm{Im} A_{-1}.$$

Here, c is some explicit constant and V' is the complementary space to V with respect to W (so $\dim V' < \dim V$ here). Finally, the term $\{...\}$ on the RHS is interpreted to be 0 if V' is anisotropic.

4.6. Rallis inner product formula. The Siegel-Weil formulas above and the theory of the doubling zeta integral, as completed by Yamana [95], enable one to establish the Rallis inner product formula. For the result in the first term range, we refer the reader to Yamana [95]. In the second term range, we have [16]:

Theorem 4.3. *Suppose that*

$$\dim W < \dim V \leq 2 \dim W \quad \textit{and} \quad r \leq \dim W$$

so that we are either in the second term range or the convergent range, depending on whether $\dim V \leq \dim W + r$ *or not. Let* π *be a cuspidal representation of* $\mathrm{U}(W)$ *and consider its global theta lift* $\Theta(\pi)$ *to* $\mathrm{U}(V)$.

(i) *Assume that* $\Theta(\pi)$ *is cuspidal. Then for* $\phi_1, \phi_2 \in \omega_{\psi,V,W}$ *and* $f_1, f_2 \in \pi$,

$$\begin{aligned}&\langle \theta(\phi_1, f_1), \theta(\phi_2, f_2) \rangle \\ =&[E:F] \cdot \mathrm{Val}_{s=s_{V,W}} \left(L(s + \frac{1}{2}, \pi \times \chi_V) \cdot Z^*(s, \phi_1 \otimes \overline{\phi_2}, f_1, f_2) \right),\end{aligned}$$

where $s_{V,W} = (\dim V - \dim W)/2 > 0$, $L(s, \pi \times \chi_V)$ *is the standard L-function of* π, *and* $Z^*(s, -)$ *denotes the normalized doubling zeta integral.*

(ii) *Assume further that for all places* v *of* F, *the local theta lift* $\Theta_{n,r}(\pi_v)$ *is nonzero. Then* $L(s + \frac{1}{2}, \pi \times \chi_V)$ *is holomorphic at* $s = s_{V,W}$, *so that*

$$\langle \theta(\phi_1, f_1), \theta(\phi_2, f_2) \rangle = [E:F] \cdot L(s_{V,W} + \frac{1}{2}, \pi \times \chi_V) \cdot Z^*(s_{V,W}, \phi_1 \otimes \overline{\phi_2}, f_1, f_2).$$

4.7. Nonvanishing of global theta lifts. As a consequence, we have the following local-global criterion for the nonvanishing of global theta lifts.

Theorem 4.4. *Assume the same conditions on* (V, W) *as in Theorem 4.3. Let* π *be a cuspidal representation of* $\mathrm{U}(W)$ *and consider its global theta lift* $\Theta(\pi)$ *to* $\mathrm{U}(V)$. *Assume that* $\Theta(\pi)$ *is cuspidal.*

(i) *If* $\Theta(\pi)$ *is nonzero, then*

  (a) *for all places* v, $\Theta(\pi_v) \neq 0$, *and*
  (b) $L(s_{V,W} + \frac{1}{2}, \pi \times \chi_V) \neq 0$ *i.e. nonzero holomorphic.*

(ii) *The converse to (i) holds if* $K_v = k_v \times k_v$ *for all archimedean places* v *of* k.

More generally, under the conditions (a) and (b) in (i), there is a Hermitian space V' *over* K *such that*

- $V' \otimes_k k_v \cong V \otimes_k k_v$ *for every finite or complex place of* k;
- *the global theta lift* $\Theta'(\pi)$ *of* π *to* $\mathrm{U}(V')$ *is nonzero.*

The reason for not having the converse to (i) in general is that, if $K_v/k_v = \mathbb{C}/\mathbb{R}$, we do not know the equivalence of the nonvanishing of the local theta correspondence and that of the normalised doubling zeta integral on an appropriate submodule of its domain.

5. Variations and extensions.

In this section, we want to mention some extensions of the theory of theta correspondence which have been pursued in the last 20 years.

5.1. Exceptional theta correspondence. There is no reason to confine oneself to dual pairs in the symplectic group. One could consider dual pairs in any connected reductive group G. For theta correspondence, however, one also needs the analog of the Weil representation. It turns out that the Weil representation is the "smallest" infinite-dimensional representation of the metaplectic group. This suggests that one should consider the analogous smallest representation of $G(F)$. Such a representation is called a minimal representation of $G(F)$.

There is a series of work devoted to the construction and classification of minimal representations of an arbitrary $G(F)$. Of these, one might mention various papers of Kazhdan, Savin and Torasso [36, 37, 80, 83]. In the global case, the automorphic realisation of the minimal representations have been constructed, largely using residues of Eisenstein series [22].

With the theory of minimal representations in place, one can start to consider theta correspondence. While the setup is the same as classical theta correspondence, one key difference is that one does not know the analog of the Howe duality conjecture for exceptional theta correspondence; in particular, one does not know the analog of Theorem 3.1. For work on exceptional theta correspondences, we may mention a series of papers by Savin and various collaborators [25, 30, 31, 57].

5.2. Singular theta lifting of Borcherds. In his 1994 ICM address [10], Borcherds described a singular theta lifting for classical dual pairs. In classical global theta correspondence, one integrates the theta kernel against cusp forms, and there is no issue with convergence. In Borcherd's context, one is trying to lift functions which blow up exponentially at the cusps, and Borcherds gave a regularisation of such a singular theta integral. An example of such a function is the classical j-function on the upper half plane. This theory is so far not representation theoretic in nature, but it allows Borcherds to construct many beautiful examples of automorphic forms which possess infinite product expansion [11], analogous to the classical η-function.

5.3. Arithmetic theta lifting of Kudla. Since the mid-1990's, Kudla [40] has pursued an arithmetic version of the theory of theta correspondence. This provides a lifting of automorphic forms to classes in the arithmetic Chow group of a Shimura variety. One goal of Kudla's program is to establish an arithmetic version of the Rallis inner product formula in the equal rank case, which involves the central derivative of the standard L-function instead of the central value. This will require an arithmetic Siegel-Weil formula. A low rank example was established in the book [45] of Kudla-Rapoport-Yang. In his PhD thesis [53, 54], Y. F. Liu formulated such a conjectural arithmetic Rallis inner product formula in the context of unitary groups of arbitrary rank.

5.4. Geometric theta lifting of Lysenko. The theory of theta correspondence was almost single-handedly extended to the framework of the Geometric Langlands Program by S. Lysenko [55, 56]. Together with V. Lafforgue [46, 47], the theory of minimal representations was also suitably geometrized.

6. Local Langlands correspondence

In the rest of this report, we will discuss a number of applications of theta correspondence. The first such application is to the local Langlands conjecture (LLC). Unlike the earlier sections, we will no longer restrict ourselves to unitary groups.

6.1. LLC for $\mathbf{GSp_4}$. In [18] and [20], the local theta correspondence was used to establish the LLC for the group $\mathrm{GSp}_4(F)$ and its non-split inner form, where F is a p-adic field. This uses an extension of the theta correspondence from the setting of isometry dual pairs to the setting of similitude dual pairs.

Let us briefly explain how theta correspondence is used in the proof. Let W be the 4-dimensional symplectic space, V the split quadratic space of dimension 6 and trivial discriminant, and V' the anisotropic quadratic space of dimension 4 and trivial discriminant. These quadratic spaces belong to the two different related Witt towers, and we consider the similitude theta correspondence for $\mathrm{GSp}(W) \times \mathrm{GO}(V)$ and $\mathrm{GSp}(W) \times \mathrm{GO}(V')$. By the dichotomy statement in Corollary 3.5, we deduce that each $\pi \in \mathrm{Irr}(\mathrm{GSp}(W))$ has nonzero theta lift to exactly one of $\mathrm{GO}(V)$ or $\mathrm{GO}(V')$. Using this, one deduces an injection

$$\mathrm{Irr}(\mathrm{GSp}(W)) \hookrightarrow \mathrm{Irr}(\mathrm{GSO}(V)) \bigsqcup \mathrm{Irr}(\mathrm{GSO}(V')).$$

Now one notes that

$$\begin{cases} \mathrm{GSO}(V') \cong (\mathrm{GL}_2(F) \times D^\times)/\{(t, t^{-1}) : t \in F^\times\} \\ \mathrm{GSO}(V) \cong (\mathrm{GL}_4(F) \times F^\times)/\{(t, t^{-2}) : t \in F^\times\}, \end{cases}$$

where D is the quaternion division F-algebra. In particular, the LLC is known for these two groups, so one may assign L-parameters to representations of $\mathrm{GSp}_4(F)$.

6.2. LLC for $\mathbf{G_2}$. We now describe an ongoing work of the author with G. Savin on the LLC for the split exceptional group of type G_2 using the exceptional theta correspondence. Quite amazingly, it turns out that a similar strategy as in the GSp_4 case can be implemented.

More precisely, one has the two dual pairs

$$\mathrm{G}_2 \times \mathrm{PB}^\times \subset \mathrm{E}_6^{\mathrm{B}} \quad \text{and} \quad \mathrm{G}_2 \times \mathrm{PGSp}_6 \subset \mathrm{E}_7$$

where B denotes a degree 3 division F-algebra, $\mathrm{E}_6^{\mathrm{B}}$ is an inner form of type E_6 and F-rank 2 and E_7 is the split group of this type. Consider the local theta correspondence for these two dual pairs, a key result is the following analog of dichotomy and the Howe duality conjecture:

Theorem 6.1. *Each $\pi \in \mathrm{Irr}(\mathrm{G}_2)$ has nonzero theta lift to exactly one of $\mathrm{PB}^\times$ or PGSp_6. Moreover, the nonzero $\theta(\pi)$ is irreducible. In particular, one has an injection*

$$\mathrm{Irr}(\mathrm{G}_2) \hookrightarrow \mathrm{Irr}(PB^\times) \sqcup \mathrm{Irr}(\mathrm{PGSp}_6).$$

By the Jacquet-Langlands correspondence and the LLC for PGL_3 and Sp_6 (due to Arthur), one may then hope to assign L-parameters to $\pi \in \mathrm{Irr}(\mathrm{G}_2)$. This theorem will play a key role in our ongoing work to establish the full LLC for G_2.

7. Gross-Prasad conjecture

One typical application of theta correspondence is that it relates certain periods on one member of a dual pair with certain periods on the other member. One such family of periods which has attracted much attention recently is the Gross-Prasad (GP) periods, which was considered by Gross and Prasad in the context of the special orthogonal groups in two papers [23, 24] some twenty years ago. They formulated precise conjectures for the nonvanishing of the GP periods. In a recent paper [13], these conjectures were extended to arbitrary classical groups. For ease of exposition, we shall consider the case of unitary groups.

7.1. GP periods. Let V_{n+1} be a Hermitian space of dimension $n+1$ over E and W_n a skew-Hermitian space of dimension n over E. Let $V_n \subset V_{n+1}$ be a nondegenerate subspace of codimension 1, so that we have a natural inclusion $\mathrm{U}(V_n) \hookrightarrow \mathrm{U}(V_{n+1})$. In particular, if we set

$$G_n = \mathrm{U}(V_n) \times \mathrm{U}(V_{n+1}) \quad \text{or} \quad \mathrm{U}(W_n) \times \mathrm{U}(W_n)$$

and

$$H_n = \mathrm{U}(V_n) \quad \text{or} \quad \mathrm{U}(W_n),$$

then we have a diagonal embedding $\Delta : H_n \hookrightarrow G_n$.

In the Hermitian case, one is interested in determining $\dim_{\mathbb{C}} \mathrm{Hom}_{\Delta H_n}(\pi, \mathbb{C})$ for $\pi \in \mathrm{Irr}(G_n)$. We shall call this the *Bessel* case of the GP conjecture. Indeed, what we have described is a special case: the general Bessel case deals with a pair of Hermitian spaces $V' \subset V$ such that $\dim V/V'$ is odd.

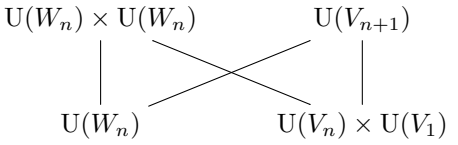
In the skew-Hermitian case, the restriction problem requires another piece of data: a Weil representation ω_{ψ,χ,W_n} of $\mathrm{U}(W_n)$, where χ is a character of $E^\times$ such that $\chi|_{F^\times} = \omega_{E/F}$. Then one is interested in determining $\dim_{\mathbb{C}} \mathrm{Hom}_{\Delta H_n}(\pi, \omega_{\psi,\chi,W_n})$. We shall call this the *Fourier-Jacobi* case (FJ) of the GP conjecture. As before, the general FJ case deals with a pair of skew-Hermitian spaces $W' \subset W$ such that $\dim W/W'$ is even. To unify notation, we shall let $\nu = \mathbb{C}$ or ω_{ψ,χ,W_n} in the respective cases.

7.2. Gross-Prasad conjecture. It was shown in [5] and [81] that the above Hom spaces have dimension at most 1. Thus the main issue is to determine when the Hom space is nonzero. The Gross-Prasad conjecture gives an answer for this issue, formulated in the framework of the local Langlands correspondence. It can be loosely stated as follows:

1. Given a generic L-parameter ϕ for G_n there is a unique η such that the representation $\pi(\phi,\eta)$ satisfies $\mathrm{Hom}_{\Delta H_n}(\pi(\phi,\eta),\nu) \neq 0$.
2. There is a precise recipe, in terms of local ϵ-factor for the distinguished character η.

In a stunning series of papers [87], [88], [89], [90], Waldspurger has established the Bessel case of the GP conjecture for special orthogonal groups in the case of tempered L-parameters; the case of general generic L-parameters is then dealt with by Mœglin-Waldspurger [64]. Beuzart-Plessis [7], [8], [9] has since extended Waldspurger's techniques to settle the Bessel case of the GP conjecture for unitary groups in the tempered case.

7.3. Theta correspondence. Now the Bessel and Fourier-Jacobi cases of the GP conjecture are related by the local theta correspondence. More precisely, there is a see-saw diagram



and the associated see-saw identity reads:

$$\mathrm{Hom}_{\mathrm{U}(W_n)}(\Theta_{\psi,\chi,V_n,W_n}(\sigma)\otimes\omega_{\psi,\chi,V_1,W_n},\pi)\cong\mathrm{Hom}_{\mathrm{U}(V_n)}(\Theta_{\psi,\chi,V_{n+1},W_n}(\pi),\sigma)$$

for $\pi\in\mathrm{Irr}(\mathrm{U}(W_n))$ and $\sigma\in\mathrm{Irr}(\mathrm{U}(V_n))$. Hence the left-hand side of the see-saw identity concerns the Fourier-Jacobi case (FJ) whereas the right-hand side concerns the Bessel case (B). It is thus apparent that precise knowledge of the local theta correspondence for unitary groups of (almost) equal rank will give the precise relation of (FJ) to (B).

In particular, as a consequence of the proof of Prasad's conjecture in Theorem 3.9, the FJ case of the GP conjecture was verified in [15] . Hence one has:

Theorem 7.1. *Assume the LLC for unitary groups. Then both the Bessel and FJ cases of the GP conjecture hold.*

8. Shimura-Waldspurger correspondence

We will conclude by returning to the Shimura-Waldspurger (SW) correspondence for Mp_2, which in some sense initiated many of the developments discussed in this paper. In particular, we will discuss its extension to Mp_{2n}.

8.1. Local SW correspondence. Let F be a nonarchimedean local field. Let W be the $2n$-dimensional symplectic vector space, and let V^+ and V^- be the two $2n+1$-dimensional quadratic spaces with trivial discriminant, with V^+ split. Then one may consider the theta correspondence for $\mathrm{Mp}(W)\times\mathrm{O}(V^\epsilon)$. As a consequence of Theorem 3.6, the following was shown in [17]:

Theorem 8.1. *Fix a nontrivial additive character ψ of F. The theta correspondence with respect to ψ gives a bijection*

$$\mathrm{Irr}_\epsilon(\mathrm{Mp}W))\longleftrightarrow\mathrm{Irr}(\mathrm{SO}(V^+))\sqcup\mathrm{Irr}(\mathrm{SO}(V^-)),$$

where we consider genuine representations of $\mathrm{Mp}(W)$ on the LHS. Assuming the LLC for $\mathrm{SO}(V^\pm)$, one then inherits an LLC for $\mathrm{Mp}(W)$. Moreover, this LLC satisfies a list of expected properties which characterise it uniquely.

When F is archimedean, the analogous theorem was obtained by Adams-Barbasch [4] some 20 years ago, and described in Adams' 1994 ICM talk [2].

8.2. Global SW correspondence. Now assume that we are working over a number field k. It is natural to attempt to use the global theta correspondence to obtain a precise description

of the automorphic discrete spectrum of $\mathrm{Mp}(W_{\mathbb{A}})$. For readers familiar with Waldspurger's work [84, 85] in the case when $\dim W = 2$, it will be apparent that there is an obstruction to this approach: the global theta lift $\Theta(\pi)$ of a cuspidal representation π of $\mathrm{Mp}(W_{\mathbb{A}})$ or $\mathrm{SO}(V_{\mathbb{A}})$ may be 0 and it is nonzero precisely when $L(1/2, \pi) \neq 0$.

This obstruction already occurs when $\dim W = 2$, and was not easy to overcome. Waldspurger had initially alluded to results of Flicker proved by the trace formula. Nowadays, one could appeal to a result of Friedberg-Hoffstein [12], stating that if $\epsilon(1/2, \pi) = 1$, then there exists a quadratic Hecke character χ such that $L(1/2, \pi \times \chi) \neq 0$. When $\dim W > 2$, however, the analogous analytic result does not seem to be forthcoming and may be very hard. We are going to suggest a new approach in the higher rank case, but before that, we would like to describe the analog of Arthur's conjecture for Mp_{2n}.

8.3. Arthur's conjecture for $\mathbf{Mp_{2n}}$. For a fixed additive automorphic character ψ, one expects that

$$L^2_{disc} = \bigoplus_{\Psi} L^2_{\Psi,\psi}$$

where

$$\Psi = \bigoplus_i \Psi_i = \bigoplus_i \Pi_i \boxtimes S_{r_i}$$

is a global discrete A-parameter for Mp_{2n}; it is also an A-parameter for SO_{2n+1}. Here, S_{r_i} is the r_i-dimensional representation of $\mathrm{SL}_2(\mathbb{C})$ and Π_i is a cuspidal representation of GL_{n_i} such that

$$\begin{cases} L(s, \Pi_i, \wedge^2) \text{ has a pole at } s = 1, \text{ if } r_i \text{ is odd;} \\ L(s, \Pi_i, \mathrm{Sym}^2) \text{ has a pole at } s = 1, \text{ if } r_i \text{ is even.} \end{cases}$$

Moreover, we have $\sum_i n_i r_i = 2n$ and the summands Ψ_i are mutually distinct.

For a given Ψ, one inherits the following additional data:

- for each v, one inherits a local A-parameter

$$\Psi_v = \bigoplus_i \Psi_{i,v} = \bigoplus_i \Pi_{i,v} \boxtimes S_{r_i}.$$

  By the LLC for GL_N, we may regard each $\Pi_{i,v}$ as an n_i-dimensional representation of the Weil-Deligne group WD_{k_v}. Hence, we may regard Ψ_v as a $2n$-dimensional representation of $WD_{k_v} \times \mathrm{SL}_2(\mathbb{C})$.

- one has a "global component group"

$$A_\Psi = \bigoplus_i \mathbb{Z}/2\mathbb{Z} \cdot a_i$$

  which is a $\mathbb{Z}/2\mathbb{Z}$-vector space equipped with a distinguished basis indexed by the Ψ_i's. Similarly, for each v, we have the local component group A_{Ψ_v} which is defined as the component group of the centralizer of the image of Ψ_v, thought of as a representation of $WD_{k_v} \times \mathrm{SL}_2(\mathbb{C})$. There is a natural diagonal map

$$\Delta : A_\Psi \longrightarrow \prod_v A_{\Psi_v}.$$

- For each v, one has a local A-packet associated to Ψ_v and ψ_v:

$$\Pi_{\Psi_v,\psi_v} = \{\sigma_{\eta_v} : \eta_v \in \mathrm{Irr}(A_{\Psi_v})\},$$

consisting of unitary representations (possibly zero, possibly reducible) of $\mathrm{Mp}_{2n}(k_v)$ indexed by the set of irreducible characters of A_{Ψ_v}. On taking tensor products of these local A-packets, we obtain a global A-packet

$$A_{\Psi,\psi} = \{\sigma_\eta : \eta = \otimes_v \eta_v \in \mathrm{Irr}(\prod_v A_{\Psi_v})\}$$

consisting of abstract unitary representations $\sigma_\eta = \otimes_v \sigma_{\eta_v}$ of $\mathrm{Mp}_{2n}(\mathbb{A})$ indexed by the irreducible characters $\eta = \otimes_v \eta_v$ of $\prod_v A_{\Psi_v}$.

- Arthur has attached to Ψ a quadratic character (possibly trivial) ϵ_Ψ of A_Ψ, This character plays an important role in the multiplicity formula for the automorphic discrete spectrum of SO_{2n+1}. For Mp_{2n}, we need to define a modification of ϵ_Ψ.

  More precisely, consider the L-parameter $\Phi_\Psi = \bigoplus_i \Phi_{\Psi_i}$ associated to Ψ, with

$$\Phi_{\Psi_i} = \bigoplus_{k=0}^{r_i-1} \Pi_i \cdot |-|^{(r_i-1-2k)/2}.$$

  Then define $\eta_\Psi \in \mathrm{Irr} A_\Psi$ by

$$\eta_\Psi(a_i) = \epsilon(1/2, \Phi_{\Psi_i}) = \begin{cases} \epsilon(1/2, \Pi_i), \text{ if } L(s, \Pi_i, \wedge^2) \text{ has a pole at } s = 1; \\ 1, \text{ if } L(s, \Pi_i, \mathrm{Sym}^2) \text{ has a pole at } s = 1. \end{cases}$$

  The modified quadratic character of A_Ψ in the metaplectic case is

$$\tilde{\epsilon}_\Psi = \epsilon_\Psi \cdot \eta_\Psi.$$

We can now state the conjecture.

Conjecture 8.2 (Arthur Conjecture for Mp_{2n}). *There is a decomposition*

$$L^2_{disc}(\mathrm{Mp}_{2n}) = \bigoplus_\Psi L^2_{\Psi,\psi}$$

where the sum runs over equivalence classes of discrete A-parameters of Mp_{2n}. *For each such* Ψ,

$$L^2_{\Psi,\psi} \cong \bigoplus_{\eta \in \mathrm{Irr}(\prod_v A_{\Psi_v}) : \Delta^*(\eta) = \tilde{\epsilon}_\Psi} \sigma_\eta$$

8.4. A new approach. In an ongoing work, we are developing a new approach for the Arthur conjecture described above. Namely, by results of Arthur [6], one now has a classification of the automorphic discrete spectrum of SO_{2r+1} for all r. Instead of trying to construct the automorphic discrete spectrum of Mp_{2n} by theta lifting from SO_{2n+1}, one could attempt to use theta liftings from SO_{2r+1} for $r \geq n$. Let us illustrate this in the case when $\dim W = 2$.

Let π be a cuspidal representation of $\mathrm{PGL}_2(\mathbb{A}) = \mathrm{SO}(V_{\mathbb{A}}^+)$. Then π gives rise to a near equivalence class in the automorphic discrete spectrum of Mp_2. If $L(1/2, \pi) \neq 0$, this near equivalence class can be exhausted by the global theta lifts of π and its Jacquet-Langlands transfer to inner forms of PGL_2. When $L(1/2, \pi) = 0$, we consider the A-parameter

$$\psi = \pi \boxtimes S_1 \oplus 1 \boxtimes S_2 \quad \text{for } \mathrm{SO}_5.$$

This is a so-called Saito-Kurokawa A-parameter. By Arthur, ψ indexes a near equivalence class in the automorphic discrete spectrum of SO_5. In a well-known paper [72], Piatetski-Shapiro gave a construction of the Saito-Kurokawa representations by theta lifting from Mp_2, using Waldspurger's results as initial data. However, *one can turn the table around*.

Namely, taking the Saito-Kurokawa near equivalence classes as given by Arthur, one can consider their theta lift back to Mp_2. By the Rallis inner product formula, such a theta lift is nonzero if the partial L-function

$$L^S(s, \Phi_\psi) = L^S(s, \pi) \cdot \zeta(s + \frac{1}{2}) \cdot \zeta(s - \frac{1}{2})$$

has a pole at $s = 3/2$, or equivalently if $L^S(3/2, \pi) \neq 0$. Now this is certainly much easier to ensure than the nonvanishing at $s = 1/2$! In this way, one can construct the desired near equivalence class for Mp_2 associated to π and by studying the local theta correspondence in detail, one can recover Waldspurger's results from 30 years ago.

References

[1] J. Adams, *L-functoriality for dual pairs*, Astérisque **171-172** (1989), 85–129.

[2] ———, *Genuine representations of the metaplectic group and epsilon factors*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), 721–731, Birkhäuser, Basel, 1995.

[3] J. Adams and D. Barbasch, *Reductive dual pair correspondence for complex groups*, J. Funct. Anal. **132** (1995), no. 1, 1–42.

[4] ———, *Genuine representations of the metaplectic group*, Compositio Math 113 (1998), no. 1, 23-66.

[5] A. Aizenbud, D. Gourevitch, S. Rallis, and G. Schiffmann, *Multiplicity one theorems*, Ann. of Math. **172** (2010), 1407–1434.

[6] J. Arthur, *The endoscopic classification of representations: orthogonal and symplectic groups*, Colloquium Publications **61**, American Mathematical Society, 2013.

[7] R. Beuzart-Plessis, *La conjecture locale de Gross-Prasad pour les représentations tempérées des groupes unitaires*, arXiv:1205.2987.

[8] ———, *Expression d'un facteur epsilon de paire par une formule intégrale*, arXiv: 1212.1082.

[9] ———, *Endoscopie et conjecture raffinée de Gan-Gross-Prasad pour les groupes unitaires*, arXiv:1212.0951.

[10] Borcherds, R. E., *Automorphic forms on* $\mathrm{O}_{s+2,2}(\mathbb{R})^+$ *and generalized Kac-Moody algebras*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zurich, 1994), 744-752, Birkhäuser, Basel, 1995.

[11] ———, *Automorphic forms on* $\mathrm{O}_{s+2,2}(\mathbb{R})^+$ *and infinite products*, Invent. Math. **120** (1995), no. 1, 161–213.

[12] S. Friedberg and J. Hoffstein, *Nonvanishing theorems for automorphic L-functions on GL(2)*, Ann. of Math. (2) **142** (1995), no. 2, 385-423.

[13] W. T. Gan, B. H. Gross, and D. Prasad, *Symplectic local root numbers, central critical L-values, and restriction problems in the representation theory of classical groups*, Astérisque **346** (2012), 1–109.

[14] W. T. Gan and A. Ichino, *Formal degrees and local theta correspondence*, Invent. Math. **195** (2014), no. 3, 509–672.

[15] W. T. Gan and A. Ichino, *Gross-Prasad conjecture and local theta correspondence*, preprint.

[16] W. T. Gan, Y. Qiu, and S. Takeda, *The regularized Siegel-Weil formula (the second term identity) and the Rallis inner product formula*, Invent. Math. (2014), DOI:10.1007/s00222-014-0509-0.

[17] W. T. Gan and G. Savin, *Representations of metaplectic groups I: epsilon dichotomy and local Langlands correspondence*, Compos. Math. **148** (2012), 1655–1694.

[18] W. T. Gan and S. Takeda, *The local Langlands conjecture for* $\mathrm{GSp}(4)$, Ann. of Math. **173** (2011), 1841–1882.

[19] W. T. Gan and S. Takeda, *On the Howe duality conjecture in classical theta correspondence*, preprint, arXiv:1405.2626.

[20] W. T. Gan and W. Tantono, *The local Langlands conjecture for* GSp_4 *II: the case of inner forms*, American J. of Math. 136, no. 3 (2014), 761-805.

[21] D. Ginzburg, D.H. Jiang, and D. Soudry, *Poles of L-functions and theta liftings for orthogonal groups*, J. Inst. Math. Jussieu **8** (2009), no. 4, 693-741.

[22] D. Ginzburg, S. Rallis, and D. Soudry, *On the automorphic theta representation for simply-laced groups*, Israel Journal of Math. **100** (1997), 61-116.

[23] B. H. Gross and D. Prasad, *On the decomposition of a representation of* SO_n *when restricted to* SO_{n-1}, Canad. J. Math. **44** (1992), 974–1002.

[24] B. H. Gross and D. Prasad, *On irreducible representations of* $\mathrm{SO}_{2n+1} \times \mathrm{SO}_{2m}$, Canad. J. Math. **46** (1994), 930–950.

[25] B. Gross and G. Savin, *Motives with Galois group of type* G_2*: an exceptional theta-correspondence,* Compositio Math. **114** (1998), no. 2, 153–217.

[26] Z. Gong and L. Grenié, *An inequality for local unitary theta correspondence*, Ann. Fac. Sci. Toulouse Math. **20** (2011), 167–202.

[27] M. Harris, S. S. Kudla, and W. J. Sweet, Jr., *Theta dichotomy for unitary groups*, J. Amer. Math. Soc. **9** (1996), 941–1004.

[28] R. Howe, *Transcending classical invariant theory*, J. Amer. Math. Soc. **2** (1989), 535–552.

[29] R. Howe and Piatetski-Shapiro, *A counterexample to the "generalized Ramanujan conjecture" for (quasi-) split groups*, Proceddings of Symp. Pure Math. Vol. 33 (1979), 315–322.

[30] J.-S. Huang, K. Magaard, and G. Savin, *Unipotent representations of G_2 arising from the minimal representation of D_4^E*, Journal Reine Angew Math. **500** (1998), 65–81.

[31] J.-S. Huang and P. Pandžić and G. Savin, *New dual pair correspondences*, Duke Math. J. **82** (1996), 447–471.

[32] A. Ichino, *On the regularized Siegel-Weil formula*, J. Reine Angew. Math. **539** (2001), 201–234.

[33] _____, *A regularized Siegel-Weil formula for unitary groups*, Math. Z. 247 (2004), 241–277.

[34] _____, *On the Siegel-Weil formula for unitary groups*, Math. Z. 255 (2007), no. 4, 721–729.

[35] T. Ikeda, *On the residue of the Eisenstein series and the Siegel-Weil formula*, Compositio Math. 103 (1996), 183–218.

[36] D. Kazhdan, *The minimal representation of D_4*, in *Operator algebras, unitary representations, enveloping algebras and invariant theory; in honor of Jacques Dixmier*, Progress in Math. 92 (1990), Birkhauser, 125–158.

[37] D. Kazhdan and G. Savin, *The smallest representation of simply-laced groups*, in *Israel Math. Conference Proceedings, Piatetski-Shapiro Festschrift*, Vol. 2 (1990), 209–233.

[38] S. S. Kudla, *On the local theta-correspondence*, Invent. Math. **83** (1986), 229–255.

[39] _____, *Splitting metaplectic covers of dual reductive pairs*, Israel J. Math. **87** (1994), 361–401.

[40] _____, *Derivatives of Eisenstein series and arithmetic geometry*, Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002), 173-183, Higher Ed. Press, Beijing, 2002.

[41] S. Kudla and S. Rallis, *On the Weil-Siegel Formula*, Journal Reine Angew. Math. 387 (1988), 1–68.

[42] _____, *On the Weil-Siegel Formula II: Isotropic Convergent Case*, J. Reine Angew. Math. 391 (1988), 65–84.

[43] _____, *A regularized Siegel-Weil formula: the first term identity*, Ann. Math. **140** (1994), 1–80.

[44] _____, *On first occurrence in the local theta correspondence, Automorphic representations, L-functions and applications: progress and prospects*, Ohio State Univ. Math. Res. Inst. Publ. **11**, de Gruyter, Berlin, 2005, pp. 273–308.

[45] S. S. Kudla, M. Rapoport, and T. H. Yang, *Modular forms and special cycles on Shimura curves*, Annals of Mathematics Studies, 161. Princeton University Press, Princeton, NJ, 2006. x+373 pp.

[46] V. Lafforgue and S. Lysenko, *Geometric Weil representation: local field case*, Compos. Math. 145 (2009), no. 1, 56–88.

[47] _____, *Geometrizing the minimal representations of even orthogonal groups*, Represent. Theory 17 (2013), 263–325.

[48] E. Lapid and S. Rallis, *On the local factors of representations of classical groups, Automorphic representations, L-functions and applications: progress and prospects*, Ohio State Univ. Math. Res. Inst. Publ. **11**, de Gruyter, Berlin, 2005, pp. 309–359.

[49] J.-S. Li, *Singular unitary representations of classical groups*, Invent. Math. **97** (1989), 237–255.

[50] _____, *Nonvanishing theorems for the cohomology of certain arithmetic quotients*, J. Reine Angew. Math. **428** (1992), 177–217.

[51] _____, *Automorphic forms with degenerate Fourier coefficients*, American Journal of Math. 119 (1997), 523–578.

[52] J.-S. Li, B. Sun, and Y. Tian, *The multiplicity one conjecture for local theta correspondences*, Invent. Math. **184** (2011), 117–124.

[53] Y. F. Liu, *Arithmetic theta lifting and L-derivatives for unitary groups I*, Algebra Number Theory **5** (2011), no. 7, 849–921.

[54] _____, *Arithmetic theta lifting and L-derivatives for unitary groups II*, Algebra Number Theory **5** (2011), no. 7, 923–1000.

[55] S. Lysenko, *Moduli of metaplectic bundles on curves and theta-sheaves*, Ann. Sci. École Norm. Sup. (4) **39** (2006), no. 3, 415–466.

[56] _____, *Geometric theta-lifting for the dual pair* SO_{2m}, Sp_{2n}, Ann. Sci. École Norm. Supér. (4) **44** (2011), no. 3, 427-493.

[57] K. Magaard and G. Savin, *Exceptional* Θ*-correspondences I*, Compositio Math. 107 (1997), 89–123.

[58] A. Mínguez, *Correspondance de Howe explicite: paires duales de type II*, Ann. Sci. Éc. Norm. Supér. **41** (2008), 717–741.

[59] _____, *The conservation relation for cuspidal representations*, Math. Ann. **352** (2012), 179–188.

[60] C. Mœglin, *Correspondance de Howe pour les paires reductives duales: quelques calculs dans le cas archimedien*, J. Funct. Anal. 85 (1989), no. 1, 1–85.

[61] _____, *Représentations quadratiques unipotentes pour les groupes classiques* p*-adiques*, Duke Math. Journal 84 (1996), 267–332.

[62] C. Mœglin, *Conjecture d'Adams pour la correspondance de Howe et filtration de Kudla*, Arithmetic geometry and automorphic forms, Adv. Lect. Math. **19**, Int. Press, Somerville, MA, 2011, pp. 445–503

[63] C. Mœglin, M.-F. Vignéras, and J.-L. Waldspurger, *Correspondances de Howe sur un corps p-adique*, Lecture Notes in Mathematics **1291**, Springer-Verlag, Berlin, 1987.

[64] C. Mœglin and J.-L. Waldspurger, *La conjecture locale de Gross-Prasad pour les groupes spéciaux orthogonaux: le cas général*, Astérisque **347** (2012), 167–216.

[65] C. P. Mok, *Endoscopic classification of representations of quasi-split unitary groups*, Mem. Amer. Math. Soc., to appear.

[66] G. Muić, *Howe correspondence for discrete series representations; the case of* $(\mathrm{Sp}(n), \mathrm{O}(V))$, J. Reine Angew. Math. **567** (2004), 99–150.

[67] ———, *On the structure of the full lift for the Howe correspondence of* $(\mathrm{Sp}(n), \mathrm{O}(V))$ *for rank-one reducibilities*, Canad. Math. Bull. **49** (2006), 578–591.

[68] ———, *On the structure of theta lifts of discrete series for dual pairs* $(\mathrm{Sp}(n), \mathrm{O}(V))$, Israel J. Math. **164** (2008), 87–124

[69] ———, *Theta lifts of tempered representations for dual pairs* $(\mathrm{Sp}_{2n}, \mathrm{O}(V))$, Canadian J. Math. **60** No. 6 (2008), 1306-1335.

[70] G. Muić and G. Savin, *Symplectic-orthogonal theta lifts of generic discrete series*, Duke Math. J. **101** (2000), no. 2, 317–333.

[71] S.-Y. Pan, *Depth preservation in local theta correspondence*, Duke Math. J. **113** (2002), no. 3, 531–592.

[72] I. Piatetski-Shapiro, *On the Saito-Kurokawa lifting*, Invent. Math. **71** (1983), 309–338.

[73] I. I. Piatetski-Shapiro and S. Rallis, *L-functions for the classical groups*, Explicit constructions of automorphic L-functions, Lecture Notes in Mathematics **1254**, Springer-Verlag, Berlin, 1987, pp. 1–52.

[74] D. Prasad, *On the local Howe duality correspondence*, Int. Math. Res. Not. (1993), 279–287.

[75] D. Prasad, *Theta correspondence for unitary groups*, Pacific J. Math. (2000), 427–438.

[76] S. Rallis, *Injectivity properties of liftings associated to Weil representations*, Compositio Math. **52** (1984), no. 2, 139–169.

[77] ———, *On the Howe duality conjecture*, Compositio Math. **51** (1984), 333–399.

[78] ———, *L-functions and the oscillator representation*, Lecture Notes in Mathematics, 1245. Springer-Verlag, Berlin, 1987. xvi+239 pp.

[79] B. Roberts, *Tempered representations and the theta correspondence*, Canadian Journal of Mathematics **50** (1998), 1105-1118.

[80] G. Savin, *Dual pair* $G_J \times PGL_2$*:* G_J *is the automorphism group of the Jordan algebra* J, Invent. Math. **118** (1994), 141–160.

[81] B. Sun, *Multiplicity one theorems for Fourier-Jacobi models*, Amer. J. Math. **134** (2012), 1655–1678.

[82] B. Sun and C.-B. Zhu, *Conservation relations for local theta correspondence*, arXiv: 1204.2969.

[83] P. Torasso, *Methode des orbites de Kirrilov-Duflo et representations minimales des groupes simples sur un corps local de caracteristisque nulle*, Duke Math. Journal **90** (1997), 261–378.

[84] J.-L. Waldspurger, *Correspondance de Shimura*, J. Math. Pures et Appl. **59** (1980), 1-133.

[85] _____, *Correspondances de Shimura et quaternions*, Forum Math. **3** (1991), no. 3, 219–307.

[86] J.-L. Waldspurger, *Démonstration d'une conjecture de dualité de Howe dans le cas p-adique, $p \neq 2$*, Festschrift in honor of I. I. Piatetski-Shapiro on the occasion of his sixtieth birthday, Part I, Israel Math. Conf. Proc. **2**, Weizmann, Jerusalem, 1990, pp. 267–324.

[87] _____, *Une formule intégrale reliée à la conjecture locale de Gross-Prasad*, Compos. Math. **146** (2010), 1180–1290.

[88] _____, *Une formule intégrale reliée à la conjecture locale de Gross-Prasad, 2ème partie: extension aux représentations tempérées*, Astérisque **346** (2012), 171–312.

[89] _____, *Calcul d'une valeur d'un facteur ϵ par une formule intégrale*, Astérisque **347** (2012), 1–102.

[90] _____, *La conjecture locale de Gross-Prasad pour les représentations tempérées des groupes spéciaux orthogonaux*, Astérisque **347** (2012), 103–165.

[91] A. Weil, *Sur certains groupes d'opérateurs unitaires*, Acta Math. **111** (1964), 143–211.

[92] _____, *Sur la formule de Siegel dans la theorie des groupes classiques*, Acta Math. **113** (1965) 1–87.

[93] S. Yamana, *On the Siegel-Weil formula: the case of singular forms*, Compositio Math. **147** (2011), 1003–1021.

[94] _____, *On the Siegel-Weil formula for quaternionic unitary groups*, Amer. J. Math. **135** (2013), no. 5, 1383–432.

[95] _____, *L-functions and theta correspondence for classical groups*, to appear in Invent. Math.

Mathematics Department, National University of Singapore, Block S17, 10 Lower Kent Ridge Road, Singapore 119076

E-mail: matgwt@nus.edu.sg

Automorphic Galois representations and the cohomology of Shimura varieties

Michael Harris

Abstract. The first part of this report describes the class of representations of Galois groups of number fields that have been attached to automorphic representations. The construction is based on the program for analyzing cohomology of Shimura varieties developed by Langlands and Kottwitz. Using p-adic methods, the class of Galois representations obtainable in this way can be expanded slightly; the link to cohomology remains indispensable at present. It is often possible to characterize the set of Galois representations that can be attached to automorphic forms, using the modularity lifting methods initiated by Wiles a bit over 20 years ago. The report mentions some applications of results of this kind. The second part of the report explains some recent results on critical values of automorphic L-functions, emphasizing their relation to the motives whose ℓ-adic realizations were discussed in the first part.

Mathematics Subject Classification (2010). Primary 11F80; Secondary 11F70, 11G18, 11F67.

Keywords. Galois representation, Shimura variety, special values of L-functions.

1. Introduction

Algebraic number theory has benefited immeasurably over the past four decades from the applications of the methods and results of the Langlands program to the study of Galois representations attached to automorphic forms. Yet Galois representations do not figure prominently in Langlands's original conjectures, apart from the complex Galois representations that are the object of the Artin conjecture. There seems to be no completely precise statement in the literature of a *Langlands reciprocity conjecture* – a bijection between representations of Galois groups with values in the ℓ-adic points of reductive groups, subject to certain natural restrictions (including a version of irreducibility), and of automorphic representations of related reductive groups – although number theorists believe there should be such a conjecture and have a general idea of how it should go. The best general account of this question is still contained in the expanded version [69] of Taylor's 2002 ICM talk.

The first objective of the present survey is to describe the results in the direction of reciprocity obtained since the publication of [69]. Construction of the correspondence in one direction – from automorphic representations to Galois representations – has progressed considerably, even in directions that could not have been expected ten years ago. All of the Galois representations associated to automorphic representations have been constructed, either directly or by p-adic interpolation, using the cohomology of Shimura varieties. This source of Galois representations has been or soon will be exhausted, and new methods will

Proceedings of the International Congress of Mathematicians, Seoul, 2014

need to be invented in order to find the Galois representations attached to automorphic representations that cannot be related in any way to cohomology of Shimura varieties, notably the representations of Galois groups of number fields that are not totally real nor CM.

Little was known at the time of [69] regarding the converse direction, the problem of proving that a given Galois representation ρ is attached to automorphic forms, when $\dim \rho > 2$. Now there is a mature theory of *automorphy lifting theorems*, in the spirit of the results developed by Wiles for his proof of Fermat's Last Theorem, applying in all dimensions. The attempt to complete this theory represents one of the most active branches of algebraic number theory, and is largely responsible for the rapid growth of interest in the p-adic local Langlands program.

Let K be a number field. The Galois group $\Gamma_K := Gal(\overline{\mathbb{Q}}/K)$ acts on the p-adic étale cohomology of an algebraic variety or motive M defined over K, and this action determines the L-function $L(s, M)$. Theoretical considerations guarantee that the p-adic Galois representations on the cohomology of most algebraic varieties cannot be realized in the cohomology of Shimura varieties; for example, the cohomology of a generic hypersurface cannot be obtained in this way. Present methods, therefore, cannot prove the analytic continuation of $L(s, M)$ for most motives arising from geometry. When the Galois representation is attached to an automorphic form, on the other hand, then so is $L(s, M)$, and this implies analytic (or at least meromorphic) continuation of the latter. Moreover, the conjectures concerned with the values at integer points of $L(s, M)$ (of Deligne, Beilinson, or Bloch-Kato) can be studied with the help of automorphic forms. Everything one knows in the direction of the Birch-Swinnerton-Dyer Conjecture, for example, has been proved by means of this connection. There has been a great deal of activity in this direction as well, especially in connection with the growth of the "relative" theory of automorphic forms (the relative trace formula and conjectures of Gan-Gross-Prasad, Ichino-Ikeda, and Sakellaridis-Venkatesh). The second part of this paper reviews some of the recent results on special values of L-functions.

The conjectures on special values of complex L-functions are accompanied by conjectures on the existence of p-adic analytic functions interpolating their normalized special values. The article concludes with a few speculative remarks about automorphic p-adic L-functions.

2. Automorphic forms and Galois representations

2.1. Construction of automorphic Galois representations. Class field theory classifies abelian extensions of a number field K in terms of the the structure of the idèle class group $GL(1, K)\backslash GL(1, \mathbb{A}_K)$. In doing so it also identifies 1-dimensional representations of Γ_K with continuous characters of the idèle class group. Non-abelian class field theory can be traced back to the 1950s, when Eichler and Shimura realized that 2-dimensional ℓ-adic Galois representations could be attached to classical cusp forms that are eigenvalues of the Hecke algebra. A conjectural classification of n-dimensional ℓ-adic Galois representations, in terms of the Langlands program, was formulated in Taylor's 2002 ICM talk (cf. [69]). We review this conjecture quickly. For any finite set S of places of K, let $\Gamma_{K,S}$ be the Galois group of the maximal extension of K unramified outside S. Taylor adopts the framework of Fontaine and Mazur, who restrict their attention in [25] to continuous representations $\rho : \Gamma_K \to GL(n, \overline{\mathbb{Q}}_\ell)$ satisfying the following two axioms:

1. ρ factors through $\Gamma_{K,S}$ for some finite set S of places of K (usually containing the primes dividing ℓ);
2. For all primes v of K of residue characteristic ℓ, the restriction of ρ to a decomposition group $G_v \subset \Gamma_K$ at v is *de Rham* in the sense of Fontaine.

A ρ satisfying these two conditions is either called *geometric* or *algebraic*, depending on the context. Condition (1) guarantees that, at all but finitely many primes v of K, the restriction ρ_v of ρ to a decomposition group G_v is determined up to equivalence, and up to semisimplification, by the characteristic polynomial $P_v(\rho, T)$ of the conjugacy class $\rho(Frob_v) \in GL(n, \overline{\mathbb{Q}}_\ell)$. One of the Fontaine-Mazur conjectures implies that there is a number field E such that all $P_v(\rho, T)$ have coefficients in E; by choosing an embedding $\iota : E \hookrightarrow \mathbb{C}$ we may thus define $P_v(\rho, T)$ as a polynomial of degree n in $\mathbb{C}[T]$ with non-vanishing constant term. The set of such polynomials is in bijection with the set of (equivalence classes of) irreducible smooth representations Π_v of $GL(n, K_v)$ that are *spherical*: the space of vectors in Π_v that are invariant under the maximal compact subgroup $GL(n, \mathcal{O}_v) \subset GL(n, K_v)$, where $\mathcal{O}_v$ is the ring of integers in K_v, is non-trivial and necessarily one-dimensional. We let $\Pi_v(\rho)$ be the spherical representation corresponding to $P_v(\rho, T)$.

An irreducible representation $\Pi_v(\rho)$ of $GL(n, K_v)$ can be attached to ρ for primes $v \in S$ as well. If v is not of residue characteristic ℓ, the restriction of ρ to G_v gives rise by a simple procedure to an n-dimensional representation $WD(\rho, v)$ of the Weil-Deligne group WD_v at v. The local Langlands correspondence [41, 43] is a bijection between n-dimensional representations of WD_v and irreducible smooth representations of $GL(n, K_v)$, and we obtain $\Pi_v(\rho)$ using this bijection. If v divides ℓ, condition (2) allows us to define $WD(\rho, v)$ by means of Fontaine's D_{pst} functor. Fontaine's construction also provides a set of Hodge-Tate numbers $HT(\rho, v)$ for each archimedean prime v. This datum, together with the action of a complex conjugation c_v in a decomposition group G_v when v is a real prime, defines an n-dimensional representation ρ_v of the local Weil group W_v, and thus an irreducible $(\mathfrak{g}_v, U_v)$-module $\Pi_v(\rho)$, where $\mathfrak{g}_v$ is the (complexified) Lie algebra of $G(K_v)$ and U_v is a maximal compact subgroup of $G(K_v)$. We let $\Pi(\rho)$ denote the restricted direct product (with respect to the $GL(n, \mathcal{O}_v)$-invariant vectors at finite primes outside S) of the $\Pi_v(\rho)$, as v ranges over all places of K.

If v is an archimedean place of K, the Harish-Chandra homomorphism identifies the center $Z(\mathfrak{g}_v)$ with the symmetric algebra of a Cartan subalgebra $\mathfrak{t}_v \subset \mathfrak{g}_v$. The maximal ideals of $Z(\mathfrak{g}_v)$ are in bijection with linear maps $Hom(\mathfrak{t}, \mathbb{C})$. The *infinitesimal character* of an irreducible $(\mathfrak{g}_v, U_v)$-module Π_v is the character defining the action of $Z(\mathfrak{g}_v)$ on Π_v; its kernel is a maximal ideal of $Z(\mathfrak{g}_v)$, and thus determines a linear map $\lambda_{\Pi_v} \in Hom(\mathfrak{t}_v, \mathbb{C})$. In [17], Clozel defines an irreducible $(\mathfrak{g}_v, U_v)$-module Π_v to be *algebraic* if λ_{Π_v} belongs to the lattice in $Hom(\mathfrak{t}_v, \mathbb{C})$ spanned by the highest weights of finite-dimensional representations. Denote by $|\bullet|_v$ the v-adic absolute value, $|\bullet|_{\mathbf{A}}$ the adele norm. The following corresponds to Conjectures 3.4 and 3.5 of [69].

Conjecture 2.1.

(1) *Let $\rho : \Gamma_K \to GL(n, \overline{\mathbb{Q}}_\ell)$ be an irreducible geometric Galois representation. Then the local component*

$$\Pi_v(\rho)\left(\frac{1-n}{2}\right) := \Pi_v(\rho) \otimes |\bullet|_v^{\frac{1-n}{2}} \circ \det$$

is algebraic at each archimedean prime v of K, and the representation $\Pi_v(\rho)$ of $GL(n, \mathbf{A}_K)$ occurs in the space of cusp forms on $GL(n, K)\backslash GL(n, \mathbf{A}_K)$.

(2) *Conversely, let Π be a cuspidal automorphic representation of $GL(n, \mathbf{A}_K)$. Suppose $\Pi_v(\frac{1-n}{2})$ is algebraic for every archimedean place v of K. Then for each prime ℓ, there exists an irreducible geometric n-dimensional representation*

$$\rho_{\ell,\Pi} : \Gamma_K \ \rightarrow \ GL(n, \overline{\mathbb{Q}}_\ell)$$

such that

$$\Pi\left(\frac{1-n}{2}\right) := \Pi \otimes |\bullet|_{\mathbf{A}}^{\frac{1-n}{2}} \circ \det \ \xrightarrow{\sim} \ \Pi(\rho_{\ell,\Pi}).$$

The Galois representations $\rho_{\ell,\Pi}$ are called *automorphic*.[1] Quite a lot is known about this conjecture when K is either a CM field or a totally real field, almost exclusively in the *regular* case, when λ_Π is the infinitesimal character of an irreducible finite-dimensional representation of $G(K_v)$ for all archimedean v. Let S be a finite set of primes of K, let ρ be an n-dimensional ℓ-adic representation of Γ_K, and say that Π and ρ_ℓ correspond away from S if $\Pi_v = \Pi_v(\rho)$ for $v \notin S$. The following theorem represents the current state of knowledge regarding part (b) of Conjecture 2.1; part (a) will be treated in the next section. In its details it may already be obsolete by the time of publication.

Theorem 2.2. *Let K be a CM field or a totally real field. Let Π be a cuspidal automorphic representation of $GL(n, \mathbf{A}_K)$. Suppose Π_v is algebraic and regular for every archimedean place v of K.*

(a) *Let S be the set of finite primes at which Π is ramified. If ℓ is a rational prime, let $S(\ell)$ denote the union of S with the set of primes of K dividing ℓ. For each prime ℓ, there exists a completely reducible geometric n-dimensional representation*

$$\rho_{\ell,\Pi} : \Gamma_K \ \rightarrow \ GL(n, \overline{\mathbb{Q}}_\ell)$$

such that $\Pi(\frac{1-n}{2})$ and $\rho_{\ell,\Pi}$ correspond away from $S(\ell)$.

(b) *Suppose Π is* polarized*, in the following sense:*

(1) *If K is a CM field,*

$$\Pi^\vee \ \xrightarrow{\sim} \ \Pi^c,$$

where ^{c} denotes the action of complex conjugation acting on K

(2) *If K is totally real,*

$$\Pi^\vee \ \xrightarrow{\sim} \ \Pi \otimes \omega$$

for some Hecke character ω of $GL(1, \mathbf{A}_K)$.

Here $^\vee$ denotes contragredient. Then there is a compatible family of n-dimensional representations $\rho_{\ell,\Pi}$ satisfying (b) of 2.1. Moreover, $\rho_{\ell,\Pi}$ is de Rham, in the sense of Fontaine, at all primes v dividing ℓ.

[1] When G is a reductive algebraic group, Buzzard and Gee have conjectured a correspondence between automorphic representations of G that satisfy an algebraicity condition at archimedean places and compatible systems of ℓ-adic representations with values in the Langlands L-group of G [9]. The relation of this conjecture with Conjecture 2.1 is a bit subtle; two different algebraicity conditions are relevant to the conjecture.

2.1.1. *p*-adic approximation. To forestall certain kinds of cognitive dissonance, we switch from ℓ-adic to p-adic representations in this section. Part (b) of Theorem 2.2 has been proved over the course of several decades by a number of people. For general n-dimensional representations, the most relevant references are [17, 41, 48] for work before the proof by Laumon and Ngô of the Fundamental Lemma; and [16, 19, 20, 51, 56, 62, 64] for results based on the Fundamental Lemma. I refer the reader to the discussion in [35], and take this opportunity to insist on the centrality of Labesse's results in [51] and earlier papers, which are inexplicably omitted from some accounts.[2]

Under the polarization hypothesis of case (b), most $\rho_{p,\Pi}$ are realized in the cohomology of Shimura varieties $S(G)$ attached to appropriate unitary groups G. Some important representations are nevertheless missing when n is even. To complete the proof of (b), the missing representations are constructed by *p-adic approximation*. One needs to show that Π is in some sense the limit of a sequence of Π_i that do satisfy the strong regularity hypothesis[3] For $n = 2$ two approximation methods had been applied: Wiles used the ideas due to Hida, while Taylor obtained the most complete results by adapting ideas of Ribet. In the intervening years, the theory of *eigenvarieties*, which originated in the work of Coleman and Mazur, had been developed to define p-adic famiies of automorphic forms in a very general setting. Chenevier's thesis [14] generalized the approximation method of Wiles to attach p-adic Galois representations of dimension $n > 2$ to non-ordinary Π, using eigenvarieties. Its extension in the book [6] with Bellaïche, and the subsequent article [15] were almost sufficient to construct the missing $\rho_{p,\Pi}$ as the limit of ρ_{p,Π_i} as above. The final steps in the construction, and the proofs of most of the local properties of 2.1, were carried out in [16], using a descent argument introduced by Blasius and Ramakrishnan in [8] and extended by Sorensen in [65]. The remaining local properties – determination of local ℓ-adic and p-adic monodromy of $\rho_{p,\Pi}$ were not known when [35] was written; they were obtained in most cases in [4] and completed in [12, 13].

Part (a) of Theorem 2.2 is much more recent. The first result of this type was obtained for $GL(2)$ over imaginary quadratic fields by Taylor in [67], following his joint work [40] with Soudry and the author; this was extended to general CM fields by Mok [54]. The proof of part (a) in [38] starts with an old idea of Clozel. Let K be a CM field and let $K^+ \subset K$ be the fixed field under complex conjugation. Let G_n be the unitary group of a $2n$-dimensional hermitian space over K, and assume G_n is quasi split. Then G_n, viewed by restriction of scalars as an algebraic group over $\mathbb{Q}$, contains a maximal parabolic subgroup P_n with Levi factor isomorphic to $R_{K/\mathbb{Q}}GL(n)_K$. Let $S(n, K)$ be the locally symmetric space attached to $GL(n, \mathbf{A}_K)$. Since K is a CM field, $S(n, K)$ is not an algebraic variety, and therefore its ℓ-adic cohomology does not carry a representation of any Galois group. If Π is a cuspidal automorphic representation of $GL(n, \mathbf{A}_K)$ that is polarized, then the twisted trace formula attaches to Π a collection (an L-packet) of automorphic representations of the unitary group G mentioned above; thus Π transfers to the cohomology of the $S(G)$, and this is where the Galois representation is realized (in nearly all cases).

When Π is not polarized, one uses the theory of Eisenstein series for the parabolic group

[2]Although complete base change from unitary groups remains to be established (the quasi-split case has recently been treated in [55]), Labesse proved the basic properties in the case of cohomological representations, without which the proof of Theorem 2.2 would have been impossible.

[3]Strictly speaking, the limits discussed here are taken relative to the Zariski topology on appropriate eigenvarieties, so the term "p-adic limit" would not be quite appropriate. In many cases the missing representations can indeed be obtained as actual limits in the p-adic topology, but as far as I know these cases have not been given an intrinsic characterization.

P_n to attach a family $E(s,\Pi)$ of automorphic representations of G_n, with $s \in \mathbb{C}$. Up to twisting Π by a positive integral power of the norm, we may assume $E(s,\Pi)$ is regular at 0 and write $E(\Pi) = E(0,\Pi)$. Then $E(\Pi)$ is also cohomological and (for nearly all positive integral twists) defines a non-trivial class in the cohomology of the Shimura variety $S(G_n)$ attached to (the unitary similitude group of) G_n. The realization in p-adic étale cohomology of this Eisenstein class then defines a p-adic Galois representation. However, it is easy to see that the semisimplification of this representation is a sum of abelian characters, and therefore it cannot be used to construct the desired $\rho_{p,\Pi}$.

Some years later, Skinner (and independently Urban) revived Clozel's idea by suggesting that $E(\Pi)$ might be realized as the limit in a p-adic family of a sequence of *cuspidal* cohomological automorphic representations τ_i of G_n. One then considers the collection of $2n$-dimensional representations ρ_{p,τ_i}. The symbol $\chi_{E(\Pi)} = \lim_i tr\rho_{p,\tau_i}$ then makes sense as a $\overline{\mathbb{Q}}_p$-valued function on $\Gamma_{K,S}$ for appropriate S, and because it is the limit of traces of genuine representations it defines a $2n$-dimensional *pseudorepresentation*. The latter notion is an abstraction of the invariance properties of the character of a representation, first constructed in the 2-dimensional case by Wiles, then defined by Taylor in general using results (especially results of Procesi) from invariant theory. Taylor's theory implies that $\chi_{E(\Pi)}$ is the character of a unique $2n$-dimensional representation, and by varying Π among its abelian twists it can be shown by elementary methods that $\chi_{E(\Pi)}$ breaks up as the sum of two n-dimensional pieces, one of which is the $\rho_{p,\Pi}$ of Theorem 2.2.

The hard part is to obtain $E(\Pi)$ as the limit of cuspidal τ_i. What this means is that the eigenvalues of Hecke operators at primes at which Π is unramified are p-adic limits of the corresponding Hecke eigenvalues on τ_i. In [38] this is achieved by realizing $E(\Pi)$ in a p-adic cohomology theory that satisfies a short list of desirable properties. The most important properties are (i) the global cohomology is computed as the hypercohomology in the (rigid) Zariski topology of the de Rham complex and (ii) the cohomology has a weight filtration, characterized by the eigenvalues of an appropriate Frobenius operator. The cohomology theory chosen in [38] is a version of Berthelot's rigid cohomology (generalizing Monsky-Washnitzer cohomology). This is calculated on the complement, in the *minimal* (Baily-Borel) compactification $S(G_n)^*$ of $S(G_n)$, of the vanishing locus of lifts (modulo increasing powers of p) of the Hasse invariant. This complement is affinoid and therefore by (i) the cohomology can be computed by a complex whose terms are spaces of p-adic modular forms, in the sense of Katz. By analyzing the finiteness properties of this complex, and using the density of genuine holomorphic modular forms in the space of p-adic modular forms, [38] writes $E(\Pi)$ as the limit of cuspidal τ_i, as required.

About a year after the results of [38] were announced, Scholze discovered a more flexible construction based on a very different cohomology theory, the p-adic étale cohomology of *perfectoid spaces*. The topological constructions in [38] can in principle also lift torsion classes in the cohomology of the locally symmetric space attached to $GL(n,\mathbf{A}_K)$ to torsion classes in the cohomology of $S(G_n)$, but rigid cohomology cannot detect torsion classes. The p-adic étale cohomology of perfectoid spaces does not have this defect, and Scholze's article [61] not only gives a new and more conceptual proof of the results of [38] but applies to torsion classes as well. Thus Scholze proved a long-standing conjecture, first formulated by Ash in [2], that has greatly influenced subsequent speculation on p-adic representations of general Galois groups. The reader is referred to Scholze's article in the current proceedings for more information about his results.

Restrictions on Galois representations on the cohomology of Shimura varieties. In part (b) of 2.2 the proof of the deepest local properties of the (polarized) $\rho_{p,\Pi}$ at primes dividing p were proved by finding representations closely related to $\rho_{p,\Pi}$ (the images under tensor operations) directly in the cohomology of Shimura varieties. When Π is not polarized, the $\rho_{p,\Pi}$ are still constructed in [38] and [61] by a limiting process , starting from a family of ρ_{p,Π_i} of geometric origin, but there is every reason to believe (see below) that the $\rho_{p,\Pi}$ and its images under tensor operations will almost never be obtained in the cohomology of Shimura varieties, and although they are expected to be geometric no one has the slightest idea where they might arise in the cohomology of algebraic varieties.

Room for improvement. The infinitesimal character $\lambda_\Pi \in Hom(\mathfrak{t}_v, \mathbb{C})$ is regular provided it is orthogonal to no roots of $\mathfrak{t}_v$ in $\mathfrak{g}_v$; in other words, if it is contained in the interior of a Weyl chamber. The regularity hypothesis in Theorem 2.2 can sometimes be relaxed to allow non-degenerate limits of discrete series, whose infinitesimal characters lie on one or more walls of a Weyl chamber. The first result of this type is the Deligne-Serre theorem which attaches (Artin) representations of $\Gamma_{\mathbb{Q}}$ to holomorphic modular forms of weight 1. This has recently been generalized by Goldring [28] to representations of $GL(n)$ obtained by base change from holomorphic limits of discrete series of unitary groups.

2.2. Reciprocity. Number theorists can't complain of a shortage of Galois representations. The étale cohomology of algebraic varieties over a number field K provides an abundance of ℓ-adic representations of Γ_K satisfying the two Fontaine-Mazur axioms. One of the Fontaine-Mazur conjectures predicts that any irreducible representation of Γ_K satistying these axioms is equivalent to a constituent of ℓ-adic cohomology of some (smooth projective) variety V over K. The reciprocity Conjecture 2.1 (a) has been tested almost exclusively for ρ arising from geometry in this way. The paradigmatic case in which $K = \mathbb{Q}$ and V is an elliptic curve was discussed in the ICM talks of Wiles (in 1994) and Taylor (in 2002).[4] The Fontaine-Mazur conjecture itself has been solved in almost all 2-dimensional cases when $K = \mathbb{Q}$ for ρ that take complex conjugation to a matrix with determinant -1. Two different proofs have been given by Kisin and Emerton; both of them take as their starting point the solution by Khare and Wintenberger of Serre's conjecture on 2-dimensional modular representations of $\Gamma_{\mathbb{Q}}$. All of these results are discussed in a number of places, for example in [24, 46, 47]. I will therefore concentrate on results valid in any dimension n.

Let $\rho : \Gamma_K \to GL(n, \mathcal{O})$ be a continuous representation with coefficients in an ℓ-adic integer ring $\mathcal{O}$ with maximal ideal $\mathfrak{m}$ and residue field k; let $\bar{\rho} : \Gamma_K \to GL(n, k)$ denote the reduction of ρ modulo $\mathfrak{m}$. We say ρ is *residually automorphic* if $\bar{\rho} \xrightarrow{\sim} \bar{\rho}_{\ell,\Pi}$ for some cuspidal automorphic representation Π of $GL(n, \mathbf{A}_K)$. The method for proving reciprocity initiated by Wiles consists in proving theorems of the following kind:

Theorem 2.3 (Modularity Lifting Theorem, prototypical statement). *Suppose $\bar{\rho}$ is residually automorphic. Then every lift of $\bar{\rho}$ to characteristic zero that satisfies axioms (1) and (2) of Fontaine-Mazur, as well as*

(1) *a polarization condition;*

(2) *conditions on the size of the image of $\bar{\rho}$ (typically including the hypothesis that $\bar{\rho}$ is*

[4]The nomenclature associated with the conjecture in this particular case, which predates the Fontaine-Mazur conjecture, is a matter of considerable sociological and philosophical interest.

absolutely irreducible); and

(3) *ramification conditions at primes dividing ℓ (typically including a regularity hypothesis)*

is automorphic. In particular, if ρ itself satisfies conditions (1), (2), *and* (3), *then ρ is automorphic.*

The method for proving such theorems is called the *Taylor-Wiles method* or the *Taylor-Wiles-Kisin method*, depending on context, and is named after its inventors in the setting when $n = 2$. The first theorems of this kind for arbitrary n were proved in [21, 70]. Together with the results of [39] they imply the Sato-Tate theorem for elliptic curves over $\mathbb{Q}$ with non-integral j-invariant (see below). Subsequent improvements have allowed for less restrictive conditions in (2) and (3). The following theorem of Barnet-Lamb, Gee, Geraghty, and Taylor [3] represents the current state of the art.

Theorem 2.4 (Modularity Lifting Theorem). *Let K be a CM field with totally real subfield K^+, and let $c \in Gal(K/K^+)$ denote complex conjugation. Let ρ be as in 2.3. Suppose $\ell \geq 2(n+1)$ and K does not contain a primitive ℓ-th root of* 1. *Suppose ρ satisfies axioms (1) and (2) of Fontaine-Mazur, as well as*

(1) *$\rho^c \xrightarrow{\sim} \rho^\vee \otimes \mu$, where μ is an ℓ-adic character of Γ_{K^+} such that $\mu(c_v) = -1$ for every complex conjugation c_v;*

(2) *The restriction of $\bar{\rho}$ to $\Gamma_{K(\zeta_\ell)}$ is absolutely irreducible; and*

(3) *For any prime v of K dividing ℓ the restriction ρ_v of ρ to the decomposition group Γ_v is potentially diagonalizable and is HT-*regular*: ρ_v has n distinct Hodge-Tate weights.*

Suppose ρ is residually automorphic. Then ρ is automorphic.

Remark 2.5. This is not the most general statement – there is a version of this theorem when K is totally real, and condition (2) can be replaced by adequacy.

Remark 2.6. The first novelty is the simplification of condition (2) on the image of $\bar{\rho}$: Thorne showed in [72] that the Taylor-Wiles-Kisin method works when the image of $\bar{\rho}$ is what he called *adequate*, and this condition is implied by the irreducibility condition (2) as long as $\ell \geq 2(n+1)$. The second novelty in 2.4 is the notion of potential diagonalizability. This is roughly the requirement that, after a finite base change, ρ_v, for v dividing ℓ, is crystalline and can be deformed in a moduli space of crystalline representations to a sum of characters. It is known that ρ_v in the Fontaine-Laffaille range (the setting of [21, 70]) and ordinary ρ_v (the setting of [5, 27]) are potentially diagonalizable, but the condition is more general. In particular, it is preserved under finite ramified base change, which allows for considerable flexibility.

2.3. Potential automorphy. The need to assume residual automorphy places important restrictions on the application of theorems on the model of 2.3 to reciprocity. For some applications, however, it is enough to know that a given ρ is *potentially* residually automorphic: that ρ becomes residually automorphic after base change to an unspecified totally real or CM Galois extension K'/K. One can then often use a modularity lifting theorem to prove that $\rho \mid_{\Gamma_{K'}}$ is automorphic, in other words that ρ is *potentially automorphic*. If ρ is attached to

a motive M, then $L(s,\rho) = L(s,M)$ is given by an Euler product that converges absolutely in some right half-plane. An application of Brauer's theorem on induced characters then implies that $L(s,\rho)$ has a *meromorphic* continuation to the entire plane, and moreover (by a theorem due to Shahidi and to Jacquet-Piatetski-Shapiro-Shalika) that $L(s,\rho)$ has no zeroes down to the right-hand edge of the critical strip.

Potential automorphy was introduced by Taylor in [68] in order to prove a potential version of the Fontaine-Mazur conjecture for 2-dimensional Galois representations. The method was generalized to higher dimensions in [39] and in subsequent work of Barnet-Lamb. The idea is the following. A theorem of the form 2.3 can be applied to an ℓ-adic ρ that is residually automorphic. But it can also be applied if $\rho = \rho_\ell$ is a member of a compatible family $\{\rho_{\ell'}\}$ of ℓ'-adic representations, where ℓ' varies over all primes, provided at least one ρ_{ℓ_1} in the family is known to be residually automorphic. It thus suffices to find a motive M of rank n such that

Hypothesis 2.7. *$\bar\rho_{\ell,M} \simeq \bar\rho$ and $\bar\rho_{\ell_1,M}$ is known a priori to be residually automorphic for some $\ell_1 \neq \ell$.*

Typically one assumes $\bar\rho_{\ell',M}$ is induced from an algebraic Hecke character. The motives used in [39] are the invariants M_t, under a natural group action, in the middle-dimensional cohomology of the $n-1$-dimensional hypersurfaces X_t with equation (depending on t, with $t^{n+1} \in \mathbb{P}^1 \setminus \{0,1,\infty\}$)

$$f_t(X_0,\dots,X_n) = (X_0^{n+1} + \cdots + X_n^{n+1}) - (n+1)tX_0\dots X_n = 0 \qquad (2.1)$$

This *Dwork family* of hypersurfaces was known to physicists for their role in the calculations that led to the formulation of the mirror symmetry conjectures [11]; and they were known to number theorists because Dwork had studied their cohomology in connection with p-adic periods.

The isomorphism class of X_t depends on t^{n+1} and one sees that their cohomology defines a hypergeometric local system over $\mathbb{P}^1 \setminus \{0,1,\infty\}$. Properties of this local system proved by a number of people, are used, together with a "local-global principle" due to Moret-Bailly, to find a t over a totally real (or CM) Galois extension K'/K such that M_t satisfies Hypothesis 2.7.

Applying the method of potential automorphy is not always automatic. One has to satisfy the conditions of Moret-Bailly's theorem as well as conditions (1), (2), and (3) of 2.3. More details can be found in [35] (which was written, however, before the simplifications of [72] and [3]). Here are a few applications:

Theorem 2.8. *Let $K = \mathbb{Q}$ and let Π be a cuspidal holomorphic automorphic representation of $GL(2)_{\mathbb{Q}}$ (attached to an elliptic modular form of weight $k \geq 2$, say) to which one can associate a compatible family of 2-dimensional ℓ-adic representations $\rho_{\ell,\Pi}$. Suppose Π is not obtained by automorphic induction from a Hecke character of an imaginary quadratic field. Then $Sym^n \rho_{\ell,\Pi}$ is potentially automorphic for all $n \geq 1$.*

This theorem was proved first when $k = 2$ in [21, 39, 70], assuming Π_v is a Steinberg representation for some v. This hypothesis was dropped, and was generalized to all k in [5]. It follows from the arguments of Serre in [63] and from the non-vanishing of $L(s, Sym^n\rho)$ mentioned above, that this implies the *Sato-Tate conjecture* for elliptic modular forms [5, 21, 39, 70]:

Theorem 2.9. *Let f be an elliptic modular newform of weight k for $\Gamma_0(N)$ (for some N), and assume the ℓ-adic Galois representations $\rho_{\ell,f}$ attached to f are not dihedral. For any prime p not dividing N, let $a_p(f)$ denote the eigenvalue of the normalized Hecke operator at p on f. Let $\tilde{a}_p(f) = a_p(f)/2p^{\frac{k-1}{2}}$, which is known to be a real number in the interval $[-1,1]$. As p varies, the $\tilde{a}_p(f)$ are equidistributed in $[-1,1]$ for the measure $\sqrt{1-t^2}dt$.*

In particular, if E is an elliptic curve over $\mathbb{Q}$ without complex multiplication, and $1+p-a_p(E)$ is the number of points of E over $\mathbb{F}_p$, then the numbers $a_p(E)/2p^{\frac{1}{2}}$ are equidistributed in in $[-1,1]$ for the measure $\sqrt{1-t^2}dt$.

The hypothesis that f has trivial nebentypus (is a form for $\Gamma_0(N)$) is unnecessary and was only included to allow for a simple statement. A version of 2.8 for Hilbert modular forms was proved by Barnet-Lamb, Gee, and Geraghty, and they derived the corresponding version of Theorem 2.9. All of these results were subsumed in the following theorem of Patrikis and Taylor [59], a strengthening of one of the main theorems of [3]:

Theorem 2.10. *Let K be totally real (resp. CM) and let $\{r_\lambda\}$ be a weakly compatible family of λ-adic representations of Γ_K (where λ runs over finite places of a number field M). Assume the r_λ are pure of fixed weight w (the Frobenius eigenvalues at an unramified place of norm q are Weil $q^{\frac{w}{2}}$-numbers); that they are HT-regular; and that they satisfy an appropriate polarization condition. Then there is a finite totally real (resp. CM) Galois extension K'/K over which the family becomes automorphic.*

The Hodge-Tate multiplicities of n-dimensional ℓ-adic representations realized on the cohomology of the Dwork family are at most 1; moreover, n has to be even, and each Hodge-Tate weight between 0 and $n-1$ occurs. Griffiths transversality implies that such a condition is inevitable when Hodge structures vary in non-trivial families. This appears to restrict the applicability of the Dwork family to proving potential automorphy. However, it was observed in [34], and more generally in [5], that it suffices to prove that a given $\rho_{\ell,\Pi}$ becomes automorphic after tensoring with the Galois representation obtained by induction from an automorphic Galois character attached to a Hecke character of an appropriate cyclic CM extension K'/K. This observation was applied in the proof of 2.9 and more systematically in [3], in both cases in order to replace the given Hodge-Tate weights of ρ by the set of weights adapted to the cohomology of the Dwork family.

Remark 2.11. Let f be as in Theorem 2.9 and Π the associated automorphic representation. Theorem 2.9 is equivalent to the assertion that, as p varies over primes unramified for $\rho_{\ell,\Pi}$, the conjugacy classes of $\rho_{\ell,f}(Frob_p)$, normalized so that all eigenvalues have complex absolute value 1, are equidistributed in the space of conjugacy classes of $SU(2)$. A version of the Sato-Tate conjecture can be formulated for a general motive M; $SU(2)$ is replaced by the derived subgroup of the compact real form of the Mumford-Tate group $MT(M)$ of M. In order to prove this conjecture for more complicated $MT(M)$ one would have to be able to prove the corresponding generalization of Theorem 2.8, with the symmetric powers replaced by the full set of equivalence classes of irreducible representations σ of $MT(M)^{der}$. But even if the ℓ-adic representation $\rho_{\ell,M}$ attached to M is HT-regular, $\sigma \circ \rho_{\ell,M}$ is generally not HT-regular, and thus cannot be obtained by Theorem 2.2. Thus one has no way to start proving potential automorphy of $\sigma \circ \rho_{\ell,M}$ once $MT(M)^{0,der}$ is of rank greater than 1.

2.3.1. p-adic realization of very general Galois representations. It was mentioned above that the proof of 2.2 is completed by a p-adic approximation argument. One says more gen-

erally that a p-adic representation $\rho : \Gamma_{K,S} \to GL(n, \overline{\mathbb{Q}}_p)$ for some S is *p-adically automorphic* if $\rho = \lim_i \rho_i$ (for example, in the sense of pseudo-representations, where the limit can be in the Zariski or in the p-adic topology), where each ρ_i is an automorphic Galois representation of $\Gamma_{K,S}$. The theory of eigenvarieties shows that p-adically automorphic Galois representations vary in p-adic analytic families. The representations $\rho_{p,\Pi}$ of 2.2 are HT-regular because Π is cohomological, but analytic families of p-adically automorphic Galois representations can specialize to representations that are Hodge-Tate but not regular, and to representations that are not Hodge-Tate at all.

One can ask whether a given ρ is p-adically automorphic. There are discrete obstructions; for example the set of ramified primes is finite in any p-adic family. There are also sign obstructions. The 2-dimensional Galois representations $\rho_{\ell,f}$ attached to an elliptic modular form f are *odd*: $\det \rho_{\ell,f}(c) = -1$ when c is complex conjugation. In other words, no representation ρ for which $\det \rho_{\ell,f}(c) = 1$ can be obtained in the cohomology of a Shimura variety. The signature of complex conjugation is constant on p-adic analytic families of Galois representations, and therefore represents an obstruction to realizing such an *even* representation as a p-adically automorphic representation.

However, the direct sum of two even representations does not necessarily have such a sign obstruction. Similar discrete invariants characterize p-adically automorphic Galois representations in higher dimension, but they can be made to vanish upon taking appropriate direct sums. Say ρ is *p-adically stably automorphic* if $\rho \oplus \rho'$ is p-adically automorphic for some ρ'. One knows what this means if K is a totally real or CM field. If not, let $K_0 \subset K$ be the maximal totally real or CM subfield, and say a p-adic representation ρ is *p-adically stably automorphic* if $\rho \oplus \rho'$ is the restriction to Γ_K of a p-adically automorphic representation of Γ_{K_0}.

Question 2.12. *Is every p-adic representation of Γ_K that satisfies the Fontaine-Mazur axioms stably p-adically automorphic?*

The main theorem of [30] states, roughly, that *every* p-adic representation of Γ_K is "stably potentially residually automorphic," where the reader is invited to guess what that means.

One can often define analytic or geometric invariants of p-adic families by interpolation of their specializations to automorphic points. Thus one defines p-adic L-functions or Galois cohomology (Selmer groups) of p-adic families. Specializations to points not known to be automorphic (e.g., because they are not HT-regular) define invariants of the corresponding Galois representations.

2.3.2. Prospects for improvement.

(a) Condition (1) in Theorem 2.4 corresponds to the polarization condition in (b) of Theorem 2.2. At present no one knows how to remove this condition and thus to prove the reciprocity conjecture for all representations constructed in Theorem 2.2 (see, however, the articles [10] of Calegari and Geraghty and [31] of Hansen). Removing condition (1) is sufficient, and probably necessary, to show that the $\rho_{\ell,\Pi}$ of Theorem 2.2 are irreducible for (almost) all ℓ.

(b) Although we have seen that substitutes can be found for residual irreducibility in applications to compatible families, it remains a major obstacle for many applications. In addition to the argument applied in Skinner-Wiles for 2-dimensional representations of $\Gamma_{\mathbb{Q}}$, Thorne has recently found a new method based on level raising [73].

(c) The article [70] replaces the very deep questions regarding congruences between automorphic forms of different levels ("level-raising", which an earlier version of [21] proposed to solve by generalizing Ihara's Lemma on congruences between elliptic modular forms) by a careful study of the singularities of certain varieties of tame representations of local Galois groups. But this comes at the cost of losing control of nilpotents in the deformation rings. In particular, current methods cannot classify liftings of $\bar{\rho}$ to rings in which ℓ is nilpotent. This may be important if one wants to extend the results of this section to the torsion representations constructed by Scholze.

(d) Dieulefait has expanded on the ideas used by Khare and Wintenberger to prove the Serre conjecture and has proved some astonishing results. For example, he has proved base change of elliptic modular forms to any totally real extension [23]. The methods of [46] and of [23] do not assume residual automorphy but actually prove it in the cases they consider. It is not yet known whether or not these methods can be applied in higher dimension.

(e) The authors of [3] ask whether every potentially crystalline representation is potentially diagonalizable. An affirmative answer would expand the range of their methods. The regularity hypothesis of Condition (3) seems insuperable for the moment. At most one can hope to prove reciprocity for representations like those constructed in [28], with Hodge-Tate multiplicities at most 2. The recent proof by Pilloni and Stroh of the Artin conjecture for (totally odd) 2-dimensional complex representations of Γ_K, when K is totally real, is the strongest result known in this direction. As long as one has no method for constructing automorphic Galois representations with Hodge-Tate multiplicities 3 or greater, the reciprocity question for such representations will remain inaccessible.

3. Critical values of automorphic L-functions

3.1. Critical values and automorphic motives. Let M be a (pure) motive of rank n over a number field K, with coefficients in a number field E. By restriction of scalars we can and will regard M as a motive of rank $n[K : \mathbb{Q}]$ over $\mathbb{Q}$. The values at integer points of the L-function $L(s, M)$ are conjectured to contain deep arithmetic information about M. If, for example, $M = M(A)$ is the motive attached to the cohomology in degree 1 of an abelian variety A, then the value, or more generally the first non-vanishing derivative, of $L(s, M(A))$ at $s = 1$ is predicted by the Birch-Swinnerton-Dyer conjecture. This is the only *critical value* of $L(s, M(A))$, in the sense of Deligne (the importance of critical values had previously been noted by Shimura). Deligne formulated his conjecture on critical values in one of his contributions to the 1977 Corvallis conference. We follow Deligne in working with motives for absolute Hodge cycles; thus M is by definition a collection of compatible realizations in the cohomology of smooth projective algebraic varieties. The realization in ℓ-adic cohomology gives the Galois representation $\rho_{\ell,M}$ on an ℓ-adic vector space M_ℓ, and therefore determines $L(s, M)$. Extension of scalars from $\mathbb{Q}$ to $\mathbb{C}$ makes M a motive over $\mathbb{C}$, whose cohomology is thus a direct factor of the cohomology of a complex manifold, whose topological cohomology is a $\mathbb{Q}$-vector space called M_B (Betti realization). Complex conjugation on the points of $M(\mathbb{C})$ acts on M_B as an involution F_∞. As a motive over $\mathbb{Q}$, M also has the algebraic de Rham cohomology, a $\mathbb{Q}$-vector space M_{dR} with a decreasing

Hodge filtration $\dots F^q M_{dR} \subset F^{q-1} M_{dR} \dots$ by $\mathbb{Q}$-subspaces. For any integer m let $M(m)$ denote the Tate twist $M \otimes \mathbb{Q}(m)$. Hodge theory defines comparison isomorphisms

$$I(m)_{M,\infty} : M(m)_B \otimes \mathbb{C} \xrightarrow{\sim} M(m)_{dR} \otimes \mathbb{C}.$$

This isomorphism does not respect the rational structures on the two sides. By restricting $I(m)_{M,\infty}$ to the $+1$-eigenspace of F_∞ in $M(m)_B$ and then projecting on a certain quotient $M(m)_{dR}/F^q M(m)_{dR} \otimes \mathbb{C}$, one defines an isomorphism between two complex vector spaces of dimension roughly half that of M, provided $M(m)$ is *critical* in Deligne's sense. The determinant of this isomorphism, calculated in rational bases of the two sides, is the *Deligne period* $c^+_{\mathbb{Q}}(M(m))$. It is a determinant of a matrix of integrals of rational differentials in M_{dR} over rational homology cycles, and is well defined up to $\mathbb{Q}^\times$-multiples. More generally, if M is a *motive with coefficients in* a number field E – in other words, if there are actions of E on each of the vector spaces M_B, M_{dR}, M_ℓ, compatible with the comparison isomorphisms – then there is a Deligne period $c^+_E(M(m))$ well-defined up to $E^\times$-multiples; moreover, $L(s, M)$ then defines an element of $E \otimes \mathbb{C}$, as in [22]. In the following discussion we will drop the subscript and just write $c^+(M(m))$ for the Deligne period with coefficients.

We call $s = m$ a *critical value* of $L(s, M)$ if $M(m)$ is critical. The set of critical m can be read off from the Gamma factors in the (conjectural) functional equation of $L(s, M)$ ([22], Definition 1.3). When $M = M(A)$, $s = 1$ is the only critical value. Deligne's conjecture is the assertion that

Conjecture 3.1 ([22]). *If m is a critical value of the motive M with coefficients in E, then*

$$L(m, M)/c^+(M(m)) \in E^\times.$$

Beilinson's conjectures express the non-critical integer values of $L(s, M)$ at non-critical integers in terms of the motivic cohomology (higher algebraic K-theory) of M. Automorphic methods give very little information about non-critical values of the L-functions of motives that can be related to automorphic forms, and this survey has nothing to say about them. On the other hand, the de Rham realizations of the motives that arise in the cohomology of Shimura varieties are given explicitly in terms of automorphic forms. One can therefore state versions of Deligne's conjecture for certain of these motives entirely in the language of automorphic forms.[5] The literature on special values of L-functions is vast and a book-length survey is long overdue. Automorphic versions of Deligne's conjecture represent a relatively small segment of the literature that is still too extensive for treatment in the space of this article. The proofs are generally quite indirect, not least because one can rarely write down M_B in terms of automorphic forms. When M is realized in the cohomology (with coefficients) of a Shimura variety $S(G)$, one can occasionally define non-trivial classes in M_B by projecting onto M the cycles defined by Shimura subvarieties $S(G') \subset S(G)$. Integrating differential forms on $S(G) \times S(G)$ over the diagonal cycle $S(G)$ amounts to computing a

[5] Strictly speaking, Deligne's conjecture only makes sense in the setting of a theory of motives that is the subject of very difficult conjectures. For example, one expects that if M and M' are motives such that the triples $(M_B, M_{dR}, I(m)_{M,\infty})$ and $(M'_B, M'_{dR}, I(m)_{M',\infty})$ are isomorphic, then M and M' are isomorphic as motives. This would follow from the Hodge conjecture. Similarly, one assumes that $L(s, M) = L(s, M')$ implies that $M \simeq M'$; this would follow from the Tate conjecture.

Blasius's proof of Deligne's conjecture for L-functions of Hecke characters of CM fields is carried out within the framework of motives for absolute Hodge cycles. It is practically the only authentically motivic result known in this direction.

cohomological cup product. In this brief account we limit our attention to a class of motives whose Deligne periods can be factored as products of cup products of this kind.

Suppose K is a CM field. As explained in 2.1.1, most of the representations $\rho_{\ell,\Pi}$ of Γ_K are realized in the cohomology of a Shimura varieties $S(G)$ attached to unitary groups G. Along with the n-dimensional Galois representation this construction yields a candidate for the rank n motive $M(\Pi)$. Originally $M(\Pi)$ is defined over K; one obtains a motive $RM(\Pi) = R_{K/\mathbb{Q}}M(\Pi)$ by restriction of scalars to $\mathbb{Q}$, taking into account the theorem of Borovoi and Milne on conjugation of Shimura varieties (the Langlands conjecture). The spaces $RM(\Pi)_{dR}$ and $RM(\Pi)_B$ satisfy analogues of conditions (1) and (3) of Theorem 2.4. The regularity condition (3) implies there is a set of integers $q_1 < q_2 < \cdots < q_n$ such that $\dim_E F^q RM(\Pi)_{dR}/F^{q+1}RM(\Pi)_{dR} = 1$ if and only if $q = q_i$ for some i, and the dimension is 0 otherwise. Here $E = E(\Pi)$ is the field of coefficients of $RM(\Pi)$ (more precisely, E is a finite product of number fields). We choose a non-zero $\mathbb{Q}$-rational E-basis ω_i of $F^{q_i}RM(\Pi)_{dR}/F^{q_{i+1}}RM(\Pi)_{dR}$, view ω_i as a (vector-valued) automorphic form on $G(\mathbb{Q})\backslash G(\mathbf{A})$, and let $Q_i(\Pi) =< \omega_i, \omega_i >$ denote its appropriately normalized L_2 inner product with itself.

Conjecture 3.2. *Up to multiplication by $E^\times$, each $Q_i(\Pi)$ depends only on the automorphic representation Π of $GL(n)$ and not on the realization in the cohomology of a Shimura variety.*

This conjecture is implied by the Tate conjecture. It has been verified in many cases for the (holomorphic) period $Q_1(\Pi)$. The author has partial results for general $Q_i(\Pi)$.

Given any motive M of rank n satisfying conditions (1) and (3) of 2.4 we can define invariants $Q_i(M)$ in the same way, and a determinant factor $q(M)$ (for this and what follows, see [32, 36], and section 4 of [29]). For any integer $0 \leq r \leq n$ we write

$$P_{\leq r}(M) = q(M)^{-1} \cdot \prod_{i \leq r} Q_i(M).$$

Let M' be a second motive of rank n', satisfying conditions (1) and (3) of 2.4. Then for any integer m critical for $R_{K/\mathbb{Q}}(M \otimes M')$ there is a factorization (cf. [29] (4.11)):

$$c^+(R(M \otimes M')(m)) \sim (2\pi i)^{c(m,n,n')} \prod_{r=1}^{n} P_{\leq r}(M)^{a_r} \prod_{r'=1}^{n'} P_{\leq r'}(M')^{b_{r'}} \tag{3.1}$$

where $\sim$ means that the ratio of the two sides lies in the multiplicative group of the coefficient field, $c(m,n,n')$ is an explicit polynomial in m and the dimensions, $0 \leq a_r := a(r, M, M'), b_{r'} := b(r', M, M')$ and

$$\sum_r a_r \leq n'; \quad \sum_{r'} b_{r'} \leq n.$$

Defining Π as above, there is an (ad hoc) determinant factor $q(\Pi)$, and we let

$$P_{\leq r}(\Pi) = q(\Pi)^{-1} \cdot \prod_{i \leq r} Q_i(\Pi).$$

An automorphic version of Deligne's conjecture is

Conjecture 3.3. *Let Π and Π' be cuspidal automorphic representations of $GL(n)_K$ and $GL(n')_K$, satisfying the hypotheses of Theorem 2.2 (b). Let m be a critical value of*

$$L(s, R_{K/\mathbb{Q}}(M(\Pi) \otimes M(\Pi')) = L(s - \frac{n+n'-2}{2}, \Pi \times \Pi').$$

Then

$$L(m, R_{K/\mathbb{Q}}(M(\Pi) \otimes M(\Pi'))) \sim (2\pi i)^{c(m,n,n')} \prod_{r=1}^{n-1} P_{\leq r}(\Pi)^{a_r} \prod_{r'=1}^{n-2} P_{\leq r'}(\Pi')^{b_{r'}},$$

with a_r, $b_{r'}$ as in (3.1)*.*

The integers a_r and $b_{r'}$ of (3.1) are determined purely by the relative position of the Hodge decompositions of $M_{dR} \otimes \mathbb{C}$ and $M'_{dR} \otimes \mathbb{C}$ (and don't depend on m). Suppose $M = RM(\Pi)$, $M' = RM(\Pi')$, with Π and Π' as in (3.3). The regularity hypotheses imply that there are finite-dimensional representations $W(\Pi_\infty)$ and $W'(\Pi'_\infty)$ of $GL(n)_K$ and $GL(n')_K$, respectively, such that Π_∞ and $W(\Pi_\infty)$ (resp. Π'_∞ and $W'(\Pi'_\infty)$) have the same infinitesimal characters. The a_i and $b_{i'}$ can be computed explicitly in terms of the highest weights of $W(\Pi_\infty)$ and $W'(\Pi'_\infty)$. For example, suppose $n' = n-1$ and

$$Hom_{GL(n-1,K\otimes\mathbb{C})}(W(\Pi_\infty) \otimes W(\Pi'_\infty), \mathbb{C}) \neq 0. \tag{3.2}$$

Then $a_i = b_{i'} = 1, 1 \leq i \leq n-1; 1 \leq i' \leq n-2$; $a_n = b_{n-1} = 0$.

Theorem 3.4. *Suppose K is an imaginary quadratic field. Let Π and Π' be as in 3.3. Suppose moreover that the infinitesimal characters of Π_∞ and Π'_∞ satisfy 3.2 and are sufficiently regular. Then there are constants $c'(m, \Pi_\infty, \Pi'_\infty)$ such that*

$$L(m, R_{K/\mathbb{Q}}(M(\Pi) \otimes M(\Pi')))/[c'(m, \Pi_\infty, \Pi'_\infty) \prod_{r=1}^{n-1} P_{\leq r}(\Pi) \prod_{r'=1}^{n-2} P_{\leq r'}(\Pi')] \in \overline{\mathbb{Q}} \tag{3.3}$$

for every critical value m.

This is a reinterpretation of Theorem 1.2 of [29]. There the invariants $P_{\leq r}(\Pi)$ are replaced by complex numbers $P^{(r)}(\Pi)$, which are Petersson square norms of *holomorphic* automorphic forms on unitary Shimura varieties of different signatures (and it is shown that the quotient in (3.3) lies in a specific number field). Naturally one expects the constants $c'(m, \Pi_\infty, \Pi'_\infty)$ to be powers of $2\pi i$. The Tate conjecture implies an identity between the two kinds of invariants, and this has been proved (up to unspecified archimedean factors, and up to $\overline{\mathbb{Q}}$-multiples) in [33] (and subsequent unpublished work).

The methods of [29] are based on interpreting the Rankin-Selberg integral for $GL(n) \times GL(n-1)$ as a cohomological cup product. Such arguments have been used previously by Mahnkopf and Raghuram; see [60] for the most general results in this direction. Earlier results on this problem were conditional on the conjecture that certain archimedean zeta integrals did not vanish identically. Sun's recent proof of this conjecture [66] has revived interest in the problem and one can expect rapid progress in the next few years. For general number fields one does not have the analogues of the invariants $P_{\leq r}(\Pi)$ and the results of [60] are expressed in terms of period invariants obtained by comparing the cohomological rational structure of Π with one defined by Whittaker models. The (mild) regularity hypothesis of 3.4

is required in the comparison of these Whittaker period invariants with the motivic invariants $P_{\leq r}(\Pi)$. Similar arguments should suffice to treat the cases of Conjecture 3.3 for $n' \leq n-1$ that satisfy an analogue of (3.2), for general CM fields. (The case where $n' = 1$ was treated by the author in a series of papers, starting with [32], and is used crucially in the proof of Theorem 3.4.) The full scope of the methods of [29] is not yet clear, but it is certain that it is not limited to the situation of (3.2). The identification of $c'(m, \Pi_\infty, \Pi'_\infty)$ with the invariant $(2\pi i)^{c(m,n,n-1)}$ is likely to follow from these methods as well.[6]

3.2. How general are these results? Only a restricted class of Galois representations can be obtained using the cohomology of Shimura varieties, and only those that can be realized directly in the cohomology are associated to motives that admit an automorphic interpretation. The Rankin-Selberg L-functions described in the previous section, along with a few related constructions (symmetric and exterior squares and adjoint L-functions), seem to be the only ones whose critical values can be analyzed by automorphic methods. Raghuram's results in [60] apply only under the hypothesis (3.2). It should be straightforward to generalize his methods to pairs Π, Π' where Π is cuspidal and Π' is an essentially tempered cohomological Eisenstein series, as in [29] (or earlier work of Mahnkopf). If Raghuram's results could be extended to cases where neither Π nor Π' is cuspidal, then the hypothesis (3.2) would be superfluous (in Theorem 3.4 as well).

A motivic analysis of critical values of Rankin-Selberg L-functions, as in Theorem 3.4, has thus far only been carried out for CM fields. Bhagwat has proved an analogue of the relation (3.1) when $K = \mathbb{Q}$, following earlier work of Yoshida (see the appendix to [60]) and similar factorizations must hold for totally real fields. As far as I know, no one has proposed automorphic interpretations of the terms that occur in Bhagwat's factorization. For Π satisfying the polarization condition as in (b) of Theorem 2.2 it should be possible to interpret some of them as periods of motives realized in the cohomology of Shimura varieties attached to special orthogonal groups of signature $(2, n)$. In the absence of a polarization condition, Shimura varieties seem to be of no help.

3.3. Exact formulas for the central critical value. The conjectures of Bloch-Kato and Fontaine-Perrin-Riou give exact formulas for special values of motivic L-functions. The algebraic quotients $L(m, M)/c^+(M(m))$ and their generalizations to non-critical values are expressed explicitly as products of local and global algebraic factors defined in terms of Galois cohomology. For the central critical value these expressions generalize the Birch-Swinnerton-Dyer conjecture for the value at $s = 1$ of $L(s, M(A))$, in the notation of the previous section.

Beginning with the thesis of Waldspurger, exact formulas have also been found for certain central values of automorphic L-functions. The conjecture of Ichino-Ikeda, and its version for unitary groups formulated by N. Harris, [42, 45] give exact formulas for central values in the framework of the Gan-Gross-Prasad conjectures [26]. In what follows K is a CM field. We change notation and let Π denote a cuspidal automorphic representation of $GL(n)_K$ that descends to a (cuspidal) L-packet $P_{\Pi,V}$ of a given $G = U(V)$, viewed as group over K^+, with $\dim V = n$. Similarly, Π' is an automorphic representation of $GL(n-1)_K$ obtained by base change from a (cuspidal) L-packet $P_{\Pi',V'}$ of $G' = U(V')$. It

[6] Note added in proof. This has now been carried out, at least when the coefficients are sufficiently regular, by Lin Jie.

is assumed that V' embeds in V as a non-degenerate hermitian subspace of codimension 1. For any $\pi \in P_{\Pi,V}$ and $\pi' \in P_{\Pi',V'}$, the pairing

$$I = \pi \otimes \pi' \ \to \ \mathbb{C} : \ \ f \otimes f' \mapsto \int_{G'(K^+)\backslash G'(\mathbf{A})} f(g')f'(g')dg', \ f \in \pi, \ f' \in \pi' \qquad (3.4)$$

is invariant under the diagonal action of $G'(\mathbf{A})$. One of the Gan-Gross-Prasad conjectures asserts that the space of such invariant pairings is of dimension 1 for exactly one pair (V, V') and one pair $(\pi, \pi') \in P_{\Pi,V} \times P_{\Pi',V'}$, and that the lucky pair is identified by a complicated formula involving root numbers. The non-archimedean part of this conjecture has been proved by R. Beuzart-Plessis, following the method used by Waldspurger to solve the analogous conjecture for special orthogonal groups [7, 74]. Thus if one fixes a non-trivial pairing $B : \pi \otimes \pi' \ \to \ \mathbb{C}$, the pairing I defined in 3.4 is a multiple of B. The Ichino-Ikeda Conjecture can be seen as a determination of this multiple. In the statement of the conjecture, the superscript $^\vee$ denotes contragredient; all integrals are taken with respect to Tamagawa measure.

Conjecture 3.5 ([45]). *Let $f \in \pi$, $f' \in \pi'$, $f^\vee \in \pi^\vee$, $f'^{,\vee} \in \pi'^{,\vee}$, and suppose all four vectors are factorizable. Then*

$$\frac{I(f, f') \cdot I(f^\vee, f'^{,\vee})}{< f, f^\vee >_2 < f', f'^{,\vee} >_2} = 2^{-r} \prod_{v \in S} Z_v(f, f', f^\vee, f'^{,\vee}) \cdot \Delta \cdot \frac{L(\frac{1}{2}, \Pi \times \Pi')}{L(1, \pi, Ad)L(1, \pi', Ad)}.$$

Here $< \bullet, \bullet >_2$ are the L_2 pairings, the factor 2^{-r} is trivial when Π and Π' are cuspidal but not in general, S is the set of ramified primes for π, π', and the chosen vectors, including archimedean primes, the Z_v for $v \in S$ are normalized integrals of matrix coefficients attached to the data, Δ is a special value of a finite product of abelian L-functions (the L-function of the Gross motive), the numerator on the right-hand side is the Rankin-Selberg product for $GL(n) \times GL(n-1)$, and the factors in the denominator are the Langlands L-functions for G and G' attached to the adjoint representations of their L-groups.

Here and elsewhere, $L(s, \bullet)$ denotes the non-archimedean Euler product. The L-functions in the right-hand side are given the unitary normalization. Thus the completed L-function $\Lambda(s) = L_\infty(s, \Pi \times \Pi') \cdot L(s, \Pi \otimes \Pi')$ in the numerator of the right-hand side always satisfies $\Lambda(s) = \pm\Lambda(1-s)$. When Π and Π' satisfy (b) of 2.2, however, there is a second (motivic) normalization as well, in which the value $s = \frac{1}{2}$ is replaced by an integer value, and all the values of L-functions that occur in the right-hand side are critical.

Conjecture 3.5 is of no interest when the sign is -1, because the numerator vanishes trivially. When the L-function is motivic, there have been proposals for an arithmetic substitute for the conjecture in this case, with $L(\frac{1}{2}, \bullet)$ replaced by its derivative at $s = \frac{1}{2}$, along the lines of the Gross-Zagier conjecture and subsequent work. When the sign is $+1$, the conjecture refines the global Gan-Gross-Prasad conjecture, which asserts that $L(\frac{1}{2}, \Pi \times \Pi') = 0$ if and only if the pairing I of 3.4 is trivial.

When $L(\frac{1}{2}, \Pi \times \Pi') \neq 0$, Conjecture 3.5 gives an exact expression for its value, provided one can make good choices of the test vectors $f, f', f^\vee, f'^{,\vee}$ and can control the local zeta integrals. It is natural to speculate that these zeta integrals can be interpreted in terms of local Galois cohomological information, and that when Π and Π' are attached to motives, the expressions on the two sides of Conjecture 3.5 can be matched termwise with corresponding expressions in the Bloch-Beilinson and Bloch-Kato conjectures. The local factor

$Z_v(f, f', f^\vee, f'^{,\vee})$ is the integral of the matrix coefficient of π_v attached to the pair $(f_v, f_v^\vee)$ against the matrix coefficient of π'_v attached to $(f'_v, f'^{,\vee}_v)$. The following question is deliberately vague.

Question 3.6. *For any given pair of local (ramified) representations π_v, π'_v, is there a quadruple $f_v, f'_v, f_v^\vee, f'^{,\vee}_v$ such that the local zeta integral $Z_v(f, f', f^\vee, f'^{,\vee})$ exactly equals the local Galois-cohomological factor in the Bloch-Kato conjecture?*

As explained in [37], the expressions on the left-hand side are algebraic multiples of invariants called *Gross-Prasad periods* that depend only on Π and Π', provided the test vectors are chosen to be rationally normalized (with respect to coherent cohomology). The denominators are closely related to the $P_{\leq r}$ defined above. Combining Conjecture 3.5 with Conjecture 3.3, one gets conjectural expressions for the Gross-Prasad periods as well in terms of $P_{\leq r}(\pi)$ and $P_{\leq r'}(\pi')$; see [37], Conjecture 5.16.

In order to compare the local terms of Conjecture 3.5 with the Galois-cohomological data of the Bloch-Kato conjecture, *integral* normalizations of the test vectors are needed. It is well known, however, that even the module of elliptic modular forms with integral modular Fourier coefficients is not spanned by Hecke eigenfunctions. This is the phenomenon of *congruences* between Hecke eigenvalues for different automorphic representations, which is the subject of theorems of the form 2.3, and it is no less relevant to automorphic representations of groups other than $GL(2)$.

3.3.1. Adjoint L-functions. The denominator of the Ichino-Ikeda formula is relevant to the problem of integral normalization of test vectors. The point $s = 1$ is the only critical value of the adjoint L-functions that occur there. Suppose π has an associated motive $M(\Pi) = M(\pi)$. Then for any prime ℓ, the Bloch-Kato conjecture identifies the ℓ-adic valuation of the quotient of $L(1, \pi, Ad)$ by an (integrally normalized) Deligne period with the order of a Galois cohomology group that is supposed to count the number of ℓ-adic deformations of the residual Galois representation $\bar\rho_{\ell,\pi}$. When $n = 2$ and K is totally real, a version of this conjecture has been proved by Diamond-Flach-Guo and Dimitrov, combining the methods of Theorem 2.4 with the results of [44].

Hida's paper [44] was the starting point for his theory of families of modular forms, and was the first to establish a relation between the critical value of the adjoint L-function and congruences between modular forms. In dimension $n > 2$, the special cases of the Ichino-Ikeda conjecture proved by Wei Zhang in [75] are used in [29] to relate the Whittaker period of a Π satisfying (b) of Theorem 2.2 to $L(1, \pi, Ad)$, up to rational multiples. One hopes this provides a starting point for determining $L(1, \pi, Ad)$ up to units in number fields, as required by the Bloch-Kato conjecture.

3.4. Two speculative remarks on automorphic p-adic L-functions.

Remark 3.7. Deligne's conjecture is the starting point of the construction of p-adic L-functions. The algebraic values on the left-hand side of the identify in 3.1, suitably normalized, are predicted to extend analytically whenever M and m vary in p-adic families. The literature is vast but fragmentary, and the author's ongoing project with Eischen, Li, and Skinner will only add one (rather bulky) fragment to the collection when it is finished. Current plans are limited to *ordinary* (Hida) families, but ultimately one expects the method to extend to completely general families. In particular, such p-adic L-functions could be

specialized to the "very general" p-adic representations of 2.3.1. Moreover, using Brauer induction, one could even attach a p-adic L-functions to a motivic Galois representation $\rho_{p,M}$ that is potentially p-adically automorphic. Although such a function would have no obvious connection to the complex L-function of M, it could conceivably be related to the Galois cohomology of $\rho_{p,M}$.

Remark 3.8. One can study the behavior of the right-hand side of Conjecture 3.5 when Π and Π' vary in p-adic families. Given the right choice of data in the local zeta integrals at primes dividing p, the result should be a p-adic meromorphic function of Π and Π'. Can this function be constructed directly on the left-hand side of the identity?

Acknowledements. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 290766 (AAMOT). I thank Matt Emerton, Toby Gee, Mark Kisin, and Gaëtan Chenevier for their careful reading of earlier versions of this manuscript.

References

[1] J. Arthur, *The Endoscopic Classification of Representations: Orthogonal and Symplectic Groups*, Colloquium Publications, Providence: AMS (2013).

[2] A. Ash, *Galois representations attached to mod p cohomology of* $GL(n, \mathbb{Z})$, Duke Math. J. **65** (1992), 235–255.

[3] T. Barnet-Lamb, T. Gee, D. Geraghty, and R. Taylor, *Potential automorphy and change of weight*, Ann. of Math., in press.

[4] _____, *Local-global compatibility for l=p: I*, Ann. de Math. de Toulouse **21** (2012), 57–92. ; II, Ann. Math. ENS, in press.

[5] T. Barnet-Lamb, D. Geraghty, M. Harris, and R. Taylor, *A family of Calabi-Yau varieties and potential automorphy II*, Proc. RIMS, **47** (2011), 29–98.

[6] J. Bellaïche, and G. Chenevier, *Families of Galois representations and Selmer groups*, Astérisque **324** (2009).

[7] R. Beuzart-Plessis, *La conjecture locale de Gross-Prasad pour les représentations tempérées des groupes unitaires*, Thesis, Paris (2013).

[8] D. Blasius and D. Ramakrishnan, *Maass forms and Galois representations. Galois groups over* $\mathbb{Q}$, MSRI Publ. **16**, New York: Springer (1989) 33–77.

[9] K. Buzzard and T. Gee, *The conjectural connections between automorphic representations and Galois representations*, Proceedings of the LMS Durham Symposium 2011 (to appear).

[10] F. Calegari and D. Geraghty, *Modularity Lifting Theorems beyond the Taylor-Wiles Method. I, II*, preprints (2012).

[11] P. Candelas, X. C. De La Ossa, P. S. Green, and L. Parkes. *A pair of Calabi-Yau manifolds as an exactly soluble superconformal theory*, Nuclear Physics B **359** (1991) 21–74.

[12] A. Caraiani, *Local-global compatibility and the action of monodromy on nearby cycles*, Duke Math. J. **161** (2012) 2311–2413.

[13] ———, *Monodromy and local-global compatibility for $\ell = p$*, preprint (2012).

[14] G. Chenevier, *Familles p-adiques de formes automorphes pour GL_n*, J. Reine Angew. Math. **570** (2004) 143–217.

[15] ———, *Une application des variétés de Hecke des groupes unitaires*, preprint (2009), available at http://www.math.polytechnique.fr/chenevier/pub.html.

[16] G. Chenevier and M. Harris, *Construction of automorphic Galois representations II*, Cambridge Math. J. **1** (2013) 53–73.

[17] L. Clozel, *Motifs et formes automorphes: applications du principe de fonctorialité*, L. Clozel and J.S. Milne, eds., Automorphic forms, Shimura varieties, and L–functions I, (Ann Arbor 1988), Perspectives in Math. **10** (1990) 77–159.

[18] ———, *Représentations Galoisiennes associées aux représentations automorphes autoduales de GL(n)*, Publ. Math. I.H.E.S. **73** (1991) 97–145.

[19] L. Clozel, M. Harris, and J.-P. Labesse, *Endoscopic transfer*, in On the stabilization of the trace formula, Stab. Trace Formula Shimura Var. Arith. Appl., 1, Int. Press, Somerville, MA (2011), 475–496.

[20] ———, *Construction of automorphic Galois representations*, I, in On the stabilization of the trace formula, Stab. Trace Formula Shimura Var. Arith. Appl., 1, Int. Press, Somerville, MA (2011), 497–527.

[21] L. Clozel, M. Harris, and R. Taylor, *Automorphy for some ℓ-adic lifts of automorphic mod l Galois representations*, Publ. Math. I.H.E.S. **108** (2008), 1–181.

[22] P. Deligne, *Valeurs de fonctions L et périodes d'intégrales*, With an appendix by N. Koblitz and A. Ogus in: Proc. Sympos. Pure Math., Vol. XXXIII, part II, AMS, Providence, R.I., (1979), pp. 313–346.

[23] L. Dieulefait, *Langlands base change for GL(2)*, Ann. of Math. **176** (2012) 1015–1038.

[24] M. Emerton, *Local-global compatibility in the p-adic Langlands programme for $GL_2/\mathbb{Q}$*, preprint March 23, 2011.

[25] J.-M. Fontaine and B. Mazur, *Geometric Galois Representations*, in Elliptic curves, modular forms, and Fermat's last theorem (Hong Kong 1993), Internat. Press, Cambridge MA, (1995), 41–78.

[26] W. T. Gan, B. Gross, and D. Prasad, *Symplectic local root numbers, central critical L-values and restriction problems in the representation theory of classical groups*, Astérisque **346**, (2012), 1–110.

[27] D. Geraghty, *Modularity lifting theorems for ordinary Galois representations*, Harvard University Thesis, (2010).

[28] W. Goldring, *Galois Representations Associated to Holomorphic Limits of Discrete Series I*, Unitary Groups, Compositio Math., in press.

[29] H. Grobner and M. Harris, *Whittaker periods, motivic periods, and special values of tensor product L-functions*, preprint (2013).

[30] R. Guralnick, M. Harris, and N. M. Katz, *Automorphic realization of residual Galois representations*, J. Eur. Math. Soc. **12** (2010), 915–937.

[31] D. Hansen, *Minimal modularity lifting for GL_2 over an arbitrary number field*, Math. Res. Letters (in press).

[32] M. Harris, *L-functions and periods of polarized regular motives*, J. Reine Angew. Math. **483** (1997), 75–161.

[33] ———, *Cohomological automorphic forms on unitary groups, II period relations and values of L- functions*, in Harmonic Analysis, Group Representations, Automorphic Forms and Invariant Theory, **12**, Lecture Notes Series, Institute of Mathematical Sciences, National University of Singapore, (2007), 89–150.

[34] ———, *Potential automorphy of odd-dimensional symmetric powers of elliptic curves and applications, Algebra, arithmetic, and geometry*, in honor of Yu. I. Manin. Vol. II, Progr. Math., vol. 270, Birkhaüser Boston Inc., Boston, MA, 2009, 1–21.

[35] ———, *Arithmetic applications of the Langlands program*, Japanese J. Math., 3rd series **5**, (2010), 1–71.

[36] ———, *L-functions and periods of adjoint motives*, Algebra and Number Theory **7** (2013), 117–155.

[37] ———, *Beilinson-Bernstein Localization over $\mathbb{Q}$ and Periods of Automorphic Forms*, IMRN **2013** (2013), 2000–2053.

[38] M. Harris, K.-W. Lan, R. Taylor, and J. Thorne, *On the Rigid Cohomology of Certain Shimura Varieties*, preprint (2013).

[39] M. Harris, N. Shepherd-Barron, and R. Taylor, *A family of Calabi-Yau varieties and potential automorphy*, Ann. of Math. **171** (2010), 779–813.

[40] M. Harris, D. Soudry, and R. Taylor, *ℓ-adic representations associated to modular forms over imaginary quadratic fields. I. Lifting to $GSp4(\mathbb{Q})$*, Invent. Math. **112** (1993), 377–411.

[41] M. Harris and R. Taylor, *The geometry and cohomology of some simple Shimura varieties*, Ann. of Math. Studies **151** (2001).

[42] R. N. Harris, *The Refined Gross-Prasad Conjecture for Unitary Groups*, Int. Math. Res. Not. (2012) doi: 10.1093/imrn/rns219.

[43] G. Henniart, *Une preuve simple des conjectures de Langlands pour $GL(n)$ sur un corps p-adique*, Inventiones Math. **139** (2000), 439–455.

[44] H. Hida, *Congruence of cusp forms and special values of their zeta functions*, Invent. Math. **63** (1981), 225–261.

[45] A. Ichino and T. Ikeda, *On the periods of automorphic forms on special orthogonal groups and the Gross-Prasad conjecture*, Geom. Funct. Anal. **19** (2010), 1378–1425.

[46] C. Khare and J.-P. Wintenberger, *Serre's modularity conjecture*, Proceedings of the International Congress of Mathematicians, Volume II, Hindustan Book Agency, New Delhi, (2010), 280–293.

[47] M. Kisin, *The Fontaine-Mazur conjecture for GL_2*, J. AMS. **22** (2009), 641–690.

[48] R. Kottwitz, *On the λ-adic representations associated to some simple Shimura varieties*, Invent. Math. **108** (1992), 653–665.

[49] ———, *Shimura varieties and λ-adic representations*, in L. Clozel and J.S. Milne, eds., Automorphic forms, Shimura varieties, and L–functions I, (Ann Arbor 1988), Perspectives in Math. **10** (1990), 161–209.

[50] ———, *Points on some Shimura varieties over finite fields*, J. AMS. **5** (1992), 373–444.

[51] J.-P. Labesse, *Changement de base CM et séries discrètes*, in: On the Stabilization of the Trace Formula, Vol. I, eds., L. Clozel, M. Harris, J.-P. Labesse, and B.-C. Ngô, International Press, Boston, MA, 2011, pp. 429–470.

[52] R. P. Langlands, *Automorphic representations, motives, and Shimura varieties. Ein Märchen*, Proc. Symp. Pure Math. **33**, Part 2, (1979), 205–246.

[53] R. P. Langlands and D. Ramakrishnan, eds., *The zeta functions of Picard modular surfaces*, Université de Montréal, Centre de Recherches Mathématiques, Montreal (1992).

[54] C. P. Mok, *Galois representations attached to automorphic forms on GL_2 over CM fields*, Compositio Math., in press.

[55] ———, *Endoscopic classification of representations of quasi-split unitary groups*, Memoirs AMS, in press.

[56] S. Morel, *On the cohomology of certain non-compact Shimura varieties*, Annals of Mathematics Studies **173** (2010).

[57] Ngô B.-C, *Le lemme fondamental pour les algèbres de Lie*, Publ. Math. I.H.E.S. **111** (2010), 1–169.

[58] V. Pilloni, B. Stroh, *Surconvergence, ramification, et modularité*, Preprint (2013).

[59] S. Patrikis and R. Taylor, *Automorphy and irreducibility of some l-adic representations*, preprint (2012).

[60] A. Raghuram, *Critical values of Rankin-Selberg L-functions for $GL_n \times GL_{n-1}$ and the symmetric cube L-functions for GL_2*, preprint (2013).

[61] P. Scholze, *On torsion in the cohomology of locally symmetric varieties*, Preprint, Bonn (2013).

[62] P. Scholze and S.-W. Shin, *On the cohomology of compact unitary group Shimura varieties at ramified split places*, J. AMS. **26** (2013), 261–294.

[63] J.-P. Serre, *Abelian ℓ-adic representations and elliptic curves*, Redwood City: Addison-Wesley (1989).

[64] S.-W. Shin, *Galois representations arising from some compact Shimura varieties*, Ann. of Math. **173** (2011), 1645–1741.

[65] C. Sorensen, *A patching lemma, to appear in Paris book project*, volume 2, online at http://www.math.ucsd.edu/~csorense/.

[66] B.-Y. Sun, *The nonvanishing hypothesis at infinity for Rankin-Selberg convolutions*, arXiv:1307.5357.

[67] R. Taylor, *Galois representations associated to Siegel modular forms of low weight*, Duke Math. J. **63** (1991), 281–332.

[68] ———, *Remarks on a conjecture of Fontaine and Mazur*, J. Inst Math. Jussieu **1** (2002), 1–19.

[69] ———, *Galois Representations*, Annales de la Faculte des Sciences de Toulouse **13** (2004), 73–119.

[70] ———, *Automorphy for some ℓ-adic lifts of automorphic mod ℓ Galois representations, II*, Publ. Math. IHES. **108** (2008), 1–181.

[71] R. Taylor and T. Yoshida, *Compatibility of local and global Langlands correspondences*, J. AMS. **20** (2007), 467-493.

[72] J. Thorne, *On the automorphy of l-adic Galois representations with small residual image*, J. Inst. Math. Jussieu **11** (2012), 855–920.

[73] ———, *Automorphy lifting for residually reducible l-adic Galois representations*, Preprint, online at http://www.math.harvard.edu/~thorne/.

[74] J.-L. Waldspurger, *La conjecture locale de Gross-Prasad pour les représentations tempérées des groupes spéciaux orthogonaux*, Astérisque **347** (2012), 103–165.

[75] W. Zhang, *Automorphic period and the central value of Rankin-Selberg L-function*, preprint (2013)

Univ Paris Diderot, Sorbonne Paris Cité, UMR 7586, Institut de Mathématiques de Jussieu-Paris Rive Gauche, Case 247, 4 place Jussieu F-75005, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR 7586, IMJ-PRG, F-75005 Paris, France; CNRS, UMR7586, IMJ-PRG, F-75013 Paris, France; Department of Mathematics, Columbia University, New York, NY 10027, USA

E-mail: harris@math.jussieu.fr; harris@math.columbia.edu

The ternary Goldbach problem

Harald Andrés Helfgott

Abstract. The ternary Goldbach conjecture, or three-primes problem, states that every odd number n greater than 5 can be written as the sum of three primes. The conjecture, posed in 1742, remained unsolved until now, in spite of great progress in the twentieth century. In 2013 – following a line of research pioneered and developed by Hardy, Littlewood and Vinogradov, among others – the author proved the conjecture. In this, as in many other additive problems, what is at issue is really the proper usage of the limited information we possess on the distribution of prime numbers. The problem serves as a test and whetting-stone for techniques in analysis and number theory – and also as an incentive to think about the relations between existing techniques with greater clarity. We will go over the main ideas of the proof. The basic approach is based on the circle method, the large sieve and exponential sums. For the purposes of this overview, we will not need to work with explicit constants; however, we will discuss what makes certain strategies and procedures not just effective, but efficient, in the sense of leading to good constants. Still, our focus will be on qualitative improvements.

Mathematics Subject Classification (2010). Primary 11P32.

Keywords. Analytic number theory, additive problems, prime numbers.

The question we will discuss, or one similar to it, seems to have been first posed by Descartes, in a manuscript published only centuries after his death [14, p. 298]. Descartes states: "Sed & omnis numerus par fit ex uno vel duobus vel tribus primis" ("But also every even number is made out of one, two or three prime numbers.") This statement comes in the middle of a discussion of sums of polygonal numbers, such as the squares.

Statements on sums of primes and sums of values of polynomials (polygonal numbers, powers n^k, etc.) have since shown themselves to be much more than mere curiosities – and not just because they are often very difficult to prove. Whereas the study of sums of powers can rely on their algebraic structure, the study of sums of primes leads to the realization that, from several perspectives, the set of primes behaves much like the set of integers – and that this is truly hard to prove.

If, instead of the primes, we had a random set of odd integers S whose density – an intuitive concept that can be made precise – equaled that of the primes, then we would expect to be able to write every odd number as a sum of three elements of S, and every even number as the sum of two elements of S. We would have to check by hand whether this is true for small odd and even numbers, but it is relatively easy to show that, after a long enough check, it would be very unlikely that there would be any exceptions left among the infinitely many cases left to check.

The question, then, is in what sense we need the primes to be like a random set of integers; in other words, we need to know what we can prove about the regularities of the

Proceedings of the International Congress of Mathematicians, Seoul, 2014

distribution of the primes. This is one of the main questions of analytic number theory; progress on it has been very slow and difficult. Thus, the real question is how to use well the limited information we do have on the distribution of the primes.

1. History and new developments

The history of the conjecture starts properly with Euler and his close friend, Christian Goldbach, both of whom lived and worked in Russia at the time of their correspondence – about a century after Descartes' isolated statement. Goldbach, a man of many interests, is usually classed as a serious amateur; he seems to have awakened Euler's passion for number theory, which would lead to the beginning of the modern era of the subject [71, Ch. 3, §IV]. In a letter dated June 7, 1742 – written partly in German, partly in Latin – Goldbach made a conjectural statement on prime numbers, and Euler rapidly reduced it to the following conjecture, which, he said, Goldbach had already posed to him: every positive integer can be written as the sum of at most three prime numbers.

We would now say "every integer greater than 1", since we no long consider 1 to be a prime number. Moreover, the conjecture is nowadays split into two:

- the *weak*, or ternary, Goldbach conjecture states that every odd integer greater than 5 can be written as the sum of three primes;
- the *strong*, or binary, Goldbach conjecture states that every even integer greater than 2 can be written as the sum of two primes.

As their names indicate, the strong conjecture implies the weak one (easily: subtract 3 from your odd number n, then express $n-3$ as the sum of two primes).

The strong conjecture remains out of reach. A short while ago – the first complete version appeared on May 13, 2013 – the present author proved the weak Goldbach conjecture.

Main Theorem. *Every odd integer greater than* 5 *can be written as the sum of three primes.*

The proof is contained in the preprints [28], [27], [29]. It builds on the great progress towards the conjecture made in the early 20th century by Hardy, Littlewood and Vinogradov. In 1937, Vinogradov proved [67] that the conjecture is true for all odd numbers n larger than some constant C. (Hardy and Littlewood had shown the same under the assumption of the Generalized Riemann Hypothesis, which we shall have the chance to discuss later.)

It is clear that a computation can verify the conjecture only for $n \leq c$, c a constant: computations have to be finite. What can make a result coming from analytic number theory be valid only for $n \geq C$?

An analytic proof, generally speaking, gives us more than just existence. In this kind of problem, it gives us more than the possibility of doing something (here, writing an integer n as the sum of three primes). It gives us a rigorous estimate for the number of ways in which this *something* is possible; that is, it shows us that this number of ways equals

$$\text{main term} + \text{error term}, \tag{1.1}$$

where the main term is a precise quantity $f(n)$, and the error term is something whose absolute value is at most another precise quantity $g(n)$. If $f(n) > g(n)$, then (1.1) is nonzero, i.e., we will have shown that the existence of a way to write our number as the sum of three primes.

(Since what we truly care about is existence, we are free to weigh different ways of writing n as the sum of three primes however we wish – that is, we can decide that some primes "count" twice or thrice as much as others, and that some do not count at all.)

Typically, after much work, we succeed in obtaining (1.1) with $f(n)$ and $g(n)$ such that $f(n) > g(n)$ asymptotically, that is, for n large enough. To give a highly simplified example: if, say, $f(n) = n^2$ and $g(n) = 100n^{3/2}$, then $f(n) > g(n)$ for $n > C$, where $C = 10^4$, and so the number of ways (1.1) is positive for $n > C$.

We want a moderate value of C, that is, a C small enough that all cases $n \leq C$ can be checked computationally. To ensure this, we must make the error term bound $g(n)$ as small as possible. This is our main task. A secondary (and sometimes neglected) possibility is to rig the weights so as to make the main term $f(n)$ larger in comparison to $g(n)$; this can generally be done only up to a certain point, but is nonetheless very helpful.

As we said, the first unconditional proof that odd numbers $n \geq C$ can be written as the sum of three primes is due to Vinogradov. Analytic bounds fall into several categories, or stages; quite often, successive versions of the same theorem will go through successive stages.

1. An *ineffective* result shows that a statement is true for some constant C, but gives no way to determine what the constant C might be. Vinogradov's first proof of his theorem (in [67]) is like this: it shows that there exists a constant C such that every odd number $n > C$ is the sum of three primes, yet gives us no hope of finding out what the constant C might be.[1] Many proofs of Vinogradov's result in textbooks are also of this type.
2. An *effective*, but not explicit, result shows that a statement is true for some unspecified constant C in a way that makes it clear that a constant C could in principle be determined following and reworking the proof with great care. Vinogradov's later proof ([68], translated in [69]) is of this nature. As Chudakov [8, §IV.2] pointed out, the improvement on [67] given by Mardzhanishvili [41] already had the effect of making the result effective.[2]
3. An *explicit* result gives a value of C. According to [8, p. 201], the first explicit version of Vinogradov's result was given by Borozdkin in his unpublished doctoral dissertation, written under the direction of Vinogradov (1939): $C = \exp(\exp(\exp(41.96)))$. Such a result is, by definition, also effective. Borodzkin later [2] gave the value $C = e^{e^{16.038}}$, though he does not seem to have published the proof. The best – that is, smallest – value of C known before the present work was that of Liu and Wang [40]: $C = 2 \cdot 10^{1346}$.
4. What we may call an *efficient* proof gives a reasonable value for C – in our case, a value small enough that checking all cases up to C is feasible.

How far were we from an efficient proof? That is, what sort of computation could ever be feasible? The number of picoseconds since the beginning of the universe is less than 10^{30}, whereas the number of protons in the observable universe is currently estimated at

[1] Here, as is often the case in ineffective results in analytic number theory, the underlying issue is that of *Siegel zeros*, which are believed not to exist, but have not been shown not to; the strongest bounds on (i.e., against the existence of) such zeros are ineffective, and so are all of the many results using such estimates.

[2] The proof in [41] combined the bounds in [67] with a more careful accounting of the effect of the single possible Siegel zero within range.

$\sim 10^{80}$. This means that even a parallel computer the size of the universe could never perform a computation requiring 10^{110} steps, even if it ran for the age of the universe. Thus, $C = 2 \cdot 10^{1346}$ is too large.

I gave a proof with $C = 10^{29}$ in May 2013. Since D. Platt and I had verified the conjecture for all odd numbers up to $n \leq 8.8 \cdot 10^{30}$ by computer [31], this established the conjecture for all odd numbers n.

(In December 2013, C was reduced to 10^{27} [29]. The verification of the ternary Goldbach conjecture up to $n \leq 10^{27}$ can be done in a home computer over a weekend. All must be said: this uses the verification of the binary Goldbach conjecture for $n \leq 4 \cdot 10^{18}$ [46], which itself required computational resources far outside the home-computing range. Checking the conjecture up to $n \leq 10^{27}$ was not even the main computational task that needed to be accomplished to establish the Main Theorem – that task was the finite verification of zeros of L-functions in [48], a general-purpose computation that should be useful elsewhere. We will discuss the procedure at the end of the article.)

What was the strategy of [27–29]? The basic framework is the one pioneered by Hardy and Littlewood for a variety of problems – namely, the *circle method*, which, as we shall see, is an application of Fourier analysis over $\mathbb{Z}$. (There are other, later routes to Vinogradov's result; see [21, 24] and especially the recent work [57], which avoids using anything about zeros of L-functions inside the critical strip.) Vinogradov's proof, like much of the later work on the subject, was based on a detailed analysis of exponential sums, i.e., Fourier transforms over $\mathbb{Z}$. So is the proof that we will sketch.

At the same time, the distance between $2 \cdot 10^{1346}$ and 10^{27} is such that we cannot hope to get to 10^{27} (or any other reasonable constant) by fine-tuning previous work. Rather, we must work from scratch, using the basic outline in Vinogradov's original proof and other, initially unrelated, developments in analysis and number theory (notably, the large sieve). Merely improving constants will not do; we must do qualitatively better than previous work (by non-constant factors) if we are to have any chance to succeed. It is on these qualitative improvements that we will focus.

* * *

It is only fair to review some of the progress made between Vinogradov's time and ours. Here we will focus on results; later, we will discuss some of the progress made in the techniques of proof. For a fuller account up to 1978, see R. Vaughan's ICM lecture notes on the ternary Goldbach problem [65].

In 1933, Schnirelmann proved [56] that every integer $n > 1$ can be written as the sum of at most K primes for some unspecified constant K. (This pioneering work is now considered to be part of the early history of additive combinatorics.) In 1969, Klimov gave an explicit value for K (namely, $K = 6 \cdot 10^9$); he later improved the constant to $K = 115$ (with G. Z. Piltay and T. A. Sheptickaja) and $K = 55$. Later, there were results by Vaughan [63] ($K = 27$), Deshouillers [15] ($K = 26$) and Riesel-Vaughan [54] ($K = 19$).

Ramaré showed in 1995 that every even number $n > 1$ can be written as the sum of at most 6 primes [51]. In 2012, Tao proved [58] that every odd number $n > 1$ is the sum of at most 5 primes.

There have been other avenues of attack towards the strong conjecture. Using ideas close to those of Vinogradov's, Chudakov [9, 10], Estermann [19] and van der Corput [62] proved (independently from each other) that almost every even number (meaning: all elements of a subset of density 1 in the even numbers) can be written as the sum of two primes. In 1973, J.-

R. Chen showed [4] that every even number n larger than a constant C can be written as the sum of a prime number and the product of at most two primes ($n = p_1+p_2$ or $n = p_1+p_2p_3$). Incidentally, J.-R. Chen himself, together with T.-Z. Wang, was responsible for the best bounds on C (for ternary Goldbach) before Lui and Wang: $C = \exp(\exp(11.503)) < 4 \cdot 10^{43000}$ [6] and $C = \exp(\exp(9.715)) < 6 \cdot 10^{7193}$ [7].

Matters are different if one assumes the Generalized Riemann Hypothesis (GRH). A careful analysis [18] of Hardy and Littlewood's work [23] gives that every odd number $n \geq 1.24 \cdot 10^{50}$ is the sum of three primes if GRH is true. According to [18], the same statement with $n \geq 10^{32}$ was proven in the unpublished doctoral dissertation of B. Lucke, a student of E. Landau's, in 1926. Zinoviev [72] improved this to $n \geq 10^{20}$. A computer check ([16]; see also [55]) showed that the conjecture is true for $n < 10^{20}$, thus completing the proof of the ternary Goldbach conjecture under the assumption of GRH. What was open until now was, of course, the problem of giving an unconditional proof.

2. The circle method: Fourier analysis on $\mathbb{Z}$

It is common for a first course on Fourier analysis to focus on functions over the reals satisfying $f(x) = f(x+1)$, or, what is the same, functions $f : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$. Such a function (unless it is fairly pathological) has a Fourier series converging to it; this is just the same as saying that f has a Fourier transform $\widehat{f} : \mathbb{Z} \to \mathbb{C}$ defined by $\widehat{f}(n) = \int_{\mathbb{R}/\mathbb{Z}} f(\alpha)e(-\alpha n)d\alpha$ and satisfying $f(\alpha) = \sum_{n\in\mathbb{Z}} \widehat{f}(n)e(\alpha n)d\alpha$ (*Fourier inversion theorem*).

In number theory, we are especially interested in functions $f : \mathbb{Z} \to \mathbb{C}$. Then things are exactly the other way around: provided that f decays reasonably fast as $n \to \pm\infty$ (or becomes 0 for n large enough), f has a Fourier transform $\widehat{f} : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$ defined by $\widehat{f}(\alpha) = \sum_n f(n)e(-\alpha n)$ and satisfying $f(n) = \int_{\mathbb{R}/\mathbb{Z}} \widehat{f}(\alpha)e(\alpha n)$. (Highbrow talk: we already knew that $\mathbb{Z}$ is the Fourier dual of $\mathbb{R}/\mathbb{Z}$, and so, of course, $\mathbb{R}/\mathbb{Z}$ is the Fourier dual of $\mathbb{Z}$.) "Exponential sums" (or "trigonometrical sums", as in the title of [69]) are sums of the form $\sum_n f(\alpha)e(-\alpha n)$; the "circle" in "circle method" is just a name for $\mathbb{R}/\mathbb{Z}$.

The study of the Fourier transform $\widehat{f}$ is relevant to additive problems in number theory, i.e., questions on the number of ways of writing n as a sum of k integers of a particular form. Why? One answer could be that $\widehat{f}$ gives us information about the "randomness" of f; if f were the characteristic function of a random set, then $\widehat{f}(\alpha)$ would be very small outside a sharp peak at $\alpha = 0$. We can also give a more concrete and immediate answer. Recall that, in general, the Fourier transform of a convolution equals the product of the transforms; over $\mathbb{Z}$, this means that for the additive convolution

$$(f * g)(n) = \sum_{\substack{m_1,m_2\in\mathbb{Z} \\ m_1+m_2=n}} f(m_1)g(m_2),$$

the Fourier transform satisfies the simple rule

$$\widehat{f * g}(\alpha) = \widehat{f}(\alpha) \cdot \widehat{g}(\alpha).$$

We can see right away from this that $(f * g)(n)$ can be non-zero only if n can be written as $n = m_1 + m_2$ for some m_1, m_2 such that $f(m_1)$ and $g(m_2)$ are non-zero. Similarly,

$(f * g * h)(n)$ can be non-zero only if n can be written as $n = m_1 + m_2 + m_3$ for some m_1, m_2, m_3 such that $f(m_1)$, $f_2(m_2)$ and $f_3(m_3)$ are all non-zero. This suggests that, to study the ternary Goldbach problem, we define $f_1, f_2, f_3 : \mathbb{Z} \to \mathbb{C}$ so that they take non-zero values only at the primes.

Hardy and Littlewood defined $f_1(n) = f_2(n) = f_3(n) = 0$ for n non-prime (and also for $n \leq 0$), and $f_1(n) = f_2(n) = f_3(n) = (\log n)e^{-n/x}$ for n prime (where x is a parameter to be fixed later). Here the factor $e^{-n/x}$ is there to provide "fast decay", so that everything converges; as we will see later, Hardy and Littlewood's choice of $e^{-n/x}$ (rather than some other function of fast decay) comes across in hindsight as being very clever, though not quite best-possible. (Their "choice" was, to some extent, not a choice, but an artifact of their version of the circle method.) The term $\log n$ is there for technical reasons – in essence, it makes sense to put it there because a random integer around n has a chance of about $1/(\log n)$ of being prime.

We can see that $(f_1 * f_2 * f_3)(n) \neq 0$ if and only if n can be written as the sum of three primes. Our task is then to show that $(f_1 * f_2 * f_3)(n)$ (i.e., $(f * f * f)(n)$) is non-zero for every n larger than a constant $C \sim 10^{27}$. Since the transform of a convolution equals a product of transforms,

$$(f_1 * f_2 * f_3)(n) = \int_{\mathbb{R}/\mathbb{Z}} \widehat{f_1 * f_2 * f_3}(\alpha) e(\alpha n) d\alpha = \int_{\mathbb{R}/\mathbb{Z}} (\widehat{f_1}\widehat{f_2}\widehat{f_3})(\alpha) e(\alpha n) d\alpha. \quad (2.1)$$

Our task is thus to show that the integral $\int_{\mathbb{R}/\mathbb{Z}} (\widehat{f_1}\widehat{f_2}\widehat{f_3})(\alpha) e(\alpha n) d\alpha$ is non-zero.

As it happens, $\widehat{f}(\alpha)$ is particularly large when α is close to a rational with small denominator. Moreover, for such α, it turns out we can actually give rather precise estimates for $\widehat{f}(\alpha)$. Define $\mathfrak{M}$ (called the set of *major arcs*) to be a union of narrow arcs around the rationals with small denominator:

$$\mathfrak{M} = \bigcup_{q \leq r} \bigcup_{\substack{a \bmod q \\ (a,q)=1}} \left(\frac{a}{q} - \frac{1}{qQ}, \frac{a}{q} + \frac{1}{qQ}\right),$$

where Q is a constant times x/r, and r will be set later. We can write

$$\int_{\mathbb{R}/\mathbb{Z}} (\widehat{f_1}\widehat{f_2}\widehat{f_3})(\alpha) e(\alpha n) d\alpha = \int_{\mathfrak{M}} (\widehat{f_1}\widehat{f_2}\widehat{f_3})(\alpha) e(\alpha n) d\alpha + \int_{\mathfrak{m}} (\widehat{f_1}\widehat{f_2}\widehat{f_3})(\alpha) e(\alpha n) d\alpha, \quad (2.2)$$

where $\mathfrak{m}$ is the complement $(\mathbb{R}/\mathbb{Z}) \setminus \mathfrak{M}$ (called *minor arcs*).

Now, we simply do not know how to give precise estimates for $\widehat{f}(\alpha)$ when α is in $\mathfrak{m}$. However, as Vinogradov realized, one can give reasonable upper bounds on $|\widehat{f}(\alpha)|$ for $\alpha \in \mathfrak{m}$. This suggests the following strategy: show that

$$\int_{\mathfrak{m}} |\widehat{f_1}(\alpha)||\widehat{f_2}(\alpha)||\widehat{f_3}(\alpha)| d\alpha < \int_{\mathfrak{M}} \widehat{f_1}(\alpha)\widehat{f_2}(\alpha)\widehat{f_3}(\alpha) e(\alpha n) d\alpha. \quad (2.3)$$

By (2.1) and (2.2), this will imply immediately that $(f_1 * f_2 * f_3)(n) > 0$, and so we will be done.

3. The major arcs $\mathfrak{M}$

3.1. What do we really know about L-functions and their zeros? Before we start, let us give a very brief review of basic analytic number theory (in the sense of, say, [13]). A *Dirichlet character* $\chi : \mathbb{Z} \to \mathbb{C}$ of modulus q is a character of $(\mathbb{Z}/q\mathbb{Z})^*$ lifted to $\mathbb{Z}$. (In other words, $\chi(n) = \chi(n+q)$, $\chi(ab) = \chi(a)\chi(b)$ for all a, b and $\chi(n) = 0$ for $(n,q) \neq 1$.) A *Dirichlet L-series* is defined by

$$L(s,\chi) = \sum_{n=1}^{\infty} \chi(n) n^{-s}$$

for $\Re(s) > 1$, and by analytic continuation for $\Re(s) \leq 1$. (The Riemann zeta function $\zeta(s)$ is the L-function for the trivial character, i.e., the character χ such that $\chi(n) = 1$ for all n.) Taking logarithms and then derivatives, we see that

$$-\frac{L'(s,\chi)}{L(s,\chi)} = \sum_{n=1}^{\infty} \Lambda(n) n^{-s}, \tag{3.1}$$

where Λ is the *von Mangoldt function* ($\Lambda(n) = \log p$ if n is some prime power p^α, $\alpha \geq 1$, and $\Lambda(n) = 0$ otherwise).

Dirichlet introduced his characters and L-series so as to study primes in arithmetic progressions. In general, and after some work, (3.1) allows us to restate many sums over the primes (such as our Fourier transforms $\widehat{f}(\alpha)$) as sums over the zeros of $L(s,\chi)$. A *non-trivial zero* of $L(s,\chi)$ is a zero of $L(s,\chi)$ such that $0 < \Re(s) < 1$. (The other zeros are called trivial because we know where they are, namely, at negative integers and, in some cases, also on the line $\Re(s) = 0$. In order to eliminate all zeros on $\Re(s) = 0$ outside $s = 0$, it suffices to assume that χ is *primitive*; a primitive character modulo q is one that is not induced by (i.e., not the restriction of) any character modulo $d|q$, $d < q$.)

The Generalized Riemann Hypothesis for Dirichlet L-functions is the statement that, for every Dirichlet character χ, every non-trivial zero of $L(s,\chi)$ satisfies $\Re(s) = 1/2$. Of course, the Generalized Riemann Hypothesis (GRH) – and the Riemann Hypothesis, which is the special case of χ trivial – remains unproven. Thus, if we want to prove unconditional statements, we need to make do with partial results towards GRH. Two kinds of such results have been proven:

- **Zero-free regions.** Ever since the late nineteenth century (Hadamard, de la Vallée-Poussin) we have known that there are hourglass-shaped regions (more precisely, of the shape $\frac{c}{\log t} \leq \sigma \leq 1 - \frac{c}{\log t}$, where c is a constant and where we write $s = \sigma + it$) outside which non-trivial zeros cannot lie. Explicit values for c are known [35, 36, 42]. There is also the Vinogradov-Korobov region [39, 70], which is broader asymptotically but narrower in most of the practical range (see [20], however).
- **Finite verifications of GRH.** It is possible to (ask the computer to) prove small, finite fragments of GRH, in the sense of verifying that all non-trivial zeros of a given finite set of L-functions with imaginary part less than some constant H lie on the critical line $\Re(s) = 1/2$. Such verifications go back to Riemann, who checked the first few zeros of $\zeta(s)$. Large-scale, rigorous computer-based verifications are now a possibility.

Most work in the literature follows the first alternative, though [58] did use a finite verification of RH (i.e., GRH for the trivial character). Unfortunately, zero-free regions seem

too narrow to be useful for the ternary Goldbach problem. Thus, we are left with the second alternative.

In coordination with the present work, Platt [48] verified that all zeros s of L-functions for characters χ with modulus $q \leq 300000$ satisfying $\Im(s) \leq H_q$ lie on the line $\Re(s) = 1/2$, where

- $H_q = 10^8/q$ for q odd, and
- $H_q = \max(10^8/q, 200 + 7.5 \cdot 10^7/q)$ for q even.

This was a medium-large computation, taking a few hundreds of thousands of core-hours on a parallel computer. It used *interval arithmetic* for the sake of rigor; we will later discuss what this means.

The choice to use a finite verification of GRH, rather than zero-free regions, had consequences on the manner in which the major and minor arcs had to be chosen. As we shall see, such a verification can be used to give very precise bounds on the major arcs, but also forces us to define them so that they are narrow and their number is constant. To be precise: the major arcs were defined around rationals a/q with $q \leq r$, $r = 300000$; moreover, as will become clear, the fact that H_q is finite will force their width to be bounded by $c_0 r/qx$, where c_0 is a constant (say $c_0 = 8$).

3.2. Estimates of $\widehat{f}(\alpha)$ for α in the major arcs. Recall that we want to estimate sums of the type $\widehat{f}(\alpha) = \sum f(n)e(-\alpha n)$, where $f(n)$ is something like $(\log n)\eta(n/x)$ for n equal to a prime, and 0 otherwise; here $\eta : \mathbb{R} \to \mathbb{C}$ is some function of fast decay, such as Hardy and Littlewood's choice, $\eta(t) = e^{-t}$. Let us modify this just a little – we will actually estimate

$$S_\eta(\alpha, x) = \sum \Lambda(n)e(\alpha n)\eta(n/x), \tag{3.2}$$

where Λ is the von Mangoldt function (as in (3.1)) . The use of α rather than $-\alpha$ is just a bow to tradition, as is the use of the letter S (for "sum"); however, the use of $\Lambda(n)$ rather than just plain $\log p$ does actually simplify matters.

The function η here is sometimes called a *smoothing function* or simply a *smoothing*. It will indeed be helpful for it to be smooth on $(0, \infty)$, but, in principle, it need not even be continuous. (Vinogradov's work implicitly uses, in effect, the "brutal truncation" $1_{[0,1]}(t)$, defined to be 1 when $t \in [0, 1]$ and 0 otherwise; that would be fine for the minor arcs, but, as it will become clear, it is a bad idea as far as the major arcs are concerned.)

Assume α is on a major arc, meaning that we can write $\alpha = a/q + \delta/x$ for some a/q (q small) and some δ (with $|\delta|$ small). We can write $S_\eta(\alpha, x)$ as a linear combination

$$S_\eta(\alpha, x) = \sum_\chi c_\chi S_{\eta,\chi}\left(\frac{\delta}{x}, x\right) + \text{tiny error term}, \tag{3.3}$$

where

$$S_{\eta,\chi}\left(\frac{\delta}{x}, x\right) = \sum \Lambda(n)\chi(n)e(\delta n/x)\eta(n/x). \tag{3.4}$$

In (3.3), χ runs over primitive Dirichlet characters of moduli $d|q$, and c_χ is small ($|c_\chi| \leq \sqrt{d}/\phi(q)$).

To estimate the sums $S_{\eta,\chi}$, we will use L-functions, together with one of the most common tools of analytic number theory, the Mellin transform. This transform is essentially a

Laplace transform with a change of variables, and a Laplace transform, in turn, is a Fourier transform taken on a vertical line in the complex plane. For f of fast enough decay, the Mellin transform $F = Mf$ of f is given by

$$F(s) = \int_0^\infty f(t) t^s \frac{dt}{t};$$

we can express f in terms of F by the *Mellin inversion formula*

$$f(t) = \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} F(s) t^{-s} ds$$

for any σ within an interval. We can thus express $e(\delta t)\eta(t)$ in terms of its Mellin transform F_δ and then use (3.1) to express $S_{\eta,\chi}$ in terms of F_δ and $L'(s,\chi)/L(s,\chi)$; shifting the integral in the Mellin inversion formula to the left, we obtain what is known in analytic number theory as an *explicit formula*:

$$S_{\eta,\chi}(\delta/x, x) = [\widehat{\eta}(-\delta)x] - \sum_\rho F_\delta(\rho) x^\rho + \text{tiny error term}.$$

Here the term between brackets appears only for χ trivial. In the sum, ρ goes over all non-trivial zeros of $L(s,\chi)$, and F_δ is the Mellin transform of $e(\delta t)\eta(t)$. (The tiny error term comes from a sum over the trivial zeros of $L(s,\chi)$.) We will obtain the estimate we desire if we manage to show that the sum over ρ is small.

The point is this: if we verify GRH for $L(s,\chi)$ up to imaginary part H, i.e., if we check that all zeroes ρ of $L(s,\chi)$ with $|\Im(\rho)| \le H$ satisfy $\Re(\rho) = 1/2$, we have $|x^\rho| = \sqrt{x}$. In other words, x^ρ is very small (compared to x). However, for any ρ whose imaginary part has absolute value greater than H, we know next to nothing about its real part, other than $0 \le \Re(\rho) \le 1$. (Zero-free regions are notoriously weak for $\Im(\rho)$ large; we will not use them.) Hence, our only chance is to make sure that $F_\delta(\rho)$ is very small when $|\Im(\rho)| \ge H$.

This has to be true for both δ very small (including the case $\delta = 0$) and for δ not so small ($|\delta|$ up to $c_0 r/q$, which can be large because r is a large constant). How can we choose η so that $F_\delta(\rho)$ is very small in both cases for $\tau = \Im(\rho)$ large?

The method of *stationary phase* is useful as an exploratory tool here. In brief, it suggests (and can sometimes prove) that the main contribution to the integral

$$F_\delta(t) = \int_0^\infty e(\delta t)\eta(t) t^s \frac{dt}{t} \tag{3.5}$$

can be found where the phase of the integrand has derivative 0. This happens when $t = -\tau/2\pi\delta$ (for $\operatorname{sgn}(\tau) \ne \operatorname{sgn}(\delta)$); the contribution is then a moderate factor times $\eta(-\tau/2\pi\delta)$. In other words, if $\operatorname{sgn}(\tau) \ne \operatorname{sgn}(\delta)$ and δ is not too small ($|\delta| \ge 8$, say), $F_\delta(\sigma + i\tau)$ behaves like $\eta(-\tau/2\pi\delta)$; if δ is small ($|\delta| < 8$), then F_δ behaves like F_0, which is the Mellin transform $M\eta$ of η. Here is our goal, then: the decay of $\eta(t)$ as $|t| \to \infty$ should be as fast as possible, and the decay of the transform $M\eta(\sigma + i\tau)$ should also be as fast as possible.

This is a classical dilemma, often called the *uncertainty principle* because it is the mathematical fact underlying the physical principle of the same name: you cannot have a function η that decreases extremely rapidly and whose Fourier transform (or, in this case, its Mellin transform) also decays extremely rapidly. What does "extremely rapidly" mean here? It

means (as Hardy himself proved) "faster than any exponential e^{-Ct}". Thus, Hardy and Littlewood's choice $\eta(t) = e^{-t}$ seems essentially optimal at first sight.

However, it is not optimal. We can choose η so that $M\eta$ decreases exponentially (with a constant C somewhat worse than for $\eta(t) = e^{-t}$), but η decreases faster than exponentially. This is a particularly appealing possibility because it is $t/|\delta|$, and not so much t, that risks being fairly small. (To be explicit: say we check GRH for characters of modulus q up to $H_q \sim 50 \cdot c_0 r/q \geq 50|\delta|$. Then we only know that $|\tau/2\pi\delta| \gtrsim 8$. So, for $\eta(t) = e^{-t}$, $\eta(-\tau/2\pi\delta)$ may be as large as e^{-8}, which is not negligible. Indeed, since this term will be multiplied later by other terms, e^{-8} is simply not small enough. On the other hand, we can assume that $H_q \geq 200$ (say), and so $M\eta(s) \sim e^{-(\pi/2)|\tau|}$ is completely negligible, and will remain negligible even if we replace $\pi/2$ by a somewhat smaller constant.)

We shall take $\eta(t) = e^{-t^2/2}$ (that is, the Gaussian). This is not the only possible choice, but it is in some sense natural. It is easy to show that the Mellin transform F_δ for $\eta(t) = e^{-t^2/2}$ is a multiple of what is called a *parabolic cylinder function* $U(a, z)$ with imaginary values for z. There are plenty of estimates on parabolic cylinder functions in the literature – but mostly for a and z real, in part because that is one of the cases occuring most often in applications. There are some asymptotic expansions and estimates for $U(a, z)$, a, z, general, due to Olver (see, e.g., [47]), but unfortunately they come without fully explicit error terms for a and z within our range of interest. (The same holds for [59].)

In the end, using the *saddle-point method*, I derived bounds for the Mellin transform F_δ of $\eta(t)e(\delta t)$ with $\eta(t) = e^{-t^2/2}$: for $s = \sigma + i\tau$ with $\sigma \in [0, 1]$ and $|\tau| \geq \max(100, 4\pi^2|\delta|)$,

$$|F_\delta(s)| + |F_\delta(1-s)| \leq 4.226 \cdot \begin{cases} e^{-0.1065\left(\frac{\tau}{\pi\delta}\right)^2} & \text{if } |\tau| < \frac{3}{2}(\pi\delta)^2, \\ e^{-0.1598|\tau|} & \text{if } |\tau| \geq \frac{3}{2}(\pi\delta)^2. \end{cases} \tag{3.6}$$

Similar bounds hold for σ in other ranges, thus giving us (similar) estimates for the Mellin transform F_δ for $\eta(t) = t^k e^{-t^2/2}$ and σ in the critical range $[0, 1]$.

A moment's thought shows that we can also use (3.6) to deal with the Mellin transform of $\eta(t)e(\delta t)$ for any function of the form $\eta(t) = e^{-t^2/2}g(t)$ (or, more generally, $\eta(t) = t^k e^{-t^2/2}g(t)$), where $g(t)$ is any *band-limited function*. By a band-limited function, we could mean a function whose Fourier transform is compactly supported; while that is a plausible choice, it turns out to be better to work with functions that are band-limited with respect to the Mellin transform – in the sense of being of the form

$$g(t) = \int_{-R}^{R} h(r) t^{-ir} dr,$$

where $h : \mathbb{R} \to \mathbb{C}$ is supported on a compact interval $[-R, R]$, with R not too large (say $R = 200$).

After deriving an explicit formula general enough to work with all the weights $\eta(t)$ we have discussed, and once we consider the input provided by Platt's finite verification of GRH up to H_q, we obtain simple bounds for different weights. For $\eta(t) = e^{-t^2/2}$, $x \geq 10^8$, χ a primitive character of modulus $q \leq r = 300000$, and any $\delta \in \mathbb{R}$ with $|\delta| \leq 4r/q$, we obtain

$$S_{\eta,\chi}\left(\frac{\delta}{x}, x\right) = I_{q=1} \cdot \widehat{\eta}(-\delta)x + E \cdot x, \tag{3.7}$$

where $I_{q=1} = 1$ if $q = 1$, $I_{q=1} = 0$ if $q \neq 1$, and

$$|E| \leq 5.281 \cdot 10^{-22} + \frac{1}{\sqrt{x}} \left(\frac{650400}{\sqrt{q}} + 112 \right). \tag{3.8}$$

Here $\widehat{\eta}$ stands for the Fourier transform from $\mathbb{R}$ to $\mathbb{R}$ normalized as follows:

$$\widehat{\eta}(t) = \int_{-\infty}^{\infty} e(-xt)\eta(x)dx$$

Thus, $\widehat{\eta}(-\delta)$ is just $\sqrt{2\pi}e^{-2\pi^2\delta^2}$ (self-duality of the Gaussian).

This is one of the main results of [27]. Similar bounds are also proven there for $\eta(t) = t^2e^{-t^2/2}$, as well as for a weight of type $\eta(t) = te^{-t^2/2}g(t)$, where $g(t)$ is a band-limited function, and also for a weight η defined by a multiplicative convolution. The conditions on q ($q \leq r = 300000$) and δ are what we expected from the outset.

Thus concludes our treatment of the major arcs. This is arguably the easiest part of the proof; it was actually what I left for the end, as I was fairly confident it would work out.

4. The minor arcs $\mathfrak{m}$

4.1. Qualitative goals and main ideas. What kind of bounds do we need? What is there in the literature?

We wish to obtain upper bounds on $|S_\eta(\alpha, x)|$ for some weight η and any $\alpha \in \mathbb{R}/\mathbb{Z}$ not very close to a rational with small denominator. Every α is close to some rational a/q; what we are looking for is a bound on $|S_\eta(\alpha, x)|$ that decreases rapidly when q increases.

Moreover, we want our bound to decrease rapidly when δ increases, where $\alpha = a/q + \delta/x$. In fact, the main terms in our bound will be decreasing functions of $\max(1, |\delta|/8) \cdot q$. (Let us write $\delta_0 = \max(2, |\delta|/4)$ from now on.) This will allow our bound to be good enough outside narrow major arcs, which will get narrower and narrower as q increases – that is, precisely the kind of major arcs we were presupposing in our major-arc bounds.

It would be possible to work with narrow major arcs that become narrower as q increases simply by allowing q to be very large (close to x), and assigning each angle to the fraction closest to it. This is the common procedure. However, this makes matters more difficult, in that we would have to minimize at the same time the factors in front of terms x/q, $x/\sqrt{q}$, etc., and those in front of terms q, $\sqrt{qx}$, and so on. (These terms are being compared to the trivial bound x.) Instead, we choose to strive for a direct dependence on δ throughout; this will allow us to cap q at a much lower level, thus making terms such as q and $\sqrt{qx}$ negligible.

How good must our bounds be? Since the major-arc bounds are valid only for $q \leq r = 300000$ and $|\delta| \leq 4r/q$, we cannot afford even a single factor of $\log x$ (or any other function tending to ∞ as $x \to \infty$) in front of terms such as $x/\sqrt{q|\delta_0|}$: a factor like that would make the term larger than the trivial bound x for $q|\delta_0|$ equal to a constant (r, say) and x very large. Apparently, there was no such "log-free bound" with explicit constants in the literature, even though such bounds were considered to be in principle feasible, and even though previous work ([5, 11, 12, 58]) had gradually decreased the number of factors of $\log x$. (In limited ranges for q, there were log-free bounds without explicit constants; see [11, 53]. The estimate in [69, Thm. 2a, 2b] was almost log-free, but not quite. There were

also bounds [3, 37] that used L-functions, and thus were not really useful in a truly minor-arc regime.)

It also seemed clear that a main bound proportional to $(\log q)^2 x/\sqrt{q}$ (as in [58]) was too large. At the same time, it was not really necessary to reach a bound of the best possible form that could be found through Vinogradov's basic approach, namely

$$|S_\eta(\alpha, x)| \leq C\frac{x\sqrt{q}}{\phi(q)}. \tag{4.1}$$

Such a bound had been proven by Ramaré [53] for q in a limited range and C non-explicit; later, in [50] Ramaré broadened the range to $q \leq x^{1/48}$ and gave an explicit value for C, namely, $C = 13000$. Such a bound is a notable achievement, but, unfortunately, it is not useful for our purposes. Rather, we will aim at a bound whose main term is bounded by a constant around 1 times $x(\log \delta_0 q)/\sqrt{\delta_0 \phi(q)}$; this is slightly worse asymptotically than (4.1), but it is much better in the delicate range of $\delta_0 q \sim 300000$.

* * *

We see that we have several tasks. One of them is the removal of logarithms: we cannot afford a single factor of $\log x$, and, in practice, we can afford at most one factor of $\log q$. Removing logarithms will be possible in part because of the use of efficient techniques (the large sieve for sequences with prime support) but also because we will be able to find cancellation at several places in sums coming from a combinatorial identity (namely, Vaughan's identity). The task of finding cancellation efficiently (that is, with good constants) is particularly delicate. Bounding a sum such as $\sum_n \mu(n)$ efficiently is harder than estimating a sum such as $\sum_n \Lambda(n)$ equally well, even though we are used to thinking of these problems as equivalent.

We have said that our bounds will improve as $|\delta|$ increases. This dependence on δ will be secured in different ways at different places. Sometimes δ will appear as an argument, as in $\widehat{\eta}(-\delta)$; for η piecewise continuous with $\eta' \in L_1$, we know that $|\widehat{\eta}(t)| \to 0$ as $|t| \to \infty$. Sometimes we will obtain a dependence on δ by using several different rational approximations to the same $\alpha \in \mathbb{R}$. Lastly, we will obtain a good dependence on δ in bilinear sums by supplying a scattered input to a large sieve.

If there is a main moral to the argument, it lies in the close relation between the circle method and the large sieve. The circle method rests on the estimation of an integral involving a Fourier transform $\widehat{f} : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$; as we will later see, this leads naturally to estimating the ℓ_2-norm of $\widehat{f}$ on subsets (namely, unions of arcs) of the circle $\mathbb{R}/\mathbb{Z}$. The large sieve can be seen as an approximate discrete version of Plancherel's identity, which states that $|\widehat{f}|_2 = |f|_2$.

Both in this section and in §5, we shall use the large sieve in part so as to use the fact that some of the functions we work with have prime support, i.e., are non-zero only on prime numbers. There are ways to use prime support to improve the output of the large sieve. In §5, these techniques will be refined and then translated to the context of the circle method, where f has (essentially) prime support and $|\widehat{f}|^2$ must be integrated over unions of arcs. The main point is that the large sieve is not being used as a black box; rather, we can adapt ideas from (say) the large-sieve context and apply them to the circle method.

Lastly, there are the benefits of a continuous η. Hardy and Littlewood already used a continuous η; this was abandoned by Vinogradov, presumably for the sake of simplicity.

The idea that smooth weights η can be superior to sharp truncations is now commonplace. As we shall see, using a continuous η is helpful in the minor-arcs regime, but not as crucial there as for the major arcs. We will not use a smooth η; we will prove our estimates for any continuous η that is piecewise C_1, and then, towards the end, we will choose to use the same weight $\eta = \eta_2$ as in [58], in part because it has compact support, and in part for the sake of comparison. The moral here is not quite the common dictum "always smooth", but rather that different kinds of smoothing can be appropriate for different tasks; in the end, we will show how to coordinate different smoothing functions η.

4.2. Combinatorial identities. Generally, since Vinogradov, a treatment of the minor arcs starts with a combinatorial identity expressing $\Lambda(n)$ (or the characteristic function of the primes) as a sum of two or more convolutions. (In this section, by a convolution $f * g$, we will mean the *Dirichlet convolution* $(f * g)(n) = \sum_{d|n} f(d)g(n/d)$, i.e., the multiplicative convolution on the semigroup of positive integers.)

In some sense, the archetypical identity is

$$\Lambda = \mu * \log,$$

but it will not usually do: the contribution of $\mu(d)\log(n/d)$ with d close to n is too difficult to estimate precisely. There are alternatives: for example, there is Selberg's identity

$$\Lambda(n)\log n = \mu * \log^2 - \Lambda * \Lambda, \tag{4.2}$$

or the generalization of this to $\Lambda(n)(\log n)^k = \mu * \log^{k+1} - \dots$ (Bomberi-Selberg). Another useful (and very simple) identity was that used by Daboussi's [12].

The proof of Vinogradov's three-prime result was simplified substantially in [64] by the introduction of *Vaughan's identity*:

$$\Lambda(n) = \mu_{\le U} * \log - \Lambda_{\le V} * \mu_{\le U} * 1 + 1 * \mu_{>U} * \Lambda_{>V} + \Lambda_{\le V}, \tag{4.3}$$

where we are using the notation

$$f_{\le W} = \begin{cases} f(n) & \text{if } n \le W, \\ 0 & \text{if } n > W, \end{cases} \qquad f_{>W} = \begin{cases} 0 & \text{if } n \le W, \\ f(n) & \text{if } n > W. \end{cases}$$

Of the resulting sums ($\sum_n (\mu_{\le U} * \log)(n) e(\alpha n)\eta(n/x)$, etc.), the first three are said to be of *type I*, *type I* (again) and *type II*; the last sum, $\sum_{n \le V} \Lambda(n)$, is negligible.

One of the advantages of Vaughan's identity is its flexibility: we can set U and V to whatever values we wish. Its main disadvantage is that it is not "log-free", in that it seems to impose the loss of two factors of $\log x$: if we sum each side of (4.3) from 1 to x, we obtain $\sum_{n \le x} \Lambda(n) \sim x$ on the left side, whereas, if we bound the sum on the right side without the use of cancellation, we obtain a bound of $x(\log x)^2$. Of course, we will obtain some cancellation from the phase $e(\alpha n)$, but that is not enough.

As was pointed out in [58], it is possible to get a factor of $(\log q)^2$ instead of a factor of $(\log x)^2$ in the type II sums by setting U and V appropriately. A factor of $(\log q)^2$ is still too large in practice, and there are also the factors of $\log x$ in type I sums. Vinogradov had already managed to get an essentially log-free result (by a rather difficult procedure) in [69, Ch. IX]. The result in [11] is log-free. Unfortunately, the explicit result in [12] – the study of which encouraged me at the beginning of the project – is not. For a while, I worked with

the Bombieri-Selberg identity with $k = 2$. Ramaré obtained a log-free bound in [53] using the Diamond-Steinig identity, which is related to Bombieri-Selberg.

In the end, I decided to use Vaughan's identity. This posed a challenge: to obtain cancellation in Vaughan's identity at every possible step, beyond the cancellation given by the phase $e(\alpha n)$. It is clear that the presence of the Möbius function μ should give, in principle, some cancellation; we will show how to use it to obtain as much cancellation as we need.

4.3. Type I sums. There are two type I sums, namely,

$$\sum_{m\le U} \mu(m) \sum_n (\log n) e(\alpha m n) \eta\left(\frac{mn}{x}\right) \tag{4.4}$$

and

$$\sum_{v\le V} \Lambda(v) \sum_{u\le U} \mu(u) \sum_n e(\alpha v u n) \eta\left(\frac{vun}{x}\right). \tag{4.5}$$

In either case, $\alpha = a/q + \delta/x$, where q is larger than a constant r and $|\delta/x| \le 1/qQ_0$ for some $Q_0 > \max(q, \sqrt{x})$. For the purposes of this exposition, we will set it as our task to estimate the slightly simpler sum

$$\sum_{m\le D} \mu(m) \sum_n e(\alpha m n) \eta\left(\frac{mn}{x}\right), \tag{4.6}$$

where D can be U or UV or something else less than x.

Why can we consider this simpler sum without omitting anything essential? It is clear that (4.4) is of the same kind as (4.6). The inner double sum in (4.5) is just (4.6) with αv instead of α; this enables us to estimate (4.5) by means of (4.6) for q small, i.e., the more delicate case. If q is not small, then the approximation $\alpha v \sim av/q$ may not be accurate enough. In that case, we collapse the two outer sums in (4.5) into a sum $\sum_n (\Lambda_{\le V} * \mu_{\le U})(n)$, and treat all of (4.5) much as we will treat (4.6); since q is not small, we can afford to bound $(\Lambda_{\le V} * \mu_{\le U})(n)$ trivially (by $\log n$) in the less sensitive terms.

Let us first outline Vinogradov's procedure for bounding type I sums. Just by summing a geometric series, we get $\left|\sum_{n\le N} e(\alpha n)\right| \le \min(N, c/\{\alpha\})$, where c is a constant and $\{\alpha\}$ is the distance from α to the nearest integer. Vinogradov splits the outer sum in (4.6) into sums of length q. When m runs on an interval of length q, the angle am/q runs through all fractions of the form b/q; due to the error δ/x, αm could be close to 0 for two values of n, but otherwise $\{\alpha m\}$ takes values bounded below by $1/q$, $2/q$, etc. Thus

$$\left|\sum_{y<m\le y+q} \mu(m) \sum_{n\le N} e(\alpha m n)\right| \le \sum_{y<m\le y+q} \left|\sum_{n\le N} e(\alpha m n)\right| \le \frac{2N}{m} + 2cq \log eq \tag{4.7}$$

for any $y \ge 0$.

There are several ways to improve this. One is simply to estimate the inner sum more precisely; this was already done in [12]. One can also define a smoothing function η, as in (4.6); it is easy to get

$$\left|\sum_{n\le N} e(\alpha n) \eta\left(\frac{n}{x}\right)\right| \le \min\left(x|\eta|_1 + \frac{|\eta'|_1}{2}, \frac{|\eta'|_1}{2|\sin(\pi\alpha)|}, \frac{|\widehat{\eta''}|_\infty}{4x(\sin \pi\alpha)^2}\right).$$

Except for the third term, this is as in [58]. We could also choose carefully which bound to use for each m; surprisingly, this gives an improvement – in fact, an important one, for m large. However, we still get a term proportional to N/m as in (4.7), and this contributes about $(x \log x)/q$ to the sum (4.6), thus giving us an estimate that is not log-free.

What we have to do, naturally, is to take out the terms with $q|m$ for m small. We obtain a log-free bound for the sum over the terms with $m \leq M = \min(D, Q/2)$ with $q \nmid m$, since αm is then never too close to 0. For $m \leq M$ divisible by q, we can estimate the inner sum in (4.6) by the Poisson summation formula; writing $m = aq$, we get a main term

$$\frac{x\mu(q)}{q} \cdot \widehat{\eta}(-\delta) \cdot \sum_{\substack{a \leq M/q \\ (a,q)=1}} \frac{\mu(a)}{a}, \tag{4.8}$$

where (a, q) stands for the greatest common divisor of a and q. It is clear that we have to get cancellation over μ here. There is an elegant elementary argument [22] showing that the absolute value of the sum in (4.8) is at most 1. We need to gain one more log, however. This was done by Ramaré [49].

What shall we do for $m > Q/2$? We can always give a bound

$$\sum_{y < m \leq y+q} \min\left(A, \frac{C}{|\sin \pi\alpha n|^2}\right) \leq 3A + \frac{4q}{\pi}\sqrt{AC} \tag{4.9}$$

for y arbitrary; since AC will be of constant size, $(4q/\pi)\sqrt{AC}$ is pleasant enough, but the contribution of $3A \sim 3|\eta|_1 x/y$ seems lethal (it adds a multiple of $(x \log x)/q$ to the total) and at first sight unavoidable: the values of m for which αm is close to 0 no longer correspond to the congruence class $m \equiv 0 \bmod q$, and thus cannot be taken out.

The solution is to switch approximations. (The idea of using different approximations to the same α is neither new nor recent in the general context of the circle method: see [66, §2.8, Ex. 2]. What may be new is its use to clear a hurdle in type I sums.) What does this mean? If α were exactly, or almost exactly, a/q, then there would be no other very good approximations in a reasonable range. However, note that we can *define* $Q = \lfloor x/|\delta q| \rfloor$ for $\alpha = a/q + \delta/x$, and still have $|\alpha - a/q| \leq 1/qQ$. If δ is very small, Q will be larger than $2D$, and there will be no terms with $Q/2 < m \leq D$ to worry about.

What happens if δ is not very small? We know that, for any Q', there is an approximation a'/q' to α with $|\alpha - a'/q'| \leq 1/q'Q'$ and $q' \leq Q'$. However, for $Q' > Q$, we know that a'/q' cannot equal a/q: by the definition of Q, the approximation a/q is not good enough, i.e., $|\alpha - a/q| \leq 1/qQ'$ does not hold. Since $a/q \neq a'/q'$, we see that $|a/q - a'/q'| \geq 1/qq'$, and, if we take $Q' \geq (1+\epsilon)Q$, this implies that q' is relatively large ($q' \geq (\epsilon/(1+\epsilon))Q$).

Thus, for $m > Q/2$, the solution is to apply (4.9) with a'/q' instead of a/q. The contribution of A fades into insignificance: for the first sum over a range $y < m \leq y + q'$, $y \geq Q/2$, it contributes at most $x/(Q/2)$, and all the other contributions of A sum up to at most a constant times $(x \log x)/q'$.

Proceeding in this way, we obtain a total bound for (4.6) whose main terms are proportional to

$$\frac{1}{\phi(q)} \frac{x}{\log \frac{x}{q}} \min\left(1, \frac{1}{\delta^2}\right), \quad \frac{2}{\pi}\sqrt{|\widehat{\eta''}|_\infty} \cdot D \quad \text{and} \quad q \log \max\left(\frac{D}{q}, q\right), \tag{4.10}$$

with good, explicit constants. The first term – usually the largest one – is precisely what we needed: it is proportional to $(1/\phi(q))x/\log x$ for q small, and decreases rapidly as $|\delta|$ increases.

4.4. Type II, or bilinear, sums. We must now bound

$$S = \sum_m (1 * \mu_{>U})(m) \sum_{n>V} \Lambda(n) e(\alpha mn) \eta(mn/x).$$

At this point it is convenient to assume that η is the Mellin convolution of two functions. The *multiplicative* or *Mellin convolution* on $\mathbb{R}^+$ is defined by

$$(\eta_0 *_M \eta_1)(t) = \int_0^\infty \eta_0(r)\eta_1\left(\frac{t}{r}\right)\frac{dr}{r}.$$

Tao [58] takes $\eta = \eta_2 = \eta_1 *_M \eta_1$, where η_1 is a brutal truncation, viz., the function taking the value 2 on $[1/2, 1]$ and 0 elsewhere. We take the same η_2, in part for comparison purposes, and in part because this will allow us to use off-the-shelf estimates on the large sieve. (Brutal truncations are rarely optimal in principle, but, as they are very common, results for them have been carefully optimized in the literature.) Clearly

$$S = \int_V^{X/U} \sum_m \left(\sum_{\substack{d>U\\ d|m}} \mu(d)\right) \eta_1\left(\frac{m}{x/W}\right) \cdot \sum_{n\geq V} \Lambda(n) e(\alpha mn) \eta_1\left(\frac{n}{W}\right)\frac{dW}{W}. \tag{4.11}$$

By Cauchy-Schwarz, the integrand is at most $\sqrt{S_1(U,W)S_2(V,W)}$, where

$$\begin{aligned} S_1(U,W) &= \sum_{\frac{x}{2W} < m \leq \frac{x}{W}} \left|\sum_{\substack{d>U\\ d|m}} \mu(d)\right|^2, \\ S_2(V,W) &= \sum_{\frac{x}{2W} \leq m \leq \frac{x}{W}} \left|\sum_{\max\left(V, \frac{W}{2}\right) \leq n \leq W} \Lambda(n) e(\alpha mn)\right|^2. \end{aligned} \tag{4.12}$$

We must bound $S_1(U,W)$ by a constant times x/W. We are able to do this – with a good constant. (A careless bound would have given a multiple of $(x/U)\log^3(x/U)$, which is much too large.) First, we reduce $S_1(U,W)$ to an expression involving an integral of

$$\sum_{\substack{r_1\leq x\\ (r_1,r_2)=1}} \sum_{r_2\leq x} \frac{\mu(r_1)\mu(r_2)}{\sigma(r_1)\sigma(r_2)}. \tag{4.13}$$

We can bound (4.13) by the use of bounds on $\sum_{n\leq t}\mu(n)/n$, combined with the estimation of infinite products by means of approximations to $\zeta(s)$ for $s \to 1^+$. After some additional manipulations, we obtain a bound for $S_1(U,W)$ whose main term is at most $(3/\pi^2)(x/W)$ for each W, and closer to $0.22482x/W$ on average over W.

(This is as good a point as any to say that, throughout, we can use a trick in [58] that allows us to work with odd values of integer variables throughout, instead of letting m or n range over all integers. Here, for instance, if m and n are restricted to be odd, we obtain a bound of $(2/\pi^2)(x/W)$ for individual W, and $0.15107x/W$ on average over W.)

Let us now bound $S_2(V, W)$. This is traditionally done by Linnik's dispersion method. However, it should be clear that the thing to do nowadays is to use a large sieve, and, more specifically, a large sieve for primes. In order to take advantage of prime support, we use Montgomery's inequality ([33, 43]; see the expositions in [44, pp. 27–29] and [34, §7.4]) combined with Montgomery and Vaughan's large sieve with weights [45, (1.6)], following the general procedure in [45, (1.6)]. We obtain a bound of the form

$$\frac{\log W}{\log \frac{W}{2q}} \left(\frac{x}{4\phi(q)} + \frac{qW}{\phi(q)} \right) \frac{W}{2} \tag{4.14}$$

on $S_2(V, W)$, where, of course, we can also choose *not* to gain a factor of $\log W/2q$ if q is close to or greater than W.

It remains to see how to gain a factor of $|\delta|$ in the major arcs, and more specifically in $S_2(V, W)$. To explain this, let us step back and take a look at what the large sieve is. Given a civilized function $f : \mathbb{Z} \to \mathbb{C}$, Plancherel's identity tells us that

$$\int_{\mathbb{R}/\mathbb{Z}} \left| \widehat{f}\,(\alpha) \right|^2 d\alpha = \sum_n |f(n)|^2.$$

The large sieve can be seen as an approximate, or statistical, version of this: for a "sample" of points $\alpha_1, \alpha_2, \dots, \alpha_k$ satisfying $|\alpha_i - \alpha_j| \geq \beta$ for $i \neq j$, it tells us that

$$\sum_{1 \leq j \leq k} \left| \widehat{f}\,(\alpha_j) \right|^2 \leq (X + \beta^{-1}) \sum_n |f(n)|^2, \tag{4.15}$$

assuming that f is supported on an interval of length X.

Now consider $\alpha_1 = \alpha, \alpha_2 = 2\alpha, \alpha_3 = 3\alpha \dots$. If $\alpha = a/q$, then the angles $\alpha_1, \dots, \alpha_q$ are well-separated, i.e., they satisfy $|\alpha_i - \alpha_j| \geq 1/q$, and so we can apply (4.15) with $\beta = 1/q$. However, $\alpha_{q+1} = \alpha_1$. Thus, if we have an outer sum of length $L > q$ – in (4.12), we have an outer sum of length $L = x/2W$ – we need to split it into $\lceil L/q \rceil$ blocks of length q, and so the total bound given by (4.15) is $\lceil L/q \rceil (X + q) \sum_n |f(n)|^2$. Indeed, this is what gives us (4.14), which is fine, but we want to do better for $|\delta|$ larger than a constant.

Suppose, then, that $\alpha = a/q + \delta/x$, where $|\delta| > 8$, say. Then the angles α_1 and α_{q+1} are not identical: $|\alpha_1 - \alpha_{q+1}| \leq q|\delta|/x$. We also see that α_{q+1} is at a distance at least $q|\delta|/x$ from $\alpha_2, \alpha_3, \dots \alpha_q$, provided that $q|\delta|/x < 1/q$. We can go on with $\alpha_{q+2}, \alpha_{q+3}, \dots$, and stop only once there is overlap, i.e., only once we reach α_m such that $m|\delta|/x \geq 1/q$. We then give all the angles $\alpha_1, \dots, \alpha_m$ – which are separated by at least $q|\delta|/x$ from each other – to the large sieve at the same time. We do this $\lceil L/m \rceil \leq \lceil L/(x/|\delta|q) \rceil$ times, and obtain a total bound of $\lceil L/(x/|\delta|q) \rceil (X + x/|\delta|q) \sum_n |f(n)|^2$, which, for $L = x/2W$, $X = W/2$, gives us about

$$\left(\frac{x}{4Q} \frac{W}{2} + \frac{x}{4} \right) \log W$$

provided that $L \geq x/|\delta|q$ and, as usual, $|\alpha - a/q| \leq 1/qQ$. This is very small compared to the trivial bound $\lesssim xW/8$.

What happens if $L < x/|\delta q|$? Then there is never any overlap: we consider all angles α_i, and give them all together to the large sieve. The total bound is $(W^2/4 + xW/2|\delta|q) \log W$. If $L = x/2W$ is smaller than, say, $x/3|\delta q|$, then we see clearly that there are non-intersecting swarms of α_i around the rationals a/q. We can thus save a factor of $\log$ (or rather $(\phi(q)/q) \log (W/|\delta q|)$) by applying Montgomery's inequality, which operates by strewing displacements of the given angles (or, here, the swarms) around the circle to the extent possible while keeping everything well-separated. In this way, we obtain a bound of the form

$$\frac{\log W}{\log \frac{W}{|\delta|q}} \left(\frac{x}{|\delta|\phi(q)} + \frac{q}{\phi(q)} \frac{W}{2} \right) \frac{W}{2}.$$

Compare this to (4.14); we have gained a factor of $|\delta|/4$, and so we use this estimate when $|\delta| > 4$. (In [28], the criterion is $|\delta| > 8$, but, since there we have $2\alpha = a/q + \delta/x$, the value of δ there is twice what it is here; this is a consequence of working with sums over the odd integers, as in [58].)

* * *

We have succeeded in eliminating all factors of $\log$ we came across. The only factor of $\log$ that remains is $\log x/UV$, coming from the integral $\int_V^{x/U} dW/W$. Thus, we want UV to be close to x, but we cannot let it be too close, since we also have a term proportional to $D = UV$ in (4.10), and we need to keep it substantially smaller than x. We set U and V so that UV is $x/\sqrt{q \max(4, |\delta|)}$ or thereabouts.

In the end, after some work, we obtain the main result in [28]. We recall that $S_\eta(\alpha, x) = \sum_n \Lambda(n) e(\alpha n) \eta(n/x)$ and $\eta_2 = \eta_1 *_M \eta_1 = 4 \cdot 1_{[1/2,1]} * 1_{[1/2,1]}$.

Theorem 4.1. *Let $x \geq x_0$, $x_0 = 2.16 \cdot 10^{20}$. Let $2\alpha = a/q + \delta/x$, $q \leq Q$, $\gcd(a, q) = 1$, $|\delta/x| \leq 1/qQ$, where $Q = (3/4)x^{2/3}$. If $q \leq x^{1/3}/6$, then*

$$|S_\eta(\alpha, x)| \leq \frac{R_{x,\delta_0 q} \log \delta_0 q + 0.5}{\sqrt{\delta_0 \phi(q)}} \cdot x + \frac{2.5x}{\sqrt{\delta_0 q}} + \frac{2x}{\delta_0 q} \cdot L_{x,\delta_0,q} + 3.2x^{5/6}, \tag{4.16}$$

where $\delta_0 = \max(2, |\delta|/4)$,

$$\begin{aligned} R_{x,t} &= 0.27125 \log \left(1 + \frac{\log 4t}{2 \log \frac{9x^{1/3}}{2.004t}} \right) + 0.41415 \\ L_{x,\delta,q} &= \frac{\log \delta^{\frac{7}{4}} q^{\frac{13}{4}} + \frac{80}{9}}{\phi(q)/q} + \log q^{\frac{80}{9}} \delta^{\frac{16}{9}} + \frac{111}{5}. \end{aligned} \tag{4.17}$$

If $q > x^{1/3}/6$, then

$$|S_\eta(\alpha, x)| \leq 0.2727 x^{5/6} (\log x)^{3/2} + 1218 x^{2/3} \log x.$$

The factor $R_{x,t}$ is small in practice; for typical "difficult" values of x and $\delta_0 x$, it is less than 1. The crucial things to notice in (4.16) are that there is no factor of $\log x$, and that, in the main term, there is only one factor of $\log \delta_0 q$. The fact that δ_0 helps us as it grows is precisely what enables us to take major arcs that get narrower and narrower as q grows.

5. Integrals over the major and minor arcs

So far, we have sketched (§3) how to estimate $S_\eta(\alpha, x)$ for α in the major arcs and η based on the Gaussian $e^{-t^2/2}$, and also (§4) how to bound $|S_\eta(\alpha, x)|$ for α in the minor arcs and $\eta = \eta_2$, where $\eta_2 = 4 \cdot 1_{[1/2,1]} *_M 1_{[1/2,1]}$. We now must show how to use such information to estimate integrals such as the ones in (2.3).

We will use two smoothing functions η_+, η_*; in the notation of (2.2), we set $f_1 = f_2 = \Lambda(n)\eta_+(n/x)$, $f_3 = \Lambda(n)\eta_*(n/x)$, and so we must give a lower bound for

$$\int_{\mathfrak{M}} (S_{\eta_+}(\alpha, x))^2 S_{\eta_*}(\alpha, x) e(-\alpha n) d\alpha \tag{5.1}$$

and an upper bound for

$$\int_{\mathfrak{m}} \left|S_{\eta_+}(\alpha, x)\right|^2 S_{\eta_*}(\alpha, x) e(-\alpha n) d\alpha \tag{5.2}$$

so that we can verify (2.3).

The traditional approach to (5.2) is to bound

$$\begin{aligned} \int_{\mathfrak{m}} (S_{\eta_+}(\alpha, x))^2 S_{\eta_*}(\alpha, x) e(-\alpha n) d\alpha &\le \int_{\mathfrak{m}} \left|S_{\eta_+}(\alpha, x)\right|^2 d\alpha \cdot \max_{\alpha \in \mathfrak{m}} \widehat{\eta_*}(\alpha) \\ &\le \sum_n \Lambda(n)^2 \eta_+^2\left(\frac{n}{x}\right) \cdot \max_{\alpha \in \mathfrak{m}} S_{\eta_*}(\alpha, x). \end{aligned} \tag{5.3}$$

Since the sum over n is of the order of $x \log x$, this is not log-free, and so cannot be good enough; we will later see how to do better. Still, this gets the main shape right: our bound on (5.2) will be proportional to $|\eta_+|_2^2 |\eta_*|_1$. Moreover, we see that η_* has to be such that we know how to bound $|S_{\eta_*}(\alpha, x)|$ for $\alpha \in \mathfrak{m}$, while our choice of η_+ is more or less free, at least as far as the minor arcs are concerned.

What about the major arcs? In order to do anything on them, we will have to be able to estimate both $\eta_+(\alpha)$ and $\eta_*(\alpha)$ for $\alpha \in \mathfrak{M}$. Once we do this, we will obtain that the main term of (5.1) is an infinite product (independent of the smoothing functions), times x^2, times

$$\int_0^\infty \int_0^\infty \eta_+(t_1)\eta_+(t_2)\eta_*\left(\frac{n}{x} - (t_1 + t_2)\right) dt_1 dt_2. \tag{5.4}$$

In other words, we want to maximize (or nearly maximize) the expression on the right of (5.4) divided by $|\eta_+|_2^2 |\eta_*|_1$.

One way to do this is to let η_* be concentrated on a small interval $[0, \epsilon)$. Then the right side of (5.4) is approximately

$$|\eta_*|_1 \cdot \int_0^\infty \eta_+(t)\eta_+\left(\frac{n}{x} - t\right) dt. \tag{5.5}$$

To maximize this, we should make sure that $\eta_+(t) \sim \eta_+(n/x - t)$. We set $x \sim n/2$, and see that we should define η_+ so that it is supported on $[0, 2]$ and symmetric around $t = 1$, or nearly so; this will maximize the ratio of (5.5) to $|\eta_+|_2^2 |\eta_*|_1$.

We should do this while making sure that we will know how to estimate $S_{\eta_+}(\alpha, x)$ for $\alpha \in \mathfrak{M}$. We know how to estimate $S_\eta(\alpha, x)$ very precisely for functions of the form

$\eta(t) = g(t)e^{-t^2/2}$, $\eta(t) = g(t)te^{-t^2/2}$, etc., where $g(t)$ is band-limited. We will work with a function η_+ of that form, chosen so as to be very close (in ℓ_2 norm) to a function $\eta_\circ$ that is in fact supported on $[0, 2]$ and symmetric around $t = 1$.

We choose

$$\eta_\circ(t) = \begin{cases} t^2(2-t)^3 e^{-(t-1)^2/2} & \text{if } t \in [0,2], \\ 0 & \text{if } t \notin [0,2]. \end{cases}$$

This function is obviously symmetric ($\eta_\circ(t) = \eta_\circ(2-t)$) and vanishes to high order at $t = 0$, besides being supported on $[0, 2]$.

We set $\eta_+(t) = h_R(t)te^{-t^2/2}$, where $h_R(t)$ is an approximation to the function

$$h(t) = \begin{cases} t^2(2-t)^3 e^{t-\frac{1}{2}} & \text{if } t \in [0,2] \\ 0 & \text{if } t \notin [0,2]. \end{cases}$$

We just let $h_R(t)$ be the inverse Mellin transform of the truncation of Mh to an interval $[-iR, iR]$, or, what is the same,

$$h_R(t) = \int_0^\infty h(ty^{-1}) F_R(y) \frac{dy}{y},$$

where $F_R(t) = \sin(R \log y)/(\pi \log y)$ (the Dirichlet kernel with a change of variables); since the Mellin transform of $te^{-t^2/2}$ is regular at $s = 0$, the Mellin transform $M\eta_+$ will be holomorphic in a neighborhood of $\{s : 0 \le \Re(s) \le 1\}$, even though the truncation of Mh to $[-iR, iR]$ is brutal. Set $R = 200$, say. By the fast decay of $Mh(it)$ and the fact that the Mellin transform M is an isometry, $|(h_R(t) - h(t))/t|_2$ is very small, and hence so is $|\eta_+ - \eta_\circ|_2$, as we desired.

But what about the requirement that we be able to estimate $S_{\eta_*}(\alpha, x)$ for both $\alpha \in \mathfrak{m}$ and $\alpha \in \mathfrak{M}$?

Generally speaking, if we know how to estimate $S_{\eta_1}(\alpha, x)$ for some $\alpha \in \mathbb{R}/\mathbb{Z}$ and we also know how to estimate $S_{\eta_2}(\alpha, x)$ for all other $\alpha \in \mathbb{R}/\mathbb{Z}$, where η_1 and η_2 are two smoothing functions, then we know how to estimate $S_{\eta_3}(\alpha, x)$ for all $\alpha \in \mathbb{R}/\mathbb{Z}$, where $\eta_3 = \eta_1 *_M \eta_2$, or, more generally, $\eta_*(t) = (\eta_1 *_M \eta_2)(\kappa t)$, $\kappa > 0$ a constant. This is a simple exercise in exchanging the order of integration and summation:

$$\begin{aligned} S_{\eta_*}(\alpha, x) &= \sum_n \Lambda(n) e(\alpha n) (\eta_1 *_M \eta_2)\left(\kappa \frac{n}{x}\right) \\ &= \int_0^\infty \sum_n \Lambda(n) e(\alpha n) \eta_1(\kappa r) \eta_2\left(\frac{n}{rx}\right) \frac{dr}{r} = \int_0^\infty \eta_1(\kappa r) S_{\eta_2}(rx) \frac{dr}{r}, \end{aligned}$$

and similarly with η_1 and η_2 switched.

Now that we have chosen our smoothing weights η_+ and η_*, we have to estimate the major-arc integral (5.1) and the minor-arc integral (5.2). What follows can actually be done for general η_+ and η_*; we could have left our particular choice of η_+ and η_* for the end.

Estimating the major-arc integral (5.1) may sound like an easy task, since we have rather precise estimates for $S_\eta(\alpha, x)$ ($\eta = \eta_+, \eta_*$) when α is on the major arcs; we could just replace $S_\eta(\alpha, x)$ in (5.1) by the approximation given by (3.3) and (3.7). It is, however, more efficient to express (5.1) as the sum of the contribution of the trivial character (a sum of

integrals of $(\widehat{\eta}(-\delta)x)^3$, where $\widehat{\eta}(-\delta)x$ comes from (3.7)), plus a term of the form

$$(\text{maximum of } \sqrt{q}\cdot E(q) \text{ for } q\le r)\cdot \int_{\mathfrak{M}} \left|S_{\eta_+}(\alpha,x)\right|^2 d\alpha,$$

where $E(q)=E$ is as in (3.8), plus two other terms of the same form. As usual, the major arcs $\mathfrak{M}$ are the arcs around rationals a/q with $q\le r$. We will soon discuss how to bound the integral of $\left|S_{\eta_+}(\alpha,x)\right|^2$ over arcs around rationals a/q with $q\le s$, s arbitrary. Here, however, it is best to estimate the integral over $\mathfrak{M}$ using the estimate on $S_{\eta_+}(\alpha,x)$ from (3.3) and (3.7); we obtain a great deal of cancellation, with the effect that, for χ non-trivial, the error term in (3.8) appears only when it gets squared, and thus becomes negligible.

The contribution of the trivial character has an easy approximation, thanks to the fast decay of $\widehat{\eta_\circ}$. We obtain that the major-arc integral (5.1) equals a main term $C_0 C_{\eta_\circ,\eta_*} x^2$, where

$$C_0 = \prod_{p\mid n}\left(1-\frac{1}{(p-1)^2}\right)\cdot \prod_{p\nmid n}\left(1+\frac{1}{(p-1)^3}\right),$$

$$C_{\eta_\circ,\eta_*} = \int_0^\infty\int_0^\infty \eta_\circ(t_1)\eta_\circ(t_2)\eta_*\left(\frac{n}{x}-(t_1+t_2)\right)dt_1dt_2,$$

plus several small error terms. We have already chosen $\eta_\circ$, η_* and x so as to (nearly) maximize $C_{\eta_\circ,\eta_*}$.

It is time to bound the minor-arc integral (5.2). As we said in §5, we must do better than the usual bound (5.3). Since our minor-arc bound (4.16) on $|S_\eta(\alpha,x)|$, $\alpha\sim a/q$, decreases as q increases, it makes sense to use partial summation together with bounds on

$$\int_{\mathfrak{m}_s}|S_{\eta_+}(\alpha,x)|^2 = \int_{\mathfrak{M}_s}|S_{\eta_+}(\alpha,x)|^2 d\alpha - \int_{\mathfrak{M}}|S_{\eta_+}(\alpha,x)|^2 d\alpha,$$

where $\mathfrak{m}_s$ denotes the arcs around a/q, $r<q\le s$, and $\mathfrak{M}_s$ denotes the arcs around all a/q, $q\le s$. We already know how to estimate the integral on $\mathfrak{M}$. How do we bound the integral on $\mathfrak{M}_s$?

In order to do better than the trivial bound $\int_{\mathfrak{M}_s}\le\int_{\mathbb{R}/\mathbb{Z}}$, we will need to use the fact that the series (3.2) defining $S_{\eta_+}(\alpha,x)$ is essentially supported on prime numbers. Bounding the integral on $\mathfrak{M}_s$ is closely related to the problem of bounding

$$\sum_{q\le s}\sum_{\substack{a \bmod q\\(a,q)=1}}\left|\sum_{n\le x} a_n e(a/q)\right|^2 \tag{5.6}$$

efficiently for s considerably smaller than $\sqrt{x}$ and a_n supported on the primes $\sqrt{x}<p\le x$. This is a classical problem in the study of the large sieve. The usual bound on (5.6) (by, for instance, Montgomery's inequality) has a gain of a factor of $2e^\gamma(\log s)/(\log x/s^2)$ relative to the bound of $(x+s^2)\sum_n|a_n|^2$ that one would get from the large sieve without using prime support. Heath-Brown proceeded similarly to bound

$$\int_{\mathfrak{M}_s}|S_{\eta_+}(\alpha,x)|^2 d\alpha \lesssim \frac{2e^\gamma\log s}{\log x/s^2}\int_{\mathbb{R}/\mathbb{Z}}|S_{\eta_+}(\alpha,x)|^2 d\alpha. \tag{5.7}$$

This already gives us the gain of $C(\log s)/\log x$ that we absolutely need, but the constant C is suboptimal; the factor in the right side of (5.7) should really be $(\log s)/\log x$, i.e., C should be 1. We cannot reasonably hope to do better than $2(\log s)/\log x$ in the minor arcs due to what is known as the *parity problem* in sieve theory. As it turns out, Ramaré [52] had given general bounds on the large sieve that were clearly conducive to better bounds on (5.6), though they involved a ratio that was not easy to bound in general.

I used several careful estimations (including [51, Lem. 3.4]) to reduce the problem of bounding this ratio to a finite number of cases, which I then checked by rigorous computation. This approach gave a bound on (5.6) with a factor of size close to $2(\log s)/\log x$. (This solves the large-sieve problem for $s \le x^{0.3}$; it would still be worthwhile to give a computation-free proof for all $s \le x^{1/2-\epsilon}$, $\epsilon > 0$.) It was then easy to give an analogous bound for the integral over $\mathfrak{M}_s$, namely,

$$\int_{\mathfrak{M}_s} |S_{\eta_+}(\alpha, x)|^2 d\alpha \lesssim \frac{2 \log s}{\log x} \int_{\mathbb{R}/\mathbb{Z}} |S_{\eta_+}(\alpha, x)|^2 d\alpha,$$

where $\lesssim$ can easily be made precise by replacing $\log s$ by $\log s + 1.36$ and $\log x$ by $\log x + c$, where c is a small constant. Without this improvement, the main theorem would still have been proved, but the required computation time would have been multiplied by a factor of considerably more than $e^{3\gamma} = 5.6499\ldots$.

What remained then was just to compare the estimates on (5.1) and (5.2) and check that (5.2) is smaller for $n \ge 10^{27}$. This final step was just bookkeeping. As we already discussed, a check for $n < 10^{27}$ is easy. Thus ends the proof of the main theorem.

6. Some remarks on computations

There were two main computational tasks: verifying the ternary conjecture for all $n \le C$, and checking the Generalized Riemann Hypothesis for modulus $q \le r$ up to a certain height.

The first task was not very demanding. Platt and I verified in [31] that every odd integer $5 < n \le 8.8 \cdot 10^{30}$ can be written as the sum of three primes. (In the end, only a check for $5 < n \le 10^{27}$ was needed.) We proceeded as follows. Oliveira e Silva, Herzog and Pardi [46]) had already checked that the binary Goldbach conjecture is true up to $4 \cdot 10^{18}$. Given that, all we had to do was to construct a "prime ladder", that is, a list of primes from 3 up to $8.8 \cdot 10^{30}$ such that the difference between any two consecutive primes in the list is at least 4 and at most $4 \cdot 10^{18}$. (This is a known strategy: see [55].) Then, for any odd integer $5 < n \le 8.8 \cdot 10^{30}$, there is a prime p in the list such that $4 \le n - p \le 4 \cdot 10^{18} + 2$. (Choose the largest $p < n$ in the ladder, or, if n minus that prime is 2, choose the prime immediately under that.) By [46] (and the fact that $4 \cdot 10^{18} + 2$ equals $p + q$, where $p = 2000000000000001301$ and $q = 1999999999999998701$ are both prime), we can write $n - p = p_1 + p_2$ for some primes p_1, p_2, and so $n = p + p_1 + p_2$.

Building a prime ladder involves only integer arithmetic, that is, computer manipulation of integers, rather than of real numbers. Integers are something that computers can handle rapidly and reliably. We look for primes for our ladder only among a special set of integers whose primality can be tested deterministically quite quickly (Proth numbers: $k \cdot 2^m + 1$, $k < 2^m$). Thus, we can build a prime ladder by a rigorous, deterministic algorithm that can be (and was) parallelized trivially.

The second computation is more demanding. It consists in verifying that, for every L-function $L(s,\chi)$ with χ of conductor $q \le r = 300000$ (for q even) or $q \le r/2$ (for q odd), all zeroes of $L(s,\chi)$ such that $|\Im(s)| \le H_q = 10^8/q$ (for q odd) and $|\Im(s)| \le H_q = \max(10^8/q, 200 + 7.5 \cdot 10^7/q$ (for q even) lie on the critical line. This was entirely Platt's work; my sole contribution was to request computer time. In fact, he went up to conductor $q \le 200000$ (or twice that for q even); he had already gone up to conductor 100000 in his PhD thesis. The verification took, in total, about 400000 core-hours (i.e., the total number of processor cores used times the number of hours they ran equals 400000; nowadays, a top-of-the-line processor typically has eight cores). In the end, since I used only $q \le 150000$ (or twice that for q even), the number of hours actually needed was closer to 160000; since I could have made do with $q \le 120000$ (at the cost of increasing C to 10^{29} or 10^{30}), it is likely, in retrospect, that only about 80000 core-hours were needed.

Checking zeros of L-functions computationally goes back to Riemann (who did it by hand for the special case of the Riemann zeta function). It is also one of the things that were tried on digital computers in their early days (by Turing [61], for instance; see the exposition in [1]). One of the main issues to be careful about arises whenever one manipulates real numbers via a computer: generally speaking, a computer cannot store an irrational number, and so one cannot say: "computer, give me the sine of that number" and expect a precise result. What one should do is to say: "computer, I am giving you an interval $I = [a/2^k, b/2^k]$; give me an interval $I' = [c/2^\ell, d/2^\ell]$, preferably very short, such that $\sin(I) \subset I'$". This is called interval arithmetic; it is arguably the easiest way to do floating-point computations rigorously.

Processors do not do this natively, and if interval arithmetic is implemented purely on software, computations can be slowed down by a factor of about 100. Fortunately, there are ways of running interval-arithmetic computations partly on hardware, partly on software. Platt has his own library, but there are others online (e.g. PROFIL/BIAS [38]).

Lastly, there were several relatively minor computations embedded in [27–29]. A typical computation was a rigorous version of a "proof by graph" ("the maximum of a function f is clearly less than 4 because I can see it on the screen"). There is a standard way to do this (see, e.g., [60, §5.2]); essentially, the bisection method combines naturally with interval arithmetic. Yet another computation (and not a very small one) was that involved in verifying a large-sieve inequality in an intermediate range (as we discussed in §5).

It may be interesting to note that one of the inequalities used to estimate (4.13) was proven with the help of automatic quantifier elimination [32]. Proving this inequality was a very minor task, both computationally and mathematically; in all likelihood, it is feasible to give a human-generated proof. Still, it is nice to know from first-hand experience that computers can nowadays (pretend to) do something other than just perform numerical computations – and that this is true even in current mathematical practice.

Acknowledgements. Thanks are due to J. Brandes and R. Vaughan for a discussion on a possible ambiguity in the Latin word in [14, p. 298]. Descartes' statement is mentioned (with a translation much like the one given here) in Dickson's *History* [17, Ch. XVIII]. Parts of the present article are based on a previous expository note by the author. The first version of the note appeared online, in English, in an informal venue [30]; later versions were published in Spanish ([25], translated by M. A. Morales and the author, and revised with the help of J. Cilleruelo and M. Helfgott) and French ([26], translated by M. Bilu and revised by the author). Many individuals and organizations should be thanked for their

generous help towards the work summarized here; an attempt at a full list can be found in the acknowledgments sections of [27–29]. Thanks are also due to J. Brandes, K. Gong, R. Heath-Brown, Z. Silagadze, R. Vaughan and T. Wooley, for help with historical questions.

References

[1] A. R. Booker, *Turing and the Riemann hypothesis*, Notices Amer. Math. Soc. **53** (2006), no. 10, 1208–1211.

[2] K. G. Borodzkin, *On the problem of I. M. Vinogradov's constant*, Proc. Third All-Union Math. Conf., vol. 1, Izdat. Akad. Nauk SSSR, Moscow, 1956, p. 3 (Russian).

[3] Y. Buttkewitz, *Exponential sums over primes and the prime twin problem*, Acta Math. Hungar. **131** (2011), no. 1-2, 46–58.

[4] J. R. Chen, *On the representation of a larger even integer as the sum of a prime and the product of at most two primes*, Sci. Sinica **16** (1973), 157–176.

[5] ———, *On the estimation of some trigonometrical sums and their application*, Sci. Sinica Ser. A **28** (1985), no. 5, 449–458.

[6] J. R. Chen and T. Z. Wang, *On the Goldbach problem*, Acta Math. Sinica **32** (1989), no. 5, 702–718.

[7] ———, *The Goldbach problem for odd numbers*, Acta Math. Sinica (Chin. Ser.) **39** (1996), no. 2, 169–174.

[8] N. G. Chudakov, *Introduction to the theory of Dirichlet L-functions*, OGIZ, Moscow-Leningrad, 1947 (Russian).

[9] N.G. Chudakov, *On the Goldbach problem*, C. R. (Dokl.) Acad. Sci. URSS, n. Ser. **17** (1937), 335–338 (French).

[10] ———, *On the density of the set of even numbers which are not representable as the sum of two odd primes*, Izv. Akad. Nauk SSSR Ser. Mat. 2 (1938), 25–40.

[11] H. Daboussi, *Effective estimates of exponential sums over primes*, Analytic number theory, Vol. 1 (Allerton Park, IL, 1995), Progr. Math., vol. 138, Birkhäuser Boston, Boston, MA, 1996, pp. 231–244.

[12] H. Daboussi and J. Rivat, *Explicit upper bounds for exponential sums over primes*, Math. Comp. **70** (2001), no. 233, 431–447 (electronic).

[13] H. Davenport, *Multiplicative number theory*, Markham Publishing Co., Chicago, Ill., 1967, Lectures given at the University of Michigan, Winter Term.

[14] R. Descartes, *Œuvres de Descartes publiées par Charles Adam et Paul Tannery sous les auspices du Ministère de l'Instruction publique. Physico-mathematica. Compendium musicae. Regulae ad directionem ingenii. Recherche de la vérité. Supplément à la correspondance. X.*, Paris: Léopold Cerf. IV u. 691 S. 4°, 1908.

[15] J.-M. Deshouillers, *Sur la constante de Šnirel′man*, Séminaire Delange-Pisot-Poitou, 17e année: (1975/76), Théorie des nombres: Fac. 2, Exp. No. G16, Secrétariat Math., Paris, 1977, p. 6.

[16] J.-M. Deshouillers, G. Effinger, H. te Riele, and D. Zinoviev, *A complete Vinogradov 3-primes theorem under the Riemann hypothesis*, Electron. Res. Announc. Amer. Math. Soc. **3** (1997), 99–104.

[17] L. E. Dickson, *History of the theory of numbers. Vol. I: Divisibility and primality.*, Chelsea Publishing Co., New York, 1966.

[18] G. Effinger, *Some numerical implications of the Hardy and Littlewood analysis of the 3-primes problem*, Ramanujan J. **3** (1999), no. 3, 239–280.

[19] T. Estermann, *On Goldbach's Problem: Proof that Almost all Even Positive Integers are Sums of Two Primes*, Proc. London Math. Soc. **S2-44** (1937), no. 4, 307–314.

[20] K. Ford, *Vinogradov's integral and bounds for the Riemann zeta function*, Proc. London Math. Soc. (3) **85** (2002), no. 3, 565–633.

[21] J. Friedlander and H. Iwaniec, *Asymptotic sieve for primes*, Ann. of Math. (2) **148** (1998), no. 3, 1041–1065.

[22] A. Granville and O. Ramaré, *Explicit bounds on exponential sums and the scarcity of squarefree binomial coefficients*, Mathematika **43** (1996), no. 1, 73–107.

[23] G. H. Hardy and J. E. Littlewood, *Some problems of 'Partitio numerorum'; III: On the expression of a number as a sum of primes*, Acta Math. **44** (1922), no. 1, 1–70.

[24] D. R. Heath-Brown, *The ternary Goldbach problem*, Rev. Mat. Iberoamericana **1** (1985), no. 1, 45–59.

[25] H. Helfgott, *La conjetura débil de Goldbach*, Gac. R. Soc. Mat. Esp. **16** (2013), no. 4, 709–726.

[26] H. A. Helfgott, *La conjecture de Goldbach ternaire*, Preprint. To appear in Gaz. Math.

[27] ———, *Major arcs for Goldbach's problem*, Preprint. Available at arXiv:1203.5712.

[28] ———, *Minor arcs for Goldbach's problem*, Preprint. Available at arXiv:1205.5252.

[29] ———, *The Ternary Goldbach Conjecture is true*, Preprint.

[30] ———, *The ternary Goldbach conjecture*, 2013, Available at http://valuevar.wordpress.com/2013/07/02/the-ternary-goldbach-conjecture/.

[31] H. A. Helfgott and D. Platt, *Numerical verification of the ternary Goldbach conjecture up to up to* $8.875e30$, To appear in Experiment. Math. Available at arXiv:1305.3062.

[32] H. Hong and Ch. W. Brown, *QEPCAD B – Quantifier elimination by partial cylindrical algebraic decomposition*, May 2011, version 1.62.

[33] M. N. Huxley, *Irregularity in sifted sequences*, J. Number Theory **4** (1972), 437–454.

[34] H. Iwaniec and E. Kowalski, *Analytic number theory*, American Mathematical Society Colloquium Publications, vol. 53, American Mathematical Society, Providence, RI, 2004.

[35] H. Kadiri, *An explicit zero-free region for the Dirichlet L-functions*, Preprint. Available as arXiv:0510570.

[36] ———, *Une région explicite sans zéros pour la fonction ζ de Riemann*, Acta Arith. **117** (2005), no. 4, 303–339.

[37] A. A. Karatsuba, *Basic analytic number theory*, Springer-Verlag, Berlin, 1993, Translated from the second (1983) Russian edition and with a preface by Melvyn B. Nathanson.

[38] O. Knüppel, *PROFIL/BIAS*, February 1999, version 2.

[39] N. M. Korobov, *Estimates of trigonometric sums and their applications*, Uspehi Mat. Nauk **13** (1958), no. 4 (82), 185–192.

[40] M.-Ch. Liu and T. Wang, *On the Vinogradov bound in the three primes Goldbach conjecture*, Acta Arith. **105** (2002), no. 2, 133–175.

[41] K. K. Mardzhanishvili, *On the proof of the Goldbach-Vinogradov theorem (in Russian)*, C. R. (Doklady) Acad. Sci. URSS (N.S.) **30** (1941), no. 8, 681–684.

[42] K. S. McCurley, *Explicit zero-free regions for Dirichlet L-functions*, J. Number Theory **19** (1984), no. 1, 7–32.

[43] H. L. Montgomery, *A note on the large sieve*, J. London Math. Soc. **43** (1968), 93–98.

[44] ———, *Topics in multiplicative number theory*, Lecture Notes in Mathematics, Vol. 227, Springer-Verlag, Berlin, 1971.

[45] H. L. Montgomery and R. C. Vaughan, *The large sieve*, Mathematika **20** (1973), 119–134.

[46] T. Oliveira e Silva, S. Herzog, and S. Pardi, *Empirical verification of the even Goldbach conjecture, and computation of prime gaps, up to* $4 \cdot 10^{18}$, Accepted for publication in Math. Comp., 2013.

[47] F. W. J. Olver, *Two inequalities for parabolic cylinder functions*, Proc. Cambridge Philos. Soc. **57** (1961), 811–822.

[48] D. Platt, *Numerical computations concerning GRH*, Preprint. Available at arXiv:1305.3087.

[49] O. Ramaré, *Explicit estimates on several summatory functions involving the Moebius function*, Preprint.

[50] ———, *A sharp bilinear form decomposition for primes and Moebius function*, Preprint. To appear in Acta. Math. Sinica.

[51] ———, *On Šnirel′man's constant*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. **22** (1995), no. 4, 645–706.

[52] ———, *Arithmetical aspects of the large sieve inequality*, Harish-Chandra Research Institute Lecture Notes, vol. 1, Hindustan Book Agency, New Delhi, 2009, With the collaboration of D. S. Ramana.

[53] ———, *On Bombieri's asymptotic sieve*, J. Number Theory **130** (2010), no. 5, 1155–1189.

[54] H. Riesel and R. C. Vaughan, *On sums of primes*, Ark. Mat. **21** (1983), no. 1, 46–74.

[55] Y. Saouter, *Checking the odd Goldbach conjecture up to* 10^{20}, Math. Comp. **67** (1998), no. 222, 863–866.

[56] L. Schnirelmann, *Über additive Eigenschaften von Zahlen*, Math. Ann. **107** (1933), no. 1, 649–690.

[57] X. Shao, *A density version of the Vinogradov three primes theorem*, Duke Math. J. **163** (2014), no. 3, 489–512.

[58] Terence Tao, *Every odd number greater than* 1 *is the sum of at most five primes*, Mathematics of Computation (286) **83** (2014), 997–1038.

[59] N. M. Temme and R. Vidunas, *Parabolic cylinder functions: examples of error bounds for asymptotic expansions*, Anal. Appl. (Singap.) **1** (2003), no. 3, 265–288.

[60] W. Tucker, *Validated numerics: A short introduction to rigorous computations*, Princeton University Press, Princeton, NJ, 2011.

[61] A. M. Turing, *Some calculations of the Riemann zeta-function*, Proc. London Math. Soc. (3) **3** (1953), 99–117.

[62] J. G. van der Corput, *Sur l'hypothèse de Goldbach pour presque tous les nombres pairs*, Acta Arith. **2** (1937), 266–290 (French).

[63] R. C. Vaughan, *On the estimation of Schnirelman's constant*, J. Reine Angew. Math. **290** (1977), 93–108.

[64] ———, *Sommes trigonométriques sur les nombres premiers*, C. R. Acad. Sci. Paris Sér. A-B **285** (1977), no. 16, A981–A983.

[65] ———, *Recent work in additive prime number theory*, Proceedings of the International Congress of Mathematicians (Helsinki, 1978), Acad. Sci. Fennica, Helsinki, 1980, pp. 389–394.

[66] ———, *The Hardy-Littlewood method*, second ed., Cambridge Tracts in Mathematics, vol. 125, Cambridge University Press, Cambridge, 1997.

[67] I. M. Vinogradov, *A new method in analytic number theory*, Tr. Mat. Inst. Steklova **10** (1937), 5–122 (Russian).

[68] ———, *The method of trigonometrical sums in the theory of numbers*, Tr. Mat. Inst. Steklova **23** (1947), 3–109 (Russian).

[69] ———, *The method of trigonometrical sums in the theory of numbers*, Interscience Publishers, London and New York, 1954. Translated, revised and annotated by K. F. Roth and Anne Davenport.

[70] ———, *A new estimate of the function* $\zeta(1+it)$, Izv. Akad. Nauk SSSR. Ser. Mat. **22** (1958), 161–164.

[71] A. Weil, *Number theory: An approach through history. From Hammurapi to Legendre*, Birkhäuser Boston, Inc., Boston, MA, 1984.

[72] D. Zinoviev, *On Vinogradov's constant in Goldbach's ternary problem*, J. Number Theory **65** (1997), no. 2, 334–358.

DMA - École Normale Supérieure, 45 rue d'Ulm, F-75230 Paris, France
E-mail: harald.helfgott@ens.fr

Small gaps between primes

D. A. Goldston, J. Pintz, and C. Y. Yıldırım

Abstract. This paper describes the authors' joint research on small gaps between primes in the last decade and how their methods were developed further independently by Zhang, Maynard, and Tao to prove stunning new results on primes. We now know that there are infinitely many primes differing by at most 246, and that one can find k primes a bounded distance (depending on k) apart infinitely often. These results confirm important approximations to the Hardy–Littlewood Prime Tuples Conjecture.

Mathematics Subject Classification (2010). Primary 11N05, 11N36; Secondary 11N35.

Keywords. Hardy–Littlewood prime tuples conjecture, prime numbers, sieves, gaps between primes, twin primes.

1. History

The twin prime conjecture that n and $n+2$ are both primes for infinitely many positive integers n, may have been conceived around the time of Euclid, more than two thousand years ago. Among as yet unsolved problems in mathematics it is one of the oldest. The purpose of the present article is to give an overview of the progress in the last nine years in this subject, in particular, of the results of the authors.

As a young boy Gauss observed in 1792 or 1793 that the primes around x have an average distance $\log x$ which led him to conjecture that

$$\pi(x) := \sum_{\substack{p\le x\\ p:\text{ prime}}} 1 \ \sim \mathrm{li}\ x := \int_2^x \frac{dt}{\log t} \sim \frac{x}{\log x} \qquad (x\to\infty). \tag{1.1}$$

This conjecture was proved in 1896 (independently) by Hadamard and de la Vallée Poussin, and is now called the Prime Number Theorem.

A relevant quantity in the study of small gaps between primes is

$$\Delta := \liminf_{n\to\infty} \frac{d_n}{\log n} = \liminf_{n\to\infty} \frac{p_{n+1}-p_n}{\log n}, \tag{1.2}$$

where $\{p_i\}_{i=1}^{\infty} =: \mathcal{P}$ is the set of primes sequenced in increasing order and $d_n := p_{n+1} - p_n$. The Prime Number Theorem, (1.1), immediately implies $\Delta \leqslant 1$, so the first task concerning an upper estimation of Δ was to show an estimate of the type $\Delta < 1$. During the twentieth century there were many papers on upper estimates for Δ. First, in 1926, Hardy and

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Littlewood (unpublished, see [32]) succeeded in showing, assuming the Generalized Riemann Hypothesis (GRH), that

$$\Delta \leqslant 2/3. \tag{1.3}$$

The first unconditional bound

$$\Delta \leqslant 1 - c_1, \tag{1.4}$$

with an unspecified but explicitly calculable $c_1 > 0$, was shown by Erdős in 1940 [5] who used Brun's sieve. The next big step was made by Bombieri and Davenport [2] who removed the assumption of GRH in Hardy and Littlewood's method by using Bombieri's work [1] on the large sieve and showed that

$$\Delta \leqslant \left(2 + \sqrt{3}\right)/8 = 0.466\ldots. \tag{1.5}$$

Their method gave $\Delta \leqslant 1/2$ but they were also able to combine this with an explicit version of Erdős's [5] proof which led them to (1.5). After several smaller improvements (Huxley and others), Maier [23] succeeded in combining the matrix method he developed with the ideas of Bombieri–Davenport, Erdős and Huxley, making it possible to multiply the best known bound by $e^{-\gamma}$ (γ is Euler's constant) and reach

$$\Delta \leqslant 0.248\ldots. \tag{1.6}$$

In 2005 the authors proved (see [14]; or for a brief account §2, §3 below)

$$\Delta = 0. \tag{1.7}$$

2. Ideas behind the proofs of some results concerning small gaps between consecutive primes

We begin by recounting a number of conjectures related to the twin prime conjecture and more generally to small gaps between consecutive primes. Some of them have been known for a long time, some of them were introduced by us.

Conjecture 2.1 (Twin Prime Conjecture). $d_n = 2$ *infinitely often.*

A generalization of this was formulated in 1849 by de Polignac.

Conjecture 2.2 (De Polignac's Conjecture [29]). *For every given positive even integer h, $d_n = h$ infinitely often.*

For a further generalization we need the notion of *admissible k-tuples*.

Definition 2.3. $\mathcal{H} = \{h_i\}_{i=1}^k$ $(0 \leqslant h_1 < h_2 \cdots < h_k, h_k \in \mathbb{Z})$ is *admissible* if the h_i's do not cover all residue classes mod p for any prime p.

This is clearly a necessary condition that $n + h_i \in \mathcal{P}$ for all integers $1 \leq i \leq k$ holds for infinitely many numbers n.

Dickson formulated in 1904 the conjecture that this condition was also sufficient. Although his conjecture included linear forms of type $a_i n + b_i$ $(a_i, b_i \in \mathbb{Z})$ we will consider the special case $a_i = 1$ for all $i \in [1, k]$.

Conjecture 2.4 (Dickson's Conjecture [3]). *If $\mathcal{H}$ is admissible, then $n + h_i \in \mathcal{P}$ for all $i \in [1, k]$ holds for infinitely many values of n.*

About twenty years later, in 1923, Hardy and Littlewood formulated this in a quantitative form as

Conjecture 2.5 (Hardy–Littlewood Prime-Tuples Conjecture [20]). *If $\mathcal{H}$ is an admissible k-tuple, then*

$$\sum_{\substack{n\leqslant x\\ \{n+h_i\}_{i=1}^{k}\in\mathcal{P}^k}} 1 \sim \mathfrak{S}(\mathcal{H})\frac{x}{\log^k x}, \tag{2.1}$$

where

$$\mathfrak{S}(\mathcal{H}) := \prod_p \left(1 - \frac{\nu_{\mathcal{H}}(p)}{p}\right)\left(1 - \frac{1}{p}\right)^{-k} > 0, \tag{2.2}$$

and $\nu_{\mathcal{H}}(p)$ denotes the number of distinct residue classes modulo p occupied by the elements of $\mathcal{H}$.

Note that the relation $\mathfrak{S}(\mathcal{H}) > 0$ is equivalent to $\mathcal{H}$ being admissible.

Until now the conjectures were listed in increasing strength. We introduced a weaker form of Dickson's Conjecture:

Conjecture 2.6 (Conjecture DHL$(k, 2)$). *If $\mathcal{H}$ is an admissible k-tuple, then $n + \mathcal{H}$ contains at least two primes infinitely often.*

If the above conjecture is true for at least one admissible k-tuple, then it implies another conjecture which is a good approximation to the Twin Prime Conjecture. This we called the

Conjecture 2.7 (Bounded Gaps Conjecture). *There exists an absolute constant C such that $d_n = p_{n+1} - p_n \leqslant C$ for infinitely many n.*

A still weaker form of the Bounded Gap Conjecture is

Conjecture 2.8 (Small Gaps Conjecture). $\Delta = \liminf\limits_{n\to\infty} \dfrac{p_{n+1} - p_n}{\log p_n} = 0.$

Within the scope of our work the existence of small or bounded gaps between consecutive primes is intimately connected with the distribution of primes in arithmetic progressions. The following definition of an admissible level ϑ of primes was already known and used in sieve theory.

Definition 2.9. ϑ is called an *admissible level of distribution of primes* if for any $\varepsilon > 0$, $A > 0$ we have for any $X > 2$

$$\sum_{q\leqslant X^{\vartheta-\varepsilon}} \max_{\substack{a\\(a,q)=1}} \left| \sum_{\substack{p\equiv a(q)\\ p\leqslant X}} \log p - \frac{X}{\varphi(q)} \right| \leqslant \frac{C(A,\varepsilon)X}{(\log X)^A}, \tag{2.3}$$

where $C(A, \varepsilon)$ is an ineffective constant depending on A and ε.

The largest known level $\vartheta = 1/2$ is the celebrated Bombieri–Vinogradov [1, 38] Theorem. The strongest possibility, $\vartheta = 1$, is the Elliott–Halberstam [4] Conjecture, and more generally one can introduce

Conjecture 2.10 (Conjecture EH(ϑ)). (2.3) *is true for a fixed* $\vartheta \in (\frac{1}{2}, 1]$.

We succeeded in showing in 2005 the following result.

Theorem 2.11 ([14]). *If* EH(ϑ) *is true for some fixed* $\vartheta > 1/2$, *then* DHL$(k, 2)$ *is true for* $k > k_0(\vartheta)$ *and consequently the Bounded Gaps Conjecture is true, i.e.* $\liminf_{n\to\infty} d_n < \infty$.

Theorem 2.12 ([14]). *The Small Gaps Conjecture is true, i.e.* $\Delta = 0$.

We improved this somewhat later to

Theorem 2.13 ([15]). $\liminf_{n\to\infty} \dfrac{d_n}{(\log n)^{1/2}(\log\log n)^2} < \infty$.

Concerning the frequency of small gaps we showed

Theorem 2.14 ([17, 18]). *Given any fixed* $\eta > 0$ *the relation*

$$d_n = p_{n+1} - p_n < \eta \log n \tag{2.4}$$

holds for a positive proportion of all gaps.

One of the important ideas which yielded a proof of the Small Gaps Conjecture in [14] and which – along with the work of Y. Motohashi and J. Pintz [25] – represented an important step in the first proof of the Bounded Gaps Conjecture by Y. Zhang [39] was to attack, among the listed seven conjectures, particularly DHL$(k, 2)$. The idea was to find suitable non-negative weights a_n for $n \in [N, 2N)$ to be abbreviated later as $n \sim N$, such that a_n should be relatively large compared with $S = \sum_{n\sim N} a_n > 0$ if the set

$$n + \mathcal{H}_k = \{n + h_i\}_{i=1}^k \tag{2.5}$$

contains some (possibly several) primes. A good quantitative formulation is to consider (and try to maximize) the ratio

$$E_j = \frac{S_j}{S^*} := \frac{\sum_{n\sim N} a_n \chi_{\mathcal{P}}(n + h_j)\log(n + h_j)}{\sum_{n\sim N} a_n \log 3N}, \tag{2.6}$$

where $\chi_{\mathcal{P}}(m)$ denotes the characteristic function of primes, that is, $\chi_{\mathcal{P}}(m) = 1$ if m is prime and 0 otherwise.

The quantity

$$\alpha(\mathcal{H}_k) = \sum_{j=1}^k E_j \tag{2.7}$$

describes the (weighted) average number of primes in $n + \mathcal{H}_k$ if n runs between N and $2N$, i.e. $n \sim N$. If we succeed in obtaining for a k-tuple $\mathcal{H} = \mathcal{H}_k$ a lower bound greater than 1 for the quantity in (2.7), then DHL$(k, 2)$ is proved (at least for a single $\mathcal{H} = \mathcal{H}_k$), and from this the Bounded Gaps Conjecture follows immediately.

(i) If we start with the simple uniform choice $a_n \equiv 1$ we obtain

$$\alpha(\mathcal{H}_k) \sim \frac{k}{\log N} \quad \text{as} \quad N \to \infty, \tag{2.8}$$

which clearly tends to 0.

(ii) Choosing $a_n = 1$ if $\{n + h_i\}_{i=1}^k \in \mathcal{P}^k$ and 0 otherwise, we can seemingly reach the optimal value

$$\alpha(\mathcal{H}_k) = k \text{ unless } S = \sum_{n\sim N} a_n = 0. \tag{2.9}$$

Unfortunately, to exclude the possibility $S = S(N) = 0$ for $N > N_0$ is equivalent to the proof of Dickson's Conjecture, so we arrive at a tautology.

In the following $\mathcal{H} = \mathcal{H}_k$ will always be an admissible k-tuple, but to simplify notation we often write simply $\mathcal{H}$ instead of $\mathcal{H}_k$.

(iii) An essentially equivalent formulation of the above is to use the generalized von Mangoldt function

$$a_n = \Lambda_k\big(P_{\mathcal{H}}(n)\big) := \sum_{d|P_{\mathcal{H}}(n)} \mu(d) \left(\log \frac{P_{\mathcal{H}}(n)}{d}\right)^k, \quad P_{\mathcal{H}}(n) = \prod_{i=1}^k (n + h_i) \tag{2.10}$$

which vanishes if $P_{\mathcal{H}}(n)$ has more than k distinct prime factors. However, in this case a direct evaluation of S seems to be hopeless, since d can be as large as N^k.

(iv) It was an idea of Selberg to approximate (2.10) with the divisors cut at $R = N^c$ and accordingly use

$$\sum_{\substack{d|P_{\mathcal{H}}(n)\\ d\leqslant R}} \mu(d) \log^k \frac{R}{d}. \tag{2.11}$$

However, this might be negative.

(v) So the next idea is the weight used in the so-called k-dimensional Selberg sieve, i.e., simply the square of (2.11), namely,

$$a_{n,k} = \left(\sum_{d\leqslant R, d|P_{\mathcal{H}}(n)} \mu(d) \log^k \frac{R}{d}\right)^2. \tag{2.12}$$

In this case choosing $R \leqslant N^{1/2}L^{-A}$, $L = \log N$, $A > A_0(k)$, S can be readily evaluated. Assuming EH(ϑ), the unconditional case being EH(1/2) (the Bombieri–Vinogradov Theorem), the more difficult sum S_j can also be evaluated, but only under the stronger constraint

$$R \leqslant N^{(\vartheta-\varepsilon)/2}. \tag{2.13}$$

This yields for the crucial quantity $\alpha(\mathcal{H}_k)$ in (2.7)

$$\alpha(\mathcal{H}_k) = \vartheta - \varepsilon + O\left(\frac{1}{k}\right) \tag{2.14}$$

primes on average, which is unfortunately still less than 1 even under the strongest hypothesis $\vartheta = 1$, the original Elliott–Halberstam Conjecture.

(vi) The winning choice is if we are more modest and instead of Dickson's Conjecture approximate the situation when $\prod_{i=1}^k (n + h_i)$ has at most $k + \ell$ different prime factors where $\ell \geqslant 0$ is a free parameter. (The choice $\ell = 1$ was used earlier by Heath-Brown

[22], however, not to localize primes in $n+\mathcal{H}$ but to find n values where all components $n+h_i$ are almost primes). This means that we use (2.12) with $k+\ell$ instead of k, i.e. our choice in [14] was

$$a_{n,k+\ell} = \left(\sum_{d \leqslant R, d | P_{\mathcal{H}}(n)} \mu(d) \log^{k+\ell} \frac{R}{d} \right)^2. \tag{2.15}$$

This yielded under the condition (2.13) a gain of a factor 2, rather surprisingly. More precisely we got

$$\alpha(\mathcal{H}_k) = 2(\vartheta - \varepsilon) + O\left(\frac{\ell}{k}\right) + O\left(\frac{1}{\ell}\right). \tag{2.16}$$

Under the optimal choice $\ell = \left[\sqrt{k}/2\right]$ this meant

$$\alpha(\mathcal{H}_k) = 2(\vartheta - \varepsilon) + O\left(\frac{1}{\sqrt{k}}\right). \tag{2.17}$$

Consequently if EH(ϑ) is true for some $\vartheta > 1/2$ we obtain $\alpha(\mathcal{H}_k) > 1$ primes on average if $k > C/(\vartheta - 1/2)^2$.

In the unconditional case $\vartheta = 1/2$, this yielded Theorem 2.12 but missed the goal DHL$(k, 2)$ by a hairbreadth.

The way to see how this argument could lead to a proof of the Small Gaps Conjecture begins by observing that on average only $\left(2\varepsilon + \frac{c_1}{\sqrt{k}}\right)$ primes were "missing" to obtain more than one prime on average. Using all numbers of the form

$$n + h, \;\; h \in [1, H], \;\; H = \eta \log N \tag{2.18}$$

with an arbitrarily small but fixed $\eta > 0$ instead of only

$$n + h_i, \;\; h_i \in \mathcal{H}_k \tag{2.19}$$

we could pick up more primes so as to fill the missing part.

If in case of $h \in [1, H] \setminus \mathcal{H}_k$ we expect heuristically $n+h$ to be prime with a probability $1/\log N$, we can hope to collect

$$\eta > 2\varepsilon + \frac{c_1}{\sqrt{k}} + O\left(\frac{k}{\log N}\right) \tag{2.20}$$

primes among $n + h$ on average if $n \sim N$, $h \in [1, H] \setminus \mathcal{H}_k$.

The condition (2.20) is clearly satisfied if

$$\varepsilon < \frac{\eta}{3}, \;\; k > C_2 \eta^{-2}, \;\; N > N_0(k, \varepsilon, \eta). \tag{2.21}$$

In the original work [14] we used a result of Gallagher [11] and an averaging procedure over all $\mathcal{H}_k \subset [1, H]$ to show that the above sketched heuristic works in practice. In the next

section we use a simpler way, which avoids Gallagher's Theorem and uses a single, suitably chosen k-tuple $\mathcal{H}_k$ for all k.

We will not sketch the rather complicated procedure to show Theorem 2.13. We just mention here that it needs the investigation of k-tuples with

$$k \asymp \frac{(\log N)^{1/2}}{(\log\log N)^2}, \quad \ell \asymp \sqrt{k}. \tag{2.22}$$

In the work [27] it was shown that using a suitable polynomial $P(x)$ instead of the simple $x^{k+\ell}$ in (2.15) ($x = \log(R/d)$) one can improve Theorem 2.13 further to

Theorem 2.15 ([27]). $\liminf\limits_{n\to\infty} \dfrac{d_n}{(\log N)^{3/7}(\log\log N)^{4/7}} < \infty.$

One can raise the more general question of finding the optimal polynomial, or more generally the optimal function $P(x)$. B. J. Conrey calculated the optimal weight function, actually a Bessel-type function. Later in the work [10] an exact analysis confirmed the optimality of the Bessel-type function and the fact that it yielded instead of (2.17) the sharper estimate

$$\alpha(\mathcal{H}_k) = 2(\vartheta - \varepsilon) + O(k^{-2/3}). \tag{2.23}$$

This was, however, the same strength as the polynomial in [27] and [10] apart from the implicit constant in the above O symbol. Therefore the result in Theorem 2.15 can be considered as the limit of the original GPY method.

Concerning Theorem 2.14 the crucial idea is the fact, discovered by the second named author ([26]), and independently by Friedlander and Iwaniec [9] that the weights a_n are strongly concentrated on numbers n where all components $n + h_i$ are almost prime, more precisely for numbers n with

$$P^-\left(\prod_{i=1}^{k}(n+h_i)\right) > N^\delta, \ \ n \sim N, \tag{2.24}$$

where δ is an arbitrarily small fixed positive constant and $P^-(m)$ denotes the smallest prime factor of n. In fact, it was proved in [26] that

$$\sum_{\substack{n\sim N\\ P^-(P_{\mathcal{H}}(n))\leqslant N^\delta}} a_n \leqslant C\delta \sum_{n\sim N} a_n \tag{2.25}$$

with a constant $C = C(k)$. (The factor $C(k)\delta$ was improved to $C'k^3\delta^2$ with an absolute constant C' in [17]).

3. Sketch of the proof of Theorems 2.11 and 2.12

In the following we consider a general sieve situation when the number of residues sieved out mod p satisfies

$$\Omega_{\mathcal{H}}(p) = \Omega(p) = k \ \text{ for } \ p \nmid \Delta(\mathcal{H}) := \prod_{i>j}(h_i - h_j), \ k \text{ fixed} \tag{3.1}$$

and let $\Omega(n)$ be extended multiplicatively for all squarefree values of n. Actually we have $\Omega(p) = \Omega_{\mathcal{H}}(p) = \nu_{\mathcal{H}}(p)$. There are three possibilities:

(i) to work analytically with two complex variables (cf. [14]);

(ii) to work elementarily (cf. [13] using pure sieve methods beyond (2.3));

(iii) to work partially elementarily and partially analytically with one complex variable.

Here we will pursue the third possibility, worked out in an unpublished note of K. Soundararajan [37].

We use a somewhat more general weight function: a polynomial $P(y)$ but note that the argument would work the same for a function $P(y)$ analytic on $[0,1]$, if $P(y)$ has at least a kth order zero at 0.

First we evaluate the sum of the weights a_n, where in the following we will define

$$a_n = \left(\sum_{d\leqslant R} \mu(d) P\left(\frac{\log(R/d)}{\log R}\right)\right)^2, \tag{3.2}$$

$$S = \sum_{n\sim N} a_n \sim N \sum_{d,e\leqslant R}{}' \mu(d)\mu(e)\frac{\Omega([d,e])}{[d,e]} P\left(\frac{\log(R/d)}{\log R}\right) P\left(\frac{\log(R/e)}{\log R}\right) \tag{3.3}$$

(we ignored a negligible error of size $O(R^{2+\varepsilon})$) and $\sum'$ will always denote summation over squarefree variables).

Introducing the notation $(d,e) = u$, $d = um$, $e = un$, $(m,n) = 1$ we obtain

$$S \sim N \sum_{u\leqslant R}{}' \sum_{\substack{m,n\leqslant R/u\\(m,n)=1\\(m,u)=(n,u)=1}}{}' \frac{\mu(m)\mu(n)\Omega(u)\Omega(m)\Omega(n)}{umn} P\left(\frac{\log(R/um)}{\log R}\right) P\left(\frac{\log R/un}{\log R}\right). \tag{3.4}$$

We can rewrite the condition $(m,n) = 1$ using the relation

$$\sum_{\beta|m,\beta|n} \mu(\beta) = \begin{cases} 1 & \text{if } (m,n) = 1,\\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

as

$$S \sim N \sum_{u\leqslant R}{}' \sum_{\beta\leqslant R/u}{}' \mu(\beta)\frac{\Omega(u)\Omega^2(\beta)}{u\beta^2} \left(\sum_{\substack{m'\leqslant R/u\beta\\(m',u)=1}}{}' \frac{\mu(\beta m')\Omega(m')}{m'} P\left(\frac{\log(R/u\beta m')}{\log R}\right)\right)^2. \tag{3.6}$$

Grouping terms with the same value of $u\beta =: \gamma$ with notation $m = m'$ we have

$$S \sim N \sum_{\gamma\leqslant R}{}' \frac{\Omega(\gamma)}{\gamma} \left(\sum_{\beta|\gamma}{}' \frac{\mu(\beta)\Omega(\beta)}{\beta}\right) \left(\sum_{\substack{m\leqslant R/\gamma\\(m,\gamma)=1}}{}' \frac{\mu(m)\Omega(m)}{m} P\left(\frac{\log(R/\gamma m)}{\log R}\right)\right)^2. \tag{3.7}$$

Let us denote the inner sum by $J\left(\gamma, \frac{R}{\gamma}\right)$ where the first variable refers to the condition $(m,\gamma)=1$, the second to $m \leqslant R/\gamma$. Further let for a squarefree γ

$$G(s+1,\gamma) := \sum_{\substack{m \\ (m,\gamma)=1}}{}' \frac{\mu(m)\Omega(m)}{m^{s+1}} =: \zeta(s+1)^{-k}F(s+1,\gamma). \tag{3.8}$$

Here we have for Re $s>0$

$$F(s+1,\gamma) = \prod_p \left(1-\frac{\Omega(p)}{p^{s+1}}\right)\left(1-\frac{1}{p^{s+1}}\right)^{-k}\prod_{p|\gamma}\left(1-\frac{\Omega(p)}{p^{s+1}}\right)^{-1}. \tag{3.9}$$

Using the Taylor expansion

$$P(x) = \sum_{j=k}^{\infty} \frac{P^{(j)}(0)x^j}{j!} \tag{3.10}$$

and Perron's formula ($c>0$, arbitrary)

$$\frac{1}{2\pi i}\int_{(c)} \frac{x^s}{s^{j+1}}ds = \begin{cases} \frac{(\log x)^j}{j!} & \text{if } x \geqslant 1, \\ 0 & \text{if } 0 \leqslant x \leqslant 1 \end{cases} \qquad (j \in \mathbb{Z}^+) \tag{3.11}$$

we can rewrite $J\left(\gamma, \frac{R}{\gamma}\right)$ as

$$\begin{aligned} J\left(\gamma,\frac{R}{\gamma}\right) &= \sum_{j=k}^{\infty}\frac{P^{(j)}(0)}{(\log R)^j}\sum_{\substack{m\leqslant R/\gamma \\ (m,\gamma)=1}}{}'\frac{\mu(m)\Omega(m)}{m}\frac{1}{j!}\left(\log\frac{R/\gamma}{m}\right)^j \\ &= \sum_{j=k}^{\infty}\frac{P^{(j)}(0)}{(\log R)^j}\cdot\frac{1}{2\pi i}\int_{(c)}\sum_{\substack{m=1 \\ (m,\gamma)=1}}^{\infty}\frac{\mu(m)\Omega(m)}{m^{s+1}}\left(\frac{R}{\gamma}\right)^s\frac{ds}{s^{j+1}} \\ &= \sum_{j=k}^{\infty}\frac{P^{(j)}(0)}{(\log R)^j}\cdot\frac{1}{2\pi i}\int_{(c)}F(s+1,\gamma)\zeta(s+1)^{-k}\left(\frac{R}{\gamma}\right)^s\frac{ds}{s^{j+1}}. \end{aligned} \tag{3.12}$$

Since $F(s+1,\gamma)$ is regular for $\sigma > -\frac{1}{2}$ we can transform the line inside the zero-free region of $\zeta(s+1)$, that is, to $\sigma > 1-c/(\log(|t|+2))$, $|t| \leqslant \exp\left(\sqrt{\log R}\right)$. The integral is negligible on the new contour and so we obtain by the residue at $s=0$

$$\begin{aligned} J\left(\gamma,\frac{R}{\gamma}\right) &\sim \sum_{j=k}^{\infty}\frac{P^{(j)}(0)}{(\log R)^j}F(1,\gamma)\frac{(\log R/\gamma)^{j-k}}{(j-k)!} \\ &= \frac{F(1,\gamma)}{(\log R)^k}\sum_{\nu=0}^{\infty}\frac{P^{(\nu+k)}(0)}{\nu!}\left(\frac{\log(R/\gamma)}{\log R}\right)^{\nu} \\ &= \frac{F(1,\gamma)}{(\log R)^k}P^{(k)}\left(\frac{\log R/\gamma}{\log R}\right). \end{aligned} \tag{3.13}$$

We remark that although this argument does not work if R/γ is not large enough, that part can be shown to be negligible directly from (3.7). So we obtain

$$S \sim \frac{N}{(\log R)^{2k}} \sideset{}{'}\sum_{\gamma \leqslant R} \frac{\Omega(\gamma)}{\gamma} \prod_{p|\gamma} \left(1 - \frac{\Omega(p)}{p}\right) \cdot F(1,\gamma)^2 \left(P^{(k)}\left(\frac{\log R/\gamma}{\log R}\right)\right)^2 \tag{3.14}$$
$$\sim \frac{N}{(\log R)^{2k}} \mathfrak{S}^2(\mathcal{H}) \sideset{}{'}\sum_{\gamma \leqslant R} \frac{\Omega(\gamma)}{\gamma} \prod_{p|\gamma} \left(1 - \frac{\Omega(p)}{p}\right)^{-1} \left(P^{(k)}\left(\frac{\log R/\gamma}{\log R}\right)\right)^2.$$

Since apart from finitely many primes, for which

$$p \mid \Delta(\mathcal{H}) := \prod_{i>j} (h_i - h_j) \tag{3.15}$$

we have $\Omega(p) = k$, the behaviour of $\Omega(n)$ is similar to that of the generalized divisor function

$$\tau_k(n) = \sum_{n_1 n_2 \dots n_k = n} 1. \tag{3.16}$$

This implies (for the details see Lemma 11 of [13])

$$\sideset{}{'}\sum_{\gamma \leqslant x} \frac{\Omega(\gamma)}{\gamma} \prod_{p|\gamma} \left(1 - \frac{\Omega(p)}{p}\right)^{-1} \sim \mathfrak{S}(\mathcal{H})^{-1} \frac{(\log x)^k}{k!}. \tag{3.17}$$

The sum in (3.14) can be evaluated from (3.17) by partial summation, and we obtain

$$S \sim \frac{\mathfrak{S}(\mathcal{H}) N}{(\log R)^k (k-1)!} \int_0^1 y^{k-1} \left(P^{(k)}(1-y)\right)^2 dy. \tag{3.18}$$

Let us consider now the quantity

$$S_j = \sideset{}{'}\sum_{n \sim N} a_n \chi_{\mathcal{P}}(n + h_j) \log n, \quad h_j \in \mathcal{H}. \tag{3.19}$$

In this case (if $R < N$) the two conditions

$$n + h_j \in \mathcal{P}, \quad d \mid \prod_{i=1}^{k} (n + h_i), \quad d \leqslant R \tag{3.20}$$

and

$$n + h_j \in \mathcal{P}, \quad d \mid \prod_{\substack{i=1 \\ i \neq j}}^{k} (n + h_i), \quad d \leqslant R \tag{3.21}$$

are equivalent. So the situation is similar to (3.3) if

$$R \leqslant N^{(\vartheta - \varepsilon)/2} \tag{3.22}$$

since it is easy to see that by the condition (2.3) (which is unconditionally true with $\vartheta = 1/2$ by the Bombieri–Vinogradov Theorem) we can substitute $\chi_{\mathcal{P}}(n+h_j)\log n$ by 1. Thus we have

$$S_j \sim \sum_{d,e\leqslant R}{}' \mu(d)\mu(e)\frac{\Omega_j([d,e])}{[d,e]}P\left(\frac{\log(R/d)}{\log R}\right)P\left(\frac{\log(R/e)}{\log R}\right) \tag{3.23}$$

with the only difference that we have now $\Omega_j(p) = \Omega(p) - 1 = k-1$ if $p \nmid \Delta$. The singular series $\mathfrak{S}_j(\mathcal{H})$ is accordingly

$$\begin{aligned}\mathfrak{S}_j(\mathcal{H}) &= \prod_p \left(1 - \frac{\nu_{\mathcal{H}}(p)-1}{p-1}\right)\left(1-\frac{1}{p}\right)^{-(k-1)} \\ &= \prod_p \left(1-\frac{\nu_p(\mathcal{H})}{p}\right)\left(1-\frac{1}{p}\right)^{-k} = \mathfrak{S}(\mathcal{H}).\end{aligned} \tag{3.24}$$

So we obtain for all $j \in [1,k]$ under the stronger condition (3.22) now analogously to (3.18)

$$S_j \sim \frac{\mathfrak{S}(\mathcal{H})N}{(\log R)^{k-1}(k-2)!}\int_0^1 y^{k-2}\left(P^{(k-1)}(1-y)\right)^2 dy \tag{3.25}$$

and this gives in total for $R = N^{(\vartheta-\varepsilon)/2}$, $P^{(k-1)}(x) = Q(x)$

$$\frac{\sum_{j=1}^k S_j}{S\log 3N} \sim \frac{\log R}{\log N}k(k-1)M(Q) \sim \frac{k(k-1)(\vartheta-\varepsilon)}{2}M(Q) \tag{3.26}$$

primes on average in $\{n+h_i\}_{i=1}^k$ if n runs between N and $2N$ and the numbers n are weighted by $a_n \log n$, where

$$M(Q) = \frac{\int_0^1 y^{k-2}\big(Q(1-y)\big)^2 dy}{\int_0^1 y^{k-1}\big(Q'(1-y)\big)^2 dy}. \tag{3.27}$$

In case of the simple choice

$$P(x) = x^{k+\ell},\ \ \ell = \left[\sqrt{k}/2\right] \Leftrightarrow Q(x) = C(k,\ell)x^{\ell+1} \tag{3.28}$$

we obtain

$$\begin{aligned}M(Q) &= \frac{\int_0^1 y^{k-2}(1-y)^{2\ell+2}dy}{(\ell+1)^2\int_0^1 y^{k-1}(1-y)^{2\ell}dy} = \frac{(k-2)!(2\ell+2)!/(k+2\ell+1)!}{(\ell+1)^2(k-1)!(2\ell)!/(k+2\ell)!} \\ &= \frac{4\left(1-\frac{1}{2(\ell+1)}\right)}{(k+2\ell+1)(k-1)} \sim \frac{4\left(1-O\left(\frac{1}{\sqrt{k}}\right)\right)}{k^2}.\end{aligned} \tag{3.29}$$

By (3.26) this yields on the weighted average

$$2(\vartheta-\varepsilon)\left(1-O\left(\frac{1}{\sqrt{k}}\right)\right) \tag{3.30}$$

primes in $\{n+\mathcal{H}\}$ if $n\sim N$.

The quantity above is clearly greater than 1 if

$$\vartheta>1/2,\quad k>k_0(\vartheta), \tag{3.31}$$

which proves Theorem 2.11.

Suppose now $h_0\notin\mathcal{H}$, let $\mathcal{H}_0=\mathcal{H}\cup\{h_0\}$, and $\Omega_0(p)=\Omega_{\mathcal{H}_0}(p)$ is defined as in (3.1) with $k+1$ in place of k,

$$S_0=\sum_{n\sim N}a_n\chi_{\mathcal{P}}(n+h_0)\log n. \tag{3.32}$$

In case of $\nu_{\mathcal{H}_0}(p)=\nu_{\mathcal{H}}(p)$ we have $\Omega_0(p)=\nu_{\mathcal{H}}(p)-1$ residue classes in the sieve mod p (Ω_0 is defined as in (3.1)); if $\nu_{\mathcal{H}_0}(p)=\nu_{\mathcal{H}}(p)+1$, then $\Omega_0(p)=\nu_{\mathcal{H}}(p)$. So we have in both cases $\Omega_0(p)=\nu_{\mathcal{H}_0}(p)-1$ and $\Omega_0(p)=k$ if $p\nmid\Delta(\mathcal{H}_0)$.

This yields an analogous asymptotic to (3.18) for S_0, with $\mathcal{H}$ replaced by $\mathcal{H}_0$:

$$S_0\sim\frac{\mathfrak{S}(\mathcal{H}_0)N}{(\log R)^k(k-1)!}\int\limits_0^1 y^{k-1}\left(P^{(k)}(1-y)\right)^2dy \tag{3.33}$$

and consequently

$$\frac{S_0}{S}\sim\frac{\mathfrak{S}\left(\mathcal{H}\cup\{h_0\}\right)}{\mathfrak{S}(\mathcal{H})}\quad(\text{as } N\to\infty). \tag{3.34}$$

This relation helps us to obtain Theorem 2.12 unconditionally. Let us consider an interval of length

$$H=\eta\log N, \tag{3.35}$$

where η is an arbitrarily small fixed positive constant. Let us suppose that we can find for any k an admissible k-tuple $\mathcal{H}=\mathcal{H}_k$ such that with a fixed absolute constant $c_0>0$

$$\mathfrak{S}\left(\mathcal{H}_k\cup h_0\right)>c_0\mathfrak{S}(\mathcal{H})\ \text{ for any even } h_0. \tag{3.36}$$

In this case using only $\vartheta=1/2$, that is, the Bombieri–Vinogradov Theorem, we obtain on average

$$\frac{\sum\limits_{h=1}^{H}\sum\limits_{n\sim N}a_n\chi_{\mathcal{P}}(n+h)\log n}{\sum\limits_{n\sim N}a_n\log 3N}\geqslant(1-2\varepsilon)\left(1-O\left(\frac{1}{\sqrt{k}}\right)\right)+\frac{c_0\eta}{2}-\frac{k}{\log 3N}>1 \tag{3.37}$$

primes between n and $n+H$ if

$$k>k_0(\eta),\ \ \varepsilon<\varepsilon_0(\eta),\ \ N>N_0(\eta,k,\varepsilon). \tag{3.38}$$

In order to show the existence of $\mathcal{H}_k$ with (3.36) we can just choose

$$\mathcal{H}=\mathcal{H}_k=\left\{i\prod_{p\leqslant 2k}p\right\}_{i=1}^{k}. \tag{3.39}$$

Then we have for any even h with $\nu_p = \nu_{\mathcal{H}}(p)$

$$\begin{aligned}\frac{\mathfrak{S}(\mathcal{H}\cup h)}{\mathfrak{S}(\mathcal{H})} &\geqslant 2\prod_{2<p\leqslant 2k}\frac{1-2/p}{(1-1/p)^2}\prod_{p>2k}\frac{1-(\nu_p+1)/p}{1-(\nu_p+1)/p+\nu_p/p^2} \\ &\geqslant c_1\prod_{p>2k}\left(1+O\left(\frac{k}{p^2}\right)\right)\geqslant c_0.\end{aligned} \tag{3.40}$$

In such a way we obtain Theorem 2.12. We remark that the above proof avoids Gallagher's Theorem [11]. Another proof, also avoiding Gallagher's Theorem is given in [16] which yields some other results, like small gaps between consecutive primes in arithmetic progressions and improved upper estimates for the quantity

$$\Delta_r = \liminf_{n\to\infty}\frac{p_{n+r}-p_n}{\log p_n}. \tag{3.41}$$

4. Sketch of the proof of Theorem 2.14

The most crucial idea in the proof of Theorem 2.14 is that we will change the weights and instead of the original normalized weights (cf. (2.15)).

$$a_n = \left(\sum_{d\leq R, d\mid P_{\mathcal{H}}(n)}\mu(d)\left(\frac{\log(R/d)}{\log R}\right)^{k+\ell}\right)^2,\ P_{\mathcal{H}}(n)=\prod_{i=1}^{k}(n+h_i),\ \ell=\left[\frac{\sqrt{k}}{2}\right] \tag{4.1}$$

we will work with the new weight ($n\sim N$)

$$a'_n = \begin{cases} a_n & \text{if } P^-\big(P_{\mathcal{H}}(n)\big) > N^{\delta}, \\ 0 & \text{otherwise,}\end{cases} \tag{4.2}$$

where δ will be a fixed small positive constant with $\varepsilon<\varepsilon_0(\eta)$, $k>k_0(\eta,\varepsilon)$, $\delta<\delta_0(k,\eta,\varepsilon)$, $R=N^{(\vartheta-\varepsilon)/2}$ and we consider primes in intervals of length

$$H=\eta\log N \tag{4.3}$$

as indicated in (2.2).

As mentioned at the end of Section 2 the sum of weights $a^*_{n,\mathcal{H}}$ with $\mathcal{P}_{\mathcal{H}}(n)$ having at least one small prime divisor not exceeding N^{δ} is negligible and we have (2.25) with a constant $C=C(k)$, i.e.

$$\begin{aligned}0\leq\sum_{n\sim N}(a_n-a'_n) &= \sum_{\substack{n\sim N\\ P^-(P_{\mathcal{H}}(n))\leq N^{\delta}}} a_n \leq C\delta\sum_{n\sim N}a_n, \\ \sum_{\substack{n\sim N\\ P^-(P_{\mathcal{H}}(n))\leq N^{\delta}}} a_n\chi_{\mathcal{P}}(n+h)\log(n+h) &\leq C\delta\sum_{n\sim N}a_n\chi_{\mathcal{P}}(n+h)\log(n+h).\end{aligned} \tag{4.4}$$

These are Lemmas 4 and 5 of [26].

The other tool is Gallagher's Theorem [11], according to which for k fixed, $H \to \infty$

$$\sum_{\substack{\mathcal{H}\subset[1,H]\\ |\mathcal{H}|=k}} \mathfrak{S}(\mathcal{H}) \sim \frac{H^k}{k!}. \tag{4.5}$$

Let further (for a more detailed proof see [17] and [18])

$$\pi(n,H) := \pi(n+H) - \pi(N), \quad \Theta(n) := \begin{cases} \log n & \text{if } n \in \mathcal{P}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.6}$$

$$\Theta(n,H) := \sum_{h=1}^{H} \Theta(n+h)$$

$$M := \sum_{\substack{p_j\sim N\\ p_{j+1}-p_j\le H}} 1, \quad Q(N,H) := \sum_{\substack{n\sim N\\ \pi(n,H)>1}} 1 \le HM + O\left(Ne^{-c\sqrt{\log N}}\right), \tag{4.7}$$

and consider now instead of (3.19) the modified quantity

$$S'(h,\mathcal{H}) = \sum_{n\sim N} a'_n \Theta(n+h). \tag{4.8}$$

The substitution of a_n by a'_n will just slightly change the corresponding value of $S'(\mathcal{H})$ and $S'(h,\mathcal{H})$ respectively, to

$$S'(\mathcal{H}) = \sum_{n\sim N} a'_n = (1+O(\delta))S(\mathcal{H}), \tag{4.9}$$

$$S'(h,\mathcal{H}) = \sum_{n\sim N} a'_n \Theta(n+h) = (1+O(\delta))S(h,\mathcal{H}) \tag{4.10}$$

compared with

$$S(h,\mathcal{H}) := \sum_{n\sim N} a_n \Theta(n+h), \tag{4.11}$$

where the asymptotics for the quantity (4.11) are given in (3.25) and (3.33) respectively, and $P(x) = x^{k+\ell}$ in this section.

The crucial change is that in case of $a'_{n,\mathcal{H}} > 0$ all the prime divisors of $\mathcal{P}_{\mathcal{H}}(n)$ are at least N^δ with a fixed small δ, so by (4.1) we have a trivial estimate for it:

$$a'_n \leqslant 2^{\omega(P_{\mathcal{H}}(n))} \leqslant 2^{2k^2/\delta} \ll_{k,\delta} 1. \tag{4.12}$$

On the other hand, in this case we cannot use the simplification of Section 3, that is, to work with a suitably chosen single $\mathcal{H}_k$. Averaging over all $\mathcal{H} \subseteq [1,H]$, $|\mathcal{H}| = k$, with the abbreviations (we take the unconditional case $\vartheta = 1/2$ from now on)

$$\frac{H}{\log R} = \frac{\eta}{\left(\frac12 - \varepsilon\right)/2} = \eta', \quad \sum_{\mathcal{H}}^{(k)} = \sum_{\substack{\mathcal{H}\subset[1,H]\\ |\mathcal{H}|=k}} \tag{4.13}$$

we obtain from (3.18), using (3.28)–(3.29) and (4.5)

$$\sum_{\mathcal{H}}^{(k)} S'(\mathcal{H}) \sim (1+O(\delta))\frac{(\eta')^k NC(k,\ell)(2\ell)!}{k!(\ell+1)^2(k+2\ell)!} =: (1+O(\delta))B. \tag{4.14}$$

On the other hand, we have by (3.33) and (4.5)

$$\begin{aligned}&\sum_{\mathcal{H}}^{(k)} \sum_{\substack{n\sim N\\ h\in[1,H]\setminus\mathcal{H}}} a'_n\Theta(n+h) \qquad (4.15)\\ &\sim (k+1)\sum_{\mathcal{H}}^{(k+1)} \mathfrak{S}(\mathcal{H})\frac{NC(k,\ell)(2\ell)!(1+O(\delta))}{(\log R)^k(\ell+1)^2(k+2\ell)!} := (1+O(\delta))B\eta\log N.\end{aligned}$$

Finally, we have by (3.25) and (4.5) with $\ell = \left[\sqrt{k}/2\right]$, $\vartheta = 1/2$

$$\begin{aligned}&\sum_{\mathcal{H}}^{(k)} \sum_{h\in\mathcal{H}} \sum_{n\sim N} a'_n\Theta(n+h) \qquad (4.16)\\ &\sim (1+O(\delta))\frac{k\eta'^k NC(k,\ell)(2\ell+2)!}{k!(k+2\ell+1)!}\log R\\ &\sim (1+O(\delta))B\left(1-\frac{1}{2(\ell+1)}\right)\left(1-\frac{2\ell+1}{k+2\ell+1}\right)(1-2\varepsilon)\log N.\end{aligned}$$

Adding (4.15), (4.16) and subtracting from it (4.14) multiplied by $\log 3N$ we obtain

$$\begin{aligned}&\sum_{\mathcal{H}}^{(k)} \sum_{n\sim N} a'_n\big(\Theta(n,H)-\log 3N\big) \qquad (4.17)\\ &> B\log N\left\{(1-2\varepsilon)\left(1-\frac{C}{\sqrt{k}}\right)+\eta-1+O(\delta)\right\} > \frac{\eta}{2}B\log N\end{aligned}$$

if, as stated in the introduction of Section 4 (between (4.2) and (4.3)) we fix ε, k, δ with

$$\varepsilon < \varepsilon_0(\eta),\ \ k > k_0(\eta,\varepsilon),\ \ \delta < \delta_0(k,\eta,\varepsilon). \tag{4.18}$$

Consequently, if (4.18) holds, which we will always assume in the following, then

$$\frac{\eta}{2}B\log N < (1+o(1))\log N \sum_{\substack{n\sim N\\ \pi(n,H)>1}} \pi(n,\mathcal{H})\sum_{\mathcal{H}}^{(k)} a'_n. \tag{4.19}$$

Introducing the notation

$$T(n,H) := \sum_{\substack{\mathcal{H}\\ P^-\left(P_{\mathcal{H}}(n)\right)>N^\delta}}^{(k)} 1 \tag{4.20}$$

we have by (4.6)–(4.7), (4.12) and Cauchy's inequality

$$\eta B \ll \Big(\sum_{\substack{n\sim N\\ \pi(n,\mathcal{H})>1}} 1\Big)^{1/2}\Big(\sum_{n\sim N}\pi^2(n,H)T(n,H)^2\Big)^{1/2} \tag{4.21}$$

$$\ll \left((HM)^{1/2} + O\left(N^{1/2}e^{-c\sqrt{\log N}/2}\right)\right)\left(\sum_{n\sim N}\pi^2(n,H)T(n,H)^2\right)^{1/2}.$$

Further, we have by Selberg's sieve (Theorem 5.1 of [21] or Theorem 2 in § 2.2.2 of [19]) for any set $\mathcal{H}$ and $\delta < 1/2$

$$\sum_{\substack{n\sim N\\ P^-\left(P_{\mathcal{H}}(n)\right)>R^\delta}} 1 \leqslant \frac{|\mathcal{H}|!\mathfrak{S}(\mathcal{H})}{(\log R^\delta)^{|\mathcal{H}|}}N(1+o(1)) \quad (R,N\to\infty). \tag{4.22}$$

This implies by Gallagher's Theorem (4.5)

$$\sum_{n\sim N}\pi(n,H)^2T(n,H)^2 \ll \sum_{1\le h,h'\le H}\sum_{\mathcal{H}_1}^{(k)}\sum_{\mathcal{H}_2}^{(k)}\sum_{\substack{n\sim N\\ P^-\left(\mathcal{H}_1\cup\mathcal{H}_2\cup\{h\}\cup\{h'\}\right)>N^\delta}} 1 \tag{4.23}$$

$$\ll_k N\sum_{r=k}^{2k+2}\sum_{\mathcal{H}_0}^{(r)}\frac{\mathfrak{S}(\mathcal{H}_0)}{(\log R^\delta)^r} \ll_{k,\delta} N\sum_{r=k}^{2k+2}\left(\frac{H}{\log R}\right)^r \ll_{k,\delta} (\eta')^k N.$$

Taking into account the definition of B in (4.14) we obtain from (4.21) and (4.23)

$$\eta(\eta')^{k/2} \ll_{k,\delta} \left(\left(\frac{HM}{N}\right)^{1/2} + e^{-c\sqrt{\log N}/2}\right). \tag{4.24}$$

Consequently,

$$\frac{HM}{N} \gg_{k,\delta,\eta} 1. \tag{4.25}$$

Hence,

$$M \gg_{k,\delta,\eta} \frac{N}{\log N} \gg_{k,\delta,\eta} \pi(2N), \tag{4.26}$$

which proves Theorem 2.14.

It may be shown (see Theorem 2 of [18]) that this is sharp in the sense that the assertion does not remain true if $H = o(\log N)$. The proof uses the Selberg sieve upper bound for prime tuples and Gallagher's result (4.5).

5. Bounded gaps between primes. Zhang's theorem

We recall that in our original work (Theorem 2.11 in Section 2) we showed that EH(ϑ) for any $\vartheta > 1/2$ implies DHL$(k,2)$ for $k > k_0(\vartheta)$, consequently the Bounded Gaps Conjecture. From the proof it is trivial that the condition

$$\max_{a,(a,q)=1} \tag{5.1}$$

in (2.3) can be weakened to

$$\max_{a,(a,q)=1,P_{\mathcal{H}}(a)\equiv 0(q)} \tag{5.2}$$

if we want to show for a specific $\mathcal{H}$ that $n + \mathcal{H}$ contains at least two primes infinitely often. However, in 2008 in a joint work of Y. Motohashi and J. Pintz the following stronger form of Theorem 2.11 was proved, in which the summation in (2.3) can be reduced to smooth moduli. $P^+(n)$ will denote the largest prime factor of n.

Theorem 5.1 ([25]). *If there exist $\delta > 0$, $\vartheta > 1/2$ and an admissible k-tuple $\mathcal{H}$ with $k > k_0(\delta, \vartheta)$ such that for any $\varepsilon > 0$, $A > 0$*

$$\sum_{\substack{q \leq N^{\vartheta-\varepsilon} \\ P^+(q) \leq N^\delta}} \max_{a \atop (a,q)=1, q \mid P_{\mathcal{H}}(a)} \left| \sum_{\substack{p \equiv a(q) \\ p \sim N}} \log p - \frac{N}{\varphi(q)} \right| \leq \frac{C(A,\varepsilon)N}{\log^A N} \tag{5.3}$$

holds for $N > N_0(\mathcal{H}, \vartheta, \delta)$, then $n + \mathcal{H}$ contains at least two primes for some $n \sim N$.

Remark 5.2. Zhang proved a version of this result, and it appeared with a different proof in his work [39]. Zhang proved condition (5.3) with the explicit values

$$\vartheta = \frac{1}{2} + \frac{1}{584}, \quad \delta = \frac{1}{1168}, \tag{5.4}$$

which finally led to

Theorem 5.3 ([39]). DHL$(k, 2)$ *is true for $k \geqslant 3.5 \cdot 10^6$ and consequently*

$$\liminf_{n \to \infty} (p_{n+1} - p_n) \leqslant C = 7 \cdot 10^7.$$

His proof of (5.4) uses several deep works of Fouvry, Fouvry–Iwaniec, Bombieri–Friedlander–Iwaniec, Friedlander–Iwaniec, Heath-Brown, which are based on ideas and works of Linnik, Weil, Deligne and Birch–Bombieri concerning the estimate of Kloostermann sums.

The Polymath 8a project of T. Tao [30] introduced many improvements into this procedure (for example to apply instead of the simple weight function $P(x) = x^{k+\ell}$ the optimal Bessel function first used by Conrey, later analyzed in details in [10] together with many improvements in both the Motohashi–Pintz Theorem and in the estimation of Kloostermann sums) and obtained distribution estimates up to level $1/2 + 7/300$, and thus reached

Theorem 5.4 (Polymath 8a). DHL$(k, 2)$ *is true for $k \geqslant 632$ and consequently*

$$\liminf_{n \to \infty} (p_{n+1} - p_n) \leqslant 4680.$$

6. Bounded gaps between primes: The Maynard–Tao theorem

About half a year after the manuscript of Zhang [39], simultaneously and independently, J. Maynard [24] and in his Polymath blogs T. Tao [31] introduced another idea which led to a new, more efficient proof of the Bounded Gaps Conjecture. The main results of Maynard [24] were the following.

Theorem 6.1 (Maynard [24]). DHL$(k, 2)$ *is true for $k \geq 105$, consequently*

$$\liminf_{n \to \infty} (p_{n+1} - p_n) \leqslant 600.$$

Theorem 6.2 (Maynard [24]). *Assuming the Elliott–Halberstam Conjecture,* DHL$(k,2)$ *is true for* $k \geq 5$*, consequently*

$$\liminf_{n\to\infty}(p_{n+1}-p_n) \leqslant 12.$$

The two surprising aspects of the Maynard–Tao method were that it produced not only pairs but arbitrarily long (finite) blocks of primes in bounded intervals, and for this knowing that (2.3) holds with any fixed $\vartheta > 0$ (however small) would suffice.

The earlier known strongest result of somewhat similar nature was the much weaker one in our work [16]. It asserted for any $r > 0$

$$\Delta_r := \liminf_{n\to\infty} \frac{p_{n+r}-p_n}{\log p_n} \leq e^{-\gamma}\left(\sqrt{r}-1\right)^2. \tag{6.1}$$

Further, under the very deep Elliott–Halberstam Conjecture (see (2.3) with $\vartheta = 1$) we could show [14]

$$\Delta_2 = 0. \tag{6.2}$$

Theorem 6.3 (Maynard–Tao [24]). *We have for any* r

$$\liminf_{n\to\infty}(p_{n+r}-p_n) \ll r^3 e^{4r}. \tag{6.3}$$

The main idea of Maynard and Tao is that the weights are defined instead of

$$a_n = \left(\sum_{\substack{d\leqslant R\\ d|P_{\mathcal{H}}(n)}} \mu(d) P\left(\frac{\log R/d}{\log R}\right)\right)^2, \quad \mathcal{P}_{\mathcal{H}}(n) = \prod_{i=1}^{k}(n+h_i) \tag{6.4}$$

in the more general form

$$a_n = \left(\sum_{\substack{d_1\dots d_k\leqslant R\\ d_i|n+h_i}} \mu(d) P\left(\frac{\log d_1}{\log R},\dots,\frac{\log d_k}{\log R}\right)\right)^2, \tag{6.5}$$

where $P(t_1,\dots,t_k) := \mathbb{R}^k \to \mathbb{R}$ is a fixed piecewise differentiable function with support on $t_1+t_2+\cdots+t_k \leqslant 1$. The idea of the use of these more general weights goes back to Selberg ([36], p. 245). Similar type of weights were used by Goldston and Yıldırım [12], but due to the special choice of $P(t_1,\dots,t_k) = \prod_{i=1}^{k}(1-kt_i),\ t_i \leqslant 1/k$, this led only to the result

$$\Delta = \liminf_{n\to\infty} \frac{p_{n+1}-p_n}{\log p_n} \leqslant \frac{1}{4}. \tag{6.6}$$

We remark here that the general choice of $P\left(\frac{\log R/d}{\log R}\right)$ in Section 3 corresponds to the special case of the above with

$$P(t_1,t_2,\dots,,t_k) = \widetilde{P}(t_1+t_2+\cdots+t_k). \tag{6.7}$$

Another very interesting remark is that in order to show bounded intervals with arbitrarily long finite blocks of primes (with a bound $e^{2r/\vartheta}$ in place of e^{4r}) we do not need the value

$\vartheta = 1/2$, that is, the Bombieri–Vinogradov Theorem, just any value $\vartheta > 0$. So we obtain a numerically slightly weaker form of the existence of arbitrarily long (finite) blocks of primes in bounded invervals even by the use of the first theorem establishing a positive admissible level ϑ for the distribution of primes, due to A. Rényi [33, 34] reached in 1947–48, by the large sieve of Linnik.

Upon further work on the Maynard–Tao method in the Polymath 8b project of Tao, Theorem 6.1 has been improved to

Theorem 6.4 (Polymath 8b project). DHL$(k, 2)$ *is true for* $k \geqslant 50$*, consequently*

$$\liminf_{n\to\infty}(p_{n+1} - p_n) \leq 246.$$

7. De Polignac numbers and some conjectures of Erdős on gaps between consecutive primes

There are various 60–70 years old conjectures of Erdős on which a sharpened version of Zhang's Theorem (or that of Maynard and Tao) combined with other arguments of the second named author can give an answer. Below we give a list of them without proofs which can be found in [28]. The numerical values reflect the stage at the end of Polymath 8A.

Using an argument of the second named author (Lemma 4 in [26]) together with a more general form of the arguments of Theorem 3 of Zhang and its improvement by Tao's project, the following strengthening of Theorem 3 of Zhang can be shown. (Let $P^-(n)$ be the smallest prime factor of n.)

Theorem 7.1 ([28]). *Let* $k \geq 632$*,* $\mathcal{H}$ *an admissible* k*-tuple,* $h_i \ll \log N$*,* $N > N_0(k)$*. Then there are at least*

$$c_1(k, \mathcal{H})\frac{N}{\log^k N}$$

numbers $n \in [N, 2N)$ *such that* $n + \mathcal{H}$ *contains at least two primes and almost primes in all other components satisfying* $P^-(n + h_i) > N^{c_2(k)}$ *for* $i = 1, 2, \ldots, k$*.*

Remark 7.2. A similar version to the above-mentioned crucial Lemma 4 of [26] appears in the book Opera de Cribro of Friedlander–Iwaniec [9] published also in 2010.

Whereas the original Theorem 3 of Zhang yields only one de Polignac number, by the aid of Theorem 7.1 we can show

Theorem 7.3 ([28]). *There are infinitely many de Polignac numbers. In fact, they have a positive lower density* $> 10^{-7}$*.*

Theorem 7.4 ([28]). *There exists an ineffective* C *such that we have always at least one de Polignac number between* X *and* $X + C$ *for any* X*. (All gaps between consecutive de Polignac numbers are uniformly bounded.)*

Erdős [6] proved in 1948 the inequality

$$\liminf_{n\to\infty}\frac{d_{n+1}}{d_n} \leq 1 - c_0 < 1 + c_0 \leq \limsup\frac{d_{n+1}}{d_n} \tag{7.1}$$

with a very small positive value c_0 and conjectured that the $\liminf = 0$ and the $\limsup = \infty$.

Theorem 7.5 ([28]). $\liminf_{n\to\infty} \frac{d_{n+1}}{d_n} = 0$, $\limsup \frac{d_{n+1}}{d_n} = \infty$.

Further, we have even

$$\liminf_{n\to\infty} \frac{d_{n+1}\log n}{d_n} < \infty, \qquad \limsup_{n\to\infty} \frac{d_{n+1}}{d_n \log n} > 0. \tag{7.2}$$

In general it is difficult to show anything for three consecutive differences. However, we can show

Theorem 7.6 ([28]). $\limsup_{n\to\infty} \frac{\min(d_{n-1}, d_{n+1})}{d_n (\log n)^c} = \infty$ *with* $c = 1/632$.

Since the Prime Number Theorem implies

$$\frac{1}{N}\sum_{n=1}^{N} \frac{d_n}{\log n} = 1, \tag{7.3}$$

it is interesting to investigate the normalized distribution of the sequence d_n, $d_n/\log n$. Erdős conjectured 60 years ago that the set of limit points,

$$J = \left\{\frac{d_n}{\log n}\right\}' = [0, \infty], \tag{7.4}$$

but no finite limit point was known until 2005, when we showed $0 \in J$. (We denote by G' the set of limit points of the set G.) This was rather strange since in 1955 Erdős [7] and simultaneously Ricci [35] proved that J has positive Lebesgue measure. A partial answer to the conjecture of Erdős is

Theorem 7.7 ([28]). *There is an (ineffective) constant* c^* *such that*

$$[0, c^*] \subset J. \tag{7.5}$$

The above result raises the question whether considering a finer distribution $d_n/f(n)$ with a monotonically increasing function $f(n) \le \log n$, $f(n) \to \infty$ the same phenomenon is still true. The answer is yes.

Theorem 7.8 ([28]). *Let* $f(n) \le \log n$, $f(n) \to \infty$ *be an increasing function,*

$$J_f = \left\{\frac{d_n}{f(n)}\right\}'. \tag{7.6}$$

Then there is an (ineffective) constant c_f^* *such that*

$$[0, c_f^*] \subset J_f. \tag{7.7}$$

Zhang's theorem shows the existence of infinitely many generalized twin prime pairs with a difference at most $7 \cdot 10^7$, while the theorem of Green and Tao shows the existence of arbitrarily long (finite) arithmetic progressions in the sequence of primes. A common generalization of these two results is given below. (Let p' denote the prime following p.)

Theorem 7.9 ([28]). *There exists an even* $d \le 4680$ *with the following property. For any* k *there is a* k*-term arithmetic progression of primes such that* $p' = p + d$ *for all elements of the progression.*

Acknowledgements. The first author was supported in part by NSF Grant DMS-1104434. The second author was supported by OTKA Grants NK104183, K100291 and ERC-AdG. 321104.

References

[1] Bombieri, E., *On the large sieve*, Mathematika **12** (1965), 201–225.

[2] E. Bombieri and H. Davenport, *Small differences between prime numbers*, Proc. Roy. Soc. Ser. A **293** (1966), 1–18.

[3] L. E. Dickson, *An extension of Dirichlet's theorem on prime numbers*, Messenger of Mathematics **33** (1904), 155–161.

[4] P. D. T. A. Elliott and H. Halberstam, *A conjecture in prime number theory*, Symposia Mathematica **4** (1968), 59–72.

[5] P. Erdős, *The difference of consecutive primes*, Duke Math. J. **6** (1940), 438–441.

[6] _____, *On the difference of consecutive primes*, Bull. Amer. Math. Soc. **54** (1948), 885–889.

[7] _____, *Some problems on the distribution of prime numbers*, Teoria dei Numeri, Math Congr. Varenna, (1954), 8 pp., 1955.

[8] P. Erdős and P. Turán, *On some new questions on the distribution of prime numbers*, Bull. Amer. Math. Soc. **54** (1948), 371–378.

[9] J. Friedlander and H. Iwaniec, *Opera de cribro*, American Mathematical Society Colloquium Publications, 57, American Mathematical Society, Providence, RI, 2010.

[10] B. Farkas, J. Pintz, and S. Révész, *On the optimal weight function in the Goldston–Pintz–Yıldırım method for finding small gaps between consecutive primes*, in: Paul Turán Memorial Volume: Number Theory, Analysis and Combinatorics, pp. 75–104, de Gruyter, Berlin, 2014.

[11] P. X. Gallagher, *On the distribution of primes in short intervals*, Mathematika **23** (1976), 4–9.

[12] D. A. Goldston and C. Y. Yıldırım, *Higher correlations of divisor sums related to primes, III. Small gaps between primes*, Proc. London Math. Soc. (3) **96** (2007), 653–686.

[13] D. A. Goldston, S. W. Graham, J. Pintz, and C. Y. Yıldırım, *Small gaps between primes or almost primes*, Trans. Amer. Math. Soc. **361** (2009), no. 10, 5285–5330.

[14] D. A. Goldston, J. Pintz, and C. Yıldırım, *Primes in Tuples I*, Annals of Math. **170** (2009), 819–862.

[15] _____, *Primes in Tuples II*, Acta Math. **204** (2010), 1–47.

[16] _____, *Primes in Tuples III*, Functiones et Approximat. **35** (2006), 76–89.

[17] ———, *Primes in Tuples IV: Density of small gaps between consecutive primes*, Acta Arith. **155** (2012), No. 4, 395–417.

[18] ———, *Positive Proportion of Small Gaps Between Consecutive Primes*, Publ. Math. Debrecen **79** (2011), no. 3-4, 433–444.

[19] G. Greaves, *Sieves in Number Theory*, Springer, Berlin, Heidelberg, New York, 2001.

[20] G. H. Hardy and J. E. Littlewood, Some problems of "Partitio Numerorum" III: *On the expression of a number as a sum of primes*, Acta Math. **44** (1923), 1–70.

[21] H. Halberstam and H.-E. Richert, *Sieve Methods*, Academic Press, London, 1974.

[22] D. R. Heath-Brown, *Almost-prime k-tuples*, Mathematika **44** (1997), 245–266.

[23] H. Maier, *Small differences between prime number*, Michigan Math. J. **35** (1988), 323–344.

[24] J. Maynard, *Small gaps between primes*, preprint.

[25] Y. Motohashi and J. Pintz, *A smoothed GPY sieve*, Bull. Lond. Math. Soc. **40** (2008), no. 2, 298–310.

[26] J. Pintz, *Are there arbitrarily long arithmetic progressions in the sequence of twin primes?*, in: An Irregular Mind. Szemerédi is 70, Bolyai Soc. Math Studies, Vol. 21, Eds.: I. Bárány, J. Solymosi, pp. 525–559, Springer, 2010.

[27] ———, *Some new results on gaps between consecutive primes*, in: Paul Turán Memorial Volume: Number Theory, Analysis and Combinatorics, pp. 261–278, de Gruyter, Berlin, 2014.

[28] ———, *Polignac Numbers, Conjectures of Erdős on Gaps between Primes*, Arithmetic Progression in Primes, and the Bounded Gap Conjecture, Preprint, arXiv:1305.6289, 2013.

[29] A. de Polignac, *Six propositions arithmologiques déduites du crible d'Ératosthène*, Nouv. Ann. Math. **8** (1849), 423–429.

[30] D. H. J. Polymath, *New equidistribution estimates of Zhang type, and bounded gaps between primes*, Preprint, arXiv:1402.0811v1, 2014.

[31] ———, (in preparation); see the Polymath8b webpages.

[32] R. A. Rankin, *The difference between consecutive prime numbers. II*, Proc. Cambridge Philos. Soc. **36** (1940), 255–266.

[33] A. Rényi, *On the representation of an even number as the sum of a single prime and a single almost-prime number*, Dokl. Akad. Nauk SSSR **56** (1947), 455–458 (Russian).

[34] ———, *On the representation of an even number as the sum of a single prime and a single almost-prime number*, Izv. Akad. Nauk SSSR **12** (1948), 57–78 (Russian).

[35] G. Ricci, *Sull'andamento della differenza di numeri primi consecutivi*, Riv. Mat. Univ. Parma **5** (1954), 3–54.

[36] A. Selberg, *Collected Papers, Vol. II*, With a foreword of K. Chandresekharan. Springer, Berlin, 1991.

[37] K. Soundararajan, *Notes on Goldston–Pintz–Yıldırım* (unpublished).

[38] A. I. Vinogradov, *The density hypothesis for Dirichlet L-series*, Izv. Akad. Nauk SSSR Ser. Mat. **29** (1965), 903–934 (Russian).

[39] Y. Zhang, *Bounded gaps between primes*, Ann. of Math. (2) **179** (2014), No. 3, 1121–1174.

Department of Mathematics and Statistics, San Jose State University, San Jose, CA 95192, U.S.A.
E-mail: daniel.goldston@sjsu.edu

Rényi Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Reáltanoda u. 13–15, H-1053 Hungary
E-mail: pintz.janos@renyi.mta.hu

Boğaziçi University, Department of Mathematics, Bebek, Istanbul 34342 Turkey
E-mail: yalciny@boun.edu.tr

Some problems in analytic number theory for polynomials over a finite field

Zeev Rudnick

Abstract. The lecture explores several problems of analytic number theory in the context of function fields over a finite field, where they can be approached by methods different than those of traditional analytic number theory. The resulting theorems can be used to check existing conjectures over the integers, and to generate new ones. Among the problems discussed are: Counting primes in short intervals and in arithmetic progressions; Chowla's conjecture on the autocorrelation of the Möbius function; and the additive divisor problem.

Mathematics Subject Classification (2010). Primary 11T55; Secondary 11N05, 11N13.

Keywords. Function fields over a finite field, Chowla's conjecture, the additive divisor problem, primes in short intervals.

1. Introduction

The goal of this lecture is to explore traditional problems of analytic number theory in the context of function fields over a finite field. Several such problems which are currently viewed as intractable over the integers, have recently been addressed in the function field context with vastly different tools than those of traditional analytic number theory, and the resulting theorems can be used to check existing conjectures over the integers, and to generate new ones. The problems that I will address concern

- Counting primes in short intervals and in arithmetic progressions
- Chowla's conjecture on the autocorrelation of the Möbius function
- The twin prime conjecture
- The additive divisor problem
- The variance of sums of arithmetic functions in short intervals and arithmetic progressions.

Before describing the problems, I will briefly survey some quantitative aspects of the arithmetic of the ring of polynomials over a finite field.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. Background on arithmetic in $\mathbb{F}_q[x]$

2.1. The prime polynomial theorem. Let $\mathbb{F}_q$ be a finite field of q elements, and $\mathbb{F}_q[x]$ the ring of polynomials with coefficients in $\mathbb{F}_q$. The polynomial ring $\mathbb{F}_q[x]$ shares several qualitative properties with the ring of integers $\mathbb{Z}$, for instance having a Euclidean algorithm, hence unique factorization into irreducibles. There are also several common quantitative aspects. To set these up, I review some basics.

The units of the ring of integers are ± 1, and every nonzero integer is a multiple by a unit of a positive integer. Analogously, the units of $\mathbb{F}_q[x]$ are the nonzero scalars $\mathbb{F}_q^\times$, and every nonzero polynomial is is a multiple by a unit of a monic polynomial. The analogue of a (positive) prime is a monic irreducible polynomial. To investigate arithmetic properties of "typical" integers, one samples them uniformly in the dyadic interval $[X, 2X]$ with $X \to \infty$; likewise to investigate arithmetic properties of "typical" polynomials, one samples them uniformly from the monic polynomials $\mathcal{M}_n$ of degree n, with $\#\mathcal{M}_n = q^n \to \infty$.

The Prime Number Theorem (PNT) states that the number $\pi(x)$ of primes $p \leq x$ is asymptotically equal to

$$\pi(x) \sim \mathrm{Li}(x) := \int_2^x \frac{dt}{\log t} \sim \frac{x}{\log x}\ , \quad x \to \infty\ . \tag{2.1}$$

The Riemann Hypothesis is equivalent to the assertion that

$$\pi(x) = \mathrm{Li}(x) + O\Big(x^{1/2+o(1)}\Big)\ . \tag{2.2}$$

The Prime Polynomial Theorem asserts that the number $\pi_q(n)$ of monic irreducible polynomials of degree n is

$$\pi_q(n) = \frac{q^n}{n} + O\Big(\frac{q^{n/2}}{n}\Big)\ , \tag{2.3}$$

the implied constant absolute. This corresponds to the PNT (and to the Riemann Hypothesis) if we map $x \leftrightarrow q^n$, recalling that x is the number of positive integers up to x and q^n is the number of monic polynomials of degree n. Note that (2.3) gives an asymptotic result whenever $q^n \to \infty$; in comparison, the results described below will usually be valid only in the large finite field limit, that is n fixed and $q \to \infty$.

2.2. Cycle structure. For $f \in \mathbb{F}_q[x]$ of positive degree n, we say its cycle structure is $\lambda(f) = (\lambda_1, \dots, \lambda_n)$ if in the prime decomposition $f = \prod_\alpha P_\alpha$ (we allow repetition), we have $\#\{\alpha : \deg P_\alpha = j\} = \lambda_j$. In particular $\deg f = \sum_j j\lambda_j$. Thus we get a partition of $\deg f$, which we denote by $\lambda(f)$. For instance, $\lambda_1(f)$ is the number of roots of f in $\mathbb{F}_q$, and f is totally split in $\mathbb{F}_q[x]$ - that is $f(x) = \prod_{j=1}^n (x - a_j)$, $a_j \in \mathbb{F}_q$- if and only if $\lambda(f) = (n, 0, \dots, 0)$. Moreover f is prime if and only if $\lambda(f) = (0, 0, \dots, 0, 1)$.

The cycle structure of a permutation σ of n letters is $\lambda(\sigma) = (\lambda_1, \dots, \lambda_n)$ if in the decomposition of σ as a product of disjoint cycles, there are λ_j cycles of length j. For instance, $\lambda_1(\sigma)$ is the number of fixed points of σ, and $\sigma = I$ is the identity if and only if $\lambda(\sigma) = (n, 0, \dots)$. Moreover $\sigma \in S_n$ is an n-cycle if and only if $\lambda(\sigma) = (0, 0, \dots, 0, 1)$.

For each partition $\lambda \vdash n$, denote by $p(\lambda)$ the probability that a random permutation on n letters has cycle structure λ:

$$p(\lambda) = \frac{\#\{\sigma \in S_n : \lambda(\sigma) = \lambda\}}{\#S_n}\ . \tag{2.4}$$

Cauchy's formula for $p(\lambda)$ is

$$p(\lambda) = \prod_{j=1}^{n} \frac{1}{j^{\lambda_j} \cdot \lambda_j!} \tag{2.5}$$

In particular, the proportion of n-cycles in the symmetric group S_n is $1/n$.

The connection between cycle structures of polynomials and of permutations is by means of the following observation, a straight-forward consequence of the Prime Polynomial Theorem (2.3): Given a partition $\lambda \vdash n$, the probability that a random monic polynomial f of degree n has cycle structure λ is asymptotic, as $q \to \infty$, to the probability $p(\lambda)$ that a random permutation of n letters has that cycle structure:

$$\frac{1}{q^n} \#\{f \text{ monic}, \deg f = n : \lambda(f) = \lambda\} = p(\lambda) + O\Big(\frac{1}{q}\Big) . \tag{2.6}$$

Note that unlike the Prime Polynomial Theorem (2.3), this result (2.6) gives an asymptotic only in the large finite field limit $q \to \infty$, n fixed.

Having set up the preliminaries, I turn to discussing new results on quantitative aspects of arithmetic in $\mathbb{F}_q[x]$.

3. Asymptotics in short intervals and arithmetic progressions

3.1. Primes in short intervals. Some of the most important problems in prime number theory concern the distribution of primes in short intervals and in arithmetic progressions. According to the Prime Number Theorem, the density of primes near x is $1/\log x$. Thus one wants to know what is the number $\pi(x, H)$ of primes in an interval of length $H = H(x) \ll x$ around x:

$$\pi(x, H) := \#\{x < p \le x + H : p \text{ prime}\} . \tag{3.1}$$

We expect that for H sufficiently large,

$$\pi(x, H) \sim \frac{H}{\log x} . \tag{3.2}$$

The PNT implies that (3.2) holds for $H \approx x$, and the Riemann Hypothesis gives (3.2) for all $H > x^{1/2+o(1)}$. In 1930, Hoheisel gave an unconditional proof that (3.2) holds for all $H > x^{1-\delta}$ for any positive $\delta < 1/33,000$; this has since been improved, currently to $H > x^{7/12-o(1)}$ (Heath Brown 1988). It is believed that the result should hold for all $H > x^{\epsilon}$, for any $\epsilon > 0$, though Maier [30] showed that it does not hold for $H = (\log x)^N$ for any N; see Granville and Soundararajan [15] for a general framework for such results on irregularities of distribution and for sharper results. Selberg (1943) showed, assuming the Riemann Hypothesis, that (3.2) holds for *almost all* x provided $H/(\log x)^2 \to \infty$.

To set up an analogous problem for the polynomial ring $\mathbb{F}_q[x]$, we first need to define short intervals. For a nonzero polynomial $f \in \mathbb{F}_q[x]$, we define its norm by

$$|f| = \#\mathbb{F}_q[x]/(f) = q^{\deg f} ,$$

in analogy with the norm of a nonzero integer $0 \ne n \in \mathbb{Z}$, which is $|n| = \#\mathbb{Z}/n\mathbb{Z}$. Given a monic polynomial $A \in \mathcal{M}_n$ of degree n, and $h < n$, the "short interval" around A of diameter q^h is the set

$$I(A; h) := \{f \in \mathcal{M}_n : |f - A| \le q^h\} . \tag{3.3}$$

The number of polynomials in this "interval" is

$$H := \#I(A;h) = q^{h+1} \ . \tag{3.4}$$

We wish to count the number of prime polynomials in the interval $I(A;h)$. In the limit $q \to \infty$, Bank, Bary-Soroker and Rosenzweig [4] give an essentially optimal short interval result:

Theorem 3.1. *Fix $3 \le h < n$. Then for every monic polynomial A of degree n, the number of prime polynomials P in the interval $I(A;h) = \{f : |f - A| \le q^h\}$ about A satisfies*

$$\#\{P \text{ prime}, P \in I(A;h)\} = \frac{H}{n}\left(1 + O_n(q^{-1/2})\right) \ ,$$

the implied constant depending only on n.

For irregularities of distribution analogous to Maier's theorem in the large degree limit $n \to \infty$ (q fixed), see [36].

For other applications, we will need a version which takes into account the cycle structure:

Theorem 3.2 ([4]). *Fix $n > 1$, $3 \le h < n$ and a partition $\lambda \vdash n$. Then for any sequence of finite fields $\mathbb{F}_q$, and every monic polynomial A of degree n,*

$$\#\{f \in I(A;h) : \lambda(f) = \lambda\} = p(\lambda)H\left(1 + O_n(q^{-1/2})\right) \ ,$$

with $p(\lambda)$ as in (2.4)*,* (2.5)*, the implied constant depending only on n.*

3.2. Primes in arithmetic progressions. Dirichlet's theorem states that any arithmetic progression $n = A \bmod Q$ contains infinitely many primes provided that A and Q are coprime, and the prime number theorem in arithmetic progressions states that for fixed modulus Q, the number of such primes $p \le x$ is

$$\pi(x;Q,A) \sim \frac{\mathrm{Li}(x)}{\phi(Q)}, \quad x \to \infty \ , \tag{3.5}$$

where $\phi(Q)$ is Euler's totient function, the number of residues coprime to Q. The Generalized Riemann Hypothesis (GRH) asserts that (3.5) continues to hold for moduli as large as $Q < X^{1/2-o(1)}$. An unconditional version, for almost all $Q < x^{1/2-o(1)}$, and all $A \bmod Q$, is given by the Bombieri-Vinogradov theorem. Going beyond the GRH, the Elliott-Halberstam conjecture gives a similar statement for Q as large as $x^{1-\epsilon}$.

For $\mathbb{F}_q[x]$, it is a consequence of the Riemann Hypothesis for curves over a finite field (Weil's theorem) that given a modulus $Q \in \mathbb{F}_q[x]$ of positive degree, and a polynomial A coprime to Q, the number $\pi_q(n;Q,A)$ of primes $P = A \bmod Q$, $P \in \mathcal{M}_n$ satisfies

$$\pi_q(n;Q,A) = \frac{\pi_q(n)}{\Phi(Q)} + O(\deg Q \cdot q^{n/2}) \ ,$$

where $\Phi(Q)$ is the number of coprime residues modulo Q. For $q \to \infty$, the main term is dominant as long as $\deg Q < n/2$.

Going beyond the Riemann Hypothesis for curves, Bank, Bary-Soroker and Rosenzweig [4] show an individual asymptotic continues to hold for even larger moduli in the limit $q \to \infty$:

Theorem 3.3 ([4]). *If* $1 \le \deg Q \le n-3$ *then*

$$\pi_q(n; Q, A) = \frac{\pi_q(n)}{\Phi(Q)} \left(1 + O_n(q^{-\frac{1}{2}})\right) .$$

This should be considered as an individual version of the Elliot-Halberstam conjecture. As in the short interval case, they have a stronger result which takes into account the cycle structure.

4. Autocorrelations and twisted convolution

In this section we describe results on the autocorrelation of various classical arithmetic functions in the function field context.

4.1. Autocorrelations of the Möbius function and Chowla's conjecture. Equivalent formulations of the PNT and the Riemann Hypothesis can be given in terms of growth of partial sums of the Möbius function, defined by $\mu(n) = (-1)^k$ if n is a product of k distinct primes, and $\mu(n) = 0$ otherwise: The PNT is equivalent to nontrivial cancellation $\sum_{n \le x} \mu(n) = o(x)$, and the RH is equivalent to square-root cancellation: $\sum_{n \le x} \mu(n) = O(x^{1/2+o(1)})$.

A conjecture of Chowla on the auto-correlation of the Möbius function, asserts that given an r-tuple of distinct integers $\alpha_1, \dots, \alpha_r$ and $\epsilon_i \in \{1, 2\}$, not all even, then

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n \le N} \mu(n + \alpha_1)^{\epsilon_1} \cdot \dots \cdot \mu(n + \alpha_r)^{\epsilon_r} = 0 . \tag{4.1}$$

Note that the number of nonzero summands here, that is the number of $n \le N$ for which $n + \alpha_1, \dots, n + \alpha_r$ are all square-free, is asymptotically $\mathfrak{S}(\alpha)N$, where $\mathfrak{S}(\alpha) > 0$ if the numbers $\alpha_1, \dots, \alpha_r$ do not contain a complete system of residues modulo p^2 for every prime p, so that Chowla's conjecture (4.1) addresses non-trivial cancellation in the sum. At this time, the only known case of Chowla's conjecture (4.1) is $r = 1$ where it is equivalent with the Prime Number Theorem.

Sarnak [35] showed that Chowla's conjecture implies that $\mu(n)$ does not correlate with any "deterministic" (i. e., zero entropy) sequence. For recent studies on the correlation between $\mu(n)$ and several sequences of arithmetic functions, see [5, 8, 16, 29].

In joint work with Dan Carmon [6], we have resolved a version of Chowla's conjecture for $\mathbb{F}_q[x]$ in the limit $q \to \infty$. To formulate it, one defines the Möbius function of a nonzero polynomial $F \in \mathbb{F}_q[x]$ to be $\mu(F) = (-1)^r$ if $F = cP_1 \dots P_r$ with $0 \ne c \in \mathbb{F}_q$ and $P_1, \dots, P_r$ are distinct monic irreducible polynomials, and $\mu(F) = 0$ otherwise.

Theorem 4.1. *Fix* $r > 1$ *and assume that* $n > 1$ *and* q *is odd. Then for any choice of distinct polynomials* $\alpha_1, \dots, \alpha_r \in \mathbb{F}_q[x]$, *with* $\max \deg \alpha_j < n$, *and* $\epsilon_i \in \{1, 2\}$, *not all even,*

$$\left| \sum_{F \in \mathcal{M}_n} \mu(F + \alpha_1)^{\epsilon_1} \dots \mu(F + \alpha_r)^{\epsilon_r} \right| \ll_{r,n} q^{n - \frac{1}{2}} . \tag{4.2}$$

Thus for fixed $r, n > 1$,

$$\lim_{q\to\infty} \frac{1}{\#\mathcal{M}_n} \sum_{F\in\mathcal{M}_n} \mu(F+\alpha_1)^{\epsilon_1} \dots \mu(F+\alpha_r)^{\epsilon_r} = 0 \tag{4.3}$$

under the assumption of Theorem 4.1, giving an analogue of Chowla's conjecture (4.1).

Note that the number of square-free monic polynomials of degree n is, for $n > 1$, equal to $q^n - q^{n-1}$. Hence, given r distinct polynomials $\alpha_1, \dots, \alpha_r \in \mathbb{F}_q[x]$, with $\deg \alpha_j < n$, the number of $F \in \mathcal{M}_n$ for which all of $F(x) + \alpha_j(x)$ are square-free is $q^n + O(rq^{n-1})$ as $q \to \infty$. Thus indeed we display cancellation.

The starting point in our argument is Pellet's formula, which asserts that for the polynomial ring $\mathbb{F}_q[x]$ with q odd, the Möbius function $\mu(F)$ can be computed in terms of the discriminant $\mathrm{disc}(F)$ of $F(x)$ as

$$\mu(F) = (-1)^{\deg F} \chi_2(\mathrm{disc}(F)) \ , \tag{4.4}$$

where χ_2 is the quadratic character of $\mathbb{F}_q$. That allows us to express the LHS of (4.2) as an n-variable character sum and to estimate it by freezing all but one of the variables, and then using the Riemann Hypothesis for curves (Weil's theorem) to bound the one-variable sum. A key point is to bound the number of times when there is no cancellation in the one-variable sum.

4.2. Twin primes. It is an ancient conjecture that there are infinitely many twin primes, and a refined quantitative form, due to Hardy and Littlewood, asserts that given distinct integers $a_1, \dots, a_r$, the number $\pi(x; a_1, \dots, a_r)$ of integers $n \le x$ for which $n + a_1, \dots, n + a_r$ are simultaneously prime is asymptotically

$$\pi(x; a_1, \dots, a_r) \sim \mathfrak{S}(a_1, \dots, a_r) \frac{x}{(\log x)^r}, \quad x \to \infty \ , \tag{4.5}$$

for a certain constant $\mathfrak{S}(a_1, \dots, a_r)$, which is positive whenever there are no local congruence obstructions. Despite the striking recent breakthroughs by Zhang [37] and Maynard [31], this conjecture is still open even for $r = 2$ (twin primes).

Recently the function field version of the problem was solved. Bary-Soroker [3] proved that for given n, r then for any sequence of finite fields $\mathbb{F}_q$ of odd cardinality q, and distinct polynomials $a_1, \dots, a_r \in \mathbb{F}_q[x]$ of degree less than n, the number $\pi_q(n; a_1, \dots, a_r)$ of monic polynomials $f \in \mathbb{F}_q[x]$ of degree n such that $f + a_1, \dots, f + a_r$ are simultaneously irreducible satisfies

$$\pi_q(n; a_1, \dots, a_r) \sim \frac{q^n}{n^r}, \quad q \to \infty \ . \tag{4.6}$$

This improves on earlier results by Pollack [33] and by Bary-Soroker [2].

4.3. The additive divisor problem. The divisor function $d_r(n)$ is the number of ways of writing a positive integer n as a product of r positive integers. In particular for $r = 2$ we recover the classical divisor function $d_2(n) = \sum_{d|n} 1$. The mean value of d_r is

$$\frac{1}{x} \sum_{n\le x} d_r(n) \sim \frac{(\log x)^{r-1}}{(r-1)!}, \quad x \to \infty \ . \tag{4.7}$$

Likewise, the divisor function $d_r(f)$ for a monic polynomial $f \in \mathbb{F}_q[x]$ is defined as the number of r-tuples of monic polynomials $(a_1, \dots, a_r)$ so that $f = a_1 \cdot \dots \cdot a_r$. The mean value of d_r, when averaged over all monic polynomials of degree n, is

$$\frac{1}{q^n} \sum_{f \in \mathcal{M}_n} d_r(f) = \binom{n+r-1}{r-1} = \frac{n^{r-1}}{(r-1)!} + \dots , \tag{4.8}$$

which is a polynomial of degree $r-1$ in n.

The "additive divisor problem" (other names are "shifted divisor" and "shifted convolution") is to understand the autocorrelation of the divisor function, that is the sum (where $h \neq 0$ is fixed for this discussion)

$$D_r(X; h) := \sum_{n \leq X} d_r(n) d_r(n+h) . \tag{4.9}$$

These sums are of importance in studying the moments of the Riemann ζ-function on the critical line, see [7, 19].

For $r = 2$ (the ordinary divisor function), Ingham [18] and Estermann [10] showed that

$$\sum_{n \leq X} d_2(n) d_2(n+h) \sim X P_2(\log X; h), \quad X \to \infty \tag{4.10}$$

where $P_2(u; h)$ is a quadratic polynomial in u.

For $r \geq 3$ it is conjectured that

$$D_r(X; h) \sim X P_{2(r-1)}(\log X; h), \quad X \to \infty \tag{4.11}$$

where $P_{2(r-1)}(u; h)$ is a polynomial in u of degree $2(r-1)$, whose coefficients depend on h (and r). However, even a conjectural description of the polynomials $P_{2(r-1)}(u; h)$ is difficult to obtain, see [7, 19].

In joint work with Andrade and Bary-Soroker [1], we study a version of the additive divisor problem for $\mathbb{F}_q[x]$. We show:

Theorem 4.2. *Let $0 \neq h \in \mathbb{F}_q[x]$, and $n > \deg h$. Then for q odd,*

$$\frac{1}{q^n} \sum_{f \in \mathcal{M}_n} d_r(f) d_r(f+h) = \binom{n+r-1}{r-1}^2 + O_n(q^{-1/2}) , \tag{4.12}$$

the implied constant depending only on n.

Note that $\binom{n+r-1}{r-1}^2$ is a polynomial in n of degree $2(r-1)$ with leading coefficient $1/[(r-1)!]^2$.

4.4. About proofs. The results of this section can all be deduced from one principle (though this was not the original proof of most), namely that for a random monic polynomial $f \in \mathcal{M}_n$ of degree n, the cycle structure of f and its shift $f + \alpha$ are *independent* as $q \to \infty$. Precisely, in [1] we show that for for fixed $n > 1$, and two partitions $\lambda', \lambda'' \vdash n$, given any sequence of finite fields $\mathbb{F}_q$ of odd cardinality q, and nonzero $\alpha \in \mathbb{F}_q[x]$ of degree less than n, then

$$\lim_{q \to \infty} \frac{1}{q^n} \#\{f \in \mathcal{M}_n : \lambda(f) = \lambda', \lambda(f+\alpha) = \lambda''\} = p(\lambda') \times p(\lambda'') \tag{4.13}$$

where $p(\lambda)$, as in (2.4), (2.5), is the probability that a random permutation on n letters has cycle structure λ. This result is an elaboration of earlier work by Bary-Soroker [3] which dealt with the case of n-cycles, where $\lambda = \tilde{\lambda} = (0, \dots, 0, n)$. There is also a version allowing several distinct shifts.

To prove (4.13) we need to compute a certain Galois group: Let $\mathbb{F}$ be an algebraic closure of $\mathbb{F}_q$, $\mathbf{A} = (A_0, \dots, A_{n-1})$ be indeterminates, and

$$\mathcal{F}(\mathbf{A}, x) = x^n + A_{n-1}x^{n-1} + \cdots + A_0 \tag{4.14}$$

the generic polynomial of degree n, whose Galois group over $\mathbb{F}(\mathbf{A})$ is well-known to be the full symmetric group S_n. For nonzero $\alpha \in \mathbb{F}_q[x]$ of degree less than n, let

$$\mathcal{G}(\mathbf{A}, x) = \mathcal{F}(\mathbf{A}, x)\Big(\mathcal{F}(\mathbf{A}, x) + \alpha(x)\Big) . \tag{4.15}$$

Bary-Soroker [3] shows that for odd q, the Galois group of $\mathcal{G}$ over $\mathbb{F}(\mathbf{A})$ is the product $S_n \times S_n$, the maximal possible group. The proof requires an ingredient from the proof of Chowla's conjecture [6] discussed above.

Once we know the Galois group of $\mathcal{G}(\mathbf{A}, x)$, we apply an explicit version of Chebotarev's theorem for function fields to prove (4.13), see [1] for the details.

5. The variance of sums of arithmetic functions and matrix integrals

I now describe some results concerning the variance of sums of several arithmetic functions. A common feature is that the variance is expressed as a matrix integral.

5.1. Variance of primes in short intervals. The von Mangoldt function is defined as $\Lambda(n) = \log p$ if $n = p^k$ is a prime power, and 0 otherwise. A form of the Prime Number Theorem (PNT) is the assertion that

$$\psi(x) := \sum_{n \le x} \Lambda(n) \sim x \quad \text{as } x \to \infty . \tag{5.1}$$

To study the distribution of primes in short intervals, we define for $1 \le H \le x$,

$$\psi(x; H) := \sum_{n \in [x - \frac{H}{2}, x + \frac{H}{2}]} \Lambda(n) . \tag{5.2}$$

The Riemann Hypothesis guarantees an asymptotic formula $\psi(X; H) \sim H$ as long as $H > X^{\frac{1}{2}+o(1)}$. Goldston and Montgomery [13] studied the variance of $\psi(x; H)$, relating it to the pair correlation function of the zeros of the Riemann zeta function. The conjecture of Goldston and Montgomery, as refined by Montgomery and Soundararajan[1] [32] is that in the range $X^\epsilon < H < X^{1-\epsilon}$, as $X \to \infty$:

$$\frac{1}{X}\int_1^X |\psi(x; H) - H|^2 \, dx \sim H\Big(\log X - \log H - (\gamma + \log 2\pi)\Big) \tag{5.3}$$

with γ being Euler's constant.

With J. Keating, we prove a function field analogue of Conjecture 5.3:

[1] based on Hardy-Littlewood type heuristics

Theorem 5.1 ([26]). *For $h \le n-5$, as $q \to \infty$,*

$$\frac{1}{q^n} \sum_{A \in \mathcal{M}_n} \left| \sum_{|f-A| \le q^h} \Lambda(f) - H \right|^2 \sim H \int_{U(n-h-2)} |\operatorname{tr} U^n|^2 \, dU = H(n-h-2) \, .$$

Recall $H := \#\{f : |f-A| \le q^h\} = q^{h+1}$. Here the matrix integral is over the unitary group $U(n-h-2)$, equipped with its Haar probability measure.

5.2. Variance of primes in arithmetic progressions. A form of the Prime Number Theorem for arithmetic progression states that for a modulus Q and A coprime to Q,

$$\psi(X; Q, A) := \sum_{\substack{n \le X \\ n = A \bmod Q}} \Lambda(n) \sim \frac{X}{\phi(Q)}, \quad \text{as } X \to \infty \, . \tag{5.4}$$

In most arithmetic applications it is crucial to allow the modulus to grow with X. For very large moduli $Q > X$, there can be at most one prime in the arithmetic progression $P = A \mod Q$ so that the interesting range is $Q < X$. To study the fluctuations of $\psi(X; Q, A)$, define

$$G(X, Q) = \sum_{\substack{A \bmod Q \\ \gcd(A,Q)=1}} \left| \psi(X; Q, A) - \frac{X}{\phi(Q)} \right|^2 \, . \tag{5.5}$$

Hooley, in his ICM article [17], conjectured that under some (unspecified) conditions,

$$G(X, Q) \sim X \log Q \, . \tag{5.6}$$

Friedlander and Goldston [12] conjecture that (5.6) holds if $X^{1/2+\epsilon} < Q < X$, and further conjecture that if $X^{1/2+\epsilon} < Q < X^{1-\epsilon}$ then

$$G(X, Q) = X \left(\log Q - \left(\gamma + \log 2\pi + \sum_{p | Q} \frac{\log p}{p-1} \right) \right) + o(X) \, . \tag{5.7}$$

They show that both (5.6) (in the range $X^{1/2+\epsilon} < Q < X$) and (5.7) (in the range $X^{1/2+\epsilon} < Q < X^{1-\epsilon}$) hold assuming GRH and a strong version of the Hardy-Littlewood conjecture (4.5) on prime pairs. For $Q < X^{1/2}$ little is known. In any case, Hooley's conjecture (5.6) has not been proved in any range.

With J. Keating [26] we resolve the function-field version of Conjecture (5.6):

Theorem 5.2. *Fix $n \ge 2$. Given a sequence of finite fields $\mathbb{F}_q$ and square-free polynomials $Q(x) \in \mathbb{F}_q[x]$ with $2 \le \deg Q \le n-1$, then as $q \to \infty$,*

$$G(n; Q) \sim q^n \int_{U(\deg Q - 1)} |\operatorname{tr} U|^n dU = q^n (\deg Q - 1) \, . \tag{5.8}$$

We can compare our result (5.8) to the conjectures (5.6) and (5.7): The range $X^{1/2} < Q < X$ corresponds to $\deg Q < n < 2 \deg Q$, so that we recover the function field version of conjecture (5.6); note that (5.8) holds for all n, not just in that range. Thus we believe that Hooley's conjecture (5.6) should hold for all $Q > X^{\epsilon}$. We refer to Fiorilli's recent work [11] for a more refined conjecture in this direction.

5.3. Almost-primes. A variation on this theme was proposed by B. Rodgers [34]. Instead of primes, he considered "almost primes", that is products of two prime powers. A useful weight function for these is the generalized von Mangoldt function

$$\Lambda_2 = \Lambda * \Lambda + \deg \cdot \Lambda = \mu * \deg^2 \tag{5.9}$$

which is supported on products of two prime powers ($*$ means Dirichlet convolution). The mean value of Λ_2 over the set $\mathcal{M}_n$ of monic polynomials of degree n is

$$\frac{1}{q^n} \sum_{f \in \mathcal{M}_n} \Lambda_2(f) = n^2 - (n-1)^2 = 2n - 1 \, . \tag{5.10}$$

To count almost primes in the short intervals, set for $A \in \mathcal{M}_n$, and $1 \leq h < n$

$$\Psi_2(A; h) = \sum_{f \in I(A;h)} \Lambda_2(f) \, . \tag{5.11}$$

Rodgers showed [34] that the variance of $\Psi_2(A; h)$ is given as $q \to \infty$, for fixed n and $h \leq n - 5$, by the matrix integral

$$\operatorname{Var} \Psi_2(\bullet; h) \sim H \int_{U(n-h-2)} \left| \sum_{j=1}^{n-1} \operatorname{tr} U^j \operatorname{tr} U^{n-j} - n \operatorname{tr} U^n \right|^2 dU, \quad q \to \infty \, . \tag{5.12}$$

He shows the matrix integral to be equal to $(4(n-h-2)^3 - (n-h-2))/3$, in fact that

$$\int_{U(N)} \left| \sum_{j=1}^{n-1} \operatorname{tr} U^j \operatorname{tr} U^{n-j} - n \operatorname{tr} U^n \right|^2 dU = \sum_{d=1}^{\min(n,N)} (d^2 - (d-1)^2))^2 \, . \tag{5.13}$$

5.4. Sums of the Möbius function and the Good-Churchhouse conjecture. It is a standard heuristic to assume that the Möbius function behaves like a random variable taking values ± 1 with equal probability, and supported on the square-free integers (which have density $1/\zeta(2) = 6/\pi^2$). In particular if we consider the sums of $\mu(n)$ in blocks of length H,

$$M(x; H) := \sum_{|n-x| < H/2} \mu(n) \tag{5.14}$$

then when averaged over x, $M(x, H)$ has mean zero, and it was conjectured by Good and Churchhouse [14] in 1968 that $M(x; H)$ has variance

$$\frac{1}{X} \int_X^{2X} |M(x; H)|^2 \sim \frac{H}{\zeta(2)} \tag{5.15}$$

for $X^\epsilon < H = H(X) < X^{1-\epsilon}$. Moreover they conjectured that the normalized sums $M(x; H)/\sqrt{H/\zeta(2)}$ have asymptotically a normal distribution.

We can apply our method to evaluate the variance of sums of the Möbius function in short intervals for $\mathbb{F}_q[x]$. Set

$$\mathcal{N}_\mu(A; h) := \sum_{f \in I(A;h)} \mu(f) \, . \tag{5.16}$$

The mean value of $\mathcal{N}_\mu(A; h)$ is 0, and the variance is

Theorem 5.3 (Keating-Rudnick [27]). *If $h \le n-5$ then as $q \to \infty$,*

$$\operatorname{Var} \mathcal{N}_\mu(\bullet; h) \sim H \int_{U(n-h-2)} |\operatorname{tr} \operatorname{Sym}^n U|^2 dU = H$$

where Sym^n is the representation of the unitary group $U(N)$ on polynomials of degree n in N variables.

Theorem 5.3 is consistent with Conjecture (5.15) if we replace H by $H/\zeta_q(2)$ where $\zeta_q(2) = \sum_f 1/|f|^2$ (the sum over all monic f), which tends to 1 as $q \to \infty$.

5.5. The divisor function in short intervals. Dirichlet's divisor problem addresses the size of the remainder term $\Delta_2(x)$ in partial sums of the divisor function:

$$\Delta_2(x) := \sum_{n \le x} d_2(n) - x\Big(\log x + (2\gamma - 1)\Big) \tag{5.17}$$

where γ is the Euler-Mascheroni constant. For the higher divisor functions one defines a remainder term $\Delta_k(x)$ similarly as the difference between the partial sums $\sum_{n \le x} d_k(n)$ and a smooth term $xP_{k-1}(\log x)$ where $P_{k-1}(u)$ is a certain polynomial of degree $k-1$.

Let

$$\Delta_k(x; H) = \Delta_k(x+H) - \Delta_k(x) \tag{5.18}$$

be the remainder term for sums of d_k over short intervals $[x, x+H]$. Jutila [22], Coppola and Salerno [8], and Ivić [20, 21] show that, for $X^\epsilon < H < X^{1/2-\epsilon}$, the mean square of $\Delta_2(x, H)$ is asymptotically equal to

$$\frac{1}{X} \int_X^{2X} \Big(\Delta_2(x, H)\Big)^2 dx \sim H P_3(\log X - 2 \log H) \tag{5.19}$$

for a certain cubic polynomial P_3.

Lester and Yesha [28] showed that $\Delta_2(x, H)$, normalized to have unit mean-square using (5.19), has a Gaussian value distribution at least for a narrow range of H below $X^{1/2}$: $H = \sqrt{X}/L$, where $L = L(X) \to \infty$ with X, but $L \ll X^{o(1)}$, (see [28] for the precise statement), the conjecture being that this should hold for $X^\epsilon < H < X^{1/2-\epsilon}$ for any $\epsilon > 0$.

In joint work with J. Keating and E. Roditty-Gershon [25], we study the corresponding problem of the sum of $d_k(f)$ over short intervals for $\mathbb{F}_q[x]$. Set

$$\mathcal{N}_{d_k}(A; h) := \sum_{f \in I(A;h)} d_k(f)\,. \tag{5.20}$$

The mean value is

$$\frac{1}{q^n} \sum_{A \in \mathcal{M}_n} \mathcal{N}_{d_k}(A; h) = q^{h+1} \binom{n+k-1}{k-1}\,. \tag{5.21}$$

In analogy with (5.17), (5.18) we set

$$\Delta_k(A; h) := \mathcal{N}_{d_k}(A; h) - q^{h+1} \binom{n+k-1}{k-1}\,. \tag{5.22}$$

It can be shown that $\Delta_k(A;h) \equiv 0$ vanishes identically for $h > (1-\frac{1}{k})n - 1$. Using Theorem 3.2 [4], we can show that for all $3 \le h < n$

$$\Delta_k(A;h) \ll_{n,k} q^{h+\frac{1}{2}} \tag{5.23}$$

is smaller than the main term.

We express the mean square of $\Delta_k(A,h)$ (which is the variance of $\mathcal{N}_{d_k}(A;h)$) in terms of a matrix integral. Let $\Lambda^j : U(N) \to GL(\Lambda^j \mathbb{C}^N)$ be the exterior j-th power representation ($0 \le j \le N$). Define the matrix integrals over the group $U(N)$ of $N \times N$ unitary matrices

$$I_k(m;N) := \int_{U(N)} \Big| \sum_{\substack{j_1+\cdots+j_k=m \\ 0\le j_1,\ldots,j_k\le N}} \operatorname{tr}\Lambda^{j_1}(U)\ldots\operatorname{tr}\Lambda^{j_k}(U)\Big|^2 dU \ , \tag{5.24}$$

the integral with respect to the Haar probability measure.

By definition, $I_k(m;N) = 0$ for $m > kN$. We have a functional equation $I_k(m;N) = I_k(kN-m;N)$ and

$$I_k(m;N) = \binom{m+k^2-1}{k^2-1}, \quad m \le N \ . \tag{5.25}$$

The identity (5.25) can be proved by various means, for instance using the work of Diaconis and Gamburd [9] relating matrix integrals to counting magic squares.

Theorem 5.4 ([25]). *Let $n \ge 5$, and $h \le \min(n-5, (1-\frac{1}{k})n - 2)$. Then as $q \to \infty$,*

$$\frac{1}{q^n} \sum_{A \in \mathcal{M}_n} |\Delta_k(A;h)|^2 \sim H \cdot I_k(n; n-h-2) \ .$$

In particular for the standard divisor function ($k=2$), if $h \le n/2 - 2$ and $n \ge 8$ then

$$\frac{1}{q^n} \sum_{A \in \mathcal{M}_n} |\Delta_2(A;h)|^2 \sim H \frac{(n-2h+5)(n-2h+6)(n-2h+7)}{6} \ . \tag{5.26}$$

This is consistent with (5.19), which leads us to expect a cubic polynomial in $(n-2h)$.

6. How to compute the variance

Our results on variance described in § 5 depend on expressing the variance in terms of zeros of Dirichlet L-functions for $\mathbb{F}_q[x]$, and using recent equidistribution results of Katz [23], [24], tailor-made for this purpose. To describe how this is done, we give some background on L-functions.

6.1. Dirichlet L-functions. Let $Q(x) \in \mathbb{F}_q[x]$ be a polynomial of positive degree. A Dirichlet character modulo Q is a homomorphism $\chi : (\mathbb{F}_q[x]/(Q))^\times \to \mathbb{C}^\times$. A Dirichlet character χ is "even" if $\chi(cF) = \chi(F)$ for all $0 \ne c \in \mathbb{F}_q$, and χ is *primitive* if there is no proper divisor $Q' \mid Q$ so that $\chi(F) = 1$ whenever F is coprime to Q and $F = 1 \mod Q'$. The number of Dirichlet characters modulo Q is $\Phi(Q)$, and the number of even characters modulo Q is $\Phi^{ev}(Q) = \Phi(Q)/(q-1)$.

The L-function $\mathcal{L}(u,\chi)$ attached to χ is defined as

$$\mathcal{L}(u,\chi) = \sum_{\substack{f \text{ monic} \\ (f,Q)=1}} \chi(f) u^{\deg f} = \prod_{P \nmid Q} (1 - \chi(P) u^{\deg P})^{-1} \tag{6.1}$$

where the product, over all monic irreducible polynomials in $\mathbb{F}_q[x]$, is absolutely convergent for $|u| < 1/q$.

If $Q \in \mathbb{F}_q[x]$ is a polynomial of degree $\deg Q \geq 2$, and $\chi \neq \chi_0$ is a nontrivial character mod Q, then the L-function $\mathcal{L}(u,\chi)$ is a polynomial in u of degree at most $\deg Q - 1$. Moreover, if χ is an even character there is a "trivial" zero at $u = 1$.

For a primitive even character modulo Q, we can write

$$\mathcal{L}(u,\chi) = (1-u)\det(I - uq^{1/2}\Theta_\chi) \tag{6.2}$$

where the matrix $\Theta_\chi \in U(\deg Q - 2)$ is unitary (as follows from the Riemann Hypothesis for curves), uniquely defined up to conjugacy. It is called the unitarized Frobenius matrix of χ. Likewise, if χ is odd and primitive then $\mathcal{L}(u,\chi) = \det(I - uq^{1/2}\Theta_\chi)$ where $\Theta_\chi \in U(\deg Q - 1)$ is unitary.

Katz [24] showed that as χ varies over all primitive even characters modulo x^{N+2}, the unitarized Frobenii Θ_χ become uniformly distributed in the projectivized unitary group $PU(N)$ for $N \geq 3$ as $q \to \infty$ (and also for $N = 2$ if q is coprime to 2 and 5). Thus for any nice class function F on $U(N)$, which is invariant under the center ($F(zU) = F(u)$, z on the unit circle), we have

$$\lim_{q\to\infty} \frac{1}{\Phi_{ev}(x^{N+2})} \sum_{\substack{\chi \bmod x^{N+2} \\ \text{even primitive}}} F(\Theta_\chi) = \int_{PU(N)} F(U) dU \ . \tag{6.3}$$

6.2. Short intervals as arithmetic progressions. Our method to handle sums over short intervals $I(A;h) = \{f : |f - A| \leq q^h\}$ is to relate them to arithmetic progressions modulo x^{n-h}.

Denote by $\mathcal{P}_{\leq n}$ the set of all polynomials of degree at most n. We define a map $\theta_n : \mathcal{P}_{\leq n} \to \mathcal{P}_{\leq n}$ by

$$\theta_n(f) = x^n f(\frac{1}{x}) \tag{6.4}$$

which takes $f(x) = f_0 + f_1 x + \cdots + f_n x^n$, $n = \deg f$ to the "reversed" polynomial

$$\theta_n(f)(x) = f_0 x^n + f_1 x^{n-1} + \cdots + f_n \ . \tag{6.5}$$

Then for $B \in \mathcal{M}_{n-h-1}$, the map θ_n takes the "interval" $I(T^{h+1}B;h)$ bijectively onto the arithmetic progression $\{g \in \mathcal{P}_{\leq n} : g \equiv \theta_{n-h-1}(B) \mod x^{n-h}\}$.

6.3. A formula for the variance. The identification of short intervals with arithmetic progressions allows us to express sums of several arithmetic functions in terms of even Dirichlet characters. For the case of the von Mangoldt function, this is done in [26]. I illustrate this identification in the case of the Möbius function (Theorem 5.3): We denoted by

$\mathcal{N}_\mu(A;h) = \sum_{f\in I(A;h)} \mu(f)$. Then for $B \in \mathcal{M}_{m-h-1}$,

$$\mathcal{N}_\mu(T^{h+1}B;h) = \frac{1}{\Phi_{ev}(x^{n-h})} \sum_{\substack{\chi \bmod x^{n-h}\\ \chi\neq\chi_0 \text{ even}}} \bar\chi(\theta_{n-h-1}(B))(\mathcal{M}(n;\mu\chi) - \mathcal{M}(n-1;\mu\chi)) \tag{6.6}$$

where

$$\mathcal{M}(n;\mu\chi) = \sum_{f\in\mathcal{M}_n} \mu(f)\chi(f)\ . \tag{6.7}$$

We next express the sums $\mathcal{M}(n;\mu\chi)$ in terms of zeros of the L-function $\mathcal{L}(u,\chi)$; for χ primitive this means in terms of the unitarized Frobenius matrix Θ_χ. The connection is made by writing the generating function identity

$$\sum_{n=0}^{\infty} \mathcal{M}(n;\mu\chi)u^n = \frac{1}{\mathcal{L}(u,\chi)}\ . \tag{6.8}$$

Therefore we find that for χ primitive and even,

$$\mathcal{M}(n;\mu\chi) = \sum_{k=0}^{n} q^{k/2}\,\mathrm{tr}\,\mathrm{Sym}^k\,\Theta_\chi \tag{6.9}$$

where for $N > 1$, $\mathrm{Sym}^n : GL(N,\mathbb{C}) \to GL(\mathrm{Sym}^n\,\mathbb{C}^N)$ is the symmetric n-th power representation. Consequently we obtain

$$\mathrm{Var}\,\mathcal{N}_\mu(\bullet;h) = \frac{q^{h+1}}{\Phi_{\mathrm{ev}}(x^{n-h})} \sum_{\substack{\chi \bmod x^{n-h}\\ \chi \text{ even and primitive}}} |\,\mathrm{tr}\,\mathrm{Sym}^n\,\Theta_\chi|^2 + O(q^h)\ . \tag{6.10}$$

Using Katz's equidistribution theorem (6.3) we get

$$\lim_{q\to\infty} \frac{\mathrm{Var}(\mathcal{N}_\mu(\bullet;h))}{q^{h+1}} = \int_{PU(n-h-2)} |\mathrm{tr}\,\mathrm{Sym}^n\,U|^2\,dU\ . \tag{6.11}$$

The matrix integrals equals 1, hence we conclude that $\mathrm{Var}(\mathcal{N}_\mu(;h)) \sim q^{h+1} = H$, which is Theorem 5.3.

Acknowledgements. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 320755. I am grateful to J. Andrade, L. Bary-Soroker, J. Keating, E. Kowalski, S. Lester, E. Roditty-Gershon and K. Soundararajan for their comments on earlier versions of this survey.

References

[1] J. Andrade, L. Bary-Soroker, and Z. Rudnick, *The additive divisor problem over the rational function field*, preprint 2014.

[2] L. Bary-Soroker, *Irreducible values of polynomials*. Advances in Mathematics (2)**229** (2012), 854-74.

[3] ———, *Hardy-Littlewood tuple conjecture over large finite fields*, Int. Math. Res. Not., 1–8 (2012).

[4] E. Bank, L. Bary-Soroker, and L. Rosenzweig, *Prime polynomials in short intervals and in arithmetic progressions*, to appear in Duke Math. J., arXiv:1302.0625 [math.NT].

[5] J. Bourgain, P. Sarnak, and T. Ziegler, *Disjointness of Möbius from horocycle flows*, From Fourier analysis and number theory to radon transforms and geometry, 67–83, Dev. Math., **28**, Springer, New York, 2013.

[6] D. Carmon and Z. Rudnick, *The autocorrelation of the Mobius function and Chowla's conjecture for the rational function field*, Q J Math (1)**65** (2014), 53–61.

[7] J. B. Conrey and S. M. Gonek, *High Moments of the Riemann Zeta-Function*, Duke Math. J. **107** (2001), 577–604.

[8] G. Coppola and S. Salerno, *On the symmetry of the divisor function in almost all short intervals,* Acta Arith. **113** (2004), no. 2, 189–201.

[9] P. Diaconis and A. Gamburd, *Random matrices, magic squares and matching polynomials*, Electron. J. Combin. **11** (2004/06), no. 2, Research Paper, 2–26.

[10] T. Estermann, *Über die Darstellungen einer Zahl als Differenz von zwei Produkten*, Journal für die reine und angewandte Mathematik **164** (1931), 173–182.

[11] D. Fiorilli, *The distribution of the variance of primes in arithmetic progressions*, to appear in Int. Math. Res. Not., arXiv:1301.5663 [math.NT].

[12] J. B. Friedlander and D. A. Goldston, *Variance of distribution of primes in residue classes*. Quart. J. Math. Oxford Ser. (2) **47** (1996), no. 187, 313–336.

[13] D. A. Goldston and H. L. Montgomery, *Pair correlation of zeros and primes in short intervals*, Analytic number theory and Diophantine problems (Stillwater, OK, 1984), 183–203, Progr. Math., 70, Birkhäuser Boston, Boston, MA, 1987.

[14] I. J. Good and R. F. Churchhouse, *The Riemann Hypothesis and Pseudorandom Features of the Möbius Sequence*, Mathematics of Computation No. 104, **22** (1968), 857–861.

[15] A. Granville and K. Soundararajan, *An uncertainty principle for arithmetic sequences*, Ann. of Math. (2) **165** (2007), no. 2, 593–635.

[16] B. Green and T. Tao, *The Möbius function is strongly orthogonal to nilsequences*, Ann. of Math. (2) **175** (2012), no. 2, 541–566.

[17] C. Hooley, *The distribution of sequences in arithmetic progression*, Proc. ICM Vancouver (1974), 357–364.

[18] A. E. Ingham, *Mean-value theorems in the theory of the Riemann Zeta-function*, Proc. London Math. Soc. (2) **27** (1928), 273–300.

[19] A. Ivić, *On the ternary additive divisor problem and the sixth moment of the zeta-function*, Sieve methods, exponential sums, and their applications in number theory (Cardiff, 1995), 205–243, London Math. Soc. Lecture Note Ser., **237**, Cambridge Univ. Press, Cambridge, 1997.

[20] _____, *On the mean square of the divisor function in short intervals,* J. Théor. Nombres Bordeaux **21** (2009), 251–261.

[21] _____, *On the divisor function and the Riemann zeta-function in short intervals*, Ramanujan J. **19** (2009), 207–224.

[22] M. Jutila, *On the divisor problem for short intervals,* Studies in honour of Arto Kustaa Salomaa on the occasion of his fiftieth birthday. Ann. Univ. Turku. Ser. A I No. 186 (1984), 23–30.

[23] N. M. Katz, *On a Question of Keating and Rudnick about Primitive Dirichlet Characters with Squarefree Conductor*, Int Math Res Notices (2013) Vol. 2013, 3221–3249,

[24] _____, *Witt vectors and a question of Keating and Rudnick*, Int Math Res Notices (2013) Vol. 2013, 3613–3638.

[25] J. P. Keating, E. Roditty-Gershon, and Z. Rudnick, in preparation.

[26] J. P. Keating and Z. Rudnick, *The variance of the number of prime polynomials in short intervals and in residue classes*. Int Math Res Notices (2014) 2014 (1), 259–288.

[27] J. P. Keating and Z. Rudnick, in preparation.

[28] S. Lester and N. Yesha, *On the distribution of the divisor function and Hecke eigenvalues*, arXiv:1404.1579 [math.NT].

[29] J. Liu and P. Sarnak, *The Möbius function and distal flows*, arXiv:1303.4957[math.NT].

[30] H. Maier, *Primes in short intervals*, Michigan Math. J. **32** (1985), 221–225.

[31] J. Maynard, *Small gaps between primes*, To appear in Ann. of Math., arXiv:1311.4600 [math.NT].

[32] Montgomery, H. L. and Soundararajan, K., *Beyond pair correlation*, Paul Erdos and his mathematics, I (Budapest, 1999), 507–514, Bolyai Soc. Math. Stud., 11, Janos Bolyai Math. Soc., Budapest, 2002. arXiv:math/0003234 [math.NT].

[33] P. Pollack, *Simultaneous prime specializations of polynomials over finite fields*, in Proceedings of the London Mathematical Society. Third Series. Vol. 97 (2008), 545–67.

[34] B. Rodgers, *The covariance of almost-primes in* $\mathbb{F}_q[T]$, arXiv:1311.4905 [math.NT].

[35] P. Sarnak, *Three lectures on Möbius randomness*, (2011) available at http://www.math.ias.edu/files/wam/2011/PSMobius.pdf.

[36] F. Thorne, *Irregularities in the distributions of primes in function fields*, J. Number Theory **128** (2008), no. 6, 1784–1794.

[37] Y. Zhang, *Bounded gaps between primes*, Annals of Mathematics, Volume 179, Issue 3 (2014), 1121–1174.

Raymond and Beverly Sackler School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel

E-mail: rudnick@post.tau.ac.il

Perfectoid spaces and their applications

Peter Scholze

Abstract. We survey the theory of perfectoid spaces and its applications.

Mathematics Subject Classification (2010). Primary 14G22, 11F80; Secondary 14G20, 14C30, 14L05, 14G35, 11F03.

Keywords. Perfectoid spaces, rigid-analytic geometry, almost mathematics, p-adic Hodge theory, Shimura varieties, Langlands program.

1. Introduction

In algebraic geometry, one of the most important dichotomies is the one between characteristic 0 and positive characteristic p. Our intuition is formed from the study of complex manifolds, which are manifestly of characteristic 0, but in number theory, the most important questions are in positive or mixed characteristic. Algebraic geometry gives a framework to transport intuition from characteristic 0 to positive characteristics. However, there are also several new phenomena in characteristic p, such as the presence of the Frobenius map, which acts naturally on all spaces of characteristic p. Using the Frobenius, one can formulate the Weil conjectures, and more generally the theory of weights. This makes many results accessible over fields such as $\mathbb{F}_p((t))$, which are wide open over fields of arithmetic interest such as $\mathbb{Q}_p$. The theory of perfectoid spaces was initially designed as a means of transporting information available over $\mathbb{F}_p((t))$ to $\mathbb{Q}_p$, but has since found a number of independent applications. The purpose of this report is to give an overview of the developments in the field since perfectoid spaces were introduced in early 2011.

To study the transition between characteristic 0 and characteristic p, it is useful to look at the corresponding local fields $\mathbb{Q}_p$ and $\mathbb{F}_p((t))$:

$$\mathbb{Q}_p = \{ \sum_{n \gg -\infty} a_n p^n \mid a_n \in \{0, 1, \dots, p-1\} \} ,$$
$$\mathbb{F}_p((t)) = \{ \sum_{n \gg -\infty} a_n t^n \mid a_n \in \{0, 1, \dots, p-1\} = \mathbb{F}_p \} .$$

Although these two fields have formally 'the same' elements, the basic addition and multiplication operations are different: In $\mathbb{Q}_p$, one computes with carry, but in $\mathbb{F}_p((t))$ without carry. There are several strategies to pass from one field to the other. Let us recall the most important ones.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Letting $p \to \infty$. In model theory, one can formalize the idea that $\mathbb{Q}_p$ becomes isomorphic to $\mathbb{F}_p((t))$ as $p \to \infty$. This has the following implication: A first-order statement is true for almost all fields $\mathbb{Q}_p$ (for varying p) if and only if it is true for almost all fields $\mathbb{F}_p((t))$. The first application of this was the Ax-Kochen theorem, [4], that a homogeneous polynomial of degree d in more than d^2 variables admits a solution over $\mathbb{Q}_p$, except for a finite list of primes p (which depends only on d). In fact, the same result is true over $\mathbb{F}_p((t))$ for all p. However, there are counterexamples to the general statement over $\mathbb{Q}_p$, such as a quartic form in 18 variables over $\mathbb{Q}_2$ without a solution. More strikingly, this transfer principle is used in the proof of the fundamental lemma: Ngô, [52], has proved the fundamental lemma over $\mathbb{F}_p((t))$ (for sufficiently large p), which could then be transferred to $\mathbb{Q}_p$, if p is sufficiently large.[1]

However, this strategy cannot be used to get information about any fixed prime number p. One of the ways in which one wants to compare two fields is to compare the categories of finite extension fields. This is encapsulated by the absolute Galois group $G_K = \mathrm{Gal}(\overline{K}/K)$ of a field K, where $\overline{K}$ is some separable closure of K. If K is a local field such as $\mathbb{Q}_p$ or $\mathbb{F}_p((t))$, it comes with a decreasing ramification filtration

$$G_K \supset G_K^{(0)} = I_K \supset G_K^{(1)} = P_K \supset G_K^{(2)} \supset \dots ;$$

here, $P_K \subset I_K \subset G_K$ are the wild inertia, resp. inertia subgroups. The 'tame quotient' $G_K^{\mathrm{tame}} = G_K/G_K^{(1)} = G_K/P_K$ admits an explicit description, and P_K is a (not very explicit) pro-p-group.

Restricting ramification. From the explicit description of G_K^{tame} in the case of local fields, one knows that $G_{\mathbb{Q}_p}^{\mathrm{tame}} \cong G_{\mathbb{F}_p((t))}^{\mathrm{tame}}$ canonically. In other words, there is a canonical procedure to associate to a tame extension of $\mathbb{Q}_p$ a tame extension of $\mathbb{F}_p((t))$. This result can be strengthened if one passes to extension fields. More precisely, for any $n \geq 1$,

$$G_{\mathbb{Q}_p(p^{1/n})}/G^{(n)}_{\mathbb{Q}_p(p^{1/n})} \cong G_{\mathbb{F}_p((t))(t^{1/n})}/G^{(n)}_{\mathbb{F}_p((t))(t^{1/n})} .$$

This is a result of Deligne, [21], relying on ideas of Krasner, [49], which formalizes the idea that $\mathbb{Q}_p(p^{1/n})$ and $\mathbb{F}_p((t))(t^{1/n})$ are 'close local fields' (which get 'closer' as $n \to \infty$). Note that, again, this result plays a crucial role in the Langlands program, namely it is used (through Henniart's numerical local Langlands correspondence, [37]) in the original proof of the local Langlands correspondence for GL_n over $\mathbb{Q}_p$ by Harris-Taylor, [34].[2]

There is yet another approach, which gives a comparison of the whole Galois group.

Making things perfect(oid). Let $K = \mathbb{Q}_p(p^{1/p^\infty}) = \bigcup_m \mathbb{Q}_p(p^{1/p^m})$, which we will momentarily confuse with its completion, which has the same absolute Galois group G_K. Then a theorem of Fontaine-Wintenberger, [32], states that the absolute Galois groups of G_K and $G_{\mathbb{F}_p((t))}$ are isomorphic. This can be considered as a limit case of Deligne's theorem as one lets $n = p^m$, $m \to \infty$. Indeed, note that as $\mathbb{F}_p((t))(t^{1/p^m})$ is a purely inseparable extension of $\mathbb{F}_p((t))$, $G_{\mathbb{F}_p((t))(t^{1/p^m})} \cong G_{\mathbb{F}_p((t))}$. In fact, if one lets $K^\flat$ be the completion of

$$\mathbb{F}_p((t))(t^{1/p^\infty}) = \bigcup_m \mathbb{F}_p((t))(t^{1/p^m}) ,$$

[1] One does not need model theory to do this, as Waldspurger, [66], had earlier shown this transfer principle for large p directly.

[2] The alternative proof given in [57] avoids this argument, and gives a proof of the local Langlands correspondence for GL_n over $\mathbb{Q}_p$ which is purely in characteristic 0.

then the theorem of Fontaine-Wintenberger states equivalently that $G_K \cong G_{K^\flat}$. This theorem is one of the foundational cornerstones of p-adic Hodge theory. Moreover, it is true in a wide variety of cases: Any 'deeply ramified' extension of $\mathbb{Q}_p$ can be used in place of $\mathbb{Q}_p(p^{1/p^\infty})$.

Note that the last approach gives the cleanest result: It works for any fixed p, and produces an isomorphism of the whole Galois groups. However, it comes at the expense of passing to infinite extensions. The theory of perfectoid spaces is a generalization of this procedure to higher-dimensional objects.

2. The Fontaine-Wintenberger isomorphism

To start, let us explain the general statement of the Fontaine-Wintenberger isomorphism.[3]

Definition 2.1. A perfectoid field is a complete topological field K, whose topology comes from a nonarchimedean norm $|\cdot| : K \to \mathbb{R}_{\geq 0}$ with dense image, such that $|p| < 1$ and, letting $\mathcal{O}_K = \{x \in K \mid |x| \leq 1\}$ be the ring of integers, the Frobenius map $\Phi : \mathcal{O}_K/p \to \mathcal{O}_K/p$ is surjective.

Examples include the completions of $\mathbb{Q}_p(p^{1/p^\infty})$, $\mathbb{Q}_p(\mu_{p^\infty})$, $\overline{\mathbb{Q}}_p$ and $\mathbb{F}_p((t))(t^{1/p^\infty})$, $\overline{\mathbb{F}_p((t))}$. Note that perfectoid fields can be of characteristic 0 or p. In the first case, they contain $\mathbb{Q}_p$ naturally, as $|p| < 1$. Note that $\mathbb{Q}_p$ is not a perfectoid field (although $\mathbb{Z}_p/p = \mathbb{F}_p$ has a surjective Frobenius map), because $|\cdot| : \mathbb{Q}_p \to \mathbb{R}_{\geq 0}$ has discrete image $0 \cup p^{\mathbb{Z}} \subset \mathbb{R}_{\geq 0}$. In characteristic p, perfectoid fields are the same thing as perfect complete nonarchimedean fields.

By a construction of Fontaine, one can take any perfectoid field K, and produce a perfectoid field $K^\flat$ of characteristic p, called the tilt of K. First, one defines $\mathcal{O}_{K^\flat} = \varprojlim_\Phi \mathcal{O}_K/p$, and then defines $K^\flat$ as the fraction field of $\mathcal{O}_{K^\flat}$. It comes with a natural norm, with respect to which $\mathcal{O}_{K^\flat} \subset K^\flat$ is the ring of integers. In fact, one has the following alternative description of $K^\flat$.

Lemma 2.2. *There is a natural identification of multiplicative monoids*

$$\mathcal{O}_{K^\flat} = \varprojlim_{x \mapsto x^p} \mathcal{O}_K = \{(x^{(0)}, x^{(1)}, \ldots) \mid x^{(i)} \in \mathcal{O}_K, (x^{(i+1)})^p = x^{(i)}\}\,, \quad K^\flat = \varprojlim_{x \mapsto x^p} K\,.$$

In particular, $x \mapsto x^\sharp := x^{(0)}$ defines a multiplicative map $K^\flat \to K$, and the norm $|x|_{K^\flat} = |x^\sharp|_K$ on $K^\flat$.

The crucial input is the basic fact that if $a \cong b \mod p$, then $a^{p^n} \cong b^{p^n} \mod p^{n+1}$.

As a basic example of the tilting equivalence, the perfectoid field K which is the completion of $\mathbb{Q}_p(p^{1/p^\infty})$ tilts to the perfect nonarchimedean field $K^\flat$ which is the completion of $\mathbb{F}_p((t))(t^{1/p^\infty})$. Under the identification $K^\flat = \varprojlim_{x \mapsto x^p} K$, the element t corresponds to the sequence $(p, p^{1/p}, p^{1/p^2}, \ldots)$. In particular, $t^\sharp = p$, so in a vague sense, the map $x \mapsto x^\sharp$ is the map 'replace t by p'. However, calculating it in general involves a p-adic limit, so e.g.

$$(1+t)^\sharp = \lim_{n \to \infty} (1 + p^{1/p^n})^{p^n}\,.$$

[3]It should be noted that the original result of Fontaine-Wintenberger is quite different, at least in emphasis. The theorem as stated was only proved recently, and was noticed independently (at least) by Kedlaya-Liu and the author.

This already shows that any general theory of perfectoid objects has to be of an analytic nature.

Theorem 2.3 ([47, Theorem 3.5.6],[58, Theorem 3.7]). *Let K be a perfectoid field.*

(i) *For any finite extension L/K, L is a perfectoid field.*

(ii) *The association $L \mapsto L^\flat$ defines an equivalence between the category of finite extensions of K and the category of finite extensions of $K^\flat$.*

It is formal to deduce from part (ii) that the absolute Galois groups $G_K \cong G_{K^\flat}$ are isomorphic.

3. Untilting: Work of Fargues-Fontaine

The following question arises naturally: For a given perfectoid field L of characteristic p, in how many ways can it be untilted to a perfectoid field K of characteristic 0, $K^\flat \cong L$? The answer to this question leads naturally to the Fargues-Fontaine curve, [30, 31]. In particular, they prove the following theorem.

Theorem 3.1. *Fix a perfectoid field L of characteristic p. There is a regular noetherian scheme X_L of Krull dimension 1 (locally the spectrum of a principal ideal domain) over $\mathbb{Q}_p$ whose closed points x are in bijection with equivalence classes of pairs (K, ι), where K is a perfectoid field of characteristic 0 and $\iota : L \hookrightarrow K^\flat$ is an injection which makes $K^\flat$ a finite extension of L; here, the pairs (K, ι) and $(K, \iota \circ \Phi^n)$ are regarded as equivalent for any $n \in \mathbb{Z}$. The degree $[K^\flat : L]$ is called the degree of x. Moreover, there are (infinitely many) points of degree 1.*

In particular, one can always untilt a perfectoid field L to characteristic 0, and the ways of doing so are parametrized by a 1-dimensional object. Note that if, e.g., L is algebraically closed, then all points are of degree 1 and have algebraically closed residue field. However, the curve lives only over $\mathbb{Q}_p$, and thus is not of finite type over $\mathbb{Q}_p$. Concretely,

$$X_L = \mathrm{Proj} \bigoplus_{n \geq 0} B^+(L)^{\varphi = p^n} ,$$

where $B^+(L)$ is one of Fontaine's period rings, a certain completion of $W(\mathcal{O}_L)[\frac{1}{p}]$. A point of X_L gives rise to an ideal $I \subset W(\mathcal{O}_L)[\frac{1}{p}]$ (well-defined up to the action of Frobenius), and the corresponding perfectoid field of characteristic 0 is given by $K = W(\mathcal{O}_L)[\frac{1}{p}]/I$. This gives an explicit description of untilting in terms of Witt vectors.

The work of Fargues-Fontaine has further connections with the theory of perfectoid spaces that we cannot explain in detail here, for lack of space, cf. [29]. For the rest of this article, we will usually fix a perfectoid field K in characteristic 0, which amounts to fixing a point $\infty \in X_{K^\flat}$ of degree 1.

4. Perfectoid Spaces

Definition 4.1. A perfectoid K-algebra is a Banach K-algebra R for which the subring $R^\circ \subset R$ of powerbounded elements is a bounded subring, and such that the Frobenius map

$\Phi : R^\circ/p \to R^\circ/p$ is surjective.

The simplest example is $R = K\langle T^{1/p^\infty}\rangle$ for which $R^\circ = \mathcal{O}_K\langle T^{1/p^\infty}\rangle$ is the completion of $\mathcal{O}_K[T^{1/p^\infty}] = \bigcup_m \mathcal{O}_K[T^{1/p^m}]$. In other words, perfectoid K-algebras are algebras with 'lots of (approximate) p-power roots'. Note that perfectoid K-algebras are always quite big, e.g. nonnoetherian; also, no 'smallness' hypothesis is imposed. The mixture of completeness and nonnoetherianity might cause big trouble (as, e.g., completions of nonnoetherian algebras are not in general flat)! However, it turns out that the 'bigness' condition of surjective Frobenius forces good behaviour.

One can apply Fontaine's construction to any perfectoid K-algebra. This defines the tilting functor: Let R be a perfectoid K-algebra. Set

$$R^{\flat\circ} = \varprojlim_{\Phi} R^\circ/p = \varprojlim_{x\mapsto x^p} R^\circ$$

which is a $\mathcal{O}_{K^\flat}$-algebra, and

$$R^\flat = R^{\flat\circ} \otimes_{\mathcal{O}_{K^\flat}} K^\flat = \varprojlim_{x\mapsto x^p} R\,.$$

Proposition 4.2 ([58, Theorem 5.2]). *Fix a perfectoid field K with tilt $K^\flat$.*

(i) *For any perfectoid K-algebra R, the tilt $R^\flat$ is a perfectoid $K^\flat$-algebra with subring of powerbounded elements $R^{\flat\circ} \subset R^\flat$.*

(ii) *The functor $R \mapsto R^\flat$ defines an equivalence between the category of perfectoid K-algebras and the category of perfectoid $K^\flat$-algebras.*

Note also that for any perfectoid K-algebra R, one has a continuous multiplicative map $R^\flat \to R$, $f \mapsto f^\sharp$.

As remarked earlier, any theory of perfectoid objects has to be of an analytic nature. This reflects itself algebraically in the fact that perfectoid algebras are Banach algebras. On the level of spaces, it means that we have to work in some category of nonarchimedean analytic spaces. The classical such category is Tate's category of rigid-analytic spaces, [64], but strong finiteness assumptions are built into the foundations of this theory. There are (at least) two more recent approaches to nonarchimedean analytic spaces: Berkovich's analytic spaces, [8], and Huber's adic spaces, [40]. We choose to work with Huber's adic spaces, because we feel that it is the most natural framework; e.g., it interacts well with the theory of formal models. Moreover, one glues spaces along open subsets, which is at least technically convenient.[4]

Following Huber, we make the following definition in the perfectoid world:

Definition 4.3. A perfectoid affinoid K-algebra is a pair (R, R^+), where R is a perfectoid K-algebra, and $R^+ \subset R^\circ$ is an open and integrally closed subring.

The role of the integral subalgebra R^+ is certainly secondary, and one may safely assume that $R^+ = R^\circ$ on first reading.

Proposition 4.4. *The association $(R, R^+) \mapsto (R^\flat, R^{\flat+})$ with*

$$R^\flat = \varprojlim_{x\mapsto x^p} R\,,\ R^{\flat+} = \varprojlim_{x\mapsto x^p} R^+$$

[4] Somewhat more importantly, most examples of perfectoid spaces arise as 'inverse limits' of classical finite type spaces. Berkovich spaces are not well adapted to taking inverse limits.

defines an equivalence between perfectoid affinoid K-algebras and perfectoid affinoid $K^\flat$-algebras.

To a pair (R, R^+), Huber associates a space of continuous valuations.

Definition 4.5. A valuation on R is a map $|\cdot| : R \to \Gamma \cup \{0\}$, where Γ is a totally ordered abelian group (e.g., $\Gamma = \mathbb{R}_{>0}$, but higher-rank valuations are allowed), such that $|0| = 0$, $|1| = 1$, $|xy| = |x||y|$ and $|x + y| \leq \max(|x|, |y|)$. The valuation $|\cdot|$ is continuous if for all $\gamma \in \Gamma$, the subset $\{x \in R \mid |x| < \gamma\} \subset R$ is open.

There is an obvious notion of equivalence of valuations, and one defines $\mathrm{Spa}(R, R^+)$ as the set of equivalence classes of continuous valuations $|\cdot|$ on R such that $|R^+| \leq 1$. For a point $x \in \mathrm{Spa}(R, R^+)$, we denote by $f \mapsto |f(x)|$ the associated valuation. One may find back R^+ as

$$R^+ = \{f \in R \mid |f(x)| \leq 1 \ \forall x \in \mathrm{Spa}(R, R^+)\} .$$

One equips $\mathrm{Spa}(R, R^+)$ with the topology generated by rational subsets: For $f_1, \ldots, f_n$, $g \in R$ which generate R as an ideal, the subset

$$U(f_1, \ldots, f_n; g) = \{x \in \mathrm{Spa}(R, R^+) \mid |f_i(x)| \leq |g(x)|\} \subset \mathrm{Spa}(R, R^+)$$

is a rational subset.

Proposition 4.6 ([38, Theorem 3.5]). *The space $\mathrm{Spa}(R, R^+)$ is a spectral space. In particular, it is quasicompact, quasiseparated, and the rational subsets form a basis for the topology consisting of quasicompact open subsets, stable under finite intersections.*

Again, one finds an interesting relation under tilting.[5]

Theorem 4.7 ([58, Theorem 6.3 (i)]). *For any $x \in \mathrm{Spa}(R, R^+)$, one may define a point $x^\flat \in \mathrm{Spa}(R^\flat, R^{\flat +})$ by setting $|f(x^\flat)| := |f^\sharp(x)|$ for $f \in R^\flat$. This defines a homeomorphism $\mathrm{Spa}(R, R^+) \cong \mathrm{Spa}(R^\flat, R^{\flat +})$ preserving rational subsets.*

The proof relies on the following crucial approximation lemma.[6]

Lemma 4.8 ([58, Corollary 6.7 (i)]). *Assume K is of characteristic 0. Let $f \in R$ be any element, and fix any $\epsilon > 0$. Then there exists $g \in R^\flat$ such that for all $x \in \mathrm{Spa}(R, R^+)$,*

$$|(f - g^\sharp)(x)| \leq |p|^{1-\epsilon} \max(|f(x)|, \epsilon) .$$

This means in particular that $|f(x)| = |g^\sharp(x)|$ except if both are very small. However, $f - g^\sharp$ may be quite large if f is large.

One wants to equip the topological space $X = \mathrm{Spa}(R, R^+)$ with a structure sheaf $\mathcal{O}_X$. For this, let $U = U(f_1, \ldots, f_n; g) \subset X$ be a rational subset. Equip $R[g^{-1}]$ with the topology for which the image of $R^+[\frac{f_1}{g}, \ldots, \frac{f_n}{g}] \to R[g^{-1}]$ is open and bounded. Let $R\langle \frac{f_1}{g}, \ldots, \frac{f_n}{g} \rangle$ be the completion of $R[g^{-1}]$ with respect to this topology; it comes equipped with a natural subring

$$R\langle \frac{f_1}{g}, \ldots, \frac{f_n}{g} \rangle^+ \subset R\langle \frac{f_1}{g}, \ldots, \frac{f_n}{g} \rangle .$$

[5] A closely related result was proved earlier by Kedlaya, [46].

[6] A slightly stronger version (replacing $1 - \epsilon$ by $p/(p-1) - \epsilon$) appears in [47].

Proposition 4.9 ([39, Proposition 1.3]). *The pair*

$$(\mathcal{O}_X(U), \mathcal{O}_X^+(U)) = \left(R\langle \frac{f_1}{g}, \dots, \frac{f_n}{g} \rangle, R\langle \frac{f_1}{g}, \dots, \frac{f_n}{g} \rangle^+ \right)$$

depends only on the rational subset $U \subset X$. The map

$$\mathrm{Spa}(\mathcal{O}_X(U), \mathcal{O}_X^+(U)) \to \mathrm{Spa}(R, R^+)$$

is a homeomorphism onto U, preserving rational subsets.

The propositions of Huber so far have not used the assumption that R is perfectoid. This assumption is needed, however, to prove that $\mathcal{O}_X$ is actually a sheaf. Huber proved this when R is strongly noetherian, so e.g. if R is topologically of finite type over K. Perfectoid K-algebras are virtually never (strongly) noetherian.

Theorem 4.10 ([58, Theorem 6.3]). *Let (R, R^+) be a perfectoid affinoid K-algebra with tilt $(R^\flat, R^{\flat +})$. Let $X = \mathrm{Spa}(R, R^+)$, $X^\flat = \mathrm{Spa}(R^\flat, R^{\flat +})$. For any rational subset $U \subset X$, let $U^\flat \subset X^\flat$ be its image under the homeomorphism $X \cong X^\flat$.*

(i) *The presheaves $\mathcal{O}_X$, $\mathcal{O}_{X^\flat}$ are sheaves.*

(ii) *For any rational subset $U \subset X$, the pair $(\mathcal{O}_X(U), \mathcal{O}_X^+(U))$ is a perfectoid affinoid K-algebra, which tilts to $(\mathcal{O}_{X^\flat}(U^\flat), \mathcal{O}_{X^\flat}^+(U^\flat))$.*

(iii) *For any $i > 0$, the cohomology group $H^i(X, \mathcal{O}_X) = 0$ vanishes. In fact, $H^i(X, \mathcal{O}_X^+)$ is almost zero, i.e. killed by the maximal ideal of $\mathcal{O}_K$.*

The resulting spaces $\mathrm{Spa}(R, R^+)$ (equipped with the two sheaves of topological rings $\mathcal{O}_X$, $\mathcal{O}_X^+$) are called affinoid perfectoid spaces (over K). Objects obtained by gluing such spaces are called perfectoid spaces over K.

Corollary 4.11. *The categories of perfectoid spaces over K and over $K^\flat$ are equivalent. Here, if X tilts to $X^\flat$, then the underlying topological spaces of X and $X^\flat$ are canonically homeomorphic. Moreover, a subset $U \subset X$ is affinoid perfectoid if and only if $U^\flat \subset X^\flat$ is affinoid perfectoid. For any such U, $(\mathcal{O}_X(U), \mathcal{O}_X^+(U))$ is a perfectoid affinoid K-algebra with tilt $(\mathcal{O}_{X^\flat}(U^\flat), \mathcal{O}_{X^\flat}^+(U^\flat))$.*

For any perfectoid space X, one may define its étale site $X_{\text{ét}}$.

Theorem 4.12 ([58, Theorem 7.12, Proposition 7.13]). *Under tilting, $X_{\text{ét}} \cong X^\flat_{\text{ét}}$. Moreover, if $X = \mathrm{Spa}(R, R^+)$ is affinoid perfectoid, then $H^0(X_{\text{ét}}, \mathcal{O}_X^+) = R^+$ while $H^i(X_{\text{ét}}, \mathcal{O}_X^+)$ is almost zero for $i > 0$. In particular, $H^0(X_{\text{ét}}, \mathcal{O}_X) = R$ while $H^i(X_{\text{ét}}, \mathcal{O}_X) = 0$ for $i > 0$.*

The assertion $X_{\text{ét}} \cong X^\flat_{\text{ét}}$ is a far-reaching generalization of the Fontaine-Wintenberger isomorphism. Indeed, if we put $X = \mathrm{Spa}(K, \mathcal{O}_K)$, which tilts to $X^\flat = \mathrm{Spa}(K^\flat, \mathcal{O}_{K^\flat})$, the assertion is precisely the Fontaine-Wintenberger isomorphism. The assertion about $H^i(X_{\text{ét}}, \mathcal{O}_X^+)$ is a strengthening of Faltings's almost purity theorem, which is essentially the version of it for the finite étale site. Let us state it in our setup.

Theorem 4.13 ([58, Theorem 7.9 (iii)]). *Let R be a perfectoid K-algebra, and let S/R be finite étale. Then S is a perfectoid K-algebra, and S° is a uniformly almost finite étale R°-algebra.*

The following is an easy corollary, which gives a higher-dimensional variant of the Fontaine-Wintenberger isomorphism (for the finite étale case).

Corollary 4.14. *Let R be a perfectoid K-algebra with tilt $R^\flat$. Then tilting defines an equivalence between the categories of finite étale R-algebras and finite étale $R^\flat$-algebras.*

The almost purity theorem is interesting only in characteristic 0; in characteristic p, it is easy. Originally, Faltings proved such statements in the case of good reduction, [23], and then more generally for semistable (or more generally toric) reduction, [25]. We note that in Faltings' situation, R was the completion of an inductive limit of regular algebras. Then, by Zariski-Nagata purity, the ramification locus of S° over R° is purely of codimension 1. We know by assumption that there is no ramification in characteristic 0, as S/R is finite étale. At the generic points of R°/p, it follows from (the proof of) the Fontaine-Wintenberger result that there is *almost* no ramification. If there were none, one would get that S°/R° is finite étale. Faltings made the same argument work in the almost world. It came as a surprise that no regularity assumption is needed for the theorem.

5. Example: Projective spaces

Let K be a perfectoid field with tilt $K^\flat$. Let us consider the case of projective space. In all applications of perfectoid spaces, the hard part is to find a way to pass from objects of finite type over K to perfectoid objects. This is not possible in a canonical way, and one has to make a choice.

On $\mathbb{P}^n$, one has the map $\varphi : \mathbb{P}^n \to \mathbb{P}^n$ sending $(x_0 : \ldots : x_n)$ to $(x_0^p : \ldots : x_n^p)$. Consider $\mathbb{P}^n_K$ as an adic space over K. Then there is a perfectoid space $(\mathbb{P}^n_K)^{\mathrm{perf}}$ over K such that

$$(\mathbb{P}^n_K)^{\mathrm{perf}} \sim \varprojlim_{\varphi} \mathbb{P}^n_K \, .$$

Here, $\sim \varprojlim$, read 'being similar to the inverse limit', is a technical notion that accounts for the non-existence of inverse limits in the category of adic spaces, cf. [62, Definition 2.4.1]. Explicitly, $(\mathbb{P}^n_K)^{\mathrm{perf}}$ is glued out of $n+1$ copies of

$$\mathrm{Spa}(K\langle T_1^{1/p^\infty}, \ldots, T_n^{1/p^\infty}\rangle, \mathcal{O}_K\langle T_1^{1/p^\infty}, \ldots, T_n^{1/p^\infty}\rangle)$$

in the usual way. One can make the same construction over $K^\flat$ to get $(\mathbb{P}^n_{K^\flat})^{\mathrm{perf}}$.

Theorem 5.1 ([58, Theorem 8.5]). *The perfectoid space $(\mathbb{P}^n_K)^{\mathrm{perf}}$ tilts to $(\mathbb{P}^n_{K^\flat})^{\mathrm{perf}}$. In particular, there are homeomorphisms of topological spaces underlying the adic spaces, resp. isomorphisms of étale sites,*

$$|\mathbb{P}^n_{K^\flat}| \cong |(\mathbb{P}^n_{K^\flat})^{\mathrm{perf}}| \cong |(\mathbb{P}^n_K)^{\mathrm{perf}}| \cong \varprojlim_{\varphi} |\mathbb{P}^n_K| \, .$$

$$(\mathbb{P}^n_{K^\flat})_{\text{ét}} \cong (\mathbb{P}^n_{K^\flat})^{\mathrm{perf}}_{\text{ét}} \cong (\mathbb{P}^n_K)^{\mathrm{perf}}_{\text{ét}} \cong \varprojlim_{\varphi} (\mathbb{P}^n_K)_{\text{ét}} \, .$$

These constructions give a 'projection map'

$$\pi : \mathbb{P}^n_{K^\flat} \to \mathbb{P}^n_K$$

defined on topological spaces and étale topoi, and given by $(x_0 : \ldots : x_n) \mapsto (x_0^\sharp : \ldots : x_n^\sharp)$ in coordinates.

There are many variants to this theorem. All one needs is a 'dynamical system' (X, φ) over K such that (with respect to a suitable integral model of X) φ is a lift of Frobenius. E.g., one might take the canonical lift of an ordinary abelian variety, with its canonical lift of Frobenius. In that case, the tilt will be the perfection of the ordinary abelian variety in characteristic p. However, nothing of this sort works of curves of genus ≥ 2. Currently, there are very few explicit examples of tilting for varieties besides the cases of toric varieties and (semi-)abelian varieties. An interesting case might be the one of flag varieties.

6. Weight-monodromy conjecture

One application of the theory of perfectoid spaces is to a class of cases of the weight-monodromy conjecture. Let us briefly recall the statement, cf. [18].

Let X over $\mathbb{Q}_p$ (or a finite extension thereof) be a proper smooth variety. Fix a prime $\ell \neq p$. On the étale cohomology group $V = H^i(X_{\overline{\mathbb{Q}}_p}, \overline{\mathbb{Q}}_\ell)$, the absolute Galois group $G_{\mathbb{Q}_p}$ acts. Fix a Frobenius element $\mathrm{Frob} \in G_{\mathbb{Q}_p}$. From the Weil conjectures, [19], the Rapoport-Zink spectral sequence, [54], and de Jong's alterations, [42], the following is known about the structure of V:

(i) There is a direct sum decomposition $V = \bigoplus_{j=0}^{2i} V_j$, where all eigenvalues of Frob on V_j are Weil numbers of weight j.

(ii) There is a nilpotent operator $N : V \to V$ mapping $V_j \to V_{j-2}$, coming from the action of the pro-ℓ-inertia.

Conjecture 6.1 ([18]). *For any $j = 0, \ldots, i$, the map $N^j : V_{i+j} \to V_{i-j}$ is an isomorphism.*

This is somewhat reminiscent of the Lefschetz decomposition, and is sometimes said to be 'Mirror dual' to it. There is a similar result for projective smooth families of complex manifolds over a punctured complex disc, which is known to be true by work of Schmid, [56], and Steenbrink, [63].

Deligne proved the analogue for X over $\mathbb{F}_p((t))$ in [20].[7] Our result deduces the conjecture over $\mathbb{Q}_p$ in many cases by reduction to equal characteristic, via tilting.

Theorem 6.2 ([58]). *Let X be a geometrically connected proper smooth variety over a finite extension of $\mathbb{Q}_p$ which is a set-theoretic complete intersection in a projective smooth toric variety. Then the weight-monodromy conjecture holds true for X.*

Note that this result is new even for a smooth hypersurface in projective space. Let us note that strictly speaking, the author is not aware of any (geometrically connected projective smooth) X which does provably not satisfy this assumption. However, we can also not prove it in any reasonable generality.

7. Close local fields: Work of Hattori

Recall that the theory of perfectoid spaces developed as a generalization of the Fontaine-Wintenberger result which worked with infinite extensions of $\mathbb{Q}_p$. Hattori shows that one can,

[7] Actually, he assumed that X is already defined over a curve, but this assumption can be removed.

however, use this theory to prove generalizations of Deligne's results on close local fields. Let us state here one of his results.

For a complete discrete valuation field K with valuation v (normalized with image $\mathbb{Z} \cup \{\infty\}$) and residue characteristic p, the absolute ramification index e_K is defined as $e_K = v(p)$. In particular, $e_K = \infty$ if K is of characteristic p.

Theorem 7.1 ([35, Theorem 1.2 (ii)]). *Let K_1 and K_2 be two complete discrete valuation fields of residue characteristic p, such that the residue fields $k_1 \cong k_2$ are isomorphic. Let $j \leq \min(e_{K_1}, e_{K_2})$. Then there is an isomorphism*

$$G_{K_1}/G_{K_1}^{(j)} \cong G_{K_2}/G_{K_2}^{(j)} .$$

The main novelty is that the residue fields k_i are not assumed to be perfect. Thus, Hattori has to use the Abbes-Saito ramification filtration for complete discrete valuation fields with imperfect residue fields, [1]. This is defined in terms of geometrically connected components of certain rigid-analytic varieties. Hattori's approach is to use perfectoid spaces to compare these rigid-analytic varieties in different characteristics. For this, one has to check that connected components do not change when passing to the perfectoid world, i.e. extracting a lot of p-power roots; this uses the bound on the ramification degree and an explicit computation. Then the result follows from the homeomorphism $X \cong X^\flat$ of underlying topological spaces.

In particular, this shows that the theory of perfectoid spaces gives new information on the other approaches to changing the characteristic. We note that in the representation theory of local groups, there are Hecke algebra isomorphisms for not-too-ramified types of close local fields, mirroring the Galois story on the automorphic side, cf. [43]. It would be interesting to see if perfectoid spaces can shed new light on these Hecke algebra isomorphisms as well.

8. Rigid Motives: Work of Vezzani

Another way in which perfectoid spaces have been used to study phenomena of changing the characteristic is in relation to Ayoub's category of rigid motives, cf. [5]. Rigid motives are defined by formally repeating some constructions from $\mathbb{A}^1$-homotopy theory, working with the category of smooth rigid-analytic varieties, and replacing $\mathbb{A}^1$ by the closed unit ball. For any nonarchimedean field K and any ring Λ, one gets the resulting category of rigid motives $\mathrm{RigMot}(K, \Lambda)$ with coefficients in Λ.

The following theorem is due to Vezzani:

Theorem 8.1 ([65]). *Let K be a perfectoid field with tilt $K^\flat$. For any $\mathbb{Q}$-algebra Λ, the categories $\mathrm{RigMot}(K, \Lambda) \cong \mathrm{RigMot}(K^\flat, \Lambda)$ are canonically equivalent.*

This can be regarded as a version of the Fontaine-Wintenberger isomorphism for 'rigid motivic Galois groups'. Vezzani's strategy is to compare both categories to categories of 'perfectoid motives' which one gets from (suitable) perfectoid spaces. It is rather formal that these perfectoid motives are equivalent over K and $K^\flat$, and the task becomes to relate these to classical finite-type objects.

9. p-adic Hodge theory

The subject of p-adic Hodge theory can be regarded as a parallel to Deligne's formulation of complex Hodge theory as the interrelationship between the various cohomology theories associated with compact Kähler manifolds. Let us recall the most important results in the complex setting. Fix a compact Kähler manifold X. One has the singular cohomology $H^i(X, \mathbb{Z})$, the de Rham cohomology $H^i_{\mathrm{dR}}(X)$, and the Hodge cohomology groups $H^i(X, \Omega^j_X)$.

Theorem 9.1 (Poincaré lemma). *The inclusion $\mathbb{C} \to \Omega^\bullet_X$ is a quasi-isomorphism of sheaves of complexes. In particular,*

$$H^i(X, \mathbb{Z}) \otimes \mathbb{C} = H^i(X, \mathbb{C}) \cong H^i_{\mathrm{dR}}(X) \, .$$

Theorem 9.2 (Hodge). *The Hodge-to-de Rham spectral sequence*

$$E_1^{ij} = H^j(X, \Omega^i_X) \Rightarrow H^{i+j}_{\mathrm{dR}}(X)$$

degenerates at E_1.

Theorem 9.3 (Hodge). *There is a canonical Hodge decomposition*

$$H^i(X, \mathbb{Z}) \otimes \mathbb{C} = \bigoplus_{j=0}^{i} H^j(X, \Omega^{i-j}_X) \, .$$

Now let C be a complete and algebraically closed extension of $\mathbb{Q}_p$. For example, $C = \mathbb{C}_p$, the completion of $\overline{\mathbb{Q}}_p$. Note that C is perfectoid. Let X be a proper smooth rigid-analytic variety over C. This should be regarded as the analogue of a compact complex manifold, which is not necessarily Kähler. Prior to the author's work on the subject, all work concentrated on the case of algebraic X, but it is shown in [59] that this restriction is not necessary.

Again, one has de Rham and Hodge cohomology groups $H^i_{\mathrm{dR}}(X)$, $H^i(X, \Omega^j_X)$, defined in the same way. What replaces singular cohomology is étale cohomology $H^i_{\text{ét}}(X, \mathbb{Z}_p)$. The following result generalizes a fact well-known for algebraic varieties.

Theorem 9.4 ([59, Theorem 1.1], [60, Theorem 3.17]). *Let X be a proper rigid-analytic variety over C. Then $H^i_{\text{ét}}(X, \mathbb{Z}_p)$ is a finitely generated $\mathbb{Z}_p$-module, which vanishes for $i > 2 \dim X$.*

Properness is crucial here. In fact, already for a closed unit disc, the $\mathbb{F}_p$-cohomology is infinite-dimensional, due to the presence of Artin-Schreier covers. This is in stark contrast with the ℓ-adic case ($\ell \neq p$), where strong finiteness statements are known by work of Berkovich and Huber, [9, 40].

Before explaining the proof of the theorem, recall another result from [59].

Theorem 9.5 ([59, Theorem 1.2]). *Let U be a connected affinoid rigid-analytic variety over C. Then U is a $K(\pi, 1)$ for p-torsion coefficients. In other words, for every p-torsion local system $\mathbb{L}$ on U, the natural map*

$$H^i(X_{\text{ét}}, \mathbb{L}) \to H^i(\pi_1(X, \bar{x}), \mathbb{L}_{\bar{x}})$$

is a bijection, where $\bar{x} \in X(C)$ is a base point, and $\pi_1(X, \bar{x})$ is the profinite étale fundamental group.

There is Artin's theorem on good neighborhoods which states that a smooth algebraic variety in characteristic 0 is locally a $K(\pi, 1)$. It is interesting to note that no smallness or smoothness assumption is necessary for this result in the p-adic world. Let us briefly sketch its proof as this gives a good impression on how perfectoid spaces are used in applications to p-adic Hodge theory. Let $\tilde{U} \to U$ be 'the universal cover of U', which is the inverse limit of all finite étale covers. It is not hard to see that $\tilde{U}$ is an affinoid perfectoid space. Essentially, the existence of enough p-th roots is assured as taking p-th roots is finite étale in characteristic 0. By formal nonsense, it is enough to prove that $H^i(\tilde{U}_{\text{ét}}, \mathbb{F}_p) = 0$ for $i > 0$; we already know that $H^1(\tilde{U}_{\text{ét}}, \mathbb{F}_p) = 0$ as this parametrizes finite étale $\mathbb{F}_p$-torsors, of which there are no more. Thus, we need to prove that $H^i(\tilde{U}_{\text{ét}}, \mathbb{F}_p) = 0$ for $i > 1$.

Lemma 9.6. *Let Y be an affinoid perfectoid space. For $i > 1$, $H^i(Y_{\text{ét}}, \mathbb{F}_p) = 0$.*

Proof. By tilting, we may assume that Y is of characteristic p. Then we have the Artin-Schreier sequence $0 \to \mathbb{F}_p \to \mathcal{O}_Y \to \mathcal{O}_Y \to 0$, and the result follows from vanishing of coherent cohomology: $H^i(Y, \mathcal{O}_Y) = 0$ for $i > 0$. □

Thus, the general idea is to cover X locally by pro-étale maps from perfectoid spaces, and then use qualitative properties of perfectoid spaces, which are verified in characteristic p. For this purpose, one introduces the pro-étale site $X_{\text{proét}}$, in which X is locally perfectoid in a suitable sense.[8]

By resolution of singularities for rigid-analytic varieties, the proof of the finiteness theorem reduces to the proper smooth case; moreover, it is enough to handle the case of $\mathbb{F}_p$-coefficients. In that case, the argument is involved and makes heavy use of the full machinery of perfectoid spaces, cf. [59]. Roughly, it proceeds in two steps. First, one shows that $H^i_{\text{ét}}(X, \mathcal{O}_X^+/p)$ is almost finitely generated. This makes use of the Cartan-Serre technique of shrinking covers, and the almost vanishing of $H^i(Y_{\text{ét}}, \mathcal{O}_Y^+)$ on affinoid perfectoid spaces Y. Then one uses a variant of the Artin-Schreier sequence

$$0 \to \mathbb{F}_p \to \mathcal{O}_X^+/p \to \mathcal{O}_X^+/p \to 0$$

to deduce finiteness of $\mathbb{F}_p$-cohomology. In fact, one gets the following basic comparison result at the same time.

Theorem 9.7 ([59, Theorem 1.3], [60, Theorem 3.17]). *Let X be a proper rigid-analytic variety over C. Then the natural map*

$$H^i(X_{\text{ét}}, \mathbb{F}_p) \otimes \mathcal{O}_C/p \to H^i(X_{\text{ét}}, \mathcal{O}_X^+/p)$$

is an almost isomorphism, i.e. both the kernel and the cokernel are killed by the maximal ideal of $\mathcal{O}_C$.

This is a variant on a result of Faltings, [25, Theorem §3.8]. It forms the basic result which allows one to pass from étale cohomology to coherent cohomology (including here de Rham and Hodge cohomology). Note that the result implies the following remarkable behaviour of $M = R\Gamma(X_{\text{ét}}, \mathcal{O}_X^+)$. After inverting p, $M[p^{-1}] = R\Gamma(X_{\text{ét}}, \mathcal{O}_X) = R\Gamma(X, \mathcal{O}_X)$ is usual coherent cohomology. However, after (derived) modding out p,

$$M/p = R\Gamma(X_{\text{ét}}, \mathcal{O}_X^+/p) \cong_a R\Gamma(X_{\text{ét}}, \mathbb{F}_p) \otimes \mathcal{O}_C/p$$

[8] The idea of the pro-étale site has turned out to be quite powerful for foundational questions, even in the case of schemes. For new foundations for ℓ-adic cohomology of schemes, see [10].

is almost isomorphic to étale cohomology. In particular, $M[p^{-1}]$ lives only in degrees 0 through $\dim X$, while M itself has torsion going up until degree $2\dim X$. It also shows the full strength of the result that $H^i(Y_{\text{ét}}, \mathcal{O}_Y^+)$ is almost zero for $i > 0$, if Y is an affinoid perfectoid space: Certainly, nothing similar is true for a finite type space. It means that all the torsion in the cohomology of $\mathcal{O}_X^+$ gets killed after passing to perfectoid covers. This will be at the heart of the applications to torsion in the cohomology of locally symmetric spaces, cf. Section 15.

Let us now mention the analogues of the theorems in the complex world, stated earlier. For definiteness, we assume here that $X = X_0 \times_k C$ is the base-change of some X_0 defined over a completed discretely valued extension k of $\mathbb{Q}_p$ with perfect residue field. Moreover, we assume that X_0 is proper and smooth.

Theorem 9.8 ([59, Corollary 1.8]). *The G_k-representation $H^i_{\text{ét}}(X, \mathbb{Q}_p)$ is de Rham in the sense of Fontaine, and one has the comparison between étale and de Rham cohomology*

$$H^i_{\text{ét}}(X, \mathbb{Q}_p) \otimes_{\mathbb{Q}_p} B_{\text{dR}} \cong H^i_{\text{dR}}(X_0) \otimes_k B_{\text{dR}} .$$

In particular, $H^i_{\text{dR}}(X_0)$ is the filtered k-vector space associated with the de Rham G_k-representation $H^i_{\text{ét}}(X, \mathbb{Q}_p)$.

This is a known phenomenon in p-adic Hodge theory: To get the comparison theorems, one has to extend scalars to Fontaine's big period rings. Here, we use B_{dR}, which is a complete discrete valuation field with residue field C.

Theorem 9.9 ([59, Corollary 1.8]). *The Hodge-to-de Rham spectral sequence*

$$E_1^{ij} = H^j(X, \Omega_X^i) \Rightarrow H^{i+j}_{\text{dR}}(X)$$

degenerates at E_1.

Note that no Kähler assumption is necessary here. It is interesting to note that some non-Kähler complex manifolds have p-adic analogues, such as the Hopf surface: Divide $\mathbb{A}^2 \setminus \{(0,0)\}$ by the diagonal action of multiplication by q for some $q \in k$ with $|q| < 1$ to get a proper smooth rigid-analytic variety X. This has Hodge numbers $h^{01} = \dim H^0(X, \Omega_X^1) = 0$ while $h^{10} = \dim H^1(X, \mathcal{O}_X) = 1$, so Hodge symmetry fails. However, Hodge-to-de Rham degeneration holds true for the Hopf surface. Fortunately, Iwasawa manifolds for which the Hodge-to-de Rham degeneration fails, do not have p-adic analogues.

The next result does not need a Kähler assumption either:

Theorem 9.10 ([59, Corollary 1.8], [60, Theorem 3.20]). *There is a Hodge-Tate decomposition*

$$H^i_{\text{ét}}(X, \mathbb{Q}_p) \otimes_{\mathbb{Q}_p} C \cong \bigoplus_{j=0}^{i} H^{i-j}(X, \Omega_X^j)(-j) .$$

Here, $(-j)$ denotes a Tate twist. More generally, if X is only defined over C, there is a Hodge-Tate spectral sequence

$$E_2^{ij} = H^i(X, \Omega_X^j)(-j) \Rightarrow H^{i+j}_{\text{ét}}(X, \mathbb{Q}_p) \otimes_{\mathbb{Q}_p} C .$$

It degenerates at E_2 if X is algebraic or defined over k (and probably does in general), giving a Hodge-Tate filtration on $H^i_{\text{ét}}(X, \mathbb{Q}_p) \otimes_{\mathbb{Q}_p} C$ with associated graded pieces $H^{i-j}(X, \Omega_X^j)(-j)$.

Note the interesting differences between the Hodge-Tate spectral sequence and the Hodge-de Rham spectral sequence: It starts at E_2, and i and j are interchanged. Moreover, a Tate twist appears.

10. Relative φ-modules: Work of Kedlaya-Liu

At around the same time that the author wrote [58], Kedlaya-Liu, [47], [48], worked out closely related results[9] with the goal of constructing $\mathbb{Q}_p$-local systems on period domains, as were conjectured by Rapoport-Zink, [55]. Let us briefly recall the conjecture of Rapoport-Zink, in the case of the group GL_n.

Fix a perfect field k of characteristic p, and let V be a k-isocrystal, i.e. a $W(k)[p^{-1}]$-vector space V of finite dimension n equipped with a σ-linear isomorphism $\phi : V \to V$. Moreover, fix a 'filtration type', i.e. for each integer $i \in \mathbb{Z}$ a multiplicity $n_i \geq 0$ such that $n = \sum n_i$. The space of decreasing filtrations $\mathrm{Fil}^\bullet V \subset V$ for which $\mathrm{gr}^i V$ has dimension n_i forms naturally an algebraic variety $\mathcal{F}$ over $W(k)[p^{-1}]$; we consider $\mathcal{F}$ as an adic space over $W(k)[p^{-1}]$.

If $x \in \mathcal{F}(K)$ is a point defined over a finite extension K of $W(k)[p^{-1}]$, then, by a theorem of Colmez-Fontaine, [17], the triple $(V, \phi, \mathrm{Fil}^\bullet)$ comes from a crystalline representation $\mathbb{L}(x)$ of G_K if and only if it is weakly admissible. Weak admissibility is an analogue of a semistability condition, comparing Hodge and Newton slopes. There is a maximal open subspace $\mathcal{F}^{wa} \subset \mathcal{F}$ whose classical points are the weakly admissible points, cf. [55].

Conjecture 10.1. *For any smooth subspace $X \subset \mathcal{F}$ such that the universal filtration restricted to X satisfies Griffiths transversality, there exists a natural open subset $X^a \subset X^{wa} := X \cap \mathcal{F}^{wa}$ with the same classical points, and a $\mathbb{Q}_p$-local system $\mathbb{L}(X)$ on X^a, which gives the G_K-representation $\mathbb{L}(x)$ when passing to the fibre over any $x \in X^a(K)$.*

The original conjecture of Rapoport-Zink was more optimistic in that it conjectured the existence of $\mathbb{L}(X)$ for $X = \mathcal{F}$, and not only on subspaces where Griffiths transversality is satisfied. However, this does not fit with the p-adic Hodge theory formalism. Note that if the filtration is of 'minuscule type', meaning that $n_i \neq 0$ for at most two consecutive i, then Griffiths transversality is satisfied on all of $\mathcal{F}$. This assumption is satisfied in all cases investigated in [55], which are related to p-divisible groups.

Kedlaya announced a proof of this conjecture in [45]. Very roughly, the strategy of Kedlaya-Liu is to construct the local system locally in the pro-étale site and then glue. This reduces the problem to the perfectoid case. Moreover, now one has to construct a $\mathbb{Q}_p$-local system on the perfectoid space, or equivalently its tilt. But in characteristic p, $\mathbb{Q}_p$-local systems can be constructed from φ-modules by Artin-Schreier-Witt theory. Thus, first they build a φ-module over a relative Robba ring. Then they need to show that the locus where this φ-module is pure of slope 0 is open, and that locally on this locus, an integral structure exists. These theorems are proved in [47]; they generalize previous results of Kedlaya on slope filtrations and the existence of integral structures in the absolute setting, [44].

11. Universal covers of p-divisible groups

The following definition of a universal cover arose repeatedly in recent years, cf. e.g. [27], [30]. We identify a formal scheme S with the functor it represents on (discrete) rings, so e.g.

$$(\mathrm{Spf} \varprojlim A/I^n)(R) = \varinjlim \mathrm{Hom}(A/I^n, R) \ .$$

[9] In particular, they proved Corollary 4.14 independently.

By a commutative group G over S, we mean an fqpc sheaf of commutative groups on the category of discrete rings living over S. In other words, for any discrete ring R with an R-valued point $S(R)$, one has a commutative group $G(R)$, satisfying fpqc descent. We are particularly interested in the cases where G is an abelian variety or a p-divisible group.

Definition 11.1. Let S be a formal scheme over $\mathrm{Spf}\,\mathbb{Z}_p$, and let G/S be a commutative group. The universal cover $\tilde{G}$ of G is defined as $\tilde{G} = \varprojlim_{\times p} G$.

For example, if $G = \mathrm{Spf}\, R[[T_1, \ldots, T_d]]$ is a formal p-divisible group over a ring R, then $\tilde{G} \cong \mathrm{Spf}\, R[[T_1^{1/p^\infty}, \ldots, T_d^{1/p^\infty}]]$. In particular, if $R = \mathcal{O}_K$ is the ring of integers in a perfectoid field K, then the generic fibre $\tilde{G}_\eta$ of $\tilde{G}$ is a perfectoid space over K. If $G = \mathbb{G}_a$ is the additive group, then $\tilde{G} = 0$.

For any formal scheme S over $\mathrm{Spf}\,\mathbb{Z}_p$, we may consider the categories of universal covers of abelian varieties, resp. universal covers of p-divisible groups, over S, as full subcategories of the category of commutative groups over S.

Proposition 11.2. *Let $S' \subset S$ be a closed immersion of formal schemes defined by a topologically nilpotent ideal. Then the categories of universal covers of abelian varieties (resp. p-divisible groups) over S and S' are equivalent.*

Thus, the universal cover may be considered as a crystal on the infinitesimal site. In particular, let us fix an abelian variety or a p-divisible group G_0 over a perfect field k of characteristic p, of height h. It has a universal deformation space $S \cong \mathrm{Spf}\, W(k)[[T_1, \ldots, T_k]]$ (cf. [41]), and a universal deformation G/S. However, the universal cover $\tilde{G}$ is constant, equal to the evaluation of the crystal $\tilde{G}_0$ on the thickening $S \to \mathrm{Spec}\, k$.

Note that inside $\tilde{G}$ one has the Tate module $T_pG = \ker(\tilde{G} \to G) = \varprojlim_{\times p} G[p^n]$. If one fixes a C-valued point of the generic fibre of S, where C is an algebraically closed complete extension of $W(k)[p^{-1}]$, then $\Lambda = (T_pG)(\mathcal{O}_C) \cong \mathbb{Z}_p^h \subset \tilde{G}(\mathcal{O}_C)$ is a $\mathbb{Z}_p$-lattice. Informally, one gets back $G = \tilde{G}/T_pG$ by quotienting $\tilde{G}$ by this $\mathbb{Z}_p$-lattice. Here, $\tilde{G}$ is independent of the chosen point, but the $\mathbb{Z}_p$-lattice varies. This is reminiscent of the complex uniformization of abelian varieties: Their universal cover is constant, and different abelian varieties correspond to different $\mathbb{Z}$-lattices in the universal cover. Riemann's theorem gives a condition on when the quotient exists as an algebraic variety in terms of the existence of a polarization.

The following theorem is proved in joint work with Weinstein. We refer the reader to [62] for a more detailed discussion of this result.[10]

Theorem 11.3 ([62, Theorem D]). *Fix a p-divisible group G_0 over a perfect field k of height h and dimension d, as well as a complete and algebraically closed extension C of $W(k)[p^{-1}]$. Consider the category of lifts (G, ρ) of G_0 to $\mathcal{O}_C$ up to quasi-isogeny: Here, $G/\mathcal{O}_C$ is a p-divisible group, and $\rho : G_0 \times_k \mathcal{O}_C/p \to G \times_{\mathcal{O}_C} \mathcal{O}_C/p$ is a quasi-isogeny. Then the category of lifts (G, ρ) is equivalent to the category of $\mathbb{Z}_p$-lattices $\Lambda \cong \mathbb{Z}_p^h \subset \tilde{G}_0(\mathcal{O}_C)$ for which there exists a (necessarily unique) $h - d$-dimensional subspace $W \subset M(G_0)(\mathcal{O}_C)[p^{-1}] \cong C^h$ such that the image of Λ under the quasi-logarithm map*

$$\mathrm{qlog} : \tilde{G}_0(\mathcal{O}_C) \to M(G_0)(\mathcal{O}_C)[p^{-1}]$$

[10] One may deduce a similar result for abelian varieties by using Serre-Tate theory if one incorporates a polarization to guarantee algebraization.

lies in W, and

$$0 \to \Lambda[p^{-1}] \to \tilde{G}_0(\mathcal{O}_C) \to C^h/W \to 0$$

is exact.

This gives one analogue of Riemann's theorem on the classification of complex abelian varieties. The following theorem, again proved in joint work with Weinstein, [62], and closely related to the previous theorem, gives a different such analogue. For this, we use that any p-divisible group G over $\mathcal{O}_C$ has a Hodge-Tate filtration

$$0 \to (\operatorname{Lie} G) \otimes_{\mathcal{O}_C} C(1) \to T_p G \otimes_{\mathbb{Z}_p} C \to (\operatorname{Lie} G^*)^* \otimes_{\mathcal{O}_C} C \to 0\,,$$

which is an analogue of the Hodge-Tate filtration defined above for proper smooth varieties over C, cf. Theorem 9.10. This Hodge-Tate filtration for p-divisible groups was known previously, and is due to Faltings, [24], cf. also Fargues, [28].

Theorem 11.4 ([62, Theorem B]). *The category of p-divisible groups over $\mathcal{O}_C$ is equivalent to the category of pairs (Λ, W) where Λ is a finite free $\mathbb{Z}_p$-module, and $W \subset \Lambda \otimes_{\mathbb{Z}_p} C$ is a C-subvectorspace.*

The functor is given by $G \mapsto (T_p G, \operatorname{Lie} G \otimes_{\mathcal{O}_C} C(1))$. This is analogous to the classification of complex abelian varieties by their first singular homology, together with the Hodge filtration.

12. Lubin-Tate spaces: Work of Weinstein

Weinstein has observed that the Lubin-Tate tower at infinite level carries a natural structure as a perfectoid space. For this, fix an integer $n \geq 1$ and a p-divisible group G_0 of dimension 1 and height n. The Lubin-Tate tower at infinite level $\mathcal{M}_{G_0,\infty}$ parametrizes triples (G, ρ, α) where (G, ρ) is a deformation of G_0 up to quasi-isogeny as before, and $\alpha : \mathbb{Z}_p^n \to T_p G$ is an infinite level structure.

One may define a p-divisible group $\bigwedge G_0$ of G_0 of dimension 1 and height 1 by taking the highest exterior power of the Dieudonné module $M(G_0)$, and passing back to p-divisible groups. This uses crucially that G_0 is of dimension 1. One may construct an alternating map

$$\det : \tilde{G}_0 \otimes \ldots \otimes \tilde{G}_0 \to \widetilde{\bigwedge G_0}\,.$$

This follows from the work of Hedayatzadeh, [36], or from a result in Dieudonné theory in the joint work with Weinstein, [62]. Fix a perfectoid field K; then this gives a similar map on the generic fibre, base-changed to K:

$$\det : \tilde{G}_{0,K} \otimes \ldots \otimes \tilde{G}_{0,K} \to \widetilde{\bigwedge G_0}_K\,.$$

Inside $\widetilde{\bigwedge G_0}_K$, one has the rational Tate module $V_p(\bigwedge G_0) \subset \widetilde{\bigwedge G_0}_K$ and an exact sequence

$$0 \to V_p(\bigwedge G_0) \to \widetilde{\bigwedge G_0}_K \xrightarrow{\log} \mathbb{G}_{a,K} \to 0\,.$$

The following theorem is easy to deduce from Theorem 11.3, but was proved earlier directly by Weinstein.

Theorem 12.1 (Weinstein). *The following diagram is cartesian:*

$$\begin{array}{ccc} \mathcal{M}_{G_0,\infty} & \hookrightarrow & \left(\tilde{G}_{0,K}\right)^n \\ \downarrow & & \downarrow \det \\ V_p(\bigwedge G_0) \setminus \{0\} & \hookrightarrow & \widetilde{\bigwedge G_0}_K \end{array}$$

All intervening objects are perfectoid spaces over K, and the inclusions are locally closed (i.e., open subsets of Zariski closed subsets).

All objects in this diagram can be made completely explicit. Weinstein has used this to find explicit affinoid perfectoid subsets of $\mathcal{M}_{G_0,\infty}$ whose cohomology realizes the local Langlands correspondence for specific supercuspidal representations, cf. [12]. Recall that it is known (by the work of Harris-Taylor, [34]) that the cohomology of $\mathcal{M}_{G_0,\infty}$ realizes the local Langlands correspondence for all supercuspidal representations of $\mathrm{GL}_n(\mathbb{Q}_p)$. It is remarkable that while at any finite level, one cannot give an explicit description of the Lubin-Tate tower, it is possible to describe $\mathcal{M}_{G_0,\infty}$, together with all group actions, explicitly.

In [62], it is proved that more general Rapoport-Zink spaces become perfectoid at infinite level, and a description purely in terms of p-adic Hodge theory is given. This made it possible to prove the duality isomorphism for basic Rapoport-Zink spaces. In particular, one gets that Drinfeld and Lubin-Tate tower are isomorphic at infinite level as perfectoid spaces. This improves on earlier results of Faltings, [26], and Fargues, [28], who proved such isomorphisms, but had to struggle with formalizing them, as no category was known in which both infinite level spaces lived a priori. Their method is to work with suitable formal models; for this, new formal models have to be constructed first, which is at least technically challenging.

It was recently suggested by Rapoport-Viehmann, [53], that there should exist a theory of 'local Shimura varieties', which should relate to Rapoport-Zink spaces in the same way that general Shimura varieties relate to Shimura varieties of PEL type. The new perspective on Rapoport-Zink spaces mentioned above should make it possible to prove (parts of) their conjectures.

13. p-adic cohomology of the Lubin-Tate tower

The Lubin-Tate tower plays an important role in the Langlands program because its ℓ-adic cohomology for $\ell \neq p$ realizes the local Langlands correspondence, cf. [34]. In the emerging p-adic local Langlands program, which has taken a definitive form only for $\mathrm{GL}_2(\mathbb{Q}_p)$, cf. [13], one hopes for a similar realization of the p-adic local Langlands correspondence. However, the $\mathbb{F}_p$-cohomology of the Lubin-Tate tower is too infinite due to the presence of many Artin-Schreier covers. Still, a variant of Theorem 9.4 holds true in this context; for simplicity, we state only the version with $\mathbb{F}_p$-coefficients; a similar result holds true with $\mathbb{Z}_p$-coefficients.

Let F be a finite extension of $\mathbb{Q}_p$. Fix an admissible $\mathbb{F}_p$-representation π of $\mathrm{GL}_n(F)$. Using the Lubin-Tate tower at infinite level, which is a $\mathrm{GL}_n(F)$-torsor over $\mathbb{P}^{n-1}_{\breve{F}}$, where $\breve{F}$ denotes the completion of the maximal unramified extension of F, one gets an étale sheaf $\mathcal{F}_\pi$ on $\mathbb{P}^{n-1}_{\breve{F}}$. It is naturally $D^\times$-equivariant, and equipped with a Weil descent datum. Here, D is

the division algebra of invariant $1/n$ over F. The following theorem is work in progress of the author, and relies on the techniques of the proof of Theorem 9.4 along with the duality between Lubin-Tate and Drinfeld tower.

Theorem 13.1. *Let $C/\breve{F}$ be complete and algebraically closed. Then $H^i(\mathbb{P}^{n-1}_C, \mathcal{F}_\pi)$ is an admissible $D^\times$-representation, which vanishes for $i > 2(n-1)$, and is independent of C. The resulting functor from admissible $\mathrm{GL}_n(F)$-representations to admissible $D^\times \times G_F$-representations is compatible with some global correspondences.*

This makes it possible to pass from $\mathrm{GL}_n(F)$-representations to Galois representations in a purely local way. In the global setup, it proves that the $\mathrm{GL}_n(F)$-representation determines the local Galois group representation.

14. Shimura varieties

Fix a reductive group G over $\mathbb{Q}$ with a Shimura datum of Hodge type, giving rise to a Shimura variety S_K, $K \subset G(\mathbb{A}_f)$, over the reflex field E. There is a Hecke-equivariant compactification S^*_K, finite under the minimal compactification $S^*_K \to S^*_K$, and a flag variety $\mathcal{F}$ with G-action, such that the following are true.[11]

Theorem 14.1. *Fix a tame level $K^p \subset G(\mathbb{A}_f^p)$ and a map $E \to C$ to a complete and algebraically closed extension C of $\mathbb{Q}_p$. Let $(S^*_K)^{\mathrm{ad}}$ denote the adic space associated with $S^*_K \otimes_E C$. Then there is a perfectoid space $S^*_{K^p}$ over C such that*

$$S^*_{K^p} \sim \varprojlim_{K_p} (S^*_{K_pK^p})^{\mathrm{ad}} .$$

Moreover, there is a $G(\mathbb{Q}_p)$-equivariant Hodge-Tate period map

$$\pi_{\mathrm{HT}} : S^*_{K^p} \to \mathcal{F} .$$

The map π_{HT} is equivariant for the Hecke operators prime to p with respect to the trivial action on $\mathcal{F}$; in particular, π_{HT} contracts $G(\mathbb{A}_f^p)$-orbits. There is a cover of $\mathcal{F}$ by affinoid subsets $U \subset \mathcal{F}$ for which $\pi^{-1}_{\mathrm{HT}}(U) \subset S^*_{K^p}$ is an affinoid perfectoid subset.

The geometry of π_{HT} is very interesting. Consider the case of the modular curve. Here, $\mathcal{F} = \mathbb{P}^1$, and π_{HT} is a p-adic analogue of the embedding of the complex upper half-plane (which is a path-connected component of the inverse limit over all levels $\varprojlim_K S_K(\mathbb{C})$) into $\mathbb{P}^1(\mathbb{C})$. In both cases, the map is given by the Hodge filtration.

In the case of the modular curve, $S^*_{K^p} = S^*_{K^p}$ has a stratification into the ordinary and the supersingular locus, $S^{\mathrm{ord}}_{K^p}$ and $S^{\mathrm{ss}}_{K^p}$.[12] The flag variety is $\mathcal{F} = \mathbb{P}^1$. Then, under π_{HT}, all of $S^{\mathrm{ord}}_{K^p}$ maps into $\mathbb{P}^1(\mathbb{Q}_p)$, while the supersingular locus $S^{\mathrm{ss}}_{K^p}$ maps into Ω^2. Here, $\Omega^2 = \mathbb{P}^1 \setminus \mathbb{P}^1(\mathbb{Q}_p)$ is Drinfeld's upper half-plane, which is reminiscent of the complex upper and lower half-plane, which can be written as $\mathbb{P}^1 \setminus \mathbb{P}^1(\mathbb{R})$. It follows that π_{HT} contracts

[11] It should be possible to use the minimal compactification itself, and make $\mathcal{F}$ more explicit, but so far this has not been worked out.

[12] We regard some points of the adic space corresponding to rank-2-valuations as part of the ordinary locus which would usually be considered as part of the supersingular locus. We do so by replacing the ordinary part by its closure.

connected components of the ordinary locus to points, whereas it does something interesting on the supersingular locus.

On the ordinary locus, the map is given by the position of the canonical subgroup. On the supersingular locus, $S^{\mathrm{ss}}_{K^p}$ is a finite disjoint union of Lubin-Tate towers at infinite level (for $n = 2$); these are isomorphic to the Drinfeld tower at infinite level, which is a pro-finite étale cover of Ω^2. The composite is π_{HT}. In particular, the isomorphism between Lubin-Tate and Drinfeld tower is built into the geometry of π_{HT}.

Let us note another perspective on what the Hodge-Tate period map does. Namely, by Theorem 11.4, giving the Hodge filtration is equivalent to giving the p-divisible group. This means that the Hodge-Tate period map, on geometric points of the good reduction locus, is the map sending an abelian variety to its p-divisible group (equipped with all extra structure).

15. Torsion cohomology of locally symmetric spaces

As the final topic, we summarize the application of these ideas to the study of torsion in the cohomology of locally symmetric spaces.

Fix a reductive group G over $\mathbb{Q}$. For any (sufficiently small) compact open subgroup $K \subset G(\mathbb{A}_f)$, one has the locally symmetric space

$$Y_K = G(\mathbb{Q})\backslash (G(\mathbb{R})/K_\infty A^\circ_\infty \times G(\mathbb{A}_f)/K) \ ,$$

where $K_\infty \subset G(\mathbb{R})$ is a maximal compact subgroup, and $A_\infty \subset G(\mathbb{R})$ are the $\mathbb{R}$-valued points of the maximal $\mathbb{Q}$-split central torus, with identity component A°_∞. Fixing a tame level $K^p \subset G(\mathbb{A}_f^p)$, one defines the completed cohomology groups

$$\tilde{H}^i(K^p) = \varprojlim_n \varinjlim_{K_p} H^i(Y_{K_pK^p}, \mathbb{Z}/p^n\mathbb{Z}) \ , \tilde{H}^i_c(K^p) = \varprojlim_n \varinjlim_{K_p} H^i_c(Y_{K_pK^p}, \mathbb{Z}/p^n\mathbb{Z}) \ .$$

Also recall the cohomological degree q_0, which is 'the first interesting cohomological degree' (namely, the first one to which tempered automorphic representations of G contribute). The following conjecture was proposed by Calegari and Emerton, [14].

Conjecture 15.1. *The groups $\tilde{H}^i(K^p)$, $\tilde{H}^i_c(K^p)$ vanish for $i > q_0$.*

Concretely, this means that all cohomology classes in higher degree become infinitely p-divisible as one goes up along all levels at p. If G is a torus, the conjecture is equivalent to Leopoldt's conjecture. On the other hand, we proved the following theorem.

Theorem 15.2 ([61, Theorem I.7]). *Assume that G gives rise to a Shimura variety, so that q_0 is the (complex) dimension of the associated Shimura variety. Then Conjecture 15.1 holds true for compactly supported cohomology.*

If one establishes that also toroidal compactifications become perfectoid at infinite level, then one gets the same result for usual cohomology. Unfortunately, for all tori which give rise to Shimura varieties, the Leopoldt conjecture is trivially satisfied, as the group of units is finite.

The key to the proof is to translate everything into the setting of Shimura varieties at infinite level as perfectoid spaces. In that case, one can use the basic comparison theorem to

pass to the cohomology of $\mathcal{O}^+/p$. But at infinite level, one has almost vanishing of higher cohomology of $\mathcal{O}^+/p$ on affinoids as the space is perfectoid. This shows vanishing above the middle dimension, which is exactly the desired statement.

In fact, the same argument proves the following theorem over $\mathbb{C}$, which the author does not know how to prove directly.

Theorem 15.3. *Let $X \subset \mathbb{P}^n_{\mathbb{C}}$ be a closed subvariety of dimension d. For any $m \geq 0$, let $X_m \subset \mathbb{P}^n_{\mathbb{C}}$ be the pullback of X under the map $\mathbb{P}^n_{\mathbb{C}} \to \mathbb{P}^n_{\mathbb{C}}$ sending $(x_0 : \ldots : x_n)$ to $(x_0^{p^m} : \ldots : x_n^{p^m})$. Then, for any $i > d$,*

$$\varinjlim_m H^i(X_m, \mathbb{F}_p) = 0 \, .$$

For classes in the image of cup product with $c_1(\mathcal{O}(1))$, this follows from the fact that $c_1(\mathcal{O}(1))$ becomes infinitely p-divisible. By hard Lefschetz, this accounts for everything rationally, but it does not say anything about possible p-torsion in the cohomology.

16. Galois representations

It was conjectured since the 1970's by Grunewald that torsion in the cohomology of locally symmetric spaces gives rise to Galois representations. This conjecture was made precise by Ash, [2], and is a 'mod p analogue' of (one direction of) the global Langlands conjectures. Since then, it was numerically verified in many cases: what happens is that a Hecke eigenvalue system matches Frobenius eigenvalues of a Galois representations for the first few hundred primes. However, even in these examples, one could not prove that this happens for *all* primes.

Theorem 16.1 ([61, Theorem I.3]). *Let G be the restriction of scalars of GL_n from a totally real or CM field F. Fix any compact open subgroup $K \subset G(\mathbb{A}_f)$. Then, for any system of Hecke eigenvalues ψ appearing in $H^i(Y_K, \overline{\mathbb{F}}_p)$, there exists a (unique) continuous semisimple Galois representation*

$$\rho_\psi : G_F \to \mathrm{GL}_n(\overline{\mathbb{F}}_p)$$

such that for all but an explicit finite set of 'ramified' places v of F, the characteristic polynomial of $\rho_\psi(\mathrm{Frob}_v)$ is described by the Hecke eigenvalues.

Moreover, there is a version of this theorem for $\mathbb{Z}/p^n\mathbb{Z}$-cohomology, which in the inverse limit over n gives results for classical automorphic representations. The following result was proved earlier by Harris-Lan-Taylor-Thorne, [33], by a different method.

Theorem 16.2 ([61, Theorem I.4]). *Let π be a regular algebraic cuspidal automorphic representation of $\mathrm{GL}_n(\mathbb{A}_F)$, where F is totally real or CM. Fix an isomorphism $\mathbb{C} \cong \overline{\mathbb{Q}}_p$. Then there exists a unique continuous semisimple Galois representation*

$$\rho_{\pi,p} : G_F \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$$

such that for all but an explicit finite set of 'ramified' places v of F, the characteristic polynomial of $\rho_{\pi,p}(\mathrm{Frob}_v)$ is described by the Satake parameters.

It should be noted that in general, the cohomology of the spaces Y_K has a lot of torsion. The simplest example is the case of GL_2 over an imaginary-quadratic field in which the relevant Y_K are hyperbolic 3-manifolds. In that case, computations, as well as theoretical results, show a huge amount of torsion, cf. e.g. [7]. Therefore, the thrust of the above theorem lies in the p-torsion part of the cohomology. Moreover, recent work of Calegari-Geraghty, [15], explains how one may use sufficiently fine information about existence of Galois representations for torsion classes to prove automorphy lifting theorems for GL_n over F. Together with the strong potential automorphy machinery as in the work of Barnet-Lamb–Gee–Geraghty–Taylor, [6], this gives some hope that one can establish some potential converse results to Theorem 16.2.

Let us sketch the proof of Theorem 16.1 in the case $F = \mathbb{Q}$. Consider the Siegel moduli space S_K, $K \subset \mathrm{GSp}_{2n}(\mathbb{A}_f)$, of principally polarized abelian varieties of dimension n. From the Borel-Serre compactification, [11], it follows that the cohomology of the locally symmetric space for GL_n contributes to the cohomology of the Siegel moduli space. Note that the Borel-Serre compactification is a compactification as a real manifold with corners; this makes it possible that a purely real manifold appears in the boundary of the algebraic variety S_K. Thus, the task becomes to understand torsion in the cohomology of S_K.

Theorem 16.3 ([61, Theorem I.5]). *Let S_K, $K \subset G(\mathbb{A}_f)$, be any Shimura variety of Hodge type. Then, for any system of Hecke eigenvalues ψ appearing in $H^i_c(S_{K,\mathbb{C}}, \overline{\mathbb{F}}_p)$, there exists a cuspidal eigenform f (possibly of larger level at p, and undetermined weight) such that the Hecke eigenvalues of f are congruent to ψ modulo p.*

This produces congruences between torsion classes and classical cusp forms in large generality. Note that the classes in which we are interested start life as classes coming from the boundary; still, the theorem produces congruences to cusp forms. In particular, for non-torsion classes, it is interesting as it produces congruences between Eisenstein series and cusp forms. However, in the complementary case where S_K is proper, the theorem is also interesting as it controls all possible torsion classes. For example, it proves the existence of Galois representations for all torsion classes in $U(1, n-1)$-Shimura varieties, which is required in recent work of Emerton and Gee, [22]. The point is that one knows how to attach Galois representations to cusp forms in great generality, through the work on automorphic forms on classical groups by Arthur [3] (cf. also [51] for unitary groups) and the work of Clozel, Kottwitz and Harris-Taylor among others on the cohomology of Shimura varieties, [16, 34, 50].

To prove the theorem, one starts by using the basic comparison theorem

$$H^i_{c,\text{ét}}(S_{K,C}, \mathbb{F}_p) \otimes \mathcal{O}_C/p \cong_a H^i_{\text{ét}}(S^*_{K,C}, I^+/p)\ ,$$

where $I^+ \subset \mathcal{O}^+$ is the ideal sheaf of functions vanishing at the boundary. This variant of Theorem 9.7 is proved in [60, Theorem 3.13]. This provides a first bridge to the sheaf of cusp forms I^+, but one still has to compute cohomology on the étale site. Next, one passes to infinite level at p, and reduces to controlling $H^i_{\text{ét}}(S^*_{\overline{K}^p}, I^+/p)$. Here, $S^*_{\overline{K}^p}$ is perfectoid, so we know that $H^i_{\text{ét}}(U, I^+/p)$ is almost zero for $i > 0$ and affinoid perfectoid subsets $U \subset S^*_{\overline{K}^p}$; this is a slight variant on Theorem 4.12. This means that $H^i_{\text{ét}}(S^*_{\overline{K}^p}, I^+/p)$ can (almost) be computed by a Cech complex whose terms are the sections of I^+/p on affinoid subsets. The remaining task is to approximate these forms on U by globally defined forms (of finite level), without messing up the Hecke eigenvalues. Usually, the strategy is to multiply by a multiple

of the Hasse invariant. This kills all poles away from the ordinary locus, and works if U is the ordinary locus. However, in our case we need to do the same for a covering of all of $S^*_{\overline{K}^p}$.

The crucial property of the Hasse invariant is that it commutes with all Hecke operators prime to p. In our setup, we can use the following construction: As

$$\pi_{\mathrm{HT}} : S^*_{\overline{K}^p} \to \mathcal{F}$$

is equivariant with respect to the trivial action of the Hecke operators prime to p on $\mathcal{F}$, any function that gets pulled back from $\mathcal{F}$ will commute with all Hecke operators prime to p. The same stays true for sections of automorphic vector bundles; automorphic vector bundles come via pullback from $\mathcal{F}$.[13] In this way, one gets enough 'fake-Hasse invariants' to proceed, and prove the result.

Acknowledgements. This work was done while the author was a Clay Research Fellow.

References

[1] Abbes, A. and Saito, T., *Ramification of local fields with imperfect residue fields*, Amer. J. Math. **124** (2002), no. 5, 879–920.

[2] Ash, A., *Galois representations attached to mod p cohomology of* $\mathrm{GL}(n, \mathbb{Z})$, Duke Math. J. **65** (1992), no. 2, 235–255.

[3] Arthur, J., *The endoscopic classification of representations. Orthogonal and symplectic groups*, American Mathematical Society Colloquium Publications **61**. American Mathematical Society, Providence, RI, 2013.

[4] Ax, J. and Kochen, S., *Diophantine problems over local fields, I*, Amer. J. Math. **87** (1965), 605–630.

[5] Ayoub, J., *Motifs des variétés analytiques rigides*. http://user.math.uzh.ch/ayoub/PDF-Files/MotVarRig.pdf

[6] Barnet-Lamb, T., Gee, T., Geraghty, D., and Taylor, R., *Potential automorphy and change of weight*, to appear in Annals of Math.

[7] Bergeron, N. and Venkatesh, A., *The asymptotic growth of torsion homology for arithmetic groups*, J. Inst. Math. Jussieu **12** (2013), no. 2, 391–447.

[8] Berkovich, V., *Spectral theory and analytic geometry over non-Archimedean fields*, Mathematical Surveys and Monographs **33**. American Mathematical Society, Providence, RI, 1990.

[9] ———, *Étale cohomology for non-Archimedean analytic spaces*, Publ. Math. Inst. Hautes Études Sci. **78** (1993), 5–161.

[10] Bhatt, B. and Scholze, P., *The pro-étale topology for schemes*, arXiv:1309.1198, to appear in Proceedings of the Conference in honour of Gérard Laumon.

[13] This implies, remarkably, that automorphic vector bundles extend to the minimal compactification at infinite level; they do not extend to the minimal compactification at any finite level.

[11] Borel, A. and Serre, J.-P., *Corners and arithmetic groups. Avec un appendice: Arrondissement des variétés à coins*, par A. Douady et L. Hérault, Comment. Math. Helv. **48** (1973), 436–491.

[12] Boyarchenko, M. and Weinstein, J., *Maximal varieties and the local Langlands correspondence for* $GL(n)$, arXiv:1109.3522.

[13] Breuil, C., *The emerging* p*-adic Langlands programme*, Proceedings of the International Congress of Mathematicians. Volume II, 203–230, Hindustan Book Agency, New Delhi, 2010.

[14] Calegari, F., *Emerton, M., Completed cohomology–a survey*, Non-abelian fundamental groups and Iwasawa theory, 239–257, London Math. Soc. Lecture Note Series 393, Cambridge Univ. Press, Cambridge, 2012.

[15] Calegari, F. and Geraghty, D., *Modularity Lifting Theorems beyond the Taylor-Wiles Method, II*. arXiv:1209.6293.

[16] Clozel, L., *Représentations galoisiennes associées aux représentations automorphes autoduales de* $\mathrm{GL}(n)$, Publ. Math. Inst. Hautes Études Sci. **73** (1991), 97–145.

[17] Colmez, P., *Fontaine, J.-M., Construction des représentations* p*-adiques semi-stables*, Invent. Math. **140** (2000), no. 1, 1–43.

[18] Deligne, P., *Théorie de Hodge. I*, Actes du Congrès International des Mathématiciens (Nice, 1970), Tome 1, pp. 425–430. Gauthier-Villars, Paris, 1971.

[19] ———, *La conjecture de Weil. I*, Publ. Math. Inst. Hautes Études Sci. **43** (1974), 273–307.

[20] ———, *La conjecture de Weil. II*, Publ. Math. Inst. Hautes Études Sci. **52** (1980), 137–252.

[21] ———, *Les corps locaux de caractéristique* p*, limites de corps locaux de caractéristique* 0, Representations of reductive groups over a local field, 119–157, Travaux en Cours, Hermann, Paris, 1984.

[22] Emerton, M. and Gee, T., p*-adic Hodge-theoretic properties of étale cohomology with mod* p *coefficients, and the cohomology of Shimura varieties*, arXiv:1203.4963.

[23] Faltings, G., p*-adic Hodge theory*, J. Amer. Math. Soc. **1** (1988), no. 1, 255–299.

[24] ———, *Integral crystalline cohomology over very ramified valuation rings*, J. Amer. Math. Soc. **12** (1999), no. 1, 117–144.

[25] ———, *Almost étale extensions*, Cohomologies p-adiques et applications arithmétiques, II. Astérisque No. 279 (2002), 185–270.

[26] ———, A relation between two moduli spaces studied by V. G. Drinfeld. *Algebraic number theory and algebraic geometry*, 115–129, Contemp. Math., 300, Amer. Math. Soc., Providence, RI, 2002.

[27] ———, *Coverings of p-adic period domains*, J. Reine Angew. Math. **643** (2010), 111–139.

[28] Fargues, L., *L'isomorphisme entre les tours de Lubin-Tate et de Drinfeld et applications cohomologiques*, L'isomorphisme entre les tours de Lubin-Tate et de Drinfeld, 1–325, Progr. Math., 262, Birkhäuser, Basel, 2008.

[29] ———, *Quelques résultats et conjectures concernant la courbe*, to appear in Proceedings of the Conference in honour of Gérard Laumon.

[30] Fargues, L. and Fontaine, J.-M., *Vector bundles and p-adic Galois representations*, Fifth International Congress of Chinese Mathematicians. Part 1, 2, 77–113, AMS/IP Stud. Adv. Math., 51, pt. 1, 2, Amer. Math. Soc., Providence, RI, 2012.

[31] ———, *Courbes et fibrés vectoriels en théorie de Hodge p-adique*, http://www.math.jussieu.fr/~fargues/Courbe.pdf.

[32] Fontaine, J.-M. and Wintenberger, J.-P., *Extensions algébrique et corps des normes des extensions APF des corps locaux*, C. R. Acad. Sci. Paris Sér. A-B **288** (1979), no. 8, A441–A444.

[33] Harris, M., Lan, K.-W., Taylor, R., and Thorne, J., *On the Rigid Cohomology of Certain Shimura Varieties*, http://www.math.ias.edu/~rtaylor/rigcoh.pdf.

[34] Harris, M. and Taylor, R., *The geometry and cohomology of some simple Shimura varieties. With an appendix by Vladimir G. Berkovich*, Annals of Mathematics Studies **151**. Princeton University Press, Princeton, NJ, 2001.

[35] Hattori, S., *Ramification theory and perfectoid spaces*, arXiv:1304.5895, to appear in Compositio Math.

[36] Hedayatzadeh, H., *Exterior powers of π-divisible modules over fields*, J. Number Theory **138** (2014), 119–174.

[37] Henniart, G., *La conjecture de Langlands locale numérique pour $GL(n)$*, Ann. Sci. École Norm. Sup. (4) **21** (1988), 497–544.

[38] Huber, R., *Continuous valuations*, Math. Z. **212** (1993), no. 3, 455-477.

[39] ———, *A generalization of formal schemes and rigid analytic varieties*, Math. Z. **217** (1994), no. 4, 513–551.

[40] ———, *Étale cohomology of rigid analytic varieties and adic spaces*, Aspects of Mathematics **E30**, Friedr. Vieweg & Sohn, Braunschweig, 1996.

[41] Illusie, L., *Déformations de groupes de Barsotti-Tate (d'après A. Grothendieck)*, Seminar on arithmetic bundles: the Mordell conjecture (Paris, 1983/84), Astérisque **127** (1985), 151–198.

[42] de Jong, A. J., *Smoothness, semi-stability and alterations*, Publ. Math. Inst. Hautes Études Sci. **83** (1996), 51–93.

[43] Kazhdan, D., *Representations of groups over close local fields*, J. Analyse Math. **47** (1986), 175–179.

[44] Kedlaya, K. S., *A p-adic local monodromy theorem*, Ann. of Math. (2) **160** (2004), no. 1, 93–184.

[45] ———, *Relative p-adic Hodge theory and Rapoport-Zink period domains*, Proceedings of the International Congress of Mathematicians. Volume II, 258–279, Hindustan Book Agency, New Delhi, 2010.

[46] ———, *Nonarchimedean geometry of Witt vectors*, Nagoya Math. J. **209** (2013), 111–165.

[47] Kedlaya, K. S. and Liu, R., *Relative p-adic Hodge theory, I: Foundations*, arXiv:1301.0792.

[48] ———, *Relative p-adic Hodge theory, II: (φ, Γ)-modules*, arXiv:1301.0795.

[49] Krasner, M., *Quelques méthodes nouvelles dans la théorie des corps valués complets*, Algèbre et Théorie des Nombres, 29–39. Colloques Internationaux du Centre National de la Recherche Scientifique, no. 24, Centre National de la Recherche Scientifique, Paris, 1950.

[50] Kottwitz, R., *On the λ-adic representations associated to some simple Shimura varieties*, Invent. Math. **108** (1992), no. 3, 653–665.

[51] Mok, C. P., *Endoscopic classification of representations of quasi-split unitary groups*, arXiv:1206.0882.

[52] Ngô, B. C., *Le lemme fondamental pour les algèbres de Lie*, Publ. Math. Inst. Hautes Études Sci. **111** (2010), 1–169.

[53] Rapoport, M. and Viehmann, E., *Towards a theory of local Shimura varieties*, arXiv:1401.2849.

[54] Rapoport, M. and Zink, T., *Über die lokale Zetafunktion von Shimuravarietäten. Monodromiefiltration und verschwindende Zyklen in ungleicher Charakteristik*, Invent. Math. **68** (1982), no. 1, 21–101.

[55] ———, *Period spaces for p-divisible groups*, Annals of Mathematics Studies **141**. Princeton University Press, Princeton, NJ, 1996.

[56] Schmid, W., *Variation of Hodge structure: the singularities of the period mapping*, Invent. Math. **22** (1973), 211–319.

[57] Scholze, P., *The local Langlands correspondence for GL_n over p-adic fields*, Invent. Math. **192** (2013), no. 3, 663–715.

[58] ———, *Perfectoid spaces*, Publ. Math. Inst. Hautes Études Sci. **116** (2012), 245–313.

[59] ———, *p-adic Hodge theory for rigid-analytic varieties*, Forum Math. Pi **1** (2013), e1, 77p.

[60] ———, *Perfectoid Spaces: A survey*, Current Developments in Mathematics 2012, International Press, 2013.

[61] ———, *On torsion in the cohomology of locally symmetric varieties*, arXiv:1306.2070.

[62] Scholze, P. and Weinstein, J., *Moduli of p-divisible groups*, Cambridge Journal of Mathematics **1** (2013), no. 2, 145–237.

[63] Steenbrink, J., *Limits of Hodge structures*, Invent. Math. **31** (1975/76), no. 3, 229–257.

[64] Tate, J., *Rigid-analytic spaces*, Invent. Math. **12** (1971), 257–289.

[65] Vezzani, A., *A motivic version of the theorem of Fontaine and Wintenberger*, preprint.

[66] Waldspurger, J.-L., *Endoscopie et changement de caractéristique*, J. Inst. Math. Jussieu **5** (2006), no. 3, 423–525.

Mathematisches Institut der Universität Bonn, Endenicher Allee 60, 53115 Bonn, Germany
E-mail: scholze@math.uni-bonn.de

Stabilisation de la partie géométrique de la formule des traces tordue

J.-L. Waldspurger *

Résumé. We explain what is twisted endoscopy. We give the formulation of the geometric part of the twisted trace formula, following the works of Clozel-Labesse-Langlands and Arthur. We explain his stabilization, which is a work in progress, joint with Moeglin.

Mathematics Subject Classification (2010). AMS 11F72, 20G35, 22E30, 22E35.

Keywords. twisted endoscopy, twisted trace formula.

1. Introduction

Langlands a posé des conjectures qui classifient les représentations automorphes d'un groupe réductif connexe défini sur un corps de nombres. La forme fine de ces conjectures requiert l'usage de la théorie de l'endoscopie, elle-aussi imaginée par Langlands. En retour, cette théorie permet d'établir dans certains cas très particuliers l'existence prédite par les conjectures de correspondances entre représentations automorphes sur différents groupes. La méthode passe par la "stabilisation de la formule des traces d'Arthur-Selberg", cette formule étant l'un des outils les plus puissants de la théorie des formes automorphes. Cette stabilisation a été établie par Arthur [1–3]. Depuis les travaux de Langlands puis d'Arthur et Clozel sur le changement de base ([9, 16]), on sait qu'il est utile d'étendre la formule des traces comme la théorie de l'endoscopie à une situation "tordue", c'est-à-dire où le groupe est muni d'un automorphisme (il revient plus ou moins au même de considérer un groupe non connexe). Arthur a tiré des applications spectaculaires de cette théorie appliquée à un groupe $GL(n)$ tordu par un automorphisme extérieur non trivial, cf. [4]. Dans le cadre tordu, la formule des traces a été établie par Clozel, Labesse et Langlands et développée par Arthur. Kottwitz, Labesse et Shelstad ont élaboré après Langlands la théorie de l'endoscopie tordue. Dans un travail en voie d'achèvement, en collaboration avec Moeglin, nous stabilisons la formule des traces tordue, en suivant la méthode d'Arthur. Dans la section 2, je tente d'expliquer ce qu'est l'endoscopie tordue sur un corps de base local et j'énonce les résultats de stabilisation locaux concernant les intégrales orbitales. Dans la section 3, le corps de base devient un corps de nombres. Je décris la partie géométrique de la formule des traces tordue et j'explique sa stabilisation. Comme je l'ai dit, ce travail n'est pas encore complètement achevé et reste encore soumis à une hypothèse, à savoir la validité du lemme fondamental pour toutes les fonctions de l'algèbre de Hecke. Ngô Bao Chau a démontré ce lemme pour la fonction unité de cette algèbre [17]. Sa généralisation ne fait pas de doute et est certainement incomparablement plus facile que le théorème de Ngô Bao Chau.

* Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. La théorie locale

Cette section s'appuie sur les articles [11, 13, 14, 18]. On a donné une présentation plus personnelle de la théorie dans [19].

2.1. Espaces tordus. Soient F un corps local de caractéristique nulle et $\bar{F}$ une clôture algébrique de F. On pose $\Gamma_F = Gal(\bar{F}/F)$ et on note W_F le groupe de Weil de F. Soient G un groupe réductif connexe défini sur F et $\tilde{G}$ un espace tordu sous G. Cela signifie que $\tilde{G}$ est une variété algébrique sur F muni de deux actions de G

$$\begin{array}{ccc} G\times\tilde{G}\times G & \to & \tilde{G} \\ (g,\gamma,g') & \mapsto & g\gamma g' \end{array}$$

définies sur F, de sorte que, pour tout $\gamma\in\tilde{G}$, les applications $g\mapsto g\gamma$ et $g'\mapsto\gamma g'$ soient bijectives. Pour $\gamma\in\tilde{G}$, on note ad_γ l'automorphisme de G tel que $\gamma g' = ad_\gamma(g')\gamma$ pour tout $g'\in G$. Notons θ la restriction de ad_γ au centre $Z(G)$ de G. Elle ne dépend pas de l'élément $\gamma\in\tilde{G}$. On impose les deux conditions :

$\tilde{G}(F)\neq\emptyset$ et l'automorphisme θ de $Z(G)$ est d'ordre fini.

Outre les données G et $\tilde{G}$, on fixe un caractère ω de $G(F)$, c'est-à-dire un homomorphisme continu $\omega : G(F)\to\mathbb{C}^\times$.

Pour $\gamma\in\tilde{G}$, on note $Z_G(\gamma)$ l'ensemble de points fixes de ad_γ et G_γ la composante neutre de ce groupe. Une paire de Borel de G est un couple (B,T), où B est un sous-groupe de Borel de G et T est un sous-tore maximal de B (on ne suppose pas B et T définis sur F). Un élément $\gamma\in\tilde{G}$ est dit semi-simple s'il existe une paire de Borel (B,T) de G qui est conservée par ad_γ. Tout élément $\gamma\in\tilde{G}$ s'écrit de façon unique comme produit $\gamma = u\gamma_{ss}$ où γ_{ss} est semi-simple et u est un élément unipotent de $G_{\gamma_{ss}}$. On dit que γ est fortement régulier s'il est semi-simple et si $Z_G(\gamma)$ est commutatif.

On dit que G est quasi-déployé s'il existe une paire de Borel (B,T) définie sur F. On dit que $\tilde{G}$ est à torsion intérieure si ad_γ est un automorphisme intérieur de G pour tout $\gamma\in\tilde{G}$ (ou pour un $\gamma\in\tilde{G}$, c'est équivalent). Un tel espace tordu à torsion intérieure est isomorphe à G sur $\bar{F}$ mais pas forcément sur F. Par exemple, on peut avoir $G = SL(n)$ et $\tilde{G} = \{g\in GL(n); det(g) = d\}$, où d est un élément fixé de $F^\times$. Dans ce cas, $\tilde{G}$ est isomorphe à G sur F si et seulement si d est la puissance n-ième d'un élément de $F^\times$.

2.2. Intégrales orbitales. On note $C_c^\infty(\tilde{G}(F))$ l'espace des fonctions $f:\tilde{G}(F)\to\mathbb{C}$ dont le support est compact et qui sont lisses, c'est-à-dire localement constantes si F est non-archimédien, C^∞ si F est archimédien. Pour tout groupe réductif H défini sur F, on note par la lettre gothique $\mathfrak{h}$ son algèbre de Lie. Soit $\gamma\in\tilde{G}(F)$. Notons γ_{ss} sa partie semi-simple, posons

$$D^{\tilde{G}}(\gamma) = |det(1-ad_{\gamma_{ss}})_{|\mathfrak{g}/\mathfrak{g}_{\gamma_{ss}}}|_F,$$

où $|.|_F$ est la valeur absolue usuelle de F. Fixons des mesures de Haar sur $G(F)$ et $G_\gamma(F)$. Soit $f\in C_c^\infty(\tilde{G}(F))$. Si ω n'est pas trivial sur $G_\gamma(F)$, on pose $I^{\tilde{G}}(\gamma,\omega,f) = 0$. Si ω est trivial sur $G_\gamma(F)$, on pose

$$I^{\tilde{G}}(\gamma,\omega,f) = D^{\tilde{G}}(\gamma)^{1/2}\int_{G_\gamma(F)\backslash G(F)}\omega(g)f(g^{-1}\gamma g)\,dg.$$

On note $I(\tilde{G}(F),\omega)$ le quotient de $C_c^\infty(\tilde{G}(F))$ par le sous-espace des fonctions f telles que $I^{\tilde{G}}(\gamma,\omega,f)=0$ pour tout $\gamma\in\tilde{G}(F)$.

Supposons G quasi-déployé, $\tilde{G}$ à torsion intérieure et $\omega=1$. On supprime ω des notations. Soit γ un élément semi-simple fortement régulier de $\tilde{G}(F)$. La classe de conjugaison stable de γ est l'ensemble des $\gamma'\in\tilde{G}(F)$ tels qu'il existe $g\in G(\bar{F})$ de sorte que $\gamma'=g^{-1}\gamma g$. En général, cette classe est plus grosse que la classe de conjugaison de γ par $G(F)$: l'exemple simple $G=\tilde{G}=SL(2)$ permet de s'en convaincre. Fixons des mesures de Haar sur $G(F)$ et $G_\gamma(F)$. Pour γ' stablement conjugué à γ, les tores G_γ et $G_{\gamma'}$ sont isomorphes et la mesure sur $G_\gamma(F)$ en détermine une sur $G_{\gamma'}(F)$. Pour $f\in C_c^\infty(\tilde{G}(F))$, on pose

$$S^{\tilde{G}}(\gamma,f)=\sum_{\gamma'}I^{\tilde{G}}(\gamma',f),$$

où γ' parcourt un ensemble de représentants des classes de conjugaison par $G(F)$ dans la classe de conjugaison stable de γ. On note $SI(\tilde{G}(F))$ le quotient de $C_c^\infty(\tilde{G}(F))$ par le sous-espace des f telles que $S^{\tilde{G}}(\gamma,f)=0$ pour tout γ comme ci-dessus. C'est un quotient de $I(\tilde{G}(F))$.

Revenons au cas général. L'analyse harmonique ω-équivariante est l'étude de l'espace $I(\tilde{G}(F),\omega)$. Les travaux de ces dernières décennies sur la formule des traces ont mis en évidence l'importance plus fondamentale des espaces $SI(\tilde{G}(F))$ du cas particulier ci-dessus. La théorie de l'endoscopie tordue affirme en substance qu'étudier l'espace $I(\tilde{G}(F),\omega)$ revient à étudier les espaces $SI(\tilde{G}'_i(F))$, où $(G'_i,\tilde{G}'_i)_{i=1,\dots,n}$ est une certaine famille finie de couples déduits de $(G,\tilde{G},\omega)$ et vérifiant les hypothèses plus simples ci-dessus. Nous développerons les constructions de cette théorie dans les paragraphes suivants.

2.3. *L*-groupes. Une paire de Borel épinglée est un triplet $\mathcal{E}=(B,T,(E_\alpha)_{\alpha\in\Delta})$, où (B,T) est une paire de Borel, Δ est l'ensemble des racines simples de T dans l'algèbre de Lie $\mathfrak{u}$ du radical unipotent de B et, pour tout $\alpha\in\Delta$, E_α est un élément non nul de la droite radicielle $\mathfrak{u}_\alpha\subset\mathfrak{u}$ associée à α. Deux paires de Borel (resp. épinglées) de G sont conjuguées par un élément de G. Fixons une paire de Borel (B,T). On définit une action "quasi-déployée" de Γ_F sur T de la façon suivante. Pour tout $\sigma\in\Gamma_F$, fixons $g(\sigma)\in G$ tel que $ad_{g(\sigma)}\circ\sigma(B,T)=(B,T)$ (pour $g\in G$, ad_g est l'automorphisme $x\mapsto gxg^{-1}$ de G). On note σ_{G^*} l'automorphisme $ad_{g(\sigma)}\circ\sigma$ de T. Il ne dépend pas du choix de $g(\sigma)$ et $\sigma\mapsto\sigma_{G^*}$ est l'action cherchée. Il y a aussi un automorphisme θ de T ainsi défini : on choisit $\gamma\in\tilde{G}$ tel que ad_γ conserve (B,T) (il existe de tels γ) ; alors θ est la restriction de ad_γ à T, laquelle ne dépend pas du choix de γ.

Un groupe dual de G est un groupe réductif connexe $\hat{G}$ sur $\mathbb{C}$, muni d'une paire de Borel épinglée $\hat{\mathcal{E}}=(\hat{B},\hat{T},(\hat{E}_{\hat{\alpha}})_{\hat{\alpha}\in\hat{\Delta}})$ et d'une action algébrique de Γ_F satisfaisant aux conditions suivantes. L'action de Γ_F conserve $\hat{\mathcal{E}}$. C'est-à-dire que, pour $\sigma\in\Gamma_F$, on a $\sigma(\hat{B},\hat{T})=(\hat{B},\hat{T})$ et σ agit par une permutation $\hat{\alpha}\mapsto\sigma(\hat{\alpha})$ sur $\hat{\Delta}$ de sorte que $\sigma(\hat{E}_{\hat{\alpha}})=\hat{E}_{\sigma(\hat{\alpha})}$. Pour toute paire de Borel (B,T) de G, on se donne des isomorphismes en dualité $X_*(T)\simeq X^*(\hat{T})$, $X^*(T)\simeq X_*(\hat{T})$ (avec la notation habituelle pour ces groupes de caractères et de cocaractères), qui échangent ensembles de racines et de coracines, qui respectent les ordres sur ces ensembles définis par B et $\hat{B}$ et qui sont équivariants pour les actions de Γ_F, le tore T étant muni de l'action quasi-déployée associée à (B,T). On demande que si on remplace (B,T) par $ad_g(B,T)$, pour $g\in G$, les isomorphismes relatifs à $ad_g(B,T)$ se déduisent de ceux relatifs à (B,T) par composition avec les isomorphismes déduits de ad_g. On fixe un

tel groupe dual et on pose ${}^LG = \hat{G} \rtimes W_F$, où W_F agit sur $\hat{G}$ via l'homomorphisme naturel $W_F \to \Gamma_F$. Pour une paire de Borel (B,T), l'automorphisme θ de T se transporte en un automorphisme $\hat{\theta}$ de $\hat{T}$: si $x_* \in X_*(T)$ correspond à $\hat{x}^* \in X^*(\hat{T})$, $\theta \circ x_*$ correspond à $\hat{x}^* \circ \hat{\theta}$. Cet automorphisme ne dépend pas du choix de (B,T) et se prolonge en un automorphisme $\hat{\theta}$ de $\hat{G}$ qui conserve $\hat{\mathcal{E}}$ et commute à l'action galoisienne. On peut considérer l'ensemble $\hat{G}\hat{\theta}$ comme un espace tordu sous $\hat{G}$.

On a fixé en 2.1 un caractère ω de $G(F)$. Il convient de modifier cette donnée en fixant plutôt un élément $\mathbf{a}$ du groupe de cohomologie $H^1(W_F; Z(\hat{G}))$. Selon une construction de Langlands, cet élément détermine un caractère ω de $G(F)$. Si F est non-archimédien, la correspondance $\mathbf{a} \mapsto \omega$ est bijective. Mais, si F est archimédien, elle est seulement surjective et fixer $\mathbf{a}$ est plus précis que fixer ω.

2.4. Données endoscopiques. Une donnée endoscopique pour $(G, \tilde{G}, \mathbf{a})$ est un triplet $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$, où G' est un groupe réductif connexe quasi-déployé sur F, $\mathcal{G}'$ est un sous-groupe de ${}^LG$ et $\tilde{s}$ est un élément semi-simple de l'espace tordu $\hat{G}\hat{\theta}$, ces données vérifiant les conditions qui suivent. Notons $\hat{G}_{\tilde{s}}$ la composante neutre du commutant de $\tilde{s}$ dans $\hat{G}$. On suppose qu'il y a une suite exacte

$$1 \to \hat{G}_{\tilde{s}} \to \mathcal{G}' \to W_F \to 1$$

et que cette suite est scindée, c'est-à-dire qu'il y a un homomorphisme continu $W_F \to \mathcal{G}'$ qui est une section de la deuxième flèche. Fixons une paire de Borel épinglée de $\hat{G}_{\tilde{s}}$. Pour tout $w \in W_F$, fixons $g_w = (g(w), w) \in \mathcal{G}'$ tel que $ad_{g(w)} \circ w$ la conserve. Alors l'application qui, à w, associe l'automorphisme $ad_{g(w)} \circ w$ de $\hat{G}_{\tilde{s}}$ se quotiente ou s'étend (selon que F est archimédien ou non) en une action de Γ_F sur $\hat{G}_{\tilde{s}}$. On suppose que, muni de cette action, $\hat{G}_{\tilde{s}}$ est un groupe dual de G'. Cela nous autorise à noter $\hat{G}'$ ce groupe $\hat{G}_{\tilde{s}}$. On suppose enfin qu'il existe un cocycle $a : W_F \to Z(\hat{G})$, dont la classe est $\mathbf{a}$, de sorte que, pour tout $(g,w) \in \mathcal{G}'$, on ait l'égalité $s\hat{\theta}(g)w(s)^{-1} = a(w)g$, où on a écrit $\tilde{s} = s\hat{\theta}$ avec $s \in \hat{G}$.

Deux données endoscopiques $\mathbf{G}'_1 = (G'_1, \mathcal{G}'_1, \tilde{s}_1)$ et $\mathbf{G}'_2 = (G'_2, \mathcal{G}'_2, \tilde{s}_2)$ sont dites équivalentes s'il existe $x \in \hat{G}$ tel que $x\mathcal{G}'_1x^{-1} = \mathcal{G}'_2$ et $x\tilde{s}_1x^{-1} \in Z(\hat{G})\tilde{s}_2$. On peut toujours remplacer une donnée endoscopique par une donnée équivalente $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ de sorte que les conditions suivantes soient vérifiées : $\tilde{s} = s\hat{\theta}$, avec $s \in \hat{T}$; l'action galoisienne sur $\hat{G}'$ conserve une paire de Borel épinglée dont la paire de Borel $(\hat{B}', \hat{T}')$ sous-jacente est $(\hat{B} \cap \hat{G}', \hat{T}^{\hat{\theta},0})$, le deuxième groupe désignant la composante neutre du sous-groupe des points fixes de $\hat{\theta}$ agissant dans $\hat{T}$. Dans la suite, on ne considère que de telles données. Soient (B,T), resp. (B',T'), une paire de Borel de G, resp. G'. Des homomorphismes

$$X_*(T) \simeq X^*(\hat{T}) \overset{restriction}{\to} X^*(\hat{T}^{\hat{\theta},0}) = X^*(\hat{T}') \simeq X_*(T')$$

se déduit un homomorphisme $\xi : T \to T'$. Il se quotiente en un isomorphisme $T/(1-\theta)(T) \simeq T'$, où $1-\theta$ est l'homomorphisme $t \mapsto t\theta(t)^{-1}$ de T dans lui-même. On montre que ξ se restreint en un homomorphisme $\xi : Z(G) \to Z(G')$ qui ne dépend pas des choix de paires de Borel.

On dit que la donnée $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ est elliptique si $Z(\hat{G}')^{\Gamma_F,0} = Z(\hat{G})^{\Gamma_F,\hat{\theta},0}$, avec une notation similaire à celle ci-dessus.

2.5. L'espace endoscopique. Pour une paire de Borel épinglée $\mathcal{E}$ de G, notons $Z(\tilde{G}, \mathcal{E})$ l'ensemble des $\gamma \in \tilde{G}$ tels que ad_γ conserve $\mathcal{E}$. Notons $\mathcal{Z}(\tilde{G}, \mathcal{E})$ le quotient de $Z(\tilde{G}, \mathcal{E})$

par l'action par conjugaison de $Z(G)$. Si $\mathcal{E}'$ est une autre paire de Borel épinglée, on choisit $g \in G$ tel que $ad_g(\mathcal{E}) = \mathcal{E}'$. Alors ad_g définit un isomorphisme de $\mathcal{Z}(\tilde{G}, \mathcal{E})$ sur $\mathcal{Z}(\tilde{G}, \mathcal{E}')$ qui ne dépend pas du choix de g. On note $\mathcal{Z}(\tilde{G})$ la limite inductive des $\mathcal{Z}(\tilde{G}, \mathcal{E})$, la limite étant prise sur les paires de Borel épinglées $\mathcal{E}$ de G, les applications de transition étant les isomorphismes canoniques que l'on vient de définir. Cet objet $\mathcal{Z}(\tilde{G})$ étant canonique, il récupère une action de Γ_F. C'est un espace tordu sous $\mathcal{Z}(G) := Z(G)/(1-\theta)(Z(G))$.

Soit $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ une donnée endoscopique pour $(G, \tilde{G}, \mathbf{a})$. On définit l'espace endoscopique $\tilde{G}'$ comme le quotient de $G' \times \mathcal{Z}(\tilde{G})$ par la relation d'équivalence $(g'\xi(z), \tilde{z}) \equiv (g', z\tilde{z})$ pour tout $z \in Z(G)$. L'action galoisienne sur $G' \times \mathcal{Z}(\tilde{G})$ se descend en une action galoisienne sur $\tilde{G}'$. On a des actions à droite et à gauche de G' sur $\tilde{G}'$ qui proviennent des multiplications sur la première composante de $G' \times \mathcal{Z}(\tilde{G})$. Ainsi $\tilde{G}'$ est un espace tordu sous G' qui est à torsion intérieure. Remarquons que $\tilde{G}'(F)$ peut être vide. La construction fournit une application naturelle $\mathcal{Z}(\tilde{G}) \to \mathcal{Z}(\tilde{G}')$. Ce dernier ensemble est celui des $\delta \in \tilde{G}'$ tels que ad_δ soit l'identité.

2.6. Correspondance endoscopique. Soit $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ une donnée endoscopique pour $(G, \tilde{G}, \mathbf{a})$. Appelons diagramme un sextuplet $(\epsilon, B', T', B, T, \eta)$, où $\epsilon \in \tilde{G}'(F)$, $\eta \in \tilde{G}(F)$, (B', T'), resp. (B, T), est une paire de Borel de G', resp. G, vérifiant les conditions suivantes. Les automorphismes ad_ϵ et ad_η conservent respectivement (B', T') et (B, T) (les éléments ϵ et η sont donc semi-simples). Les tores T et T' sont définis sur F, ainsi que l'homomorphisme $\xi : T \to T'$ associé aux paires de Borel. Etendons les deux paires de Borel en des paires de Borel épinglées $\mathcal{E}$ et $\mathcal{E}'$. On peut écrire $\eta = te$, avec $t \in T$ et $e \in Z(\tilde{G}, \mathcal{E})$. L'élément e a une image naturelle e' dans $\mathcal{Z}(\tilde{G}') = Z(\tilde{G}', \mathcal{E}')$. On impose que $\epsilon = \xi(t)e'$. Cela ne dépend pas des choix d'épinglages.

Pour des éléments $\epsilon \in \tilde{G}'(F)$ et $\eta \in \tilde{G}(F)$, on dit que ces éléments se correspondent s'il existe un diagramme les joignant. On dit que $\mathbf{G}'$ est relevant s'il existe un diagramme.

Pour un élément semi-simple $\eta \in \tilde{G}$, on introduit le groupe $I_\eta = G_\eta Z(G)^\theta$. Soient $\eta, \eta' \in \tilde{G}(F)$ deux éléments semi-simples. On dit qu'ils sont stablement conjugués s'il existe $y \in G$ tel que $y^{-1}\eta y = \eta'$ et $y\sigma(y)^{-1} \in I_\eta$ pour tout $\sigma \in \Gamma_F$. Cette définition généralise celle posée en 2.2. La classe de conjugaison stable de η est l'ensemble des éléments stablement conjugués à η. Soient $\mathcal{O}'$, resp. $\mathcal{O}$, une classe de conjugaison stable d'éléments semi-simples dans $\tilde{G}'(F)$, resp. $\tilde{G}(F)$. On dit qu'elles se correspondent s'il existe $\epsilon \in \mathcal{O}'$ et $\eta \in \mathcal{O}$ qui se correspondent (si $\mathcal{O}$ est formé d'éléments fortement réguliers, c'est équivalent à ce que ϵ et η se correspondent pour tous $\epsilon \in \mathcal{O}'$ et $\eta \in \mathcal{O}$). On montre qu'à une classe $\mathcal{O}'$ correspond au plus une classe $\mathcal{O}$ et qu'inversement, l'ensemble des classes $\mathcal{O}'$ qui correspondent à une classe $\mathcal{O}$ est fini (éventuellement vide).

2.7. Transfert. Soit $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ une donnée endoscopique relevante de $(G, \tilde{G}, \mathbf{a})$. Considérons une suite exacte

$$1 \to C_1 \to G_1' \to G' \to 1,$$

où G_1' est un groupe réductif et quasi-déployé sur F et C_1 est un sous-tore central induit (c'est-à-dire que $X_*(C_1)$ possède une base permutée par l'action galoisienne). Considérons un espace tordu $\tilde{G}_1'$ sous G_1', à torsion intérieure et tel que $\tilde{G}_1'(F) \neq \emptyset$. Supposons donnée une application $\tilde{G}_1' \to \tilde{G}'$ compatible avec la projection $G_1' \to G'$. Supposons donné un plongement $\hat{\xi}_1 : \mathcal{G}' \to {}^L G_1'$ compatible aux projections sur W_F, dont la restriction à $\hat{G}' \subset \mathcal{G}'$ est un homomorphisme $\hat{G}' \to \hat{G}_1'$ dual à la projection $G_1' \to G'$. De telles données existent,

cf. [11] 2.2. Pour $(\delta_1,\gamma)\in\tilde{G}'_1(F)\times\tilde{G}(F)$, on dit que δ_1 et γ se correspondent si δ et γ se correspondent, où δ est l'image de δ_1 dans $\tilde{G}'(F)$. Kottwitz et Shelstad ont défini un facteur de transfert $\Delta_1:\tilde{G}'_1(F)\times\tilde{G}(F)\to\mathbb{C}$, cf. [11] chapitre 4 et [12]. On a $\Delta_1(\delta_1,\gamma)\neq 0$ si et seulement si δ_1 et γ se correspondent et γ est fortement régulier. Pour de tels éléments, $\Delta_1(\delta_1,\gamma)$ ne dépend que de la classe de conjugaison stable de δ_1 et on a la formule

$$\Delta_1(c_1\delta_1,g^{-1}\gamma g)=\lambda_1(c_1)^{-1}\omega(g)\Delta_1(\delta_1,\gamma)$$

pour tous $c_1\in C_1(F)$ et $g\in G(F)$, où λ_1 est un caractère de $C_1(F)$ déduit des données.

Notons $C^\infty_{c,\lambda_1}(\tilde{G}'_1(F))$ l'espace des fonctions $f_1:\tilde{G}'_1(F)\to\mathbb{C}$ qui sont lisses, à support compact modulo $C_1(F)$ et telles que $f_1(c_1\delta_1)=\lambda_1(c_1)^{-1}f_1(\delta_1)$ pour tous $c_1\in C_1(F)$ et $\delta_1\in\tilde{G}'_1(F)$. Les définitions du paragraphe 2.2 s'adaptent à cet espace, on ajoute des λ_1 en indices. Pour $f\in C^\infty_c(\tilde{G}(F))$ et $f_1\in C^\infty_{c,\lambda_1}(\tilde{G}'_1(F))$, on dit que f_1 est un transfert de f si on a l'égalité

$$S^{\tilde{G}'_1}_{\lambda_1}(\delta_1,f_1)=\sum_\gamma\Delta_1(\delta_1,\gamma)[Z_G(\gamma;F):G_\gamma(F)]^{-1}I^{\tilde{G}}(\gamma,\omega,f)$$

pour tout δ_1 dans un ensemble ouvert dense de $\tilde{G}'_1(F)$, où γ parcourt l'ensemble des éléments de $\tilde{G}(F)$ correspondant à δ_1, modulo conjugaison par $G(F)$. Pour donner un sens à cette égalité, il faut fixer des mesures de Haar sur tous les groupes intervenant. Celles sur les tores $G_\gamma(F)$ et $G'_{1,\delta_1}(F)$ doivent être reliées.

Théorème 2.1. *Tout $f\in C^\infty_c(\tilde{G}(F))$ admet un transfert $f_1\in C^\infty_{c,\lambda_1}(\tilde{G}'_1(F))$.*

Ce théorème résulte de [17] et [20] 1.5 dans le cas non archimédien. Il est dû à Shelstad [18] dans le cas archimédien. On montre que les choix de mesures de Haar disparaissent si l'on introduit pour tout groupe réductif H sur F la droite complexe $Mes(H(F))$ portée par une mesure de Haar sur $H(F)$ et si l'on considère le transfert comme une application linéaire

$$I(\tilde{G}(F),\omega)\otimes Mes(G(F))\to SI_{\lambda_1}(\tilde{G}'_1(F))\otimes Mes(G'(F)).$$

Cette application dépend des données G'_1, $\tilde{G}'_1$, C_1, $\hat{\xi}_1$, ainsi que du choix de Δ_1. Ce facteur n'est en effet pas uniquement déterminé, mais seulement à homothétie près. Considérons d'autres données G'_2,...,Δ_2 vérifiant les mêmes hypothèses. Il s'avère que l'on peut définir un isomorphisme canonique

$$C^\infty_{c,\lambda_1}(\tilde{G}'_1(F))\simeq C^\infty_{c,\lambda_2}(\tilde{G}'_2(F))$$

qui commute au transfert. C'est-à-dire que, pour $f\in C^\infty_c(\tilde{G}(F))$ et $f_1\in C^\infty_{c,\lambda_1}(\tilde{G}'_1(F))$, f_1 est un transfert de f si et seulement si l'image f_2 de f_1 par l'isomorphisme ci-dessus est un transfert de f. On peut alors définir un espace noté $C^\infty_c(\mathbf{G}')$ qui est la limite inductive des espaces $C^\infty_{c,\lambda_1}(\tilde{G}'_1(F))$, la limite étant prise sur toutes les données auxiliaires G'_1,...,Δ_1, les applications de transition étant les isomorphismes ci-dessus. Cette définition pose un problème logique, les données auxiliaires ne formant pas un ensemble, mais ce problème est facile à résoudre. Une construction analogue permet de définir des espaces $I(\mathbf{G}')$ et $SI(\mathbf{G}')$. L'application de transfert devient une application linéaire canonique

$$\begin{array}{ccc} I(\tilde{G}(F),\omega)\otimes Mes(G(F)) & \to & SI(\mathbf{G}')\otimes Mes(G'(F)) \\ \mathbf{f} & \mapsto & \mathbf{f}^{\mathbf{G}'} \end{array}$$

2.8. Les espaces de distributions. On note $D_{orb}(\tilde{G}(F),\omega)$ l'espace de formes linéaires sur $I(\tilde{G}(F),\omega)$ (que l'on peut relever en des formes linéaires sur $C_c^\infty(\tilde{G}(F))$) engendré par les intégrales orbitales $f \mapsto I^{\tilde{G}}(\gamma,\omega,f)$, cf. 2.2, quand γ décrit $\tilde{G}(F)$. Quand F est non-archimédien, c'est aussi l'espace des formes linéaires qui, relevées en des formes linéaires sur $C_c^\infty(\tilde{G}(F))$, sont supportées par un nombre fini de classes de conjugaison par $G(F)$. Il est plus canonique d'introduire comme en 2.7 la droite complexe $Mes(G(F))$ et sa duale $Mes(G(F))^*$ et de considérer les espaces $I(\tilde{G}(F),\omega)\otimes Mes(G(F))$ et $D_{orb}(\tilde{G}(F),\omega)\otimes Mes(G(F))^*$. Pour $\mathbf{f}\in I(\tilde{G}(F),\omega)\otimes Mes(G(F))$ et $\boldsymbol{\gamma}\in D_{orb}(\tilde{G}(F),\omega)\otimes Mes(G(F))^*$, on note $I^{\tilde{G}}(\boldsymbol{\gamma},\mathbf{f})$ l'évaluation de $\boldsymbol{\gamma}$ sur $\mathbf{f}$. Supposons que G soit quasi-déployé, que $\tilde{G}$ soit à torsion intérieure et que $\mathbf{a}=1$. On note $D_{orb}^{st}(\tilde{G}(F))$ le sous-espace des éléments de $D_{géom}(\tilde{G}(F))$ qui sont stables, c'est-à-dire se quotientent en une forme linéaire sur $SI(\tilde{G}(F))$. Pour $\mathbf{f}\in SI(\tilde{G}(F))\otimes Mes(G(F))$ et $\boldsymbol{\delta}\in D_{orb}^{st}(\tilde{G}(F))\otimes Mes(G(F))^*$, on note $S^{\tilde{G}}(\boldsymbol{\delta},\mathbf{f})$ l'évaluation de $\boldsymbol{\delta}$ sur $\mathbf{f}$.

Soit $\mathbf{G}'=(G',\mathcal{G}',\tilde{s})$ une donnée endoscopique relevante de $(G,\tilde{G},\mathbf{a})$. On a défini l'espace $SI(\mathbf{G}')$ comme limite inductive d'espaces $SI_{\lambda_1}(\tilde{G}'_1(F))$ construits à l'aide de données auxiliaires. Une construction analogue permet de définir un espace $D_{orb}^{st}(\mathbf{G}')$ de formes linéaires sur $SI(\mathbf{G}')$. Supposons F non-archimédien. Dualement à l'application de transfert du paragraphe précédent, on a une application linéaire

$$transfert : D_{orb}^{st}(\mathbf{G}')\otimes Mes(G'(F))^* \to D_{orb}(\tilde{G}(F),\omega)\otimes Mes(G(F))^*.$$

Le cas archimédien est plus compliqué, cf. 2.13.

2.9. Espaces de Levi. Soient P un sous-groupe parabolique de G et M une composante de Levi de P, tous deux définis sur F (on appelle M un Levi de G). Notons $\tilde{M}$ l'ensemble des $\gamma\in\tilde{G}$ tels que ad_γ conserve P et M. Supposons $\tilde{M}$ non vide. On appelle alors $\tilde{M}$ un espace de Levi de $\tilde{G}$. C'est un espace tordu sous M. On peut identifier le groupe dual $\hat{M}$ à un sous-groupe de Levi de $\hat{G}$, que l'on peut supposer standard et invariant par $\hat{\theta}$ comme par l'action galoisienne. L'élément $\mathbf{a}\in H^1(W_F;Z(\hat{G}))$ se pousse en un élément $\mathbf{a}_M\in H^1(W_F;Z(\hat{M}))$. Soit $\mathbf{M}'=(M',\mathcal{M}',\tilde{\zeta})$ une donnée endoscopique elliptique pour $(M,\tilde{M},\mathbf{a}_M)$. Dans la définition de 2.4 intervient un cocycle, ici $a_M : W_F\to Z(\hat{M})$, de classe $\mathbf{a}_M$. On voit que, quitte à remplacer $\mathbf{M}'$ par une donnée équivalente, on peut supposer que ce cocycle est à valeurs dans $Z(\hat{G})$. Sa classe dans $H^1(W_F;Z(\hat{G}))$ est alors $\mathbf{a}$. On suppose qu'il en est ainsi.

Soit $\tilde{s}\in\tilde{\zeta}Z(\hat{M})^{\Gamma_F,\hat{\theta}}/Z(\hat{G})^{\Gamma_F,\hat{\theta}}$. On montre qu'il existe une donnée endoscopique $\mathbf{G}'(\tilde{s})=(G'(\tilde{s}),\mathcal{G}'(\tilde{s}),\tilde{s})$ de $(G,\tilde{G},\mathbf{a})$ caractérisée par les propriétés suivantes : $\hat{G}'(\tilde{s})$ est la composante neutre du commutant de $\tilde{s}$ dans $\hat{G}$; $\mathcal{G}'(\tilde{s})$ est le sous-groupe de ${}^LG$ engendré par $\hat{G}'(\tilde{s})$ et $\mathcal{M}'$ (en fait, c'est simplement le produit $\hat{G}'(\tilde{s})\mathcal{M}'=\mathcal{M}'\hat{G}'(\tilde{s})$). Le groupe M' s'identifie à un Levi de $G'(\tilde{s})$ et $\tilde{M}'$ s'identifie conformément à un espace de Levi de $\tilde{G}'(\tilde{s})$. On définit une constante $i_{\tilde{M}'}(\tilde{G},\tilde{G}'(\tilde{s}))$. Si $\mathbf{G}'(\tilde{s})$ n'est pas elliptique, elle est nulle. Si $\mathbf{G}'(\tilde{s})$ est elliptique, on montre qu'il y a un homomorphisme naturel

$$Z(\hat{M})^{\Gamma_F,\hat{\theta}}/Z(\hat{G})^{\Gamma_F,\hat{\theta}}\to Z(\hat{M}')^{\Gamma_F}/Z(\hat{G}'(\tilde{s}))^{\Gamma_F}.$$

Il est surjectif et de noyau fini. Alors $i_{\tilde{M}'}(\tilde{G},\tilde{G}'(\tilde{s}))$ est l'inverse du nombre d'éléments de ce noyau.

Soient $\tilde{M}$ un espace de Levi de $\tilde{G}$ et M le Levi de G associé. On note A_M le plus grand sous-tore de $Z(M)$ déployé sur F et $A_{\tilde{M}}$ le plus grand sous-tore de A_M sur lequel

ad_γ agit trivialement pour tout $\gamma \in \tilde{M}$ (ou pour un $\gamma \in \tilde{M}$, c'est équivalent). On pose $\mathcal{A}_{\tilde{M}} = X_*(A_{\tilde{M}}) \otimes_{\mathbb{Z}} \mathbb{R}$. On doit fixer pour tout espace de Levi $\tilde{M}$ une mesure de Haar sur $\mathcal{A}_{\tilde{M}}$. De même, pour toute donnée endoscopique $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ de $(G, \tilde{G}, \mathbf{a})$ et pour tout espace de Levi $\tilde{M}'$ de $\tilde{G}'$, on doit fixer une mesure de Haar sur $\mathcal{A}_{\tilde{M}'}$. On doit imposer des conditions de compatibilité à ces divers choix. En particulier, si $\mathbf{G}'$ est elliptique, il y a un isomorphisme naturel $\mathcal{A}_{\tilde{G}} \simeq \mathcal{A}_{\tilde{G}'}$ et on impose que les mesures se correspondent par cet isomorphisme.

2.10. Intégrales orbitales pondérées. Les définitions de ce paragraphe sont dues (comme bien d'autres) à Arthur. Considérons un espace de Levi $\tilde{M}$ de $\tilde{G}$. Soient $\gamma \in \tilde{G}(F)$ et $f \in C_c^\infty(\tilde{G}(F))$. Supposons d'abord $M_\gamma = G_\gamma$. Si ω est non trivial sur $M_\gamma(F)$, on pose $J_{\tilde{M}}^{\tilde{G}}(\gamma, \omega, f) = 0$. Supposons ω trivial sur $M_\gamma(F)$. On choisit des mesures de Haar sur tous les groupes intervenant et on pose

$$J_{\tilde{M}}^{\tilde{G}}(\gamma, \omega, f) = D^{\tilde{G}}(\gamma)^{1/2} \int_{M_\gamma(F)\backslash G(F)} \omega(g) f(g^{-1}\gamma g) v_{\tilde{M}}(g)\, dg$$

où $v_{\tilde{M}}$ est un "poids" défini en [5] paragraphe 1 (à l'aide du choix d'un "bon" sous-groupe compact maximal de $G(F)$).

La définition de $J_{\tilde{M}}^{\tilde{G}}(\gamma, \omega, f)$ quand la condition $M_\gamma = G_\gamma$ n'est pas vérifiée est beaucoup plus délicate. Pour $a \in A_{\tilde{M}}(F)$ en position générale, on a $M_{a\gamma} = G_{a\gamma}$ et le terme $J_{\tilde{M}}^{\tilde{G}}(a\gamma, \omega, f)$ est défini. Plus généralement, $J_{\tilde{L}}^{\tilde{G}}(a\gamma, \omega, f)$ est défini pour tout espace de Levi $\tilde{L} \supset \tilde{M}$. Alors $J_{\tilde{M}}^{\tilde{G}}(\gamma, \omega, f)$ est défini comme la limite quand a tend vers 1 d'une certaine combinaison linéaire (à coefficients dépendant de a) de ces intégrales $J_{\tilde{L}}^{\tilde{G}}(a\gamma, \omega, f)$.

De nouveau, il est plus canonique de remplacer f par un élément de $C_c^\infty(\tilde{G}(F)) \otimes Mes(G(F))$. On s'aperçoit qu'alors les données de γ et d'une mesure sur $M_\gamma(F)$ suffisent pour définir les termes ci-dessus. Or ces données définissent aussi une intégrale orbitale sur $\tilde{M}(F)$, c'est-à-dire un élément de $D_{orb}(\tilde{M}(F), \omega) \otimes Mes(M(F))^*$. Par linéarité, on peut alors définir $J_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, \mathbf{f})$ pour $\boldsymbol{\gamma} \in D_{orb}(\tilde{M}(F), \omega) \otimes Mes(M(F))^*$ et $\mathbf{f} \in C_c^\infty(\tilde{G}(F)) \otimes Mes(G(F))$. Hormis le cas $\tilde{M} = \tilde{G}$, les applications linéaires $\mathbf{f} \mapsto J_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, \mathbf{f})$ ne sont pas ω-équivariantes, c'est-à-dire ne se quotientent pas en des applications linéaires sur $I(\tilde{G}(F), \omega) \otimes Mes(G(F))$. Arthur a défini des avatars ω-équivariants de ces applications, cf. [6] section 2. On n'en donnera pas la définition qui passe par la théorie des représentations (les objets obtenus ne sont plus, en fait, de nature "géométrique"). On note $I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, \mathbf{f})$ l'avatar ω-équivariant de $J_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, \mathbf{f})$. C'est une forme bilinéaire en $\boldsymbol{\gamma} \in D_{orb}(\tilde{M}(F), \omega) \otimes Mes(M(F))^*$ et $\mathbf{f} \in I(\tilde{G}(F), \omega) \otimes Mes(G(F))$. Comme on le verra en 3.3, ces formes bilinéaires sont les objets locaux qui interviennent dans la partie géométrique de la formule des traces.

Variante. Supposons G quasi-déployé, $\tilde{G}$ à torsion intérieure et $\mathbf{a} = 1$. Comme on l'a dit, pour γ tel que $M_\gamma \neq G_\gamma$, les termes $J_{\tilde{M}}^{\tilde{G}}(\gamma, f)$ sont définis par un procédé de limite. En fait, il y a plusieurs procédés possibles. Notons η la partie semi-simple de γ et fixons un sous-tore maximal T de G_η. Notons $\Sigma^{G_\eta}(T)$ l'ensemble des racines de T dans l'algèbre de Lie de G_η. A toute fonction $B_\eta : \Sigma^{G_\eta}(T) \to \mathbb{Q}_{>0}$ vérifiant des propriétés très restrictives (par exemple une fonction constante), on peut associer un procédé de passage à la limite. Plus globalement, supposons que, pour tout élément semi-simple $\eta \in \tilde{G}(F)$, on s'est donné une

telle fonction B_η, ces fonctions étant reliées elles-mêmes par des conditions de compatibilité. On appelle ces données un système de fonctions B. Alors on peut définir pour tout $\gamma \in \tilde{M}(F)$ un terme $J_{\tilde{M}}^{\tilde{G}}(\gamma, B, f)$. Il est égal à $J_{\tilde{M}}^{\tilde{G}}(\gamma, f)$ si $M_\gamma = G_\gamma$ mais ne l'est pas, en général, si cette condition n'est pas vérifiée. Comme ci-dessus, on en déduit des avatars invariants $I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, B, \mathbf{f})$.

2.11. Intégrales orbitales pondérées stables. Supposons le corps F non-archimédien, G quasi-déployé, $\tilde{G}$ à torsion intérieure et $\mathbf{a} = 1$. On supprime les termes $\mathbf{a}$ et $\hat{\theta}$ des notations. Soit $\tilde{M}$ un espace de Levi de $\tilde{G}$. Le triplet $\mathbf{M} = (M, {}^L M, 1)$ est une donnée endoscopique de $(M, \tilde{M})$ (la donnée "maximale"). Pour $\boldsymbol{\delta} \in D_{orb}^{st}(\tilde{M}(F)) \otimes Mes(M(F))^*$ et $\mathbf{f} \in I(\tilde{G}(F)) \otimes Mes(G(F))$, on définit un terme $S_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\delta}, \mathbf{f})$ par la formule

$$S_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\delta}, \mathbf{f}) = I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\delta}, \mathbf{f}) - \sum_{s \in Z(\hat{M})^{\Gamma_F}/Z(\hat{G})^{\Gamma_F}, s \neq 1} i_{\tilde{M}}(\tilde{G}, \tilde{G}'(s)) S_{\mathbf{M}}^{\mathbf{G}'(s)}(\boldsymbol{\delta}, \mathbf{f}^{\mathbf{G}'(s)}),$$

cf. [8] section 5. Expliquons cette formule. Le premier terme $I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\delta}, \mathbf{f})$ a déjà été défini. Pour s intervenant dans la somme, l'hypothèse $s \neq 1$ entraîne que $dim(G'(s)_{SC}) < dim(G_{SC})$, où par exemple G_{SC} est le revêtement simplement connexe du groupe dérivé de G. Supposons un instant que, dans les constructions de 2.7, on puisse choisir pour données auxiliaires $G'_1(s) = G'(s)$, $\tilde{G}'_1(s) = \tilde{G}'(s)$. En raisonnant par récurrence sur la dimension de G_{SC}, on peut supposer défini le terme $S_{\tilde{M}}^{\tilde{G}'(s)}(\boldsymbol{\delta}, \boldsymbol{\varphi})$ pour $\boldsymbol{\varphi} \in I(\tilde{G}'(s; F)) \otimes Mes(G'(F))$. Un raisonnement formel permet de s'affranchir de l'hypothèse ci-dessus et de définir en général un terme $S_{\mathbf{M}}^{\mathbf{G}'(s)}(\boldsymbol{\delta}, \boldsymbol{\varphi})$, pour $\boldsymbol{\varphi} \in I(\mathbf{G}'(s)) \otimes Mes(G'(F))$. Comme on va le voir, ce terme est stable en $\boldsymbol{\varphi}$, c'est-à-dire ne dépend que de l'image de $\boldsymbol{\varphi}$ dans $SI(\mathbf{G}'(s)) \otimes Mes(G'(F))$. On note $\mathbf{f}^{\mathbf{G}'(s)}$ le transfert de $\mathbf{f}$. C'est un élément de cet espace. Le terme $S_{\mathbf{M}}^{\mathbf{G}'(\tilde{s})}(\boldsymbol{\delta}, \mathbf{f}^{\mathbf{G}'(\tilde{s})})$ est donc bien défini.

La construction repose donc sur le théorème suivant.

Théorème 2.2. *Pour tout $\boldsymbol{\delta} \in D_{orb}^{st}(\tilde{M}(F)) \otimes Mes(M(F))^*$, la forme linéaire $\mathbf{f} \mapsto S_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\delta}, \mathbf{f})$ est stable, c'est-à-dire qu'elle se descend en une forme linéaire sur l'espace $SI(\tilde{G}(F)) \otimes Mes(G(F))$.*

Variante. Supposons donné un système de fonctions B comme en 2.10. Il s'en déduit aisément pour tout s un tel système pour chaque $\tilde{G}'(s)$. On définit alors $S_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\delta}, B, \mathbf{f})$ de la même façon que ci-dessus. Pour ce terme, le théorème est encore vérifié.

2.12. Intégrales orbitales pondérées endoscopiques. Supposons que F est non-archimédien, mais que $(G, \tilde{G}, \mathbf{a})$ est quelconque. Considérons une donnée endoscopique $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ de $(G, \tilde{G}, \mathbf{a})$. Considérons un diagramme $(\epsilon, B', T', B, T, \eta)$, où $\epsilon \in \tilde{G}'(F)$ et $\eta \in \tilde{G}(F)$ sont des éléments semi-simples. Du diagramme se déduit un homomorphisme $\mathfrak{t}^\theta \to \mathfrak{t}'$ (où $\theta = ad_\eta$). Le tore $T^{\theta,0}$ est un sous-tore maximal de G_η. Notons $\Sigma^{G_\eta}(T^{\theta,0})$ et $\Sigma^{G'_\epsilon}(T')$ les ensembles de racines de $T^{\theta,0}$ dans $\mathfrak{g}_\eta$ et de T' dans $\mathfrak{g}'_\epsilon$. On peut considérer ces racines comme des formes linéaires sur $\mathfrak{t}^\theta$, resp. $\mathfrak{t}'$. Par l'isomorphisme précédent, $\Sigma^{G'_\epsilon}(T')$ ne s'identifie pas à un sous-ensemble de $\Sigma^{G_\eta}(T^{\theta,0})$. Mais il existe une unique fonction $B_\epsilon^{\tilde{G}} : \Sigma^{G'_\epsilon}(T') \to \mathbb{Q}_{>0}$ de sorte que, pour tout $\alpha \in \Sigma^{G'_\epsilon}(T')$, $\alpha / B_\epsilon^{\tilde{G}}(\alpha)$ soit un élément de $\Sigma^{G_\eta}(T^{\theta,0})$. Plus globalement, on peut définir un système de fonctions $B^{\tilde{G}}$ sur $\tilde{G}'$, au sens de

2.10, de sorte que, pour tout diagramme comme ci-dessus, la fonction $B_\epsilon^{\tilde{G}}$ soit celle associée à ce système.

Soient $\tilde{M}$ un espace de Levi de $\tilde{G}$ et $\mathbf{M}' = (M', \mathcal{M}', \tilde{\zeta})$ une donnée endoscopique elliptique et relevante pour $(M, \tilde{M}, \mathbf{a}_M)$. Soient $\boldsymbol{\delta} \in D_{orb}^{st}(\mathbf{M}') \otimes Mes(M'(F))^*$ et $\mathbf{f} \in I(\tilde{G}(F), \omega) \otimes Mes(G(F))$. On pose

$$I_{\tilde{M}}^{\tilde{G},\mathcal{E}}(\mathbf{M}', \boldsymbol{\delta}, \mathbf{f}) = \sum_{\tilde{s} \in \tilde{\zeta} Z(\hat{M})^{\Gamma_F, \hat{\theta}} / Z(\hat{G})^{\Gamma_F, \hat{\theta}}} i_{\tilde{M}'}(\tilde{G}, \tilde{G}(\tilde{s})) S_{\mathbf{M}'}^{\mathbf{G}'(\tilde{s})}(\boldsymbol{\delta}, B^{\tilde{G}}, \mathbf{f}^{\mathbf{G}'(\tilde{s})}),$$

cf. [8] section 5. Le sens de chaque terme s'explique comme dans le paragraphe précédent.

Soit $\boldsymbol{\gamma} \in D_{orb}(\tilde{M}(F), \omega) \otimes Mes(M(F))^*$. On montre qu'il existe une famille finie $(\mathbf{M}_i')_{i=1,\dots,n}$ de données endoscopiques elliptiques et relevantes de $(M, \tilde{M}, \mathbf{a}_M)$ et, pour chaque i, un élément $\boldsymbol{\delta}_i \in D_{orb}^{st}(\mathbf{M}_i') \otimes Mes(M_i'(F))^*$, de sorte que

$$\boldsymbol{\gamma} = \sum_{i=1,\dots,n} transfert(\boldsymbol{\delta}_i).$$

Pour $\mathbf{f} \in I(\tilde{G}(F), \omega) \otimes Mes(G(F))$, posons

$$I_{\tilde{M}}^{\tilde{G},\mathcal{E}}(\boldsymbol{\gamma}, \mathbf{f}) = \sum_{i=1,\dots,n} I_{\tilde{M}}^{\tilde{G},\mathcal{E}}(\mathbf{M}_i', \boldsymbol{\delta}_i, \mathbf{f}).$$

La décomposition ci-dessus de $\boldsymbol{\gamma}$ n'est pas unique mais on montre que $I_{\tilde{M}}^{\tilde{G},\mathcal{E}}(\boldsymbol{\gamma}, \mathbf{f})$ ne dépend pas de ce choix.

Théorème 2.3. *Pour tous $\boldsymbol{\gamma} \in D_{orb}(\tilde{M}(F), \omega) \otimes Mes(M(F))^*$ et $\mathbf{f} \in I(\tilde{G}(F), \omega) \otimes Mes(G(F))$, on a l'égalité*

$$I_{\tilde{M}}^{\tilde{G},\mathcal{E}}(\boldsymbol{\gamma}, \mathbf{f}) = I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, \mathbf{f}).$$

2.13. Le cas *F* archimédien. Il y a des complications techniques lorsque F est archimédien. La première est que, pour obtenir des formules d'inversion satisfaisantes entre $\tilde{G}$ et ses données endoscopiques, on est parfois obligé d'adjoindre au couple $(G, \tilde{G})$ d'autres couples $(G_\sharp, \tilde{G}_\sharp)$, où $G_\sharp$ est une forme intérieure de G. On appelle K-espace la réunion disjointe (finie) de ces espaces $\tilde{G}_\sharp$ et, en général, on doit travailler avec de tels K-espaces, cf. [8] section 2. Cela ne modifie guère que les notations et ce n'est d'ailleurs pas utile dans le cas où G est quasi-déployé, $\tilde{G}$ est à torsion intérieure et $\mathbf{a} = 1$. Une difficulté plus sérieuse est que l'espace de distributions $D_{orb}(\tilde{G}(F), \omega) \otimes Mes(G(F))^*$ engendré par les intégrales orbitales se comporte mal par endoscopie. C'est-à-dire, considérons une donnée endoscopique $\mathbf{G}' = (G', \mathcal{G}', \tilde{s})$ de $(G, \tilde{G}, \mathbf{a})$, supposons pour simplifier que le recours à des données auxiliaires ne soit pas nécessaire et que l'on puisse définir un transfert $\mathbf{f} \mapsto \mathbf{f}^{\tilde{G}'}$ de $I(\tilde{G}(F), \omega) \otimes Mes(G(F))$ dans $SI(\tilde{G}'(F)) \otimes Mes(G'(F))$. Soit $\boldsymbol{\delta} \in D_{orb}^{st}(\tilde{G}'(F)) \otimes Mes(G'(F))^*$. Comme le montre un exemple dû à Magdy Assem, que m'a indiqué Kottwitz, il n'est pas vrai en général que l'application $\mathbf{f} \mapsto S^{\tilde{G}}(\boldsymbol{\delta}, \mathbf{f}^{\tilde{G}'})$ soit une combinaison linéaire d'intégrales orbitales. La construction de 2.12 s'évanouit si on se limite aux distributions qui sont des intégrales orbitales. La solution que l'on a retenue est de définir un espace de distributions $D_{tr-orb}(\tilde{G}(F), \omega)$ un peu plus gros que $D_{orb}(\tilde{G}(F), \omega)$, qui vérifie :

- dans la situation ci-dessus, le transfert envoie $D_{tr-orb}^{st}(\tilde{G}'(F)) \otimes Mes(G'(F))^*$ dans $D_{tr-orb}(\tilde{G}(F), \omega) \otimes Mes(G(F))^*$ (où $D_{tr-orb}^{st}(\tilde{G}'(F))$ est le sous-espace des éléments de $D_{tr-orb}(\tilde{G}'(F))$ qui sont stables) ;

- on peut définir les intégrales $I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, \mathbf{f})$ pour $\boldsymbol{\gamma} \in D_{tr-orb}(\tilde{M}(F), \omega) \otimes Mes(M(F))^*$.

On n'en donne pas la construction. En utilisant cet espace, on peut adapter les définitions de 2.11 et de 2.12. Le théorème 2.11 reste valable. Une forme modifiée du théorème 2.12 aussi.

3. La théorie globale

3.1. Espaces tordus. Soit F un corps de nombres. On note $Val(F)$ l'ensemble des places de F et, pour $v \in Val(F)$, F_v le complété de F en v. On note $\mathbb{A}$ l'anneau des adèles de F. Les définitions de 2.1 et 2.3 valent en remplaçant le corps local de ce paragraphe par le corps de nombres F. La seule différence notable est qu'au lieu de fixer un élément $\mathbf{a} \in H^1(W_F; Z(\hat{G}))$, on fixe seulement un élément $\mathbf{a} \in H^1(W_F; Z(\hat{G}))/ker^1(F; Z(\hat{G}))$, où $ker^1(F; Z(\hat{G}))$ est le noyau fini de l'homomorphisme

$$H^1(W_F; Z(\hat{G})) \to \prod_{v \in Val(F)} H^1(W_{F_v}; Z(\hat{G})).$$

On fixe des données $(G, \tilde{G}, \mathbf{a})$. De $\mathbf{a}$ se déduit un homomorphisme continu $\omega : G(\mathbb{A}) \to \mathbb{C}^\times$, trivial sur $G(F)$. Pour tout $v \in Val(F)$, les données $(G, \tilde{G}, \mathbf{a})$ se localisent en des données $(G_v, \tilde{G}_v, \mathbf{a}_v)$ sur F_v. On fixe un espace de Levi minimal $\tilde{M}_0$ de $\tilde{G}$. On note $W^{\tilde{G}}$ le quotient du normalisateur de $\tilde{M}_0$ dans $G(F)$ par son sous-groupe $M_0(F)$. Pour tout espace de Levi $\tilde{M}$, on fixe une mesure de Haar sur $\mathcal{A}_{\tilde{M}} = X_*(A_{\tilde{M}}) \otimes \mathbb{R}$ (cf. 2.9), ces mesures étant soumises à certaines conditions de compatibilité.

On fixe un ensemble fini $V_{ram} \subset Val(F)$, contenant les places archimédiennes et tel que, pour $v \in Val(F) - V_{ram}$, les données $(G_v, \tilde{G}_v, \mathbf{a}_v)$ soient "non ramifiées". On ne précisera pas ici le sens de cette expression. Elle entraîne en tout cas que, pour $v \in Val(F) - V_{ram}$, on peut fixer un sous-groupe compact hyperspécial $K_v \subset G(F_v)$ et un espace hyperspécial associé $\tilde{K}_v \subset \tilde{G}(F_v)$. On entend par là que $\tilde{K}_v$ est un sous-ensemble non vide de $\tilde{G}(F_v)$ et que, pour tout $\gamma \in \tilde{K}_v$, on a les égalités $\tilde{K}_v = K_v\gamma = \gamma K_v$. On fixe de tels objets, auxquels on impose les conditions :

- pour tout $v \in Val(F) - V_{ram}$, K_v est en "bonne position" relativement à $\tilde{M}_0$;
- pour tout $\gamma \in \tilde{G}(F)$, γ appartient à $\tilde{K}_v$ pour presque tout $v \in Val(F) - V_{ram}$ ("presque tout" signifie "sauf pour un nombre fini").

3.2. Intégrales orbitales pondérées. Soit $V \subset Val(F)$ un ensemble fini de places de F. On pose $F_V = \prod_{v \in V} F_v$. Beaucoup d'objets définis dans le cas local ont des analogues définis sur l'anneau F_V. Ils sont obtenus par produit ou tensorisation des objets sur les corps F_v pour $v \in V$. On adapte la notation en conséquence. Par exemple, $\tilde{G}(F_V) = \prod_{v \in V} \tilde{G}(F_v)$ et $I(\tilde{G}(F_V), \omega) = \otimes_{v \in V} I(\tilde{G}(F_v), \omega_v)$.

Soit $\tilde{M}$ un espace de Levi de $\tilde{G}$. Pour $\boldsymbol{\gamma} \in D_{orb}(\tilde{M}(F_V), \omega) \otimes Mes(M(F_V))^*$ et $\mathbf{f} \in I(\tilde{G}(F_V), \omega) \otimes Mes(G(F_V))$, Arthur a défini l'intégrale pondérée ω-équivariante $I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma}, \mathbf{f})$, cf. [6]. La définition ressemble à celle esquissée en 1.8, mais est de nature mi-locale, mi-globale. Elle est locale en ce sens que $\boldsymbol{\gamma}$ et $\mathbf{f}$ sont des produits d'objets locaux. Elle est globale parce que le poids $v_{\tilde{M}}$ et le processus rendant les intégrales ω-équivariantes ne font intervenir que des espaces de Levi définis sur F. Les termes obtenus sont toutefois reliés à

ceux de 1.8 par la relation suivante. On suppose $\boldsymbol{\gamma} = \otimes_{v\in V}\boldsymbol{\gamma}_v$, $\mathbf{f} = \otimes_{v\in V}\mathbf{f}_v$. D'après [6] théorème 8.1, on a

$$I_{\tilde{M}}^{\tilde{G}}(\boldsymbol{\gamma},\mathbf{f}) = \sum_{\tilde{L}^V\in\mathcal{L}(\tilde{M}_V)} d_{\tilde{M}_V}^{\tilde{G}}(\tilde{M},\tilde{L}^V)\prod_{v\in V} I_{\tilde{M}_v}^{\tilde{L}^v}(\boldsymbol{\gamma}_v,\mathbf{f}_{v,\tilde{L}^v}) \tag{3.1}$$

Expliquons cette formule. On a noté $\mathcal{L}(\tilde{M}_V)$ l'ensemble des familles $\tilde{L}^V = (\tilde{L}^v)_{v\in V}$ telles que, pour tout $v\in V$, $\tilde{L}^v$ est un espace de Levi de $\tilde{G}_v$ contenant $\tilde{M}_v$. Pour tout v, $\mathbf{f}_{v,\tilde{L}^v}$ est l'image de $\mathbf{f}_v$ dans $I(\tilde{L}^v(F_v),\omega_v)$ par l'application "terme constant" usuelle dans la théorie de l'induction. Soit $\tilde{L}^V\in\mathcal{L}(\tilde{M}_V)$. On introduit l'espace $\mathcal{A}_{\tilde{M}_V}^{\tilde{G}} = \oplus_{v\in V}(\mathcal{A}_{\tilde{M}_v}/\mathcal{A}_{\tilde{G}})$ et son sous-espace $\mathcal{A}_{\tilde{L}^V} = \oplus_{v\in V}(\mathcal{A}_{\tilde{L}^v}/\mathcal{A}_{\tilde{G}})$. L'espace $\mathcal{A}_{\tilde{M}}^{\tilde{G}} = \mathcal{A}_{\tilde{M}}/\mathcal{A}_{\tilde{G}}$ se plonge diagonalement dans $\mathcal{A}_{\tilde{M}_V}^{\tilde{G}}$, on note $\Delta_V(\mathcal{A}_{\tilde{M}}^{\tilde{G}})$ son image. Le terme $d_{\tilde{M}_V}^{\tilde{G}}(\tilde{M},\tilde{L}^V)$ est nul sauf si

$$\mathcal{A}_{\tilde{M}_V}^{\tilde{G}} = \Delta_V(\mathcal{A}_{\tilde{M}}^{\tilde{G}})\oplus\mathcal{A}_{\tilde{L}^V}^{\tilde{G}}.$$

Si cette égalité est vérifiée, $d_{\tilde{M}_V}^{\tilde{G}}(\tilde{M},\tilde{L}^V)$ est le rapport entre les mesures sur le membre de droite et celle sur le membre de gauche (ces mesures ont été fixées en 2.9 et 3.1).

3.3. La partie géométrique de la formule des traces tordue ω-équivariante. Pour tout $v\in Val(F)-V_{ram}$, notons $\mathbf{1}_{\tilde{K}_v}$ la fonction caractéristique de $\tilde{K}_v$. On note $C_c^\infty(\tilde{G}(\mathbb{A}))$ l'espace de fonctions sur $\tilde{G}(\mathbb{A})$ engendré par les fonctions $f=\otimes_{v\in Val(F)}f_v$, où $f_v\in C_c^\infty(\tilde{G}(F_v))$ pour tout v et $f_v = \mathbf{1}_{\tilde{K}_v}$ pour presque tout $v\notin V_{ram}$. Primitivement, la formule des traces tordue est une égalité $J_{géom}^{\tilde{G}}(f,\omega) = J_{spec}^{\tilde{G}}(f,\omega)$ pour $f\in C_c^\infty(\tilde{G}(\mathbb{A}))$, les deux termes dépendant d'une mesure de Haar sur $\tilde{G}(\mathbb{A})$. Elle a été établie dans [10], voir [15] pour une rédaction plus complète. Les normalisations différant selon les auteurs, nous utilisons précisément celle de [7]. On ne considère ici que la partie géométrique de cette formule. Comme toujours, nous préférons supprimer le choix de la mesure de Haar et remplacer f par $\mathbf{f}\in C_c^\infty(\tilde{G}(\mathbb{A}))\otimes Mes(G(\mathbb{A}))$. On obtient une expression $J_{géom}^{\tilde{G}}(\mathbf{f},\omega)$. Comme en 2.10, elle n'est pas ω-équivariante en $\mathbf{f}$. Fixons un ensemble fini $V\subset Val(F)$ contenant V_{ram}. On identifie $C_c^\infty(\tilde{G}(F_V))$ à un sous-espace de $C_c^\infty(\tilde{G}(\mathbb{A}))$ en identifiant $\mathbf{f}_V\in C_c^\infty(\tilde{G}(F_V))$ à $\mathbf{f}_V\otimes(\otimes_{v\notin V}\mathbf{1}_{\tilde{K}_v})$. Pour $v\notin V$, le groupe $G(F_v)$ est muni d'une mesure "canonique" pour laquelle la masse totale de K_v vaut 1. On identifie $Mes(G(F_V))$ à $Mes(G(\mathbb{A}))$ en tensorisant une mesure sur $G(F_V)$ par le produit de ces mesures canoniques pour $v\notin V$. On sait alors définir $J_{géom}^{\tilde{G}}(\mathbf{f}_V,\omega)$ pour $\mathbf{f}_V\in C_c^\infty(\tilde{G}(F_V))\otimes Mes(G(F_V))$. Arthur a transformé ce terme en une expression $I_{géom}^{\tilde{G}}(\mathbf{f}_V,\omega)$ qui est ω-équivariante, c'est-à-dire qui se descend en une forme linéaire sur $I(\tilde{G}(F_V),\omega)\otimes Mes(G(F_V))$, cf. [7] et [1] paragraphe 2. Décrivons cette expression.

Notons $\tilde{G}_{ss}$ l'ensemble des éléments semi-simples de $\tilde{G}$ et $\tilde{G}_{ss}(F)/conj$ l'ensemble des classes de conjugaison par $G(F)$ dans $\tilde{G}_{ss}(F)$ (une notation analogue sera utilisée plus loin avec F remplacé par F_V). Pour $\mathcal{O}\in\tilde{G}_{ss}(F)/conj$, on définit une certaine distribution $A^{\tilde{G}}(V,\mathcal{O},\omega)\in D_{orb}(\tilde{G}(F_V),\omega)\otimes Mes(G(F_V))^*$, qui vérifie les conditions suivantes

(i) c'est une combinaison linéaire finie d'intégrales orbitales associées aux images dans $\tilde{G}(F_V)$ d'éléments $\gamma\in\tilde{G}(F)$ dont la partie semi-simple appartient à $\mathcal{O}$;

(ii) $A^{\tilde{G}}(V,\mathcal{O},\omega) = 0$ sauf si, pour tout $v \notin V$, la classe de conjugaison dans $\tilde{G}(F_v)$ engendrée par $\mathcal{O}$ coupe $\tilde{K}_v$;

(iii) $A^{\tilde{G}}(V,\mathcal{O},\omega) = 0$ sauf si ω est trivial sur $Z(G;\mathbb{A})^\theta$ et sur $Z(G_\gamma;\mathbb{A})$ pour tout $\gamma \in \mathcal{O}$.

On note $\mathcal{L}(\tilde{M}_0)$ l'ensemble des espaces de Levi de $\tilde{G}$ contenant $\tilde{M}_0$. Pour une fonction $\mathbf{f}_V \in I(\tilde{G}(F_V),\omega) \otimes Mes(G(F_V))$, on a alors l'égalité

$$I^{\tilde{G}}_{géom}(\mathbf{f}_V,\omega) = \sum_{\tilde{M}\in\mathcal{L}(\tilde{M}_0)} |W^{\tilde{M}}||W^{\tilde{G}}|^{-1} \sum_{\mathcal{O}\in\tilde{M}_{ss}(F)/conj} I^{\tilde{G}}_{\tilde{M}}(A^{\tilde{M}}(V,\mathcal{O},\omega),\mathbf{f}_V) \qquad (3.2)$$

Pour $\mathbf{f}_V$ fixé, il n'y a qu'un nombre fini de termes non nuls. Soulignons que $I^{\tilde{G}}_{géom}(\mathbf{f}_V,\omega)$ et les distributions $A^{\tilde{M}}(V,\mathcal{O},\omega)$ dépendent des espaces $\tilde{K}_v$ pour $v \notin V$, bien que ceux-ci ne figurent pas dans la notation.

Donnons quelques précisions sur la distribution $A^{\tilde{G}}(V,\mathcal{O},\omega)$. Elle dépend de la classe de conjugaison $\mathcal{O} \in \tilde{G}_{ss}(F)/conj$ et de l'ensemble fini $V \supset V_{ram}$. Pour $\mathcal{O}$ fixé, on peut faire varier V. Pour $V \subset V'$, $A^{\tilde{G}}(V,\mathcal{O},\omega)$ et $A^{\tilde{G}}(V',\mathcal{O},\omega)$ sont reliés par une formule qui fait intervenir les intégrales orbitales pondérées (non ω-équivariantes) des fonctions $\mathbf{1}_{\tilde{K}_v}$ pour $v \in V' - V$. Considérons le cas particulier d'une classe $\mathcal{O}$ formée d'éléments fortement réguliers et elliptiques (c'est-à-dire que $\mathcal{O} \cap \tilde{M} = \emptyset$ pour tout espace de Levi $\tilde{M} \subsetneq \tilde{G}$). Supposons que cette classe vérifie les conditions (ii) et (iii). Fixons $\gamma \in \mathcal{O}$. Pour $v \notin V$, on peut d'après (ii) fixer $x_v \in G(F_v)$ tel que $x_v^{-1}\gamma x_v \in \tilde{K}_v$. On voit que l'on peut supposer $x_v = 1$ pour presque tout v. Notons x l'élement de $G(\mathbb{A})$ dont les composantes dans $G(F_v)$ sont x_v si $v \notin V$ et 1 si $v \in V$. Fixons aussi une mesure de Haar sur $G_\gamma(\mathbb{A})$. Par un procédé habituel, on définit un sous-groupe $\mathfrak{A}_{\tilde{G}} \subset A_{\tilde{G}}(\mathbb{A})$ isomorphe à $\mathcal{A}_{\tilde{G}}$ (si $F = \mathbb{Q}$, $\mathfrak{A}_{\tilde{G}}$ est la composante neutre de $A_{\tilde{G}}(\mathbb{R})$ pour la topologie réelle). Supposons V "assez grand", cette condition dépendant de $\mathcal{O}$. De même que ci-dessus, la mesure sur $G_\gamma(\mathbb{A})$ s'identifie à une mesure sur $G_\gamma(F_V)$. Soit dg une mesure de Haar sur $G(F_V)$ et $f_V \in C_c^\infty(\tilde{G}(F_V))$. Pour V assez grand, on a l'égalité

$$I^{\tilde{G}}(A^{\tilde{G}}(V,\mathcal{O},\omega), f_V \otimes dg) = [Z_G(\gamma;F) : G_\gamma(F)]^{-1}\omega(x) mes(\mathfrak{A}_{\tilde{G}}G_\gamma(F)\backslash G_\gamma(\mathbb{A}))$$
$$\int_{G_\gamma(F_V)\backslash G(F_V)} \omega(y) f_V(y^{-1}\gamma y)\, dy,$$

où dy se déduit des deux mesures fixées.

On a une application naturelle $\tilde{G}_{ss}(F)/conj \to \tilde{G}_{ss}(F_V)/conj$. Pour une classe de conjugaison $\mathcal{O}_V \in \tilde{G}_{ss}(F_V)/conj$, posons

$$A^{\tilde{G}}(\mathcal{O}_V,\omega) = \sum_{\mathcal{O}\in\tilde{G}_{ss}(F)/conj, \mathcal{O}\mapsto\mathcal{O}_V} A^{\tilde{G}}(V,\mathcal{O},\omega).$$

On peut reformuler l'égalité (3.2) en

$$I^{\tilde{G}}_{géom}(\mathbf{f}_V,\omega) = \sum_{\tilde{M}\in\mathcal{L}(\tilde{M}_0)} |W^{\tilde{M}}||W^{\tilde{G}}|^{-1} \sum_{\mathcal{O}_V\in\tilde{M}_{ss}(F_V)/conj} I^{\tilde{G}}_{\tilde{M}}(A^{\tilde{M}}(\mathcal{O}_V,\omega),\mathbf{f}_V).$$

3.4. Endoscopie. La notion de donnée endoscopique pour $(G,\tilde{G},\mathbf{a})$ se définit comme en 2.4. La seule différence est que l'on impose que le cocycle a de ce paragraphe a pour image la classe $\mathbf{a}$ dans $H^1(W_F;Z(\hat{G}))/ker^1(F;Z(\hat{G}))$. Soit $\mathbf{G}'=(G',\mathcal{G}',\tilde{s})$ une telle donnée endoscopique. On dit encore qu'elle est elliptique si $Z(\hat{G}')^{\Gamma_F,0}=Z(\hat{G})^{\Gamma_F,\hat{\theta},0}$. Il est associé à $\mathbf{G}'$ un espace tordu $\tilde{G}'$ comme en 2.5. On définit une constante $i(\tilde{G},\tilde{G}')$ de la façon suivante. Elle est nulle si $\mathbf{G}'$ n'est pas elliptique. Supposons $\mathbf{G}'$ elliptique. Notons $Aut(\mathbf{G}')$ le groupe des $x\in\hat{G}$ tels que $x\mathcal{G}'x^{-1}=\mathcal{G}'$ et $x\tilde{s}x^{-1}\in Z(\hat{G})\tilde{s}$. Notons

$$\tau(G)=|\pi_0(Z(\hat{G})^{\Gamma_F})||ker^1(F;Z(\hat{G}))|^{-1},$$

où π_0 désigne le groupe des composantes connexes. On pose

$$i(\tilde{G},\tilde{G}')=|det((1-\theta)_{|\mathcal{A}_G/\mathcal{A}_{\tilde{G}}})|^{-1}|\pi_0(Aut(\mathbf{G}'))|^{-1}\tau(G)\tau(G')^{-1}$$

$$|\pi_0(Z(\hat{G})^{\Gamma_F,0}\cap Z(\hat{G}'))|.$$

Pour tout $v\in Val(F)$, la donnée se localise en une donnée $\mathbf{G}'_v=(G'_v,\mathcal{G}'_v,\tilde{s})$ de $(G_v,\tilde{G}_v,\mathbf{a}_v)$, où $\mathbf{a}_v$ est l'image de $\mathbf{a}$ dans $H^1(W_{F_v};Z(\hat{G}))$. On dit que $\mathbf{G}'$ est relevante si $\tilde{G}'(F)\neq\emptyset$ et si, pour toute place $v\in Val(F)$, $\mathbf{G}'_v$ est relevante. Soit V un ensemble fini de places de F contenant V_{ram}. On dit que $\mathbf{G}'$ est non ramifiée hors de V si le sous-groupe d'inertie $I_v\subset W_{F_v}$ (qui est un sous-groupe de ${}^LG_v$) est contenu dans $\mathcal{G}'_v$. Si $\mathbf{G}'$ est non ramifiée hors de V, on montre que les conditions imposées à V_{ram} (que l'on n'a pas explicitées) valent aussi pour l'ensemble de places V et pour le couple $(G',\tilde{G}')$ (ou, si l'on préfère, pour le triplet obtenu en complétant ce couple par le cocycle $\mathbf{a}'$ trivial). Pour $v\in Val(F)-V$, le choix que l'on a fait de l'espace hyperspécial $\tilde{K}_v$ détermine un espace hyperspécial $\tilde{K}'_v\subset\tilde{G}'(F_v)$. Plus exactement, cet espace n'est déterminé qu'à conjugaison près par $G'_{AD}(F_v)$, où G'_{AD} est le groupe adjoint de G' mais les constructions ultérieures seront insensibles à une telle conjugaison. On fixe ainsi une famille $(\tilde{K}'_v)_{v\notin V}$, à laquelle on impose, comme il est loisible, la même condition de compatibilité globale qu'en 3.1.

Considérons une donnée endoscopique $\mathbf{G}'=(G',\mathcal{G}',\tilde{s})$ relevante et non ramifiée hors de V. On fixe des données auxiliaires G'_1, $\tilde{G}'_1$, C_1 et $\hat{\xi}_1$ comme en 2.7, ces données étant maintenant définies sur F. On leur impose, comme il est loisible, certaines conditions de non ramification hors de V. Ces conditions impliquent que l'on peut fixer pour tout $v\notin V$ un espace hyperspécial $\tilde{K}'_{1,v}\subset\tilde{G}'_1(F_v)$ qui se projette sur $\tilde{K}'_v$. On impose à ces données la même condition de compatibilité globale qu'en 3.1. En tensorisant les constructions de 2.7 sur les places $v\in V$, on définit des espaces $C^\infty_{c,\lambda_1}(\tilde{G}'_1(F_V))$, $I_{\lambda_1}(\tilde{G}'_1(F_V))$ etc..., ainsi que leurs avatars canoniques $C^\infty_c(\mathbf{G}'_V)$, $I(\mathbf{G}'_V)$ etc... Pour identifier par exemple $C^\infty_{c,\lambda_1}(\tilde{G}'_1(F_V))$ à $C^\infty_c(\mathbf{G}'_V)$, on doit choisir pour tout $v\in V$ un facteur de transfert $\Delta_{1,v}$. En fait, l'identification ne dépend que du produit $\Delta_{1,V}=\otimes_{v\in V}\Delta_{1,v}$. Un point important est que le choix que l'on vient de faire des espaces $\tilde{K}'_{1,v}$ pour $v\notin V$ détermine un tel facteur $\Delta_{1,V}$. En effet, pour tout $v\notin V$, le choix de $\tilde{K}'_{1,v}$ détermine un facteur de transfert $\Delta_{1,v}$. D'autre part, on peut définir canoniquement un facteur de transfert global, cf. [11] lemme 7.3A, [14] paragraphe IV.2. C'est-à-dire, considérons des éléments $\delta_1=(\delta_{1,v})_{v\in Val(F)}\in\tilde{G}'_1(\mathbb{A})$ et $\gamma=(\gamma_v)_{v\in Val(F)}\in\tilde{G}(\mathbb{A})$. Supposons que, pour tout v, γ_v soit fortement régulier et que $\delta_{1,v}$ et γ_v se correspondent. Imposons de plus à δ et γ une certaine condition de non ramification (impliquant que $\Delta_{1,v}(\delta_{1,v},\gamma_v)=1$ pour presque tout $v\notin V$). Alors on peut définir un facteur global $\Delta_1(\delta_1,\gamma)$. On normalise le facteur $\Delta_{1,V}$ de sorte que, pour de tels δ_1 et γ,

on ait l'égalité

$$\Delta_1(\delta_1,\gamma)=\Delta_{1,V}(\delta_{1,V},\gamma_V)\prod_{v\notin V}\Delta_{1,v}(\delta_{1,v},\gamma_v),$$

où $\delta_{1,V}=(\delta_{1,v})_{v\in V}$ et $\gamma_V=(\gamma_v)_{v\in V}$. Soit $\mathcal{O}'_V\in\tilde{G}'_{ss}(F_V)/conj$. Les constructions de 3.3 s'adaptent et on définit un élément $A^{\tilde{G}'_1}_{\lambda_1}(\mathcal{O}'_V)$. C'est une combinaison linéaire d'intégrales orbitales vues comme des formes linéaires sur $I_{\lambda_1}(\tilde{G}'_1(F_V))\otimes Mes(G'(F_V))$. A l'aide du facteur $\Delta_{1,V}$ ci-dessus, on l'identifie à un élément $A^{\mathbf{G}'}(\mathcal{O}'_V)\in D_{orb}(\mathbf{G}')\otimes Mes(G'(F))^*$. Le terme $A^{\tilde{G}'_1}_{\lambda_1}(\mathcal{O}'_V)$ dépend des choix des $\tilde{K}'_{1,v}$ mais le facteur $\Delta_{1,V}$ aussi. On voit alors que $A^{\mathbf{G}'}(\mathcal{O}'_V)$ ne dépend plus que des $\tilde{K}'_v$. On montre qu'il ne dépend pas du choix des données auxiliaires $G'_1,\dots,\hat{\xi}_1$.

L'ensemble des classes d'équivalence de données endoscopiques elliptiques, relevantes et non ramifiées hors de V est fini, on en fixe un ensemble de représentants $\mathcal{E}(\tilde{G},\mathbf{a},V)$.

3.5. Intégrales orbitales pondérées stables. On suppose G quasi-déployé, $\tilde{G}$ à torsion intérieure et $\mathbf{a}=1$. Soient $\tilde{M}$ un espace de Levi de $\tilde{G}$ et V un ensemble fini de places de F contenant V_{ram}. Pour une place archimédienne v de F, on introduit les espaces de distributions $D_{tr-orb}(\tilde{M}(F_v))$ et $D^{st}_{tr-orb}(\tilde{M}(F_v))$ évoqués en 2.13. Pour une place v non-archimédienne, on définit ces espaces comme étant simplement égaux à $D_{orb}(\tilde{M}(F_v))$, resp. $D^{st}_{orb}(\tilde{M}(F_v))$. On définit les produits tensoriels de ces espaces sur les places $v\in V$, que l'on note $D_{tr-orb}(\tilde{M}(F_V))$ et $D^{st}_{tr-orb}(\tilde{M}(F_V))$. La définition des intégrales orbitales pondérées invariantes $I^{\tilde{G}}_{\tilde{M}}(\boldsymbol{\gamma}_V,\mathbf{f}_V)$ de 3.2 s'étend au cas où $\boldsymbol{\gamma}_V$ appartient à $D_{tr-orb}(\tilde{M}(F_V))\otimes Mes(M(F_V))^*$. Cela étant, pour

$$\boldsymbol{\delta}_V\in D^{st}_{tr-orb}(\tilde{M}(F_V))\otimes Mes(M(F_V))^*\text{ et }\mathbf{f}_V\in I(\tilde{G}(F_V))\otimes Mes(G(F_V)),$$

on définit l'intégrale orbitale pondérée stable par la même formule qu'en 2.11 :

$$S^{\tilde{G}}_{\tilde{M}}(\boldsymbol{\delta}_V,\mathbf{f}_V)=I^{\tilde{G}}_{\tilde{M}}(\boldsymbol{\delta}_V,\mathbf{f}_V)-\sum_{s\in Z(\hat{M})^{\Gamma_F}/Z(\hat{G})^{\Gamma_F},s\neq 1}i_{\tilde{M}}(\tilde{G},\tilde{G}'(s))S^{\mathbf{G}'(s)}_{\mathbf{M}}(\boldsymbol{\delta}_V,\mathbf{f}^{\mathbf{G}'(s)}_V).$$

Comme dans ce paragraphe, cette définition n'est légitime que grâce au résultat suivant.

Proposition 3.1. *Pour tout élément $\boldsymbol{\delta}_V\in D^{st}_{tr-orb}(\tilde{M}(F_V))\otimes Mes(M(F_V))^*$, la forme linéaire $\mathbf{f}_V\mapsto S^{\tilde{G}}_{\tilde{M}}(\boldsymbol{\delta}_V,\mathbf{f}_V)$ est stable, c'est-à-dire se descend en une forme linéaire sur $SI(\tilde{G}(F_V))\otimes Mes(G(F_V))$.*

Cela se déduit du théorème 2.11 car on montre que les intégrales $S^{\tilde{G}}_{\tilde{M}}(\boldsymbol{\delta}_V,\mathbf{f}_V)$ s'expriment à l'aide des intégrales locales de ce paragraphe par une formule parallèle à (3.1).

3.6. Coefficients stables. On suppose G quasi-déployé, $\tilde{G}$ à torsion intérieure et $\mathbf{a}=1$. Soit V un ensemble fini de places de F contenant V_{ram}. Pour $\mathcal{O}_V\in\tilde{G}_{ss}(F_V)/conj$, on a défini en 3.3 la distribution $A^{\tilde{G}}(\mathcal{O}_V)\in D_{orb}(\tilde{G}(F_V))\otimes Mes(G(F_V))^*$. Pour une réunion finie $\mathcal{O}_V=\sqcup_{i=1,\dots,n}\mathcal{O}_{V,i}$ de telles classes, on pose $A^{\tilde{G}}(\mathcal{O}_V)=\sum_{i=1,\dots,n}A^{\tilde{G}}(\mathcal{O}_{V,i})$. Notons $\tilde{G}_{ss}(F_V)/st-conj$ l'ensemble des classes de conjugaison stable dans $\tilde{G}_{ss}(F_V)$. Pour un élément $\mathcal{O}_V$ de cet ensemble, ou pour une réunion finie de telles classes, on définit un

élément $SA^{\tilde{G}}(\mathcal{O}_V) \in D_{tr-orb}(\tilde{G}(F_V)) \otimes Mes(G(F_V))^*$ par la formule de récurrence

$$SA^{\tilde{G}}(\mathcal{O}_V) = A^{\tilde{G}}(\mathcal{O}_V) - \sum_{\mathbf{G}' \in \mathcal{E}(\tilde{G},\mathbf{a},V), G' \neq G} i(\tilde{G},\tilde{G}') transfert(SA^{\mathbf{G}'}(\mathcal{O}_V^{\tilde{G}'})).$$

Expliquons cette formule. Pour $\mathbf{G}' \in \mathcal{E}(\tilde{G},\mathbf{a},V)$, on note $\mathcal{O}_V^{\tilde{G}'}$ la réunion des classes de conjugaison stable dans $\tilde{G}'_{ss}(F_V)$ correspondant à une classe de conjugaison stable dans $\tilde{G}(F_V)$ contenue dans $\mathcal{O}_V$. Cette réunion est finie. Si $G' \neq G$, fixons des données auxiliaires G'_1,...,$(\tilde{K}'_{1,v})_{v \notin V}$ comme en 3.4. On peut supposer par récurrence sur $dim(G_{SC})$ que l'on a défini la variante $SA^{\tilde{G}'_1}_{\lambda_1}(\mathcal{O}_V^{\tilde{G}'})$ de notre distribution. Par le même procédé qu'en 3.4, elle s'identifie à un élément $SA^{\mathbf{G}'}(\mathcal{O}_V^{\tilde{G}'}) \in D_{tr-orb}(\mathbf{G}'_V) \otimes Mes(G'(F_V))^*$. On montre que cette distribution ne dépend plus des choix des espaces $\tilde{K}'_v$ pour $v \notin V$, mais seulement des $\tilde{K}_v$, lesquels sont fixés une fois pour toutes. Le théorème suivant montre que ces distributions sont stables, on peut donc les transférer.

Théorème 3.2. *Pour tout $\mathcal{O}_V \in \tilde{G}_{ss}(F_V)/st-conj$, la distribution $SA^{\tilde{G}}(\mathcal{O}_V)$ est stable.*

On voit par récurrence que la distribution $SA^{\tilde{G}}(\mathcal{O}_V)$ est supportée par les éléments de $\tilde{G}(F_V)$ dont la partie semi-simple appartient à $\mathcal{O}_V$.

3.7. La formule stable. On suppose G quasi-déployé, $\tilde{G}$ à torsion intérieure et $\mathbf{a} = 1$. Soit V un ensemble fini de places de F contenant V_{ram}. Pour $\mathbf{f}_V \in SI(\tilde{G}(F_V)) \otimes Mes(G(F_V))$, on pose

$$S^{\tilde{G}}_{géom}(\mathbf{f}_V) = \sum_{\tilde{M} \in \mathcal{L}(\tilde{M}_0)} |W^{\tilde{M}}||W^{\tilde{G}}|^{-1} \sum_{\mathcal{O}_V \in \tilde{M}_{ss}(F_V)/st-conj} S^{\tilde{G}}_{\tilde{M}}(SA^{\tilde{M}}(\mathcal{O}_V), \mathbf{f}_V).$$

On montre que, pour $\mathbf{f}_V$ fixé, il n'y a dans cette somme qu'un nombre fini de termes non nuls.

3.8. Le théorème principal. Le triplet $(G,\tilde{G},\mathbf{a})$ est ici quelconque. Soit V un ensemble fini de places contenant V_{ram}.

Théorème 3.3. *Pour tout $\mathbf{f}_V \in I(\tilde{G}(F_V),\omega) \otimes Mes(G(F_V))$, on a l'égalité*

$$I_{géom}(\mathbf{f}_V,\omega) = \sum_{\mathbf{G}' \in \mathcal{E}(\tilde{G},\mathbf{a},V)} i(\tilde{G},\tilde{G}') S^{\mathbf{G}'}_{géom}(\mathbf{f}_V^{\mathbf{G}'}).$$

La démonstration de ce théorème est très longue. En particulier, on doit utiliser la partie spectrale de la formule des traces, dont on n'a pas du tout parlé ici.

3.9. Endoscopie non standard. A plusieurs reprises, on utilise dans les preuves la méthode de descente d'Harish-Chandra. Dans le cas tordu, cette méthode appliquée à l'endoscopie fait apparaître des "triplets endoscopiques non standard". Plusieurs de nos assertions ont des contreparties pour de tels triplets. Indiquons-en une. Considérons deux groupes réductifs connexes G_1 et G_2 définis et quasi-déployés sur F. On les suppose simplement connexes. Pour $i = 1,2$, on fixe une paire de Borel (B_i,T_i) de G_i définie sur F, on note $\Sigma(T_i)$ l'ensemble des racines de T_i dans $\mathfrak{g}_i$ et $\check{\Sigma}(T_i)$ l'ensemble des coracines. On suppose donnés

un isomorphisme $j : \mathfrak{t}_1 \to \mathfrak{t}_2$ et une bijection $\tau : \Sigma(T_2) \to \Sigma(T_1)$ équivariants pour les actions galoisiennes. Il se déduit de τ une bijection $\check{\tau} : \check{\Sigma}(T_1) \to \check{\Sigma}(T_2)$ entre ensembles de coracines. On suppose que, pour tout $\check{\alpha} \in \check{\Sigma}(T_1)$, $j(\check{\alpha})$ est un multiple rationnel positif de $\check{\tau}(\check{\alpha})$ et que, pour tout $\alpha \in \Sigma(T_2)$, l'application duale j^* envoie α sur un multiple rationnel positif de $\tau(\alpha)$. L'exemple le plus frappant est le cas où $G_1 = Sp(2n)$, $G_2 = Spin(2n+1)$ et j envoie une coracine courte sur une coracine longue et une coracine longue sur deux fois une coracine courte .

Soit $v \in Val(F)$. L'isomorphisme j permet de définir une correspondance bijective entre classes de conjugaison stable semi-simples dans les algèbres de Lie $\mathfrak{g}_1(F_v)$ et $\mathfrak{g}_2(F_v)$. On peut alors définir un transfert similaire à celui de 2.7, mais au niveau des algèbres de Lie. L'analogue du "facteur de transfert" vaut 1 sur deux éléments qui se correspondent. Par l'exponentielle, on peut remonter le transfert aux groupes si on se limite à des fonctions ou des distributions à support assez proche de l'élément neutre. Pour $i = 1, 2$, on s'intéresse particulièrement à l'espace des distributions stables sur $\tilde{G}_i(F_v)$ qui sont à support unipotent. On le note $D^{st}_{unip}(\tilde{G}_i(F_v))$. Pour tout ensemble fini V de places de F, le transfert se restreint en un isomorphisme

$$D^{st}_{unip}(\tilde{G}_1(F_V)) \otimes Mes(G_1(F_V))^* \simeq D^{st}_{unip}(\tilde{G}_2(F_V)) \otimes Mes(G_2(F_V))^* \tag{3.3}$$

Soit V un ensemble fini de places de $Val(F)$, contenant les places archimédiennes et assez grand pour que, pour $v \notin V$, le triplet localisé $(G_{1,v}, G_{2,v}, j)$ vérifie une certaine condition de non ramification. Pour $i = 1, 2$, on applique les définitions des paragraphes précédents en prenant $G = G_i$, $\tilde{G} = G_i$ et $\mathbf{a} = 1$. Ainsi, pour $\mathcal{O}_V \in G_{i,ss}(F_V)/st - conj$, on dispose d'une distribution $SA^{G_i}(\mathcal{O}_V)$. Il s'avère qu'elle ne dépend pas des choix des sous-groupes hyperspéciaux hors de V (ni des espaces hyperspéciaux, qui, ici, sont forcément égaux à ces groupes). Pour $\mathcal{O}_V = \{1\}$, on note plutôt $SA^{G_i}_{unip}(V)$ cette distribution. C'est un élément de $D^{st}_{unip}(\tilde{G}_i(F_V)) \otimes Mes(G_i(F_V))^*$.

Théorème 3.4. *Les éléments $SA^{G_1}_{unip}(V)$ et $SA^{G_2}_{unip}(V)$ se correspondent par l'isomorphisme* (3.3).

Références

[1] Arthur, J., *A stable trace formula I. General expansions*, Journal of the Inst. of Math. Jussieu **1** (2002), 175–277.

[2] _____, *A stable trace formula II. Global descent*, Inventiones math. **143** (2001), 157–220.

[3] _____, *A stable trace formula III. Proof of the main theorems*, Annals of Math. **158** (2003), 769–873.

[4] _____, *The endoscopic classification of representations ; orthogonal and symplectic groups*, AMS Colloquium Publ. **61** (2013).

[5] _____, *The local behaviour of weighted orbital integrals*, Duke Math. J. **56** (1988), 223–293.

[6] ———, *The invariant trace formula I. Local theory*, Journal of the AMS **1** (1988), 323–283.

[7] ———, *The invariant trace formula II. Global theory*, Journal of the AMS **1** (1988), 501–554.

[8] ———, *On the transfer of distributions : weighted orbital integrals*, Duke Math. J. **99** (1999), 209–283.

[9] Arthur, J. and Clozel, L., *Simple algebras, base change, and the advanced theory of the trace formula*, Annals of Math. Studies **120** (1989)

[10] Clozel, L., Labesse, J.-P., and Langlands, R. P., *Friday morning seminar on the trace formula*, Institute for Advanced Study (1984)

[11] Kottwitz, R. and Shelstad, D., *Foundations of twisted endoscopy*, Astérisque **255** (1999).

[12] ———, *On splitting invariants and sign conventions in endoscopic transfer*, preprint (2012).

[13] Labesse, J.-P., *Cohomologie, stabilisation et changement de base*, Astérisque **257** (1999).

[14] ———, *Stable twisted trace formula : elliptic terms*, Journal of the Inst. of Math. Jussieu **3** (2004), 473–530.

[15] Labesse, J.-P. and Waldspurger, J.-L., *La formule des traces tordue d'après le Friday Morning Seminar*, CRM Monograph Series **31** (2013).

[16] Langlands, R.P., *Base change for $GL(2)$*, Annals of Math Studies **96** (1980).

[17] Ngô Bao Chau, *Le lemme fondamental pour les algèbres de Lie*, Publ. Math. IHES **111** (2010), 1–269.

[18] Shelstad, D., *On geometric transfer in real twisted endoscopy*, Annals of Math. **176** (2012), 1919–1985.

[19] Waldspurger, J.-L., *Stabilisation de la formule des traces tordue I : endoscopie tordue sur un corps local*, prépublication (2014).

[20] ———, *L'endoscopie tordue n'est pas si tordue*, Memoirs of the AMS **908** (2008).

Institut de mathématiques de Jussieu, 2 place Jussieu, 75005 Paris
E-mail: jean-loup.waldspurger@imj-prg.fr

∗. This paper is accepted and printed in French due to a strong request of the author.

Translation invariance, exponential sums, and Waring's problem

Trevor D. Wooley

Abstract. We describe mean value estimates for exponential sums of degree exceeding 2 that approach those conjectured to be best possible. The vehicle for this recent progress is the *efficient congruencing method*, which iteratively exploits the translation invariance of associated systems of Diophantine equations to derive powerful congruence constraints on the underlying variables. There are applications to Weyl sums, the distribution of polynomials modulo 1, and other Diophantine problems such as Waring's problem.

Mathematics Subject Classification (2010). Primary 11L15; Secondary 11P05.

Keywords. Exponential sums, Waring's problem, Hardy-Littlewood method.

1. Introduction

Although pivotal to the development of vast swathes of analytic number theory in the twentieth century, the differencing methods devised by Weyl [54] and van der Corput [15] are in many respects unsatisfactory. In particular, they improve on the trivial estimate for an exponential sum by a margin exponentially small in terms of its degree. The method introduced by Vinogradov [50, 51] in 1935, based on mean values, is rightly celebrated as a great leap forward, replacing this exponentially weak margin by one polynomial in the degree. Nonetheless, Vinogradov's methods yield bounds removed from the sharpest conjectured to hold by a margin at least logarithmic in the degree, a defect that has endured for six decades since the era in which these ideas were comprehensively analysed. In this report, we describe progress since 2010 that eliminates this defect, placing us within a whisker of establishing in full the main conjecture of the subject.

When $k, s \in \mathbb{N}$ and $\boldsymbol{\alpha} \in \mathbb{R}^k$, consider the exponential sum

$$f_k(\boldsymbol{\alpha}; X) = \sum_{1 \le x \le X} e(\alpha_1 x + \ldots + \alpha_k x^k) \tag{1.1}$$

and the mean value

$$J_{s,k}(X) = \oint |f_k(\boldsymbol{\alpha}; X)|^{2s} \, \mathrm{d}\boldsymbol{\alpha}. \tag{1.2}$$

Here, as usual, we write $e(z)$ for $e^{2\pi i z}$. Also, to save clutter, when $G : [0,1)^k \to \mathbb{C}$ is integrable, we write $\oint G(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha} = \int_{[0,1]^k} G(\boldsymbol{\alpha}) \, \mathrm{d}\boldsymbol{\alpha}$. By orthogonality, one sees that $J_{s,k}(X)$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

counts the number of integral solutions of the system of equations

$$x_1^j + \ldots + x_s^j = y_1^j + \ldots + y_s^j \quad (1 \le j \le k), \tag{1.3}$$

with $1 \le x_i, y_i \le X$ $(1 \le i \le s)$. Upper bounds for $J_{s,k}(X)$ are known collectively as *Vinogradov's mean value theorem*. We now focus discussion by recording the classical version of this theorem that emerged from the first half-century of refinements following Vinogradov's seminal paper [50] (see in particular [27, 33, 51]), culminating in the papers of Karatsuba [29] and Stechkin [42].

Theorem 1.1. *There is an absolute constant $A > 0$ having the property that, whenever s, r and k are natural numbers with $s \ge rk$, then*

$$J_{s,k}(X) \le C(k,r)X^{2s-k(k+1)/2+\Delta_{s,k}}, \tag{1.4}$$

where $\Delta_{s,k} = \frac{1}{2}k^2(1-1/k)^r$ and $C(k,r) = \min\{k^{Ask}, k^{Ak^3}\}$.

We will not concern ourselves with the dependence on s and k of constants such as $C(k,r)$ appearing in bounds for $J_{s,k}(X)$ and its allies (but see [57] for improvements in this direction). Although significant in applications to the zero-free region of the Riemann zeta function, this is not relevant to those central to this paper. Thus, implicit constants in the notation of Landau and Vinogradov will depend at most on s, k and ε, unless otherwise indicated[1].

When $k \ge 2$, the exponent $\Delta_{s,k}$ of Theorem 1.1 satisfies $\Delta_{s,k} \le k^2e^{-s/k^2}$, and so $\Delta_{s,k} = O(1/\log k)$ for $s \ge k^2(2\log k + \log\log k)$. One can refine (1.4) to obtain an asymptotic formula when s is slightly larger (see [2, Theorem 3.9], for example).

Theorem 1.2. *Let $k, s \in \mathbb{N}$ and suppose that $s \ge k^2(2\log k + \log\log k + 5)$. Then there exists a positive number $\mathfrak{C}(s,k)$ with $J_{s,k}(X) \sim \mathfrak{C}(s,k)X^{2s-k(k+1)/2}$.*

With these theorems in hand, we consider the motivation for investigating the sums $f_k(\boldsymbol{\alpha}; X)$. Many number-theoretic functions may be estimated in terms of such sums. Thus, when $\mathrm{Re}(s)$ is close to 1, estimates for the Riemann zeta function $\zeta(s)$ stem from partial summation and Taylor expansions for $\log(1+x/N)$, since

$$\sum_{N<n\le N+X} n^{-it} = N^{-it} \sum_{1\le x\le X} e\Big(-\frac{t}{2\pi}\log(1+x/N)\Big).$$

On the other hand, specialisations of $f_k(\boldsymbol{\alpha}; X)$ arise naturally in applications of interest. Indeed, work on the asymptotic formula in Waring's problem depends on the sum obtained by setting $\alpha_1 = \ldots = \alpha_{k-1} = 0$ and $\alpha_k = \beta$, namely

$$g_k(\beta; X) = \sum_{1\le x\le X} e(\beta x^k). \tag{1.5}$$

[1]Given a complex-valued function $f(t)$ and positive function $g(t)$, we use Vinogradov's notation $f(t) \ll g(t)$, or Landau's notation $f(t) = O(g(t))$, to mean that there is a positive number C for which $f(t) \le Cg(t)$ for all large enough values of t. Also, we write $f(t) \gg g(t)$ when $g(t) \ll f(t)$. If C depends on certain parameters, then we indicate this by appending these as subscripts to the notation. Also, we write $f(t) = o(g(t))$ when $f(t)/g(t) \to 0$ as $t \to \infty$. Finally, we use the convention that whenever ε occurs in a statement, then the statement holds for each fixed $\varepsilon > 0$.

Writing $R_{s,k}(n)$ for the number of representations of n as the sum of s positive integral kth powers, one finds by orthogonality that

$$R_{s,k}(n) = \int_0^1 g_k(\beta; n^{1/k})^s e(-\beta n)\,\mathrm{d}\beta.$$

The uninitiated reader will wonder why one should focus on estimates for the mean value $J_{s,k}(X)$ when many applications depend on pointwise estimates for $f_k(\boldsymbol{\alpha}; X)$. Vinogradov observed that mean value estimates suffice to obtain useful pointwise estimates for $f_k(\boldsymbol{\alpha}; X)$. To see why this is the case, note first that $|f_k(\boldsymbol{\beta}; X)|$ differs little from $|f_k(\boldsymbol{\alpha}; X)|$ provided that the latter is large, and in addition $|\beta_j - \alpha_j|$ is rather smaller than X^{-j} for each j, so that $\boldsymbol{\beta}$ lies in a small neighbourhood of $\boldsymbol{\alpha}$ having measure of order $X^{-k(k+1)/2}$. Second, one sees from (1.1) that for each integer h the sum $f_k(\boldsymbol{\alpha}; X)$ may be rewritten in the form

$$f_k(\boldsymbol{\alpha}; X) = \sum_{1-h \le x \le X-h} e\big(\alpha_1(x+h) + \ldots + \alpha_k(x+h)^k\big).$$

By estimating the tails of this sum and applying the binomial theorem to identify the coefficient of each monomial x^j, one obtains new k-tuples $\boldsymbol{\alpha}^{(h)}$ for which $f_k(\boldsymbol{\alpha}; X) = f_k(\boldsymbol{\alpha}^{(h)}; X) + O(|h|)$. These ideas combine to show that one large value $|f_k(\boldsymbol{\alpha}; X)|$ generates a collection of neighbourhoods $\mathfrak{B}(h)$, with the property that whenever $\boldsymbol{\beta} \in \mathfrak{B}(h)$, then $|f_k(\boldsymbol{\beta}; X)|$ is almost as large as $|f_k(\boldsymbol{\alpha}; X)|$. Given N disjoint such neighbourhoods over which to integrate $|f_k(\boldsymbol{\beta}; X)|^{2s}$, non-trivial estimates for $|f_k(\boldsymbol{\alpha}; X)|$ follow from the relation

$$NX^{-k(k+1)/2}|f_k(\boldsymbol{\alpha}; X)|^{2s} \ll J_{s,k}(X).$$

This circle of ideas leads to the following theorem (see [11] and [47, Theorem 5.2]).

Theorem 1.3. *Let k be an integer with $k \ge 2$, and let $\boldsymbol{\alpha} \in \mathbb{R}^k$. Suppose that there exists a natural number j with $2 \le j \le k$ such that, for some $a \in \mathbb{Z}$ and $q \in \mathbb{N}$ with $(a,q) = 1$, one has $|\alpha_j - a/q| \le q^{-2}$. Then one has*

$$f_k(\boldsymbol{\alpha}; X) \ll \Big(X^{k(k-1)/2} J_{s,k-1}(2X)(q^{-1} + X^{-1} + qX^{-j})\Big)^{1/(2s)} \log(2X).$$

To illustrate the power of this theorem, suppose that k is large, and β satisfies the condition that, whenever $b \in \mathbb{Z}$ and $q \in \mathbb{N}$ satisfy $(b,q) = 1$ and $|q\beta - b| \le X^{1-k}$, then $q > X$. By substituting the conclusion of Theorem 1.1 into Theorem 1.3, one finds that $g_k(\beta; X) \ll X^{1-\sigma(k)}$, where $\sigma(k)^{-1} = (4 + o(1))k^2 \log k$.

2. Translation invariance and a congruencing idea

A key feature of the system of equations (1.3) is *translation-dilation invariance*. Thus, the pair $\boldsymbol{x}, \boldsymbol{y}$ is an integral solution of the system

$$x_1^j + \ldots + x_t^j = y_1^j + \ldots + y_t^j \quad (1 \le j \le k), \tag{2.1}$$

if and only if, for any $\xi \in \mathbb{Z}$ and $q \in \mathbb{N}$, the pair $\boldsymbol{x}, \boldsymbol{y}$ satisfies the system

$$(qx_1 + \xi)^j + \ldots + (qx_t + \xi)^j = (qy_1 + \xi)^j + \ldots + (qy_t + \xi)^j \quad (1 \le j \le k). \tag{2.2}$$

This property ensures that $J_t(X)$ is homogeneous with respect to restriction to arithmetic progressions[2]. Let $M = X^\theta$ be a parameter to be chosen later, consider a set $\mathcal{P}$ of $\lceil k^2/\theta \rceil$ primes p with $M < p \le 2M$, and fix some $p \in \mathcal{P}$. Also, define

$$\mathfrak{f}_c(\boldsymbol{\alpha};\xi) = \sum_{\substack{1\le x\le X\\ x\equiv \xi \pmod{p^c}}} e(\alpha_1 x + \ldots + \alpha_k x^k).$$

Since $\oint |\mathfrak{f}_c(\boldsymbol{\alpha};\xi)|^{2t}\,\mathrm{d}\boldsymbol{\alpha}$ counts the number of solutions of (2.2), with $q = p^c$, for which[3] $(1-\xi)/q \le \boldsymbol{x}, \boldsymbol{y} \le (X-\xi)/q$, by translation-dilation invariance, it counts solutions of (2.1) under the same conditions on $\boldsymbol{x}$ and $\boldsymbol{y}$. Thus

$$\max_{1\le \xi\le p^c} \oint |\mathfrak{f}_c(\boldsymbol{\alpha};\xi)|^{2t}\,\mathrm{d}\boldsymbol{\alpha} \ll 1 + J_t(X/M^c). \tag{2.3}$$

Translation invariance also generates useful auxiliary congruences. Let $t = s + k$, and consider the solutions of (2.1) with $1 \le \boldsymbol{x}, \boldsymbol{y} \le X$. The number of solutions T_0 in which $x_i = x_j$ for some $1 \le i < j \le k$ may be bounded via orthogonality and Hölder's inequality, giving $T_0 \ll J_t(X)^{1-1/(2t)}$. Given a *conditioned* solution with $x_i \ne x_j$ for $1 \le i < j \le k$, there exists a prime $p \in \mathcal{P}$ with $x_i \not\equiv x_j \pmod p$ for $1 \le i < j \le k$. Let $\Xi_c(\xi)$ denote the set of k-tuples $(\xi_1, \ldots, \xi_k)$, with $1 \le \boldsymbol{\xi} \le p^{c+1}$ and $\boldsymbol{\xi} \equiv \xi \pmod{p^c}$, and satisfying the property that $\xi_i \not\equiv \xi_j \pmod{p^{c+1}}$ for $i \ne j$. Also, put

$$\mathfrak{F}_c(\boldsymbol{\alpha};\xi) = \sum_{\boldsymbol{\xi}\in \Xi_c(\xi)} \mathfrak{f}_{c+1}(\boldsymbol{\alpha};\xi_1)\ldots\mathfrak{f}_{c+1}(\boldsymbol{\alpha};\xi_k),$$

and define

$$I_{a,b}(X) = \max_{\xi,\eta} \oint |\mathfrak{F}_a(\boldsymbol{\alpha};\xi)^2 \mathfrak{f}_b(\boldsymbol{\alpha};\eta)^{2s}|\,\mathrm{d}\boldsymbol{\alpha}. \tag{2.4}$$

Then for some $p \in \mathcal{P}$, which we now fix, the number T_1 of conditioned solutions satisfies

$$T_1 \ll \oint \mathfrak{F}_0(\boldsymbol{\alpha};0) f(\boldsymbol{\alpha};X)^s f(-\boldsymbol{\alpha};X)^{s+k}\,\mathrm{d}\boldsymbol{\alpha}. \tag{2.5}$$

Thus, by Schwarz's inequality and orthogonality, one has $T_1 \ll I_{0,0}(X)^{1/2} J_t(X)^{1/2}$. By combining the above estimates for T_0 and T_1, we derive the upper bound

$$J_t(X) \ll J_t(X)^{1-1/(2t)} + I_{0,0}(X)^{1/2} J_t(X)^{1/2},$$

whence $J_t(X) \ll I_{0,0}(X)$.

By Hölder's inequality, one finds that

$$|f(\boldsymbol{\alpha};X)|^{2s} = \Big|\sum_{\eta=1}^{p} \sum_{\substack{1\le x\le X\\ x\equiv \eta \pmod p}} e(\alpha_1 x + \ldots + \alpha_k x^k)\Big|^{2s} \le p^{2s-1} \sum_{\eta=1}^{p} |\mathfrak{f}_1(\boldsymbol{\alpha};\eta)|^{2s}.$$

[2] In this section we consider k to be fixed, and hence we drop mention of k from our notations.

[3] Here we make use of slightly unconventional vector notation. Thus, we write $\boldsymbol{z} \equiv \xi \pmod q$ when $z_i \equiv \xi \pmod q$ for $1 \le i \le t$, or $a \le \boldsymbol{z} \le b$ when $a \le z_i \le b$ $(1 \le i \le t)$, and so on.

Thus, on noting the trivial relation $f(\boldsymbol{\alpha};X)=\mathfrak{f}_0(\boldsymbol{\alpha};\eta)$, one sees from (2.4) that

$$J_t(X)\ll I_{0,0}(X)\ll M^{2s}\max_{\xi,\eta}\oint|\mathfrak{F}_0(\boldsymbol{\alpha};\xi)^2\mathfrak{f}_1(\boldsymbol{\alpha};\eta)^{2s}|\,\mathrm{d}\boldsymbol{\alpha}=M^{2s}I_{0,1}(X). \tag{2.6}$$

The mean value underlying $I_{0,1}(X)$ counts the number of integral solutions of

$$\sum_{i=1}^k(x_i^j-y_i^j)=\sum_{l=1}^s\big((pu_l+\eta)^j-(pv_l+\eta)^j\big)\quad(1\le j\le k),$$

with $1\le\boldsymbol{x},\boldsymbol{y}\le X$ and $1\le p\boldsymbol{u}+\eta,p\boldsymbol{v}+\eta\le X$, in which $x_i\not\equiv x_j\pmod p$ for $i\ne j$, and similarly for $\boldsymbol{y}$. Translation invariance leads from these equations to

$$\sum_{i=1}^k\big((x_i-\eta)^j-(y_i-\eta)^j\big)=p^j\sum_{l=1}^s(u_l^j-v_l^j)\quad(1\le j\le k),$$

and hence

$$(x_1-\eta)^j+\ldots+(x_k-\eta)^j\equiv(y_1-\eta)^j+\ldots+(y_k-\eta)^j\pmod{p^j}\quad(1\le j\le k). \tag{2.7}$$

Since the x_i are distinct modulo p, Hensel's lemma shows that, for each fixed choice of $\boldsymbol{y}$, there are at most $k!p^{k(k-1)/2}$ choices for $\boldsymbol{x}\pmod{p^k}$ satisfying (2.7). An application of Cauchy's inequality shows from here that

$$I_{0,1}(X)\ll M^{k(k-1)/2}\max_{\eta}\oint\bigg(\sum_{\nu=1}^{p^k}|\mathfrak{f}_k(\boldsymbol{\alpha};\nu)|^2\bigg)^k|\mathfrak{f}_1(\boldsymbol{\alpha};\eta)|^{2s}\,\mathrm{d}\boldsymbol{\alpha}. \tag{2.8}$$

Although our notation has been crafted for later discussion of efficient congruencing, the classical approach remains visible. One applies (2.8) with $\theta=1/k$, so that $p^k>X$. Thus $|\mathfrak{f}_k(\boldsymbol{\alpha};\nu)|\le1$, and it follows from (2.6) and (2.8) that

$$J_t(X)\ll M^{2s}I_{0,1}(X)\ll M^{2s+k(k-1)/2}(M^k)^k\max_{\eta}\oint|\mathfrak{f}_1(\boldsymbol{\alpha};\eta)|^{2s}\,\mathrm{d}\boldsymbol{\alpha}.$$

It therefore follows from (2.3) that

$$J_{s+k}(X)\ll M^{2s+k(k-1)/2}X^kJ_s(X/M)\ll X^{2k}(X^{1/k})^{2s-k(k+1)/2}J_s(X^{1-1/k}).$$

This iterative relation leads from the bound $J_{k,k}(X)\ll X^k$ to the estimate presented in Theorem 1.1. Early authors, such as Vinogradov and Hua, made use of short real intervals in place of congruences, the modern shift to congruences merely adjusting the point of view from the infinite place to a finite place.

3. Lower bounds and the main conjecture

Write $T_s(X)$ for the number of diagonal solutions of (1.3) with $1\le\boldsymbol{x},\boldsymbol{y}\le X$ and $\{x_1,\ldots,x_s\}=\{y_1,\ldots,y_s\}$. Then $J_{s,k}(X)\ge T_s(X)=s!X^s+O_s(X^{s-1})$. Meanwhile, when $1\le\boldsymbol{x},\boldsymbol{y}\le X$, one has $|(x_1^j-y_1^j)+\ldots+(x_s^j-y_s^j)|\le sX^j$. Hence

$$[X]^{2s}=\sum_{|h_1|\le sX}\ldots\sum_{|h_k|\le sX^k}\oint|f_k(\boldsymbol{\alpha};X)|^{2s}e(-\alpha_1h_1-\ldots-\alpha_kh_k)\,\mathrm{d}\boldsymbol{\alpha},$$

and we deduce from the triangle inequality in combination with (1.2) that

$$X^{2s} \ll \sum_{|h_1|\le sX} \dots \sum_{|h_k|\le sX^k} J_{s,k}(X) \ll X^{k(k+1)/2} J_{s,k}(X).$$

Thus we conclude that $J_{s,k}(X) \gg X^s + X^{2s-k(k+1)/2}$, a lower bound that guides a heuristic application of the circle method towards the following conjecture.

Conjecture 3.1 (The Main Conjecture). *Suppose that s and k are natural numbers. Then for each $\varepsilon > 0$, one has $J_{s,k}(X) \ll X^{\varepsilon}(X^s + X^{2s-k(k+1)/2})$.*

We emphasise that the implicit constant here may depend on ε, s and k. The critical case of the Main Conjecture with $s = k(k+1)/2$ has special significance.

Conjecture 3.2. *When $k \in \mathbb{N}$ and $\varepsilon > 0$, one has $J_{k(k+1)/2,k}(X) \ll X^{k(k+1)/2+\varepsilon}$.*

Suppose temporarily that this critical case of the Main Conjecture holds. Then, when $s \ge k(k+1)/2$, one may apply a trivial estimate for $f_k(\boldsymbol{\alpha}; X)$ to show that

$$J_{s,k}(X) \le X^{2s-k(k+1)} \oint |f_k(\boldsymbol{\alpha}; X)|^{k(k+1)}\, \mathrm{d}\boldsymbol{\alpha} \ll X^{2s-k(k+1)/2+\varepsilon},$$

and when $s < k(k+1)/2$, one may instead apply Hölder's inequality to obtain

$$J_{s,k}(X) \le \biggl(\oint |f_k(\boldsymbol{\alpha}; X)|^{k(k+1)}\, \mathrm{d}\boldsymbol{\alpha} \biggr)^{\frac{2s}{k(k+1)}} \ll X^{s+\varepsilon}.$$

In both cases, therefore, the Main Conjecture is recovered from the critical case.

Until 2014, the critical case of the Main Conjecture was known to hold in only two cases. The case $k = 1$ is trivial. The case $k = 2$, on the other hand, depends on bounds for the number of integral solutions of the simultaneous equations

$$\left.\begin{aligned} x_1^2 + x_2^2 + x_3^2 &= y_1^2 + y_2^2 + y_3^2 \\ x_1 + x_2 + x_3 &= y_1 + y_2 + y_3 \end{aligned}\right\}, \tag{3.1}$$

with $1 \le x_i, y_i \le X$. From the identity $(a+b-c)^2 - (a^2+b^2-c^2) = 2(a-c)(b-c)$, one finds that the solutions of (3.1) satisfy $(x_1 - y_3)(x_2 - y_3) = (y_1 - x_3)(y_2 - x_3)$. From here, elementary estimates for the divisor function convey us to the bound $J_{3,2}(X) \ll X^{3+\varepsilon}$, so that Conjecture 3.2 and the Main Conjecture hold when $k = 2$. In fact, improving on earlier work of Rogovskaya [41], it was shown by Blomer and Brüdern [10] that

$$J_{3,2}(X) = \frac{18}{\pi^2} X^3 \log X + \frac{3}{\pi^2} \biggl(12\gamma - 6\frac{\zeta'(2)}{\zeta(2)} - 5 \biggr) X^3 + O(X^{5/2} \log X).$$

In particular, the factor X^{ε} cannot be removed from the statements of Conjectures 3.1 and 3.2. However, a careful heuristic analysis of the circle method reveals that when $(s,k) \ne (3,2)$, the Main Conjecture should hold with $\varepsilon = 0$. See [47, equation (7.5)] for a discussion that records precisely such a conjecture.

The classical picture of the Main Conjecture splits naturally into two parts: small s and large s. When $1 \le s \le k$, the relation $J_{s,k}(X) = T_s(X) \sim s! X^s$ is immediate

from Newton's formulae concerning roots of polynomials. Identities analogous to that above yield multiplicative relations amongst variables in the system (1.3) when $s = k + 1$. In this way, Hua [26] confirmed the Main Conjecture for $s \leq k + 1$ by obtaining the bound $J_{k+1,k}(X) \ll X^{k+1+\varepsilon}$. Vaughan and Wooley have since obtained the asymptotic formula $J_{k+1,k}(X) = T_{k+1}(X) + O(X^{\theta_k+\varepsilon})$, where $\theta_3 = \frac{10}{3}$ [48, Theorem 1.5] and $\theta_k = \sqrt{4k+5}$ $(k \geq 4)$ [49, Theorem 1]. Approximations to the Main Conjecture of the type $J_{s,k}(X) \ll X^{s+\delta_{s,k}}$, with $\delta_{s,k}$ small, can be obtained for larger values of s. Thus, on writing $\gamma = s/k$, the work of Arkhipov and Karatsuba [3] shows that permissible exponents $\delta_{s,k}$ exist with $\delta_{s,k} \ll \gamma^{3/2}k^{1/2}$, Tyrina [43] gets $\delta_{s,k} \ll \gamma^2$, and Wooley [58, Theorem 1] obtains $\delta_{s,k} = \exp(-Ak/\gamma^2)$, when $s \leq k^{3/2}(\log k)^{-1}$, for a certain positive constant A.

We turn next to large values of s. When $k \in \mathbb{N}$, denote by $H(k)$ the least integer for which the Main Conjecture for $J_{s,k}(X)$ holds whenever $s \geq H(k)$. Theorem 1.2 gives $H(k) \leq (2 + o(1))k^2 \log k$, a consequence of the classical estimate (1.4) with permissible exponent $\Delta_{s,k} = k^2e^{-s/k^2}$. In 1992, the author [56] found a means of combining Vinogradov's methods with the *efficient differencing method* (see [55], and the author's previous ICM lecture [61]), obtaining $\Delta_{s,k} \approx k^2e^{-2s/k^2}$. This yields $H(k) \leq (1+o(1))k^2 \log k$ (see [60]), halving the previous bound. Meanwhile, Hua [26, Theorem 7] has applied Weyl differencing to bound $H(k)$ for small k. We summarise the classical status of the Main Conjecture in the following theorem.

Theorem 3.3. *The Main Conjecture holds for $J_{s,k}(X)$ when:*

(i) $k = 1$ *and* 2*;*

(ii) $k \geq 2$ *and* $1 \leq s \leq k + 1$*;*

(iii) $s \geq H(k)$*, where* $H(3) = 8$*,* $H(4) = 23$*,* $H(5) = 55$*,* $H(6) = 120$*, ..., and* $H(k) = k^2(\log k + 2 \log\log k + O(1))$.

4. The advent of efficient congruencing

The introduction of the *efficient congruencing method* [62] at the end of 2010 has transformed our understanding of the Main Conjecture. Incorporating subsequent developments [20, 64], and the multigrade enhancement of the method [65–67], we can summarise the current state of affairs in the form of a theorem.

Theorem 4.1. *The Main Conjecture holds for $J_{s,k}(X)$ when:*

(i) $k = 1$*,* 2 *and* 3*;*

(ii) $1 \leq s \leq D(k)$*, where* $D(4) = 8$*,* $D(5) = 10$*,* $D(6) = 17$*,* $D(7) = 20$*, ..., and* $D(k) = \frac{1}{2}k(k + 1) - \frac{1}{3}k + O(k^{2/3})$*;*

(iii) $k \geq 3$ *and* $s \geq H(k)$*, where* $H(k) = k(k - 1)$.

As compared to the classical situation, there are three principal advances:

(a) First, the Main Conjecture holds for $J_{s,k}(X)$ in the cubic case $k = 3$ (see [67, Theorem 1.1]), so that $J_{s,3}(X) \ll X^{\varepsilon}(X^s + X^{2s-6})$. This is the first occasion, for any polynomial Weyl sum of degree exceeding 2, that the conjectural mean value estimates have been established in full, even if the underlying variables are restricted to lie in such special sets as the smooth numbers.

(b) Second, the Main Conjecture holds in the form $J_{s,k}(X) \ll X^{s+\varepsilon}$ provided that $1 \le s \le \frac{1}{2}k(k+1) - \frac{1}{3}k + O(k^{2/3})$, which as $k \to \infty$ represents 100% of the critical interval $1 \le s \le k(k+1)/2$ (see [66, Theorem 1.3]). The classical result reported in Theorem 3.3(ii) only provides such a conclusion for $1 \le s \le k+1$, amounting to 0% of the critical interval. Here, the first substantial advance was achieved by Ford and Wooley [20, Theorem 1.1], giving the Main Conjecture for $1 \le s \le \frac{1}{4}(k+1)^2$. Although Theorem 4.1(ii) comes within $\big(\frac{1}{3} + o(1)\big)\, k$ variables of proving the critical case of the Main Conjecture, it seems that a new idea is required to replace this defect by $(c + o(1))k$, for some real number c with $c < \frac{1}{3}$.

(c) Third, the Main Conjecture holds in the form $J_{s,k}(X) \ll X^{2s-k(k+1)/2+\varepsilon}$ for $s \ge k(k-1)$. The classical result reported in Theorem 3.3(iii) provides such a conclusion for $s \ge (1 + o(1))k^2 \log k$, a constraint weaker by a factor $\log k$. So far as applications are concerned, this is by far the most significant advance thus far captured by the efficient congruencing method. The initial progress [62, Theorem 1.1] shows that the Main Conjecture holds for $s \ge k(k+1)$, already within a factor 2 of the critical exponent $s = k(k+1)/2$. Subsequently, this constraint was improved first to $s \ge k^2 - 1$, and then to $s \ge k^2 - k + 1$ (see [64, Theorem 1.1] and [65, Corollary 1.2]). The further modest progress reported in Theorem 4.1(iii) was announced in [67, Theorem 1.2], and will appear in a forthcoming paper.

Prior to the advent of efficient congruencing, much effort had been spent on refining estimates of the shape $J_{s,k}(X) \ll X^{2s-k(k+1)/2+\Delta_{s,k}}$, with the permissible exponent $\Delta_{s,k}$ as small as possible (see [9, 19, 56, 58]). Of great significance for applications, efficient congruencing permits substantially sharper bounds to be obtained for such exponents than were hitherto available. Such ideas feature in [64, Theorem 1.4], and the discussion following [20, Theorem 1.2] shows that when $\frac{1}{4} \le \alpha \le 1$ and $s = \alpha k^2$, then the exponent $\Delta_{s,k} = (1 - \sqrt{\alpha})^2 k^2 + O(k)$ is permissible. Thus, in particular, the critical exponent $\Delta_{k(k+1)/2,k} = (\frac{3}{2} - \sqrt{2})k^2$ is permissible. By combining [66, Theorem 1.5] and the discussion following [65, Corollary 1.2], one arrives at the following improvement.

Theorem 4.2. *When k is large, there is a positive number $C(s) \le \frac{1}{3}$ for which*

$$J_{s,k}(X) \ll X^{(C(s)+o(1))k}\big(X^s + X^{2s-k(k+1)/2}\big).$$

When $\alpha \in [\frac{5}{8}, 1]$, moreover, one may take $C(\alpha k^2) \le \big(2 - 3\alpha + (2\alpha - 1)^{3/2}\big)/(3\alpha)$.

We finish this section by noting that Theorem 4.1(iii) permits a substantial improvement in the conclusion of Theorem 1.2.

Theorem 4.3. *Let $k, s \in \mathbb{N}$ and suppose that $s \ge k^2 - k + 1$. Then there exists a positive number $\mathfrak{C}(s,k)$ with $J_{s,k}(X) \sim \mathfrak{C}(s,k)X^{2s-k(k+1)/2}$.*

5. A sketch of the efficient congruencing method

Although complicated in detail, the ideas underlying efficient congruencing are accessible given some simplifying assumptions. In this section, we consider k to be fixed, and drop

mention of k from our notation. Let $t = (u+1)k$, where $u \geq k$ is an integer, and put $s = uk$. We define

$$\lambda_t = \limsup_{X \to \infty} (\log J_t(X))/(\log X).$$

Thus, for each $\varepsilon > 0$, one has the bound $J_t(X) \ll X^{\lambda_t + \varepsilon}$. Our goal is to establish that $\lambda_t = 2t - k(k+1)/2$, as predicted by the Main Conjecture. Define Λ via the relation $\lambda_t = 2t - \frac{1}{2}k(k+1) + \Lambda$. We suppose that $\Lambda > 0$, and seek a contradiction in order to show that $\Lambda = 0$. Our method rests on an N-fold iteration related to the approach of §2, where N is sufficiently large in terms of u, k and Λ. Let $\theta = (16k)^{-2N}$, put $M = X^\theta$, and consider a prime number p with $M < p \leq 2M$. Also, let $\delta > 0$ be small in terms of all these parameters, so that $8\delta < N(k/u)^N \Lambda\theta$.

Define the mean value

$$K_{a,b}(X) = \max_{\xi,\eta} \oint |\mathfrak{F}_a(\boldsymbol{\alpha};\xi)^2 \mathfrak{F}_b(\boldsymbol{\alpha};\eta)^{2u}| \, \mathrm{d}\boldsymbol{\alpha},$$

and introduce the normalised mean values

$$[[K_{a,b}(X)]] = \frac{K_{a,b}(X)}{(X/M^a)^{2k-k(k+1)/2}(X/M^b)^{2s}} \quad \text{and} \quad [[J_t(X)]] = \frac{J_t(X)}{X^{2t-k(k+1)/2}}.$$

Then whenever X is sufficiently large in terms of the ambient parameters, one has $[[J_t(X)]] > X^{\Lambda - \delta}$ and, when $X^{1/2} \leq Y \leq X$, we have the bound $[[J_t(Y)]] \leq Y^{\Lambda+\delta}$.

We begin by observing that an elaboration of the argument delivering (2.5) can be fashioned to replace (2.6) with the well-conditioned relation

$$J_t(X) \ll M^{2s} \max_{\xi,\eta} \oint |\mathfrak{F}_0(\boldsymbol{\alpha};\xi)^2 \mathfrak{F}_1(\boldsymbol{\alpha};\eta)^{2u}| \, \mathrm{d}\boldsymbol{\alpha} = M^{2s} K_{0,1}(X).$$

Here we have exercised considerable expedience in ignoring controllable error terms. Moreover, one may need to replace $K_{0,1}(X)$ by the surrogate $K_{0,1+h}(X)$, for a suitable integer h. An analogue of the argument leading to (2.8) yields the bound

$$K_{0,1}(X) \ll M^{k(k-1)/2} \max_{\eta} \oint \Big(\sum_{\nu=1}^{p^k} |\mathfrak{f}_k(\boldsymbol{\alpha};\nu)|^2 \Big)^k |\mathfrak{F}_1(\boldsymbol{\alpha};\eta)|^{2u} \, \mathrm{d}\boldsymbol{\alpha}.$$

By Hölder's inequality, one finds first that

$$\Big(\sum_{\nu=1}^{p^k} |\mathfrak{f}_k(\boldsymbol{\alpha};\nu)|^2 \Big)^k \leq (p^k)^{k-1} \sum_{\nu=1}^{p^k} |\mathfrak{f}_k(\boldsymbol{\alpha};\nu)|^{2k},$$

and then

$$K_{0,1}(X) \ll M^{k(k-1)/2} (M^k)^k \max_{\eta,\nu} \Big(T_1(\eta)^{1-1/u} T_2(\eta,\nu)^{1/u} \Big),$$

where

$$T_1(\eta) = \oint |\mathfrak{F}_1(\boldsymbol{\alpha};\eta)|^{2u+2} \, \mathrm{d}\boldsymbol{\alpha} \quad \text{and} \quad T_2(\eta,\nu) = \oint |\mathfrak{F}_1(\boldsymbol{\alpha};\eta)^2 \mathfrak{f}_k(\boldsymbol{\alpha};\nu)^{2s}| \, \mathrm{d}\boldsymbol{\alpha}.$$

On considering the underlying Diophantine systems, one finds that $T_1(\eta)$ may be bounded via (2.3), while $T_2(\eta,\nu)$ may be bounded in terms of $K_{1,k}(X)$. Thus

$$J_t(X) \ll M^{2s+k(k-1)/2}(M^k)^k J_t(X/M)^{1-1/u} K_{1,k}(X)^{1/u}.$$

A modicum of computation therefore confirms that

$$[[J_t(X)]] \ll [[J_t(X/M)]]^{1-1/u}[[K_{1,k}(X)]]^{1/u}. \tag{5.1}$$

The mean value underlying $K_{1,k}(X)$ counts the number of integral solutions of

$$\sum_{i=1}^{k}(x_i^j - y_i^j) = \sum_{l=1}^{s}\big((p^k u_l+\eta)^j - (p^k v_l+\eta)^j\big) \quad (1\le j\le k),$$

with $1\le \boldsymbol{x},\boldsymbol{y}\le X$ and $1\le p^k\boldsymbol{u}+\eta, p^k\boldsymbol{v}+\eta\le X$ having suitably conditioned coordinates. In particular, one has $\boldsymbol{x}\equiv\boldsymbol{y}\equiv\xi \pmod p$ but $x_i\not\equiv x_j \pmod{p^2}$ for $i\ne j$, and similarly for $\boldsymbol{y}$. Translation invariance leads from these equations to

$$\sum_{i=1}^{k}\big((x_i-\eta)^j - (y_i-\eta)^j\big) = p^{jk}\sum_{l=1}^{s}(u_l^j - v_l^j) \quad (1\le j\le k),$$

and hence to the congruences

$$(x_1-\eta)^j+\ldots+(x_k-\eta)^j \equiv (y_1-\eta)^j+\ldots+(y_k-\eta)^j \pmod{p^{jk}} \quad (1\le j\le k). \tag{5.2}$$

Since the x_i are distinct modulo p^2, an application of Hensel's lemma shows that, for each fixed choice of $\boldsymbol{y}$, there are at most $k!(p^k)^{k(k-1)/2}\cdot p^{k(k-1)/2}$ choices for $\boldsymbol{x} \pmod{p^{k^2}}$ satisfying (5.2). Here, the factor $p^{k(k-1)/2}$ reflects the fact that, even though $x_i\not\equiv x_j \pmod{p^2}$ for $i\ne j$, one has $x_i\equiv x_j \pmod p$ for all i and j. This situation is entirely analogous to that delivering (2.8) above, and thus we obtain

$$K_{1,k}(X) \ll (M^{k+1})^{k(k-1)/2}\max_{\xi,\eta}\oint\Bigg(\sum_{\substack{\nu=1\\ \nu\equiv\xi \pmod p}}^{p^{k^2}}|\mathfrak{f}_{k^2}(\boldsymbol{\alpha};\nu)|^2\Bigg)^k|\mathfrak{F}_k(\boldsymbol{\alpha};\eta)|^{2u}\,\mathrm{d}\boldsymbol{\alpha}.$$

From here, as above, suitable applications of Hölder's inequality show that

$$[[K_{1,k}(X)]] \ll [[J_t(X/M^k)]]^{1-1/u}[[K_{k,k^2}(X)]]^{1/u}. \tag{5.3}$$

By substituting this estimate into (5.1), we obtain the new upper bound

$$[[J_t(X)]] \ll \big([[J_t(X/M)]][[J_t(X/M^k)]]^{1/u}\big)^{1-1/u}[[K_{k,k^2}(X)]]^{1/u^2}.$$

By iterating this process N times, one obtains the relation

$$[[J_t(X)]] \ll \Bigg(\prod_{r=0}^{N-1}[[J_t(X/M^{k^r})]]^{1/u^r}\Bigg)^{1-1/u}[[K_{k^{N-1},k^N}(X)]]^{1/u^N}. \tag{5.4}$$

While this is a vaste oversimplification of what is actually established, it correctly identifies the relationship which underpins the efficient congruencing method.

Since $M^{k^N} < X^{1/3}$, our earlier discussion ensures that

$$[[J_t(X)]] \gg X^{\Lambda-\delta} \quad \text{and} \quad [[J_t(X/M^{k^r})]] \ll (X/M^{k^r})^{\Lambda+\delta} \quad (0 \le r \le N).$$

Meanwhile, an application of Hölder's inequality provides the trivial bound

$$[[K_{k^{N-1},k^N}(X)]] \ll \big(M^{k(k+1)/2}\big)^{k^N} X^{\Lambda+\delta}.$$

By substituting these estimates into (5.4), we deduce that

$$X^{\Lambda-\delta} \ll \bigg(X^{1/u^N} \prod_{r=0}^{N-1} \big(X/M^{k^r}\big)^{(1-1/u)/u^r}\bigg)^{\Lambda+\delta} \Big(M^{k(k+1)/2}\Big)^{(k/u)^N},$$

and hence $X^{\Lambda-\delta} \ll X^{\Lambda+\delta}(M^{\Theta})^{(k/u)^N}$, where

$$\Theta = \tfrac{1}{2}k(k+1) - (1-1/u)(\Lambda+\delta)\sum_{r=1}^{N}(u/k)^r.$$

But we have $u \ge k$, and so our hypotheses concerning N and δ ensure that

$$\Theta \le \tfrac{1}{2}k(k+1) - N(1-1/u)(\Lambda+\delta) < -\tfrac{1}{2}N\Lambda < -3(u/k)^N\delta/\theta.$$

We therefore conclude that $X^{\Lambda-\delta} \ll X^{\Lambda+\delta}M^{-3\delta/\theta} \ll X^{\Lambda-2\delta}$. This relation yields the contradiction that establishes the desired conclusion $\Lambda = 0$. We may therefore conclude that whenever $t \ge k(k+1)$, one has $J_t(X) \ll X^{2t-k(k+1)/2+\varepsilon}$.

We have sketched the proof of the Main Conjecture for $J_t(X)$ when $t \ge k(k+1)$. Theorem 4.1, which represents the latest state of play in the efficient congruencing method, goes considerably further. Two ideas underpin these advances.

First, one may sacrifice some of the power potentially available from systems of congruences such as (2.7) or (5.2) in order that the efficient congruencing method be applicable when $t < k(k+1)$. Let r be a parameter with $2 \le r \le k$, and define the generating function $\mathfrak{F}_c^{(r)}(\boldsymbol{\alpha};\xi)$ by analogy with $\mathfrak{F}_c(\boldsymbol{\alpha};\xi)$, though with r (in place of k) underlying exponential sums $\mathfrak{f}_{c+1}(\boldsymbol{\alpha};\xi_i)$. One may imitate the basic argument sketched above, with $t = (u+1)r$, to bound the analogue $K_{a,b}^{(r)}(X)$ of the mean value $K_{a,b}(X)$. In place of (5.2) one now obtains the congruences

$$(x_1-\eta)^j + \ldots + (x_r-\eta)^j \equiv (y_1-\eta)^j + \ldots + (y_r-\eta)^j \pmod{p^{jb}} \quad (1 \le j \le k). \tag{5.5}$$

For simplicity, suppose that $r \le (k-1)/2$. Then by considering the r congruence relations of highest degree here, one finds from Hensel's lemma that, for each fixed choice of $\boldsymbol{y}$, there are at most $k!$ choices for $\boldsymbol{x} \pmod{p^{(k-r)b}}$ satisfying (5.5). Although this is a weaker congruence constraint than before on $\boldsymbol{x}$ and $\boldsymbol{y}$, the cost in terms of the number of choices is smaller, and so useful estimates may nonetheless be obtained for $J_t(X)$. Ideas along these lines underpin both the work [64] of the author, and in the sharper form sketched above, that of Ford and the author [20].

The second idea conveys us to the threshold of the Main Conjecture. Again we consider the mean values $K_{a,b}^{(r)}(X)$, and for simplicity put $r = k-1$. The congruences (5.5) yield a constraint on the variables tantamount to $x_i \equiv y_i \pmod{p^{2b}}$ at little cost. Encoding this

constraint using exponential sums, and applying Hölder's inequality, one bounds $K_{a,b}^{(k-1)}(X)$ in terms of $K_{a,b}^{(k-2)}(X)$ and $K_{b,2b}^{(k-1)}(X)$. Iterating this process to successively estimate $K_{a,b}^{(k-j)}(X)$ for $j = 1, 2, \dots, k-1$, we obtain a bound for $K_{a,b}^{(k-1)}(X)$ in terms of $K_{b,jb}^{(k-1)}(X)$ $(2 \le j \le k)$ and $J_t(X/M^b)$. The heuristic potential of this idea amounts to a relation of the shape

$$[[K_{a,b}^{(k-1)}(X)]] \ll \Big(\prod_{j=2}^{k} [[K_{b,jb}^{(k-1)}(X)]]^{\phi_j} \Big) [[J_t(X/M^b)]]^{1-(k-1)/s}, \tag{5.6}$$

where the exponents ϕ_j are approximately equal to $1/s$. Again, this substantially oversimplifies the situation, since non-negligible additional factors occur. However, one discerns a critical advantage over earlier relations such as (5.3). As one iterates (5.6), one bounds $[[K_{a,b}^{(k-1)}(X)]]$ in terms of new expressions $[[K_{b,b'}^{(k-1)}(X)]]$, where the ratio b'/b is on average about $\frac{1}{2}k + 1$, as opposed to the previous ratio k. The relation (5.6) may be converted into a substitute for (5.4) of the shape

$$[[J_t(X)]] \ll \Big(\prod_{r=0}^{N-1} [[J_t(X/M^{\rho^r})]]^{1/u^r} \Big)^{1-1/u} [[K_{\rho^{N-1},\rho^N}^{(k-1)}(X)]]^{1/u^N},$$

in which ρ is close to $\frac{1}{2}k + 1$ and $t = (u+1)(k-1)$. Thus, when $u \ge \rho$, we find as before that the lower bound $[[J_t(X)]] \gg X^{\Lambda-\delta}$ is tenable only when $\Lambda = 0$, and we have heuristically established the Main Conjecture when t is only slightly larger than $k(k+1)/2$. Of course, the relation (5.6) represents an idealised situation, and the proof in detail of the results in [65, 66] contains numerous complications requiring the resolution of considerable technical difficulties.

6. Waring's problem

Investigations concerning the validity of the anticipated asymptotic formula in Waring's problem have historically followed one of two paths, associated on the one hand with Weyl, and on the other with Vinogradov. We recall our earlier notation, writing $R_{s,k}(n)$ for the number of representations of the natural number n in the shape $n = x_1^k + \dots + x_s^k$, with $\boldsymbol{x} \in \mathbb{N}^s$. A heuristic application of the circle method suggests that for $k \ge 3$ and $s \ge k+1$, one should have

$$R_{s,k}(n) = \frac{\Gamma(1+1/k)^s}{\Gamma(s/k)} \mathfrak{S}_{s,k}(n) n^{s/k-1} + o(n^{s/k-1}), \tag{6.1}$$

where

$$\mathfrak{S}_{s,k}(n) = \sum_{q=1}^{\infty} \sum_{\substack{a=1 \\ (a,q)=1}}^{q} \Big(q^{-1} \sum_{r=1}^{q} e(ar^k/q) \Big)^s e(-na/q).$$

Under modest congruence conditions, one has $1 \ll \mathfrak{S}_{s,k}(n) \ll n^{\varepsilon}$, and thus the conjectural relation (6.1) may be seen as an honest asymptotic formula (see [47, §§4.3, 4.5 and 4.6] for details). Let $\widetilde{G}(k)$ denote the least integer t with the property that, whenever $s \ge t$, the asymptotic formula (6.1) holds for all large enough n.

Leaving aside the smallest exponents $k = 1$ and 2 accessible to classical methods, the first to obtain a bound for $\widetilde{G}(k)$ were Hardy and Littlewood [21], who devised a method based on Weyl differencing to show that $\widetilde{G}(k) \le (k-2)2^{k-1} + 5$. In 1938, Hua [24] obtained a refinement based on the estimate

$$\int_0^1 |g_k(\alpha; X)|^{2^k} \, \mathrm{d}\alpha \ll X^{2^k - k + \varepsilon}, \tag{6.2}$$

in which $g_k(\alpha; X)$ is defined via (1.5), showing that $\widetilde{G}(k) \le 2^k + 1$. For small values of k, this estimate remained the strongest known for nearly half a century. Finally, Vaughan [45, 46] succeeded in wielding Hooley's Δ-functions to deduce that $\widetilde{G}(k) \le 2^k$ for $k \ge 3$. For slightly larger exponents $k \ge 6$, this bound was improved by Heath-Brown [22] by combining Weyl differencing with a novel cubic mean value estimate. His bound $\widetilde{G}(k) \le \frac{7}{8}2^k + 1$ was, in turn, refined by Boklan [8], who exploited Hooley's Δ-functions in this new setting to deduce that $\widetilde{G}(k) \le \frac{7}{8}2^k$ for $k \ge 6$.

Turning now to large values of k, the story begins with Vinogradov [50], who showed that $\widetilde{G}(k) \le 183k^9(\log k + 1)^2$, reducing estimates previously exponential in k to polynomial bounds. As Vinogradov's mean value theorem progressed to the state essentially captured by Theorem 1.1, bounds were rapidly refined to the form $\widetilde{G}(k) \le (C + o(1))k^2 \log k$, culminating in 1949 with Hua's bound [27] of this shape with $C = 4$. The connection with Vinogradov's mean value theorem is simple to explain, for on considering the underlying Diophantine systems, one finds that

$$\int_0^1 |g_k(\alpha; X)|^{2s} \, \mathrm{d}\alpha = \sum_{\boldsymbol{h}} \oint |f_k(\boldsymbol{\alpha}; X)|^{2s} e(-h_1\alpha_1 - \ldots - h_{k-1}\alpha_{k-1}) \, \mathrm{d}\boldsymbol{\alpha},$$

where the summation is over $|h_j| \le sX^j$ $(1 \le j \le k-1)$. The bound (1.4) therefore leads via the triangle inequality and (1.2) to the estimate

$$\int_0^1 |g_k(\alpha; X)|^{2s} \, \mathrm{d}\alpha \ll X^{k(k-1)/2} J_{s,k}(X) \ll X^{2s - k + \Delta_{s,k}}, \tag{6.3}$$

which serves as a surrogate for (6.2). In 1992, the author reduced the permissible value of C from 4 to 2 by applying the repeated efficient differencing method [56]. A more efficient means of utilising Vinogradov's mean value theorem to bound $\widetilde{G}(k)$ was found by Ford [18] (see also [44]), showing that $C = 1$ is permissible. Refinements for smaller values of k show that this circle of ideas surpasses the above-cited bound $\widetilde{G}(k) \le \frac{7}{8}2^k$ when $k \ge 9$ (see Boklan and Wooley [9]).

We summarise the classical state of affairs in the following theorem.

Theorem 6.1 (Classical status of $\widetilde{G}(k)$). *One has:*

(i) $\widetilde{G}(k) \le 2^k$ $(k = 3, 4, 5)$ *and* $\widetilde{G}(k) \le \frac{7}{8}2^k$ $(k = 6, 7, 8)$*;*

(ii) $\widetilde{G}(9) \le 365$*,* $\widetilde{G}(10) \le 497$*,* $\widetilde{G}(11) \le 627$*,* $\widetilde{G}(12) \le 771$*, . . . ;*

(iii) $\widetilde{G}(k) \le (1 + o(1))k^2 \log k$ $(k$ *large*$)$.

The most immediate impact of the new efficient congruencing method in Vinogradov's mean value theorem [62] was the bound $\widetilde{G}(k) \le 2k^2 + 2k - 3$, valid for $k \ge 2$. This already

supersedes the previous work presented in Theorem 6.1 when $k \ge 7$. In particular, the obstinate factor of $\log k$ is definitively removed for large values of k. Subsequent refinements [20, 63–66] have delivered further progress, especially for smaller values of k, which we summarise as follows.

Theorem 6.2 (Status of $\widetilde{G}(k)$ after efficient congruencing). *One has:*

(i) $\widetilde{G}(k) \le 2^k$ $(k = 3, 4)$;

(ii) $\widetilde{G}(5) \le 28$, $\widetilde{G}(6) \le 43$, $\widetilde{G}(7) \le 61$, $\widetilde{G}(8) \le 83$, $\widetilde{G}(9) \le 107$, $\widetilde{G}(10) \le 134$, $\widetilde{G}(11) \le 165$, $\widetilde{G}(12) \le 199$, ...;

(iii) $\widetilde{G}(k) \le (C + o(1))k^2$ (k *large*), *where* $C = 1.54079$ *is an approximation to the number* $(5 + 6\xi - 3\xi^2)/(2 + 6\xi)$, *in which* ξ *is the real root of* $6\xi^3 + 3\xi^2 - 1$.

A comparison of Theorems 6.1 and 6.2 reveals that the classical Weyl-based bounds have now been superseded for $k \ge 5$. The latest developments [65, 67] hint, indeed, at further progress even when $k = 4$. These advances for smaller values of k stem in part, of course, from the substantial progress in our new bounds for $J_{s,k}(X)$, as outlined in Theorems 4.1 and 4.2. However, an important role is also played by a novel mean value estimate for moments of $g_k(\alpha; X)$. Define the minor arcs $\mathfrak{m} = \mathfrak{m}_k$ to be the set of real numbers $\alpha \in [0, 1)$ satisfying the property that, whenever $a \in \mathbb{Z}$ and $q \in \mathbb{N}$ satisfy $(a, q) = 1$ and $|q\alpha - a| \le X^{1-k}$, then $q > X$. The argument of the proof of [63, Theorem 2.1] yields the bound

$$\int_{\mathfrak{m}} |g_k(\alpha; X)|^{2s} \, \mathrm{d}\alpha \ll X^{\frac{1}{2}k(k-1)-1} (\log X)^{2s+1} J_{s,k}(X). \tag{6.4}$$

We thus infer from Theorem 4.1(iii) that whenever $k \ge 3$ and $s \ge k(k-1)$, then

$$\int_{\mathfrak{m}} |g_k(\alpha; X)|^{2s} \, \mathrm{d}\alpha \ll X^{2s-k-1+\varepsilon}.$$

As compared to the classical approach embodied in (6.3), an additional factor X has been saved in these estimates at no cost in terms of the number of variables, and for smaller values of k this is a very substantial gain.

For large values of k, the enhancement of Ford [18] given by Ford and Wooley [20, Theorem 8.5] remains of value. When $k, s \in \mathbb{N}$, denote by $\eta(s, k)$ the least number η with the property that, whenever X is sufficiently large in terms of s and k, one has

$$J_{s,k}(X) \ll_\varepsilon X^{2s-k(k+1)/2+\eta+\varepsilon}.$$

Let $r \in \mathbb{N}$ satisfy $1 \le r \le k-1$. Then [20, Theorem 8.5] shows that whenever $s \ge r(r-1)/2$, one has

$$\int_0^1 |g_k(\alpha; X)|^{2s} \, \mathrm{d}\alpha \ll X^{2s-k+\varepsilon} \big(X^{\eta_r^*(s,k)-1/r} + X^{\eta_r^*(s,k-1)} \big),$$

where $\eta_r^*(s, w) = r^{-1}\eta(s - r(r-1)/2, w)$ for $w = k-1, k$.

Finally, we note that familiar conjectures concerning mean values of the exponential sum $g_k(\alpha; X)$ imply that one should have $\widetilde{G}(k) \le 2k + 1$ for each $k \ge 3$, and indeed it may even be the case that $\widetilde{G}(k) = k + 1$.

7. Estimates of Weyl-type, and distribution mod 1

Pointwise estimates for exponential sums appear already in the work of Weyl [54] in 1916. By applying $k-1$ Weyl-differencing steps, one bounds the exponential sum $f_k(\boldsymbol{\alpha};X)$ in terms of a new exponential sum over a linear polynomial, and this may be estimated by summing what is, after all, a geometric progression. In this way, one obtains the classical version of Weyl's inequality (see [47, Lemma 2.4]).

Theorem 7.1 (Weyl's inequality). *Let $\boldsymbol{\alpha} \in \mathbb{R}^k$, and suppose that $a \in \mathbb{Z}$ and $q \in \mathbb{N}$ satisfy $(a,q)=1$ and $|\alpha_k - a/q| \le q^{-2}$. Then one has*

$$|f_k(\boldsymbol{\alpha};X)| \ll X^{1+\varepsilon}\left(q^{-1}+X^{-1}+qX^{-k}\right)^{2^{1-k}}. \tag{7.1}$$

This provides a non-trivial estimate for $f_k(\boldsymbol{\alpha};X)$ when the leading coefficient α_k is not well-approximated by rational numbers. Consider, for example, the set $\mathfrak{m}=\mathfrak{m}_k$ defined in the preamble to (6.4). When $\alpha_k \in \mathfrak{m}$, an application of Dirichlet's theorem on Diophantine approximation shows that there exist $a \in \mathbb{Z}$ and $q \in \mathbb{N}$ with $(a,q)=1$ such that $q \le X^{k-1}$ and $|q\alpha - a| \le X^{1-k}$. The definition of $\mathfrak{m}$ then implies that $q > X$, and so Theorem 7.1 delivers the bound

$$\sup_{\alpha_k \in \mathfrak{m}} |f_k(\boldsymbol{\alpha};X)| \ll X^{1-\sigma(k)+\varepsilon}, \tag{7.2}$$

in which $\sigma(k)=2^{1-k}$. Heath-Brown's variant [22, Theorem 1] of Weyl's inequality applies mean value estimates for certain cubic exponential sums that, for $k \ge 6$, give bounds superior to (7.1) when q lies in the range $X^{5/2} < q < X^{k-5/2}$. By making use of the cubic case of the Main Conjecture in Vinogradov's mean value theorem [67], the author [68] has extended this range to $X^2 < q < X^{k-2}$.

Theorem 7.2. *Let $k \ge 6$, and suppose that $\alpha \in \mathbb{R}$, $a \in \mathbb{Z}$ and $q \in \mathbb{N}$ satisfy $(a,q)=1$ and $|\alpha - a/q| \le q^{-2}$. Then one has*

$$|g_k(\alpha;X)| \ll X^{1+\varepsilon}\Theta^{2^{-k}} + X^{1+\varepsilon}(\Theta/X)^{\frac{2}{3}2^{-k}},$$

where $\Theta = q^{-1}+X^{-3}+qX^{-k}$.

For comparison, we note that Heath-Brown [22, Theorem 1] obtains the bound

$$|g_k(\alpha;X)| \ll X^{1+\varepsilon}(X\Theta)^{\frac{4}{3}2^{-k}}.$$

Robert and Sargos [40, Théorème 4 et Lemme 7] extend these ideas when $k \ge 8$ to show that $|g_k(\alpha;X)| \ll X^{1+\varepsilon}\big(X^{17/8}\Theta'\big)^{\frac{8}{5}2^{-k}}$, in which $\Theta' = q^{-1}+X^{-4}+qX^{-k}$. See Parsell [37] for a refinement when $k=8$.

The above methods yield exponents exponentially small in k. By substituting estimates for $J_{s,k}(X)$ into the conclusion of Theorem 1.3, one obtains analogous bounds polynomial in k. Classical versions of Vinogradov's mean value theorem yield estimates of the shape (7.2) with $\sigma(k)^{-1} = (C+o(1))k^2\log k$. Thus, Linnik [33] obtained the permissible value $C = 22\,400$, and Hua [27] obtained $C=4$ in 1949. This was improved via efficient differencing [56] in 1992 to $C=2$, and subsequently the author [59] obtained $C=3/2$ by incorporating some ideas of Bombieri [11]. The latest developments in efficient congruencing yield the new exponent $\sigma(k)^{-1} = 2(k-1)(k-2)$ for $k \ge 3$, a conclusion that removes the factor $\log k$ from earlier estimates, and improves on Weyl's inequality for $k \ge 7$.

Theorem 7.3. *Let k be an integer with $k \geq 3$, and let $\boldsymbol{\alpha} \in \mathbb{R}^k$. Suppose that there exists a natural number j with $2 \leq j \leq k$ such that, for some $a \in \mathbb{Z}$ and $q \in \mathbb{N}$ with $(a,q) = 1$, one has $|\alpha_j - a/q| \leq q^{-2}$. Then one has*

$$|f_k(\boldsymbol{\alpha}; X)| \ll X^{1+\varepsilon}(q^{-1} + X^{-1} + qX^{-j})^{\sigma(k)},$$

where $\sigma(k)^{-1} = 2(k-1)(k-2)$.

This conclusion makes use of Theorem 4.1(iii), and improves slightly on [65, Theorem 11.1]. When $k \geq 6$ and α lies on a suitable subset of $\mathbb{R}$, the above-cited work of Heath-Brown may deliver estimates for $|g_k(\alpha; X)|$ superior to those stemming from Weyl's inequality, and similar comments apply to the work of Robert and Sargos, and of Parsell, when $k \geq 8$. However, Theorem 7.3 proves superior in all circumstances to the estimates of Heath-Brown when $k > 7$, and to the estimates of Robert and Sargos for all exponents k.

In many applications, it is desirable to have available estimates for $|f_k(\boldsymbol{\alpha}; X)|$ that depend on simultaneous approximations to $\alpha_1, \dots, \alpha_k$ of a given height. This is a subject to which R. C. Baker and W. M. Schmidt have made significant contributions. By exploiting such methods in combination with our new estimates for $J_{s,k}(X)$, one obtains the following conclusion (compare [65, Theorem 11.2]).

Theorem 7.4. *Let k be an integer with $k \geq 3$, and let τ and δ be real numbers with $\tau^{-1} > 4(k-1)(k-2)$ and $\delta > k\tau$. Suppose that X is sufficiently large in terms of k, δ and τ, and further that $|f_k(\boldsymbol{\alpha}; X)| > X^{1-\tau}$. Then there exist integers $q, a_1, \dots, a_k$ such that $1 \leq q \leq X^{\delta}$ and $|q\alpha_j - a_j| \leq X^{\delta - j}$ $(1 \leq j \leq k)$.*

Here, the constraint $\tau^{-1} > 4(k-1)(k-2)$ may be compared with the corresponding hypothesis $\tau^{-1} > (8 + o(1))k^2 \log k$ to be found in [5, Theorem 4.5], and the Weyl-based bound $\tau^{-1} > 2^{k-1}$ obtained in [5, Theorem 5.2] (see also [6]). The conclusion of Theorem 7.4 is superior to the latter for $k > 8$.

Bounds for exponential sums of Weyl-type may be converted into equidistribution results for polynomials modulo 1 by applying estimates of Erdős-Turán type. Write $\|\theta\|$ for the least value of $|\theta - n|$ for $n \in \mathbb{Z}$, and consider a sequence $(x_n)_{n=1}^{\infty}$ of real numbers . Then it follows from [5, Theorem 2.2], for example, that whenever $\|x_n\| \geq M^{-1}$ for $1 \leq n \leq N$, then

$$\sum_{1 \leq m \leq M} \left| \sum_{n=1}^{N} e(mx_n) \right| > \tfrac{1}{6}N.$$

By carefully exploiting this result using the methods of Baker [5], one deduces from Theorem 7.4 the following conclusion (compare [65, Theorem 11.3]).

Theorem 7.5. *When $k \geq 3$, put $\tau(k) = 1/\left(4(k-1)(k-2)\right)$. Then whenever $\boldsymbol{\alpha} \in \mathbb{R}^k$ and N is sufficiently large in terms of k and ε, one has*

$$\min_{1 \leq n \leq N} \|\alpha_1 n + \alpha_2 n^2 + \ldots + \alpha_k n^k\| < N^{\varepsilon - \tau(k)}.$$

8. Further applications

Vinogradov's mean value theorem finds application in numerous number-theoretic problems, besides those discussed in the previous two sections. We take the opportunity now to outline several applications, emphasising recent developments.

(i) *Tarry's problem.* When h, k and s are positive integers with $h \geq 2$, consider the Diophantine system

$$\sum_{i=1}^{s} x_{i1}^j = \sum_{i=1}^{s} x_{i2}^j = \ldots = \sum_{i=1}^{s} x_{ih}^j \quad (1 \leq j \leq k). \tag{8.1}$$

Let $W(k,h)$ denote the least natural number s having the property that the simultaneous equations (8.1) possess an integral solution $\boldsymbol{x}$ with

$$\sum_{i=1}^{s} x_{iu}^{k+1} \neq \sum_{i=1}^{s} x_{iv}^{k+1} \quad (1 \leq u < v \leq h).$$

The problem of estimating $W(k,h)$ was intensely investigated by E. M. Wright and L.-K. Hua (see [25, 28, 69]), the latter obtaining $W(k,h) \leq k^2(\log k + O(1))$ for $h \geq 2$. The argument of the proof of [62, Theorem 1.3] shows that $W(k,h) \leq s$ whenever one can establish the estimate $J_{s,k+1}(X) = o(X^{2s-k(k+1)/2})$. By using this criterion together with the estimates for $J_{s,k+1}(X)$ obtained via the latest efficient congruencing methods, one obtains substantial improvements in these earlier conclusions (see [65, Theorem 12.1] and [66, Theorem 12.1]).

Theorem 8.1. *When h and k are natural numbers with $h \geq 2$ and $k \geq 3$, one has $W(k,h) \leq \frac{5}{8}(k+1)^2$. Moreover, when k is large, one has $W(k,h) \leq \frac{1}{2}k(k+1)+1$.*

Although the last of these conclusions achieves the limit of current analytic approaches to bounding $W(k,h)$, explicit numerical examples are available[4] which may be applied to show that $W(k,2) = k+1$ for $1 \leq k \leq 9$ and $k = 11$.

(ii) *Sum-product theorems.* When A is a finite set of real numbers, define the sets $A + A = \{x + y : x, y \in A\}$ and $A \cdot A = \{xy : x, y \in A\}$, and more generally

$$hA = \{x_1 + \ldots + x_h : \boldsymbol{x} \in A^h\} \quad \text{and} \quad A^{(h)} = \{x_1 \ldots x_h : \boldsymbol{x} \in A^h\}.$$

A conjecture of Erdős and Szemerédi [17] asserts that for any finite set of integers A, one has $|A+A| + |A \cdot A| \gg_\varepsilon |A|^{2-\varepsilon}$. It is also conjectured that whenever A is a finite set of *real* numbers, then for each $h \in \mathbb{N}$, one should have $|hA| + |A^{(h)}| \gg_{\varepsilon,h} |A|^{h-\varepsilon}$. Chang [13] has made progress towards this conjecture by showing that when A is a finite set of integers, and $|A.A| < |A|^{1+\varepsilon}$, then $|hA| \gg_{\varepsilon,h} |A|^{h-\delta}$, where $\delta \to 0$ as $\varepsilon \to 0$. Subsequently, Bourgain and Chang [12] showed that for any $b \geq 1$, there exists $h \geq 1$ with the property that $|hA| + |A^{(h)}| \gg |A|^b$. By exploiting bounds for $W(k,h)$ of the type given by Theorem 8.1, Croot and Hart [16] have made progress toward an analogue of such conclusions for sets of real numbers.

Theorem 8.2. *Suppose that $\varepsilon > 0$ and $|A \cdot A| \leq |A|^{1+\varepsilon}$. Then there exists a number $\lambda > 0$ such that, when h is large enough in terms of ε, one has $|h(A \cdot A)| > |A|^{\lambda h^{1/3}}$.*

This conclusion (see [64, Theorem 11.5]) improves on [16, Theorem 2], where a similar result is obtained with $(h/\log h)^{1/3}$ in place of the exponent $h^{1/3}$.

(iii) *The Hilbert-Kamke problem and its brethren.* Hilbert [23] considered an extension of Waring's problem related to Vinogradov's mean value theorem. When $n_1, \ldots, n_k \in \mathbb{N}$, let $R_{s,k}(\boldsymbol{n})$ denote the number of solutions of the system

$$x_1^j + \ldots + x_s^j = n_j \quad (1 \leq j \leq k), \tag{8.2}$$

[4] See the website http://euler.free.fr/eslp/eslp.htm.

with $\boldsymbol{x} \in \mathbb{N}^s$. Put $X = \max_{1 \le j \le k} n_j^{1/j}$, and then write

$$\mathcal{J}_{s,k}(\boldsymbol{n}) = \int_{\mathbb{R}^k} \bigg(\int_0^1 e(\beta_1 \gamma + \ldots + \beta_k \gamma^k)\, \mathrm{d}\gamma \bigg)^s e(-\beta_1 n_1/X - \ldots - \beta_k n_k/X^k)\, \mathrm{d}\boldsymbol{\beta}$$

and

$$\mathcal{S}_{s,k}(\boldsymbol{n}) = \sum_{q=1}^{\infty} \sum_{\substack{1 \le a_1, \ldots, a_k \le q \\ (q, a_1, \ldots, a_k) = 1}} \big(q^{-1} f_k(\boldsymbol{a}/q; q) \big)^s\, e\left(-(a_1 n_1 + \ldots + a_k n_k)/q \right).$$

See [1, 34, 35] for an account of the analysis of this problem, and in particular for a discussion of the conditions under which real and p-adic solutions exist for the system (8.2). While the conditions $n_k^{j/k} \le n_j \le s^{1-j/k} n_k^{j/k}$ $(1 \le j \le k)$ are plainly necessary, one finds that p-adic solubility is not assured when $s < 2^k$. This classical technology gives an asymptotic formula for $R_{s,k}(\boldsymbol{n})$ provided that $s \ge (4 + o(1))k^2 \log k$. Efficient congruencing methods lead to considerable progress. The following result improves on [62, Theorem 9.2] using Theorem 4.1(iii).

Theorem 8.3. *Let $s, k \in \mathbb{N}$ and $\boldsymbol{n} \in \mathbb{N}^k$. Suppose that $X = \max n_j^{1/j}$ is sufficiently large in terms of s and k, and that the system (8.2) has non-singular real and p-adic solutions. Then whenever $k \ge 3$ and $s \ge 2k^2 - 2k + 1$, one has*

$$R_{s,k}(\boldsymbol{n}) = \mathcal{J}_{s,k}(\boldsymbol{n}) \mathcal{S}_{s,k}(\boldsymbol{n}) X^{s-k(k+1)/2} + o(X^{s-k(k+1)/2}).$$

Similar arguments apply to more general Diophantine systems. Let $k_1, \ldots, k_t$ be distinct positive integers. Suppose that $s, k \in \mathbb{N}$, and that $a_{ij} \in \mathbb{Z}$ for $1 \le i \le t$ and $1 \le j \le s$. Write

$$\phi_i(\boldsymbol{x}) = a_{i1} x_1^{k_i} + \ldots + a_{is} x_s^{k_i} \quad (1 \le i \le t),$$

and consider the Diophantine system $\phi_i(\boldsymbol{x}) = 0$ $(1 \le i \le t)$. We write $N(B; \boldsymbol{\phi})$ for the number of integral solutions of this system with $|\boldsymbol{x}| \le B$. When $L > 0$, define

$$\sigma_\infty = \lim_{L \to \infty} \int_{|\boldsymbol{\xi}| \le 1} \prod_{i=1}^{t} \max\left\{0, L(1 - L|\phi_i(\boldsymbol{\xi})|)\right\}\, \mathrm{d}\boldsymbol{\xi}.$$

Also, for each prime number p, put

$$\sigma_p = \lim_{H \to \infty} p^{H(t-s)} \mathrm{card}\{\boldsymbol{x} \in (\mathbb{Z}/p^H\mathbb{Z})^s : \phi_i(\boldsymbol{x}) \equiv 0 \pmod{p^H}\ (1 \le i \le t)\}.$$

By applying the Hardy-Littlewood method, a fairly routine application of Theorem 4.1(iii) delivers the following conclusion (compare [62, Theorem 9.1]).

Theorem 8.4. *Let s and k be natural numbers with $k \ge 3$ and $s \ge 2k^2 - 2k + 1$. Suppose that $\max k_i \le k$, and that a_{ij} $(1 \le i \le t, 1 \le j \le s)$ are non-zero integers. Suppose in addition that the system of equations $\phi_i(\boldsymbol{x}) = 0$ $(1 \le i \le t)$ has non-singular real and p-adic solutions, for each prime number p. Then*

$$N(B; \boldsymbol{\phi}) \sim \sigma_\infty \bigg(\prod_p \sigma_p \bigg) B^{s - k_1 - \ldots - k_t}.$$

(iv) *Solutions of polynomial congruences in short intervals.* There has been much activity in recent years concerning the solubility of polynomial congruences in short intervals, some of which makes use of estimates associated with Vinogradov's mean value theorem. Let $f \in \mathbb{F}_p[X]$ have degree $m \geq 3$, and let M be a positive integer with $M < p$. Denote by $I_f(M; R, S)$ the number of solutions of the congruence $y^2 \equiv f(x) \pmod p$, with $(x, y) \in [R+1, R+M] \times [S+1, S+M]$. Weil's bounds for exponential sums yield the estimate $I_f(M; R, S) = M^2 p^{-1} + O(p^{1/2}(\log p)^2)$, one that is worse than trivial for $M \leq p^{1/2}(\log p)^2$. The work of Chang et al. [14] gives estimates that remain non-trivial for significantly smaller values of M.

Theorem 8.5. *Let $f \in \mathbb{F}_p[X]$ be any polynomial of degree $m \geq 4$. Then whenever M is a positive integer with $1 \leq M < p$, we have*

$$I_f(M; R, S) \ll M^{1+\varepsilon} \left(M^3 p^{-1} + M^{3-m}\right)^{1/(2m(m-1))}.$$

This follows from [14, Theorem 4] on applying Theorem 4.1(iii). In particular, for any $\varepsilon > 0$, one finds that there exists a $\delta > 0$, depending only on ε and $\deg(f)$, such that whenever $M < p^{1/3-\varepsilon}$ and $\deg(f) \geq 4$, then $I_f(M; R, S) \ll M^{1-\delta}$.

(v) *The zero-free region for the Riemann zeta function.* We would be remiss not to mention the role of Vinogradov's mean value theorem in the proof of the widest available zero-free region for the Riemann zeta function. The sharpest estimates date from work of Vinogradov [52] and Korobov [31] in 1958 (see also [53]). Thus, there is a positive constant c_1 with the property that $\zeta(s) \neq 0$ when $s = \sigma + it$, with $\sigma, t \in \mathbb{R}$, whenever $|t| \geq 3$ and $\sigma \geq 1 - c_1(\log|t|)^{-2/3}(\log\log|t|)^{-1/3}$. More recently, Ford [19] has shown that one may take $c_1 = 1/57.54$. This, in turn leads to an effective version of the prime number theorem of the shape

$$\pi(x) = \int_2^x \frac{\mathrm{d}t}{\log t} + O\left(x \exp\left(-c_2(\log x)^{3/5}(\log\log x)^{-1/5}\right)\right),$$

where $c_2 = 0.2098$. Using currently available methods, the nature of the constant $C(k, r)$ in estimates of the shape (1.4) is significant for estimates of this type, while the precise nature of the defect in the exponent $\Delta_{s,k}$ less so. Thus, although the new estimates for $J_{s,k}(X)$ stemming from efficient congruencing have the potential to impact the numerical constants c_1 and c_2, the dependence on t and x, respectively, in the above estimates has not been affected.

9. Generalisations

Thus far, we have focused on estimates for $J_{s,k}(X)$, the number of solutions (over the ring $\mathbb{Z}$) of the translation-dilation invariant system (1.3) with $1 \leq \boldsymbol{x}, \boldsymbol{y} \leq X$. Previous authors have considered generalisations in which either the ring, or else the translation-dilation invariant system, is varied.

(i) *Algebraic number fields.* The arguments underlying the proof of Theorem 4.1 change little when the setting is shifted from $\mathbb{Z}$ to the ring of integers of a number field. When $s \geq k(k-1)$, the ensuing estimates are at most a factor X^{ε} away from the upper bound predicted by a heuristic application of the circle method. In common with Birch's use of Hua's lemma in number fields [7], our estimates are therefore robust to variation in the degree of the field extension, since Weyl-type estimates for exponential sums no longer play a significant role in applications. In forthcoming work we apply such ideas to establish the following result.

Theorem 9.1. *Let $L/\mathbb{Q}$ be a field extension of finite degree. Suppose that $d \geq 3$, $s > 2d(d-1)$ and $\boldsymbol{a} \in (L^\times)^s$. Then the hypersurface defined by $a_1x_1^d + \ldots + a_sx_s^d = 0$ satisfies weak approximation and the Hasse principle over L.*

For comparison, Birch [7, Theorem 3] gives such a conclusion only for $s > 2^d$, while the work of Körner [30] yields analogous conclusions in which the number of variables is larger, and depends also on the degree of the field extension $L/\mathbb{Q}$.

(ii) *Function fields.* Consider a finite field $\mathbb{F}_q$ of characteristic p. Let $B \in \mathbb{N}$ be large enough in terms of q, k and s, and denote by $J_{s,k}(B;q)$ the number of solutions of (1.3) with $x_i, y_i \in \mathbb{F}_q[t]$ $(1 \leq i \leq s)$ having degree at most B. When $p < k$, one can reduce (1.3) to a minimal translation-invariant system in which certain equations are omitted. We write K for the sum of the degrees of this minimal system, so that $K = k(k+1)/2$ when $p > k$, and $K < k(k+1)/2$ when $p < k$. Then, when $s \geq k(k+1)$, the efficient congruencing method adapts to give the upper bound $J_{s,k}(B;q) \ll (q^B)^{2s-K+\varepsilon}$. This and much more is contained in forthcoming work of the author joint with Y.-R. Liu, generalisations of which are described in [32].

(iii) *Multidimensional analogues.* Vinogradov's methods have been generalised to multidimensional settings by Arkhipov, Chubarikov and Karatsuba [2, 4], Parsell [36] and Prendiville [39]. Variants of the efficient congruencing method deliver much sharper conclusions in far greater generality. Let $r, s, d \in \mathbb{N}$, and consider a linearly independent system of homogeneous polynomials $\boldsymbol{F} = (F_1, \ldots, F_r)$, where $F_j(\boldsymbol{z}) \in \mathbb{Z}[z_1, \ldots, z_d]$. Suppose that for $1 \leq j \leq r$ and $1 \leq l \leq d$, the polynomial $\partial F_j/\partial z_l$ lies in $\mathrm{span}(1, F_1, \ldots, F_r)$. Such a *reduced translation-dilation invariant* system is said to have *rank* r, *dimension* d, *degree* $k = \max \deg F_j$, and *weight* $K = \sum_1^r \deg F_j$. Denote by $J_s(X;\boldsymbol{F})$ the number of integral solutions of the system of equations

$$\sum_{i=1}^{s} (F_j(\boldsymbol{x}_i) - F_j(\boldsymbol{y}_i)) = 0 \quad (1 \leq j \leq r),$$

with $1 \leq \boldsymbol{x}_i, \boldsymbol{y}_i \leq X$ $(1 \leq i \leq s)$. The work of Parsell, Prendiville and the author [38, Theorem 2.1] provides a general estimate for $J_s(X;\boldsymbol{F})$ matching the predictions of the appropriate analogue of the Main Conjecture.

Theorem 9.2. *Let $\boldsymbol{F}$ be a reduced translation-dilation invariant system of rank r, dimension d, degree k and weight K. Then $J_s(X;\boldsymbol{F}) \ll X^{2sd-K+\varepsilon}$ for $s \geq r(k+1)$.*

Reduced translation-dilation invariant systems are easy to generate by taking successive partial derivatives and reducing to a linearly independent spanning set. Thus, for example, the initial seed $x^5 + 3x^2y^3$ gives rise to just such a system

$$\boldsymbol{F} = \{x^5 + 3x^2y^3, 5x^4 + 6xy^3, x^2y^2, 10x^3 + 3y^3, xy^2, x^2y, x^2, xy, y^2, x, y\},$$

with $d = 2$, $r = 11$, $k = 5$, $K = 30$. We therefore see from Theorem 9.2 that $J_s(X;\boldsymbol{F}) \ll X^{4s-30+\varepsilon}$ for $s \geq 66$. Theorem 9.2 should be susceptible to improvement by using the ideas underlying multigrade efficient congruencing [65–67].

10. Challenges

The remarkable success of the efficient congruencing method encourages ambitious speculation concerning other potential applications, a topic we briefly explore.

(i) (*The Main Conjecture for larger s*). In Theorem 4.1, one sees that the upper bound $J_{s,k}(X) \ll X^{s+\varepsilon}$ predicted by the Main Conjecture is now known to hold for $1 \le s \le \frac{1}{2}k(k+1) - t_k$, where $t_k = \frac{1}{3}k + O(k^{2/3})$. In striking contrast, on the other side of the critical value $s = \frac{1}{2}k(k+1)$, the upper bound $J_{s,k}(X) \ll X^{2s-k(k+1)/2+\varepsilon}$ is known to hold only when $s \ge \frac{1}{2}k(k+1) + u_k$, where $u_k = \frac{1}{2}k(k-3)$. Plainly, the value of u_k is substantially larger than t_k, and an intriguing possibility is that a hitherto unseen refinement of the method might reduce u_k to a size more similar to that of t_k. This would have great significance in numerous applications.

(ii) (*Paucity*). When $k \ge 3$ and $1 \le s < \frac{1}{2}k(k+1)$, we have precise asymptotics for $J_{s,k}(X)$ only when $s \le k+1$. Since the formula $J_{s,k}(X) = T_s(X) \sim s!X^s$ is trivial for $1 \le s \le k$, the case $s = k+1$ is the only one with content. It is tempting to speculate that a suitable adaptation of efficient congruencing might confirm that $J_{s,k}(X) = T_s(X) + O(X^{s-\delta})$, for some $\delta > 0$, for some exponent $s \ge k+2$.

(iii) (*Minor arc bounds*). When $q \in \mathbb{N}$ and $\boldsymbol{a} \in \mathbb{Z}^k$, denote by $\mathfrak{M}(q,\boldsymbol{a})$ the set of points $\boldsymbol{\alpha} \in [0,1)^k$ such that $|q\alpha_j - a_j| \le X^{1-j}$ $(1 \le j \le k)$. Write $\mathfrak{M}$ for the union of the boxes $\mathfrak{M}(q,\boldsymbol{a})$ with $0 \le a_j \le q \le X$ $(1 \le j \le k)$ and $(q,a_1,\dots,a_k) = 1$, and put $\mathfrak{m} = [0,1)^k \setminus \mathfrak{M}$. The methods of §7 provide estimates of the shape $|f_k(\boldsymbol{\alpha};X)| \ll X^{1-\sigma_k+\varepsilon}$ for $\boldsymbol{\alpha} \in \mathfrak{m}$. However, when $s = k(k-1)+t$ and $t \ge 1$, our most efficient means of estimating moments of $f_k(\boldsymbol{\alpha};X)$ of order $2s$, restricted to minor arcs, proceeds by applying Theorem 4.1(iii) via the trivial bound

$$\begin{aligned}\int_{\mathfrak{m}} |f_k(\boldsymbol{\alpha};X)|^{2s}\,\mathrm{d}\boldsymbol{\alpha} &\ll \Big(\sup_{\boldsymbol{\alpha}\in\mathfrak{m}} |f_k(\boldsymbol{\alpha};X)|\Big)^{2t} \oint |f_k(\boldsymbol{\alpha};X)|^{2k(k-1)}\,\mathrm{d}\boldsymbol{\alpha}\\ &\ll X^{2s-\frac{1}{2}k(k+1)-2t\sigma_k+\varepsilon}.\end{aligned}$$

This bound is relatively weak, even when t is large. Efficient congruencing provides a possible means of deriving estimates directly for such moments, and might even lead to improvements in our lower bounds for permissible exponents σ_k.

(iv) (*Non-translation invariant systems*). The system (1.3) is translation-dilation invariant. A major desideratum is to apply a variant of efficient congruencing to systems of equations that are *not* translation invariant. The author has forthcoming work applicable to systems that are only approximately translation invariant.

References

[1] Arkhipov, G. I., *On the Hilbert-Kamke problem*, Izv. Akad. Nauk SSSR Ser. Mat. **48** (1984), no. 1, 3–52.

[2] Arkhipov, G. I., Chubarikov, V. N., and Karatsuba, A. A., *Trigonometric sums in number theory and analysis*, Walter de Gruyter, Berlin, 2004.

[3] Arkhipov, G. I. and Karatsuba, A. A., *A new estimate of an integral of I. M. Vinogradov*, Izv. Akad. Nauk SSSR Ser. Mat. **42** (1978), no. 4, 751–762.

[4] Arkhipov, G. I., Karatsuba, A. A., and Chubarikov, V. N., *Multiple Trigonometric Sums*, Trudy Mat. Inst. Steklov **151** (1980), 1–126.

[5] Baker, R. C., *Diophantine inequalites*, London Mathematical Society Monographs, vol. **1**, Oxford University Press, Oxford, 1986.

[6] ———, Correction to: "Weyl sums and Diophantine approximation" [J. London Math. Soc. (2) 25 (1982), no. 1, 25–34], J. London Math. Soc. (2) **46** (1992), no. 2, 202–204.

[7] Birch, B. J., *Waring's problem in algebraic number fields*, Proc. Cambridge Philos. Soc. **57** (1961), 449–459.

[8] Boklan, K. D., *The asymptotic formula in Waring's problem*, Mathematika **41** (1994), no. 2, 329–347.

[9] Boklan, K. D. and Wooley, T. D., *On Weyl sums for smaller exponents*, Funct. Approx. Comment. Math. **46** (2012), no. 1, 91–107.

[10] Blomer, V. and Brüdern, J., *The number of integer points on Vinogradov's quadric*, Monatsh. Math. **160** (2010), no. 3, 243–256.

[11] Bombieri, E., *On Vinogradov's mean value theorem and Weyl sums*, Automorphic forms and analytic number theory (Montreal, PQ, 1989), pp. 7–24, Univ. Montréal, Montreal, QC, 1990.

[12] Bourgain, J. and Chang, M.-C., *On the size of k-fold sum and product sets of integers*, J. Amer. Math. Soc. **17** (2004), no. 2, 473–497.

[13] Chang, M.-C., *The Erdős-Szemerédi problem on sum set and product set*, Ann. of Math. (2) **157** (2003), no. 3, 939–957.

[14] Chang, M.-C., Cilleruelo, J., Garaev, M. Z., Hernández, J., Shparlinski, I. E., and Zumulacárregui, A., *Points on curves in small boxes and applications*, Michigan Math. J. (to appear), arXiv:1111.1543.

[15] van der Corput, J. G., *Verscharfung der Abschätzungen beim Teilerproblem*, Math. Ann. **87** (1922), 39–65.

[16] Croot, E. and Hart, D., *h-fold sums from a set with few products*, SIAM J. Discrete Math. **24** (2010), no. 2, 505–519.

[17] Erdős, P. and Szemerédi, E., *On sums and products of integers*, Studies in Pure Mathematics, Birkhäuser, Basel, 1983, pp. 213–218.

[18] Ford, K. B., *New estimates for mean values of Weyl sums*, Internat. Math. Res. Notices (1995), no. 3, 155–171.

[19] ———, *Vinogradov's integral and bounds for the Riemann zeta function*, Proc. London Math. Soc. (3) **85** (2002), no. 3, 565–633.

[20] Ford, K. B. and Wooley, T. D., *On Vinogradov's mean value theorem: strongly diagonal behaviour via efficient congruencing, submitted*, arXiv:1304.6917.

[21] Hardy, G. H. and Littlewood, J. E., *Some problems of 'Partitio Numerorum': IV. The singular series in Waring's Problem and the value of the number $G(k)$*, Math. Zeit. **12** (1922), 161–188.

[22] Heath-Brown, D. R., *Weyl's inequality, Hua's inequality, and Waring's problem*, J. London Math. Soc. (2) **38** (1988), no. 2, 216–230.

[23] Hilbert, D., *Beweis für die Darstellbarkeit der ganzen Zahlen durch eine feste Anzahl n^{ter} Potenzen (Waringsches Problem)*, Math. Ann. **67** (1909), no. 3, 281–300.

[24] Hua, L.-K., *On Waring's problem*, Quart. J. Math. Oxford **9** (1938), 199–202.

[25] ———, *On Tarry's problem*, Quart. J. Math. Oxford **9** (1938), 315–320.

[26] ———, *The additive prime number theory*, Trav. Inst. Math. Stekloff, **22**, Acad. Sci. USSR, Moscow-Leningrad, 1947.

[27] ———, *An improvement of Vinogradov's mean value theorem and several applications*, Quart. J. Math. Oxford **20** (1949), 48–61.

[28] ———, *Improvement of a result of Wright*, J. London Math. Soc. **24** (1949), 157–159.

[29] Karatsuba, A. A., *The mean value of the modulus of a trigonometric sum*, Izv. Akad. Nauk SSSR Ser. Mat. **37** (1973), 1203–1227.

[30] Körner, O., *Über Mittelwerte trigonometrischer Summen und ihre Anwendung in algebraischen Zahlkörpern*, Math. Ann. **147** (1962), 205–239.

[31] Korobov, N. M., *Estimates of trigonometric sums and their applications*, Uspehi Mat. Nauk **13** (1958), no. 4 (82), 185–192.

[32] Kuo, W., Liu, Y.-R., and Zhao, X., *Multidimensional Vinogradov-type estimates in function fields*, Canad. J. Math., in press.

[33] Linnik, Yu. V., *On Weyl's sums*, Mat. Sbornik (Rec. Math.) N. S. **12** (1943), 28–39.

[34] Mit′kin, D. A., *Estimate for the number of summands in the Hilbert-Kamke problem*, Mat. Sbornik (N.S.) **129** (1986), no. 4, 549–577.

[35] ———, *Estimate for the number of summands in the Hilbert-Kamke problem*, II, Mat. Sbornik (N.S.) **132** (1987), no. 3, 345–351.

[36] Parsell, S. T., *A generalization of Vinogradov's mean value theorem*, Proc. London Math. Soc. (3) **91** (2005), no. 1, 1–32.

[37] ———, *A note on Weyl's inequality for eighth powers*, Rocky Mountain J. Math., in press.

[38] Parsell, S. T., Prendiville, S. M., and Wooley, T. D., *Near-optimal mean value estimates for multidimensional Weyl sums*, Geom. Funct. Anal. **23** (2013), no. 6, 1962–2024.

[39] Prendiville, S. M., *Solution-free sets for sums of binary forms*, Proc. London Math. Soc. (3) **107** (2013), no. 2, 267–302.

[40] Robert, O. and Sargos, P., *Un théorème de moyenne pour les sommes d'exponentielles. Application à l'inégalité de Weyl*, Publ. Inst. Math. (Beograd) (N.S.) **67** (2000), 14–30.

[41] Rogovskaya, N. N., *An asymptotic formula for the number of solutions of a system of equations*, Diophantine Approximations, Part II, Moskov. Gos. Univ., Moscow, 1986, pp. 78–84.

[42] Stechkin, S. B., *On mean values of the modulus of a trigonometric sum*, Trudy Mat. Inst. Steklov **134** (1975), 283–309.

[43] Tyrina, O. V., *A new estimate for a trigonometric integral of I. M. Vinogradov*, Izv. Akad. Nauk SSSR Ser. Mat. **51** (1987), no. 2, 363–378.

[44] Ustinov, A. V., *On the number of summands in the asymptotic formula for the number of solutions of the Waring equation*, Mat. Zametki **64** (1998), no. 2, 285–296.

[45] Vaughan, R. C., *On Waring's problem for cubes*, J. Reine Angew. Math. **365** (1986), 122–170.

[46] ———, *On Waring's problem for smaller exponents, II*, Mathematika **33** (1986), no. 1, 6–22.

[47] ———, *The Hardy-Littlewood method*, 2nd edition, Cambridge University Press, Cambridge, 1997.

[48] Vaughan, R. C. and Wooley, T. D., *On a certain nonary cubic form and related equations*, Duke Math. J. **80** (1995), no. 3, 669–735.

[49] ———, *A special case of Vinogradov's mean value theorem*, Acta Arith. **79** (1997), no. 3, 193–204.

[50] Vinogradov, I. M., *New estimates for Weyl sums*, Dokl. Akad. Nauk SSSR **8** (1935), 195–198.

[51] ———, *The method of trigonometrical sums in the theory of numbers*, Trav. Inst. Math. Stekloff **23** (1947), p.109.

[52] ———, *A new estimate of the function* $\zeta(1+it)$, Izv. Akad. Nauk SSSR. Ser. Mat. **22** (1958), 161–164.

[53] Walfisz, A. Z., *Weylsche Exponentialsummen in der neueren Zahlentheorie*, Deutscher Verlag der Wissenschaften, Berlin, 1963.

[54] Weyl, H., *Über die Gleichverteilung von Zahlen mod Eins*, Math. Ann. **77** (1916), 313–352.

[55] Wooley, T. D., *Large improvements in Waring's problem*, Ann. of Math. (2) **135** (1992), no. 1, 131–164.

[56] ———, *On Vinogradov's mean value theorem*, Mathematika **39** (1992), no. 2, 379–399.

[57] ———, *On Vinogradov's mean value theorem, II*, Michigan Math. J. **40** (1993), no. 1, 175–180.

[58] ———, *Quasi-diagonal behaviour in certain mean value theorems of additive number theory*, J. Amer. Math. Soc. **7** (1994), no. 1, 221–245.

[59] ———, *New estimates for Weyl sums*, Quart. J. Math. Oxford (2) **46** (1995), no. 1, 119–127.

[60] ———, *Some remarks on Vinogradov's mean value theorem and Tarry's problem*, Monatsh. Math. **122** (1996), no. 3, 265–273.

[61] ———, *Diophantine methods for exponential sums, and exponential sums for Diophantine problems*, Proceedings of the International Congress of Mathematicians, August 20–28, 2002, Beijing, Volume II, Higher Education Press, 2002, pp. 207–217.

[62] ———, *Vinogradov's mean value theorem via efficient congruencing*, Ann. of Math. (2) **175** (2012), no. 3, 1575–1627.

[63] ———, *The asymptotic formula in Waring's problem*, Internat. Math. Res. Notices (2012), no. 7, 1485–1504.

[64] ———, *Vinogradov's mean value theorem via efficient congruencing, II*, Duke Math. J. **162** (2013), no. 4, 673–730.

[65] ———, *Multigrade efficient congruencing and Vinogradov's mean value theorem*, submitted, arXiv:1310.8447.

[66] ———, *Approximating the main conjecture in Vinogradov's mean value theorem*, submitted, arXiv:1401.2932.

[67] ———, *The cubic case of the main conjecture in Vinogradov's mean value theorem*, submitted, arXiv:1401.3150.

[68] ———, *Mean value estimates for odd cubic Weyl sums*, submitted, arXiv:1401.7152.

[69] Wright, E. M., *The Prouhet-Lehmer problem*, J. London Math. Soc. **23** (1948), 279–285.

School of Mathematics, University of Bristol, University Walk, Clifton, Bristol BS8 1TW, UK
E-mail: matdw@bristol.ac.uk

Elementary integration of differentials in families and conjectures of Pink

Umberto Zannier

Abstract. In this short survey paper we shall consider, in particular, indefinite integrals of differentials on algebraic curves, trying to express them in *elementary terms*. This is an old-fashioned issue, for which Liouville gave an explicit criterion that may be considered a primordial example of differential algebra. Before presenting some connections with more recent topics, we shall start with an overview of the classical facts, recalling some criteria for elementary integration and relating this with issues of torsion in abelian varieties. Then we shall turn to differentials in 1-parameter algebraic families, asking for which values of the parameter we can have an elementary integral. (This had been considered already in the 80s by J. Davenport.) The mentioned torsion issues provide a connection of this with a conjecture of R. Pink in the realm of the so-called *Unlikely Intersections*. In joint work in collaboration with David Masser (still partly in progress), we have proved finiteness of the set of relevant values, under suitable necessary conditions. Here we shall give a brief account of the whole context, pointing out at the end possible links with other problems.

Mathematics Subject Classification (2010). 11G10, 11G50

Keywords. Integration, abelian varieties, torsion points, unlikely intersections, conjecture of Pink.

1. Integration in finite terms

Since the invention of integral calculus and the realization of indefinite integration as a process inverse to differentiation, a development occurred towards an 'algebra' for 'explicit' calculations in this direction. Some basic functions, like polynomials and rational functions, the exponential function e^z and its inverse $\log z$, as well as the basic trigonometric functions $\sin z, \cos z, \tan z$ and their inverses, admit indefinite integrals (i.e. a primitive)[1]which again may be expressed 'in terms of these functions'. The application of a few rules (derived from corresponding rules for differentiation) allows expressions of the same kind for other functions obtained from the former ones by rational operations and by composition. Then, a (rough) question that arose naturally, and still today presents itself very soon, already to freshmen, is: *Which other functions, constructed from the above mentioned ones (for instance), do admit an indefinite integral expressed again in similar fashion?*

To give more precise meaning to this, we may define the class of *elementary functions* as those obtained from the rational functions over $\mathbb{C}$ by a finite number of operations of the following three kinds:

(a) Algebraic operations; namely, if $f_1(z), \dots, f_d(z)$ have been already obtained, we al-

Proceedings of the International Congress of Mathematicians, Seoul, 2014

low in our class all rational functions in them and a solution $y = y(z)$ of

$$y^d + f_1(z)y^{d-1} + \ldots + f_d(z) = 0.$$

(b) Exponentiation: if $f(z)$ has been obtained, we allow $\exp f(z)$.

(c) Taking a logarithm: we allow $\log f(z)$.

Of course (b), (c) lead to trigonometrical functions as well. Some care is needed concerning e.g. the domain of definition, and also, in (a) and (c), which branch of the functions has to be taken. For instance, one may agree to consider open disks in $\mathbb{C}$ as domains, restricting possibly the disk each time an operation is performed, and choosing an arbitrary branch. In considering finite sets of functions, this procedure is legitimate, and we can even assume that our functions are meromorphic in a sufficiently small disk, so that in particular they form a field, which is moreover closed under differentiation.[2]

Needless to say, one could start with a larger class of functions, or allow other kinds of operations, like solutions of differential equations of prescribed type; however, this rapidly lead to subtle issues in differential Galois theory, and here we shall stick to the above basic special case (except for a few comments in the last section). We refer to Rosenlicht's article [29] for a self-contained detailed and clear account of this, to Risch's paper [25] for a more formal and general treatment, to Hardy's book [7] for a somewhat general treatment, however often based on examples and with few proofs, and to Ritt's [27] book for further and rather more involved issues, in several directions, and for other references.[3]

Once we have decided which is our class of elementary functions, we can say that a primitive of one such function may be expressed in *finite terms* (or in *elementary terms*, or in *closed form*) if it belongs itself to the class in question. In spite of the many indefinite integrals which can be likewise obtained, sometimes at the cost of some ingenious trick, a freshman shall soon be faced with some other ones, of quite simple functions, which he can not express in simple terms. Let us now see a few of these explicit instances which seem to defeat every attempt:

The integral $\int \frac{\mathrm{d}z}{\log z}$ is an example (coming also from Prime Number Theory); it is transformed into $\int e^z \frac{\mathrm{d}z}{z}$ by replacing z with e^z. Another one which very soon appears is the Gaussian $\int e^{-z^2}\mathrm{d}z$, coming from Probability Theory; the substitution $z^2 \leftrightarrow z$ reduces it to $\int e^{-z}\frac{\mathrm{d}z}{\sqrt{z}}$.

Integrals of algebraic differentials also are puzzling, and indeed lead to quite important issues; let us see some examples, the simplest ones coming with rational functions. It is usually explained in undergraduate courses that such a function $R(z) \in \mathbb{C}(z)$ can be integrated in finite terms, actually as a sum of a rational function and a linear combination of functions $\log(z-a)$ (at least provided we allow complex numbers). This is obtained on decomposing $R(z)$ in *partial fractions*, that is expressing it as a $\mathbb{C}$-linear combination of terms z^m and $(z-a)^{-m}$ ($m \in \mathbb{N}$).

---

[1] In fact, it would be more sensible to speak of integral, or primitive, of a *differential* $f(z)\mathrm{d}z$ rather than 'integral of a function', but we shall use also the latter terminology when it causes no confusion.

[2] This is for instance the viewpoint adopted in [29]; it is to be mentioned that this does not suffice for more abstract investigations of algebraic differential equations. In the sequel we shall not pause on such precision.

[3] Some other rather classical references in Differential Algebra are Kolchin's [10] and Kaplanski's [8] books, the latter being short but giving a very clear and useful account; a further recent and more advanced one is van der Put and Singer's treatise [24]. We point out that these books touch especially differential Galois theory and are not directly involved with our more basic context, which, in Rosenlicht's words, is in a sense 'pre-Galois'.

The same happens when we integrate a differential on a curve of genus zero, because we may parametrize the curve (and hence the differential) rationally. The best-known examples probably occur with conic curves; so, any integral $\int R(z, \sqrt{1-z^2}) \cdot \mathrm{d}z$, where $R \in \mathbb{C}(u, v)$, may be reduced to the case of rational functions through the parametrization $w = \frac{2t}{1+t^2}$, $z = \frac{1-t^2}{1+t^2}$ of the circle $w^2 + z^2 = 1$ (with inverse $t = (1-z)/w$). In particular, we find (up to a constant and up to a choice of $i = \sqrt{-1}$) the formula $\int \frac{\mathrm{d}z}{\sqrt{1-z^2}} = -i \log \frac{i+2t-it^2}{1+t^2} = -i \cdot \log(iz + \sqrt{1-z^2}) = \arcsin z$.

On the contrary, if we consider curves of higher genus, often any attempt shall fail; for instance, trying to compute the length of a lemniscate[4] leads to an integral of the form $\int \frac{\mathrm{d}z}{\sqrt{1-z^4}}$, similar only in shape to the previous one. Indeed, it seems that Fagnano tried unsuccessfuly to rationalize the integrand, on imitating the case of the circle; this failure however eventually led him to important formulae, which were subsequently recognized as duplication formulae for elliptic functions (or for points on an elliptic curve). In fact, the integral (locally) represents essentially an inverse to a Weierstrass function $\wp(z)$ associated to the elliptic curve $w^2 = 1 - z^4$ (in the same way that the former integral represents an inverse to $\sin z$). Some thirty years later Euler, with closely related investigations, arrived to general addition formulae (see the discussion in Siegel's book [33], Ch. 1). This elliptic instance (and other ones in higher genus) shall be quite relevant below.

Now, in all of these cases, the question arises on *how to prove the impossibility of an integration in finite terms?* How to convince ourselves that the failure is not merely due to a lack of ingenuity on our part?

In a sense, a fairly complete answer to this was given by Liouville, who presented a theory of integration in finite terms between 1833 and 1841. This is illustrated (in particular) in [27], but here we shall follow [29], which contains an algebraic proof of a generalization by Ostrowski (1946) of a theorem proved by Liouville's in 1835, giving a most useful criterion.

To explain this, we first recall that a differential field F is a field equipped with a derivation denoted $a \to a'$, such that $(a+b)' = a' + b'$ and $(ab)' = a'b + ab'$. The *constants* are defined as the elements $c \in F$ with $c' = 0$; they form a subfield containing 1. We agree to call a an *exponential* of b, or b a *logarithm* of a, if $b' = a'/a$.

In agreement with the above notion, we define an *elementary (differential) extension of* F to be a differential field obtained from F by a finite sequence of adjunctions of elements which are either algebraic, or exponentials or logarithms, i.e. a field $F(t_1, \ldots, t_n)$, where for each $i = 1, \ldots, n$, the element t_i is either algebraic over $F(t_1, \ldots, t_{i-1})$, or the exponential or the logarithm of an element of $F(t_1, \ldots, t_{i-1})$. We find back the previously sketched concept in case F is the field of meromorphic functions on a given disk, provided we restrict possibly the disk each time we add a new element.

With these definitions, following again [29] (see also [28]), let us state the alluded result:

Theorem 1.1 ([28, 29]). *Let F be a differential field of characteristic zero and let $\alpha \in F$. If the equation $y' = \alpha$ has a solution in some elementary differential extension field of F having the same subfield of constants, then there are constants $c_1, \ldots, c_n \in F$ and elements $u_1, \ldots, u_n, v \in F$ such that*

$$\alpha = v' + \sum_{i=1}^{n} c_i \frac{u_i'}{u_i}.$$

[4] This is the set of points in the plane such that the product of the distances from two given points has a given product; it is expressed by an equation of degree 4.

Remark 1.2.

(i) We immediately point out that the assumption about the subfield of constants is automatic in the cases of the fields of meromorphic functions mentioned above.[5] In any case, in the first place one can apply the result after appropriate enlargement of the constants, and actually the proof can be modified to show that the result remains true assuming only that the field of constants of F is algebraically closed. We omit details in this brief account; a corresponding sharper form of the theorem can be found in [25].

(ii) Note that there is an obvious converse: if α has the stated shape, then $\alpha = \beta'$, where $\beta = v + \sum c_i \log u_i$ and where 'log' is to be interpreted in the above 'differential' sense, which of course coincides with the usual one in the said context of meromorphic functions, when the derivation is $\mathrm{d}/\mathrm{d}z$; in that case $\int \alpha \cdot \mathrm{d}z = \beta$. Clearly β lies in an elementary extension of F.

In [29] one finds some nice applications of this criterion. One concerns $f(z)e^{g(z)}$ for rational functions f, g. As proved by Liouville himself, $\int f(z)e^{g(z)}\mathrm{d}z$ *is elementary if and only if there is a rational function* $a \in \mathbb{C}(z)$ *such that* $f = a' + ag'$; a proof is not too difficult starting from the theorem (with $F = \mathbb{C}(z, e^{g(z)})$). In turn, one may readily check that $\int e^{-z^2}\mathrm{d}z$ is not elementary (we find the equation $1 = a' - 2az$, with no rational function solutions: a could not have poles and would then be constant) and similarly for $\int (e^z/z)\mathrm{d}z$. Certain substitutions, e.g. as in the above examples, or integrations by parts, lead to other impossibilities. With a bit more effort, in [29] it is checked that $\int (\sin z/z)\mathrm{d}z$ is non-elementary as well; more generally, partial fraction decompositions often allows to apply the last criterion to obtain either an explicit elementary integral or an impossibility proof.

We want now to concentrate on algebraic differentials, which shall be done in the next section. Before this, let us remark that a natural question, after Liouville's theorem, is: *How to decide algorithmically if a given function has an elementary primitive? And, in the affirmative case, how to compute it?* Of course, one must give first a suitable meaning of 'given function' and 'compute'; this may be done in a reasonable way. As remarked in the book [7], at Hardy's time a general algorithm was not known, and was first announced (with some details) in Risch's short paper [26]. It is to be observed that the steps of this procedure which are probably the most subtle arise with algebraic differentials, and are related to the main context of the present article. The whole matter is discussed in detail in [6], which also raises issues which motivated part of the work with Masser considered below.

2. Integration of some algebraic differentials and a Pell equation in polynomials

In this section we shall illustrate the previous result to analyze some 'concrete' special cases of elementary integrability of algebraic differentials. (We stress that this kind of analysis is by no means new, and our contributions to the topic shall appear later.)

We have mentioned above some integrals on curves of genus zero, and also Fagnano's and Euler's attempts with integrals of the shape $\int \frac{\mathrm{d}z}{\sqrt{1-z^4}}$, which correspond to work with

[5] Note that this need not hold for instance in the case when the original field of constants is not algebraically closed; a simple example comes from $\int (z^2+1)^{-1}\mathrm{d}z$ if we work over $\mathbb{R}$, not $\mathbb{C}$.

rational differentials $\mathrm{d}z/w$ on the elliptic curve $w^2 = 1 - z^4$. Let us then consider, more generally, differentials

$$\omega = \frac{h(z)\mathrm{d}z}{w}, \qquad w^2 = f(z), \tag{2.1}$$

where $f(z)$ is a polynomial of degree $2d$, without multiple zeros, and where $h(z) \neq 0$ is a polynomial, say of degree e. For brevity, here we shall stick to these special cases.

The curve so defined is hyperelliptic. We have given for it an equation in the affine (z, w)-plane; the closure in $\mathbb{P}_2$ of such affine curve is singular (only) at infinity (for $d > 1$), but it admits a smooth projective model, which we shall denote by H; it has genus $g = d-1$.

The shape of ω is somewhat natural, also because (as we shall check) for $0 \leq e \leq d-2$ we obtain the *regular* differentials on H. Note that the Fagnano-Euler differential corresponds to $d = 2, g = 1, e = 0$. Similarly to that case, the vector-integral of a $\mathbb{C}$-basis for the regular differentials locally gives an embedding of the curve in a torus $\mathbb{C}^g/\Lambda$, corresponding to its Jacobian variety J_H; the integrals are usually called *Abelian functions*. (See for instance [11], especially Ch. 4.)

As mentioned above, H is defined by $w^2 = f(z)$ on the whole affine plane; above each point $z = \alpha \in \mathbb{C}$ it has two points $(\alpha, \pm\sqrt{f(\alpha)})$, except when $f(\alpha) = 0$, in which case the said point of the z-line is branched with index 2. Above the point $z = \infty \in \mathbb{P}_1(\mathbb{C})$, H has again two points, denoted $\infty_\pm$, corresponding to the (Laurent-Puiseux) expansions $w = \pm a_d z^d +$lower order terms, where a_d^2 is the leading coefficient of f.

Let us denote by $F = \mathbb{C}(H) = \mathbb{C}(z, w)$ the function field of H over $\mathbb{C}$. It is a differential field with respect to the derivation $\mathrm{d}/\mathrm{d}z$; note that the equation for H yields $2w\mathrm{d}w = f'(z)\mathrm{d}z$, or $2ww' = f'(z)$, which determines the derivation on the whole F. By Theorem 1.1 we find that $\int \omega$ *is an elementary function if and only if there are* $v, u_1, \ldots, u_n \in F$ *and* $c_1, \ldots, c_n \in \mathbb{C}$ *such that*

$$\omega = \mathrm{d}v + \sum_{i=1}^{n} c_i \frac{\mathrm{d}u_i}{u_i}. \tag{2.2}$$

We now want to derive from (2.2) some necessary conditions; in particular, this shall prove some cases of impossibility of the integration of ω in finite terms, and shall yield explicit integral formulae in some other cases.

Let us notice at once that in (2.2) we may assume that n is minimal, and then $c_1, \ldots, c_n$ shall be linearly independent over $\mathbb{Q}$; if not, then, by taking a suitable $\mathbb{Q}$-basis $c'_1, \ldots, c'_m$ for $\sum \mathbb{Q}c_i$, with $m < n$, we may write $c_i = \sum_j b_{ij} c'_j$ with integers b_{ij}. Hence, recalling that logarithmic differentiation $u \mapsto \mathrm{d}u/u$ sends products to sums, we find $\sum_{i=1}^n c_i \mathrm{d}u_i/u_i = \sum_{j=1}^m c'_j \mathrm{d}v_j/v_j$, where $v_j = \prod_i u_i^{b_{ij}} \in F$. However, this contradicts minimality of n.

To analyze (2.2) we study the poles of ω. We view z, w as rational functions on H. Let $p \in H(\mathbb{C})$ and suppose first that $z(p) \in \mathbb{C}$ is finite. If $f(z(p)) \neq 0$, then $w(p) \neq 0$, $z - z(p)$ is a local parameter at p, and so certainly ω is regular at p. If $f(z(p)) = 0$, then w is a local parameter at p. As noted above, we have $2w\mathrm{d}w = f'(z)\mathrm{d}z$, whence $\omega = 2h(z)\mathrm{d}w/f'(z)$. Now, f has no multiple roots, hence $f'(z(p)) \neq 0$, proving that ω is again regular at p. Hence the only possible poles of ω are $\infty_\pm$. At each of these points, $\zeta := 1/z$ is a local parameter, and we have $\mathrm{d}z = -\mathrm{d}\zeta/\zeta^2$. Hence $\omega = h(z)\mathrm{d}\zeta/w\zeta^2$. The order of z at each point at infinity is -1, and the order of w is $-d$, whence the order of ω at each such point is $d - e - 2$. In particular, we find the previous assertion that ω is everywhere regular if and

only if $e \leq d-2$.[6] (A binomial expansion for w at infinity easily shows that the residues of ω at these poles are given by certain polynomials in the coefficients of f, h, linear in the last ones.)

Now, let us first suppose that $n = 0$, i.e. no u_i appears in (2.2), so $\omega = \mathrm{d}v$ is an *exact* differential already in F. (A necessary condition for this is that ω has no residues, but if $g > 0$ this is not sufficient.) Then, since ω has no finite poles, v cannot have finite poles as well, and hence is regular on the affine part of H; since the plane equation $w^2 = f(z)$ is nonsingular at finite points, we infer that $v = a(z) + wb(z)$ with polynomials a, b. This yields $\omega = (a' + (2w)^{-1}(2fb' + f'b))\mathrm{d}z$, so $a \in \mathbb{C}$ and $2h(z) = 2f(z)b'(z) + f'(z)b(z)$. This implies also in particular $e \geq 2d-1$.

Remark 2.1. The conclusion so obtained may be rephrased by saying that the image on polynomials of the linear differential operator of the first order $\Phi := 2f(z)\frac{\mathrm{d}}{\mathrm{d}z} + f'(z)$ contains $2h(z)$. Note that, although the kernel of Φ is trivial on the said space, Φ sends a polynomial of degree m to one of degree $m + 2d - 1$, hence certainly it is not surjective and it is 'unlikely' that a 'randomly' chosen $h(z)$ lies in the image.

Now suppose that $n \geq 1$. Note that if u_i has multiplicity $m_i \in \mathbb{Z}$ at $p \in H$, then $\mathrm{d}u_i/u_i$ has a simple pole at p, with residue m_i. In particular, no differential $\mathrm{d}u_i/u_i$ can have finite zeros or poles. In fact, any $p \in H$ is (at most) a simple pole of $\sum c_i \mathrm{d}u_i/u_i$, with residue $\sum c_i m_i$. But the m_i are integers and the c_i are linearly independent over $\mathbb{Q}$, hence the residue would not be zero as soon as some $m_i \neq 0$. On the other hand, $\mathrm{d}v$ has residue 0 at p, so p is certainly a pole of ω if some $m_i \neq 0$. Since ω has no finite poles, this proves the claim. (This also proves that if ω is regular then no u_i can appear, and the previous analysis then shows that regular differentials cannot be integrated in the sought way.)

Hence each u_i which appears has divisor supported at $\{\infty_+, \infty_-\}$; however, since the divisor of a function has degree 0, we infer that each u_i which appears has divisor of the shape $m_i(\infty_+ - \infty_-)$, for an $m_i \in \mathbb{Z}$. If some u_i appears, this implies in particular that the divisor class of $\infty_+ - \infty_-$ is torsion on the Jacobian J_H of H. Let m be the exact order of torsion; then m is the minimal integer > 0 such that there is a rational function $u \in F$ with $\mathrm{div}(u) = m(\infty_+ - \infty_-)$, and it is easy to see that in this case each u_i is, up to a constant factor, an integral power of u, so that in fact we may suppose that $n = 1$, $u_1 = u$.

Remark 2.2. The equation $\omega = \mathrm{d}u/u$, with algebraic ω, to be solved with an algebraic function u, appears for instance in Baldassarri and Dwork's paper [3], where algebraic solutions of second order linear differential equations are studied, also from the algorithmic viewpoint. (See e.g. pp. 69–70, where conclusions similar to those appearing below are mentioned.)

Let us draw some consequences from these last deductions. First, the function u is regular at finite points, so is of the shape $u = x(z) + wy(z)$, for $x, y \in \mathbb{C}[z]$, $y \neq 0$. We may suppose that u has a pole of order $m > 0$ at ∞_- and no other poles, so the conjugate function (over $\mathbb{C}(z)$), denoted u^σ, has a pole of order m at ∞_+ and no other poles. This easily yields $\deg x = m, \deg y = m - d$, and the norm to $\mathbb{C}(z)$ given by uu^σ must be constant (it has no poles), hence may be assumed to be 1 upon division. So we find

$$x(z)^2 - f(z)y(z)^2 = 1, \qquad y(z) \neq 0, \tag{2.3}$$

[6] Note that this never happens when $g = 0$, but includes the Fagnano-Euler differential when $g = 1$.

namely (x, y) is a nontrivial solution of the Pell's equation $X^2 - fY^2 = 1$ over the polynomial ring $\mathbb{C}[z]$.

The relevance of the numerical Pell's equation over $\mathbb{Z}$ is well known, and we shall not pause on it here, except by recalling that the equation has always nontrivial solutions (i.e. $y \neq 0$) provided f is a positive integer, not a square. The Pell's equation over a polynomial ring is partly analogous, but rather less known, although it has also been studied since long ago, for instance in 1826 by Abel [1], just in the context of elementary integrability of certain differentials. These integrals indeed became special cases of *Abelian integrals*, in more modern terminology; so it appears that the present context is quite related to some of Abel's motivations for his most important discoveries. Here we shall describe only in small part Abel's results on this and we refer to van der Poorten and Tran's article [23] for more, and further to the writer's survey [35] (related to the content of the present article) for a description of other contexts where this equation appears.

Coming back to the above, it is readily checked that some of the steps may be reversed, and that the solvability of (2.3) amounts to $\infty_+ - \infty_-$ being torsion on J_H. It also turns out that, contrary to the numerical case, this solvability is rather 'exceptional': see Example 2.4 below and especially [35] for a detailed illustration of the meaning of this.

In general, if $(x, y) \in \mathbb{C}[z]^2$ is a solution of (2.3), let us put $u := x + wy$. After a few calculations, we find $2\mathrm{d}u/u = (2xx' - f'y^2 - 2fyy') + w((f'yx/f) + 2xy' - 2x'y)$. Differentiation of (2.3) yields $2xx' = 2fyy' + f'y^2$. On the one hand, since x, y must be coprime polynomials, this implies that y divides x', so $x' = qy$ for a $q \in \mathbb{C}[z]$, of degree $d - 1$. On the other hand, plugging into the previous formula gives $2xq = 2fy' + f'y$ and $\mathrm{d}u/u = q/w$. In particular, we find an elementary integral from a solution to the Pell's equation, a result due to Abel:

$$\int \frac{x'(z)\mathrm{d}z}{y(z)\sqrt{f(z)}} = \log\left(x(z) + \sqrt{f(z)} \cdot y(z)\right). \tag{2.4}$$

Conversely, as is easily checked, this identity (with $y \neq 0$) implies that $x(z)^2 - f(z)y(z)^2$ is constant, hence if $x, y \in \mathbb{C}[z]$, the Pell's equation (2.3) is solvable. (Also, it then turns out automatically that x'/y has to be a polynomial.)

Coming back to our previous setting, similarly to the first case we find that $\omega - \mathrm{d}v = (2w)^{-1}(2h - 2fb' - f'b))\mathrm{d}z + a'\mathrm{d}z$. If this is to be equal to $c \cdot \mathrm{d}u/u = cq/w$ for a $c \in \mathbb{C}^*$, we must have $a =$constant and $2fb' + f'b = 2h - 2cq$.

We may resume the analysis in the following criterion (see also the paper [19] with Masser):

Proposition 2.3. *Suppose that the (nonzero) differential* (2.1) *may be integrated in finite terms. Then (at least) one of the following occurs:*

(i) *The differential equation* $2\varphi' f + \varphi f' = 2h$ *has a polynomial solution* $b \in \mathbb{C}[z]$*, and then* $\deg h \geq 2d - 1$*. In this case indeed we have* $\int \omega = bw$*.*

(ii) *The Pell's equation* (2.3) *has a solution* $(x, y) \in \mathbb{C}[z]^2$*, and then the ratio* $q := x'/y$ *is a polynomial of degree* $d - 1$*; also, for some* $c \in \mathbb{C}$*, the differential equation* $2\varphi' f + \varphi f' = 2h - 2cq$ *has a polynomial solution* $b \in \mathbb{C}[z]$*, and either* h/q *is constant or again* $\deg h \geq 2d - 1$*. Now we have* $\int \omega = bw + c\log(x + wy)$*.*

Note that the Pell's equation is solvable in the Fagnano-Euler case: $(z^2)^2 - \sqrt{-1}^2(1 - z^4) = 1$, however the differential $\mathrm{d}z/w$ has no elementary integral: it is regular on H, and such differentials always escape, e.g. since $\deg h \leq d - 2$ in those cases.

We have seen that the solvability of the Pell's equation corresponds to a certain point being torsion on the Jacobian J_H; to decide whether this holds in concrete cases historically proved to be a subtle question. Although there are nowadays standard algorithms to check whether a given point on an abelian variety is torsion or not, and to find the possible torsion order[7], this matter had not been clarified until ≈1970 (see [26], [6]). In this way, if $f(z), h(z)$ are given 'explicitly', we may check effectively by the above criterion whether $\int \omega$ is elementary or not, and we may exhibit an elementary primitive when this exists: indeed, once the torsion order is known (if there is torsion), we may compute suitable x, y as in (2.3), and then (i) or (ii) gives a linear differential equation to be solved in a polynomial $b(z)$ (and a constant c in case (ii)). But the degree of a possible $b(z)$ is bounded by $\max(0, \deg h + 1 - 2d)$ and then all of this amounts to solve a given system of linear equations.

As alluded above, there are considerations showing that for a 'random' polynomial $f(z)$ of even degree ≥ 4, the Pell's equation shall have no nontrivial solutions.[8] The randomness is intended in the sense of dimensions: *The 'Pellian' polynomials of a given degree fall into a denumerable union of (algebraic) families of lower dimension compared to the family of all polynomials of that degree*.

See [35] for some details and illustrations; here we restrict to the following examples, providing evidence in this direction already for pencils of polynomials (where we remark that (ii) below is rather harder to prove than the general assertion just stated):

Example 2.4.

(i) Consider the pencil of polynomials $f_\lambda(z) = z^4 + z + \lambda$, so $H_\lambda : w^2 = f_\lambda(z)$ has genus 1. It may be proved that the Pell's equation for $f_\lambda(z)$ cannot be solved identically in λ, but that, nevertheless, the Pell's equation for f_l has nontrivial solutions for infinitely many $l \in \mathbb{C}$.[9] The ls in question appear as poles of the coordinates z, w of multiples of $\infty_+ - \infty_-$ on the elliptic curve obtained from H_λ after choosing ∞_- as an origin (note that such coordinates are algebraic functions of λ); some explicit values leading to solutions are $l = 0, 1/2, (-1 - i\sqrt{3})/4$, with minimal degree of $x(z)$ being resp. $3, 4, 4$.

In particular, all of these ls are algebraic numbers (which may be seen also on observing that otherwise the identical Pell's equation would be solvable) and moreover with bounded Weil height (as follows from a theorem of Silverman) so they are rather 'sparse'; for instance, this implies that there are only finitely many of them which are rational, or even of any fixed degree over $\mathbb{Q}$. (See [34] and the references therein for proofs of these facts.)

(ii) Consider now the pencil $f_\lambda(z) = z^6 + z + \lambda$, giving H_λ of genus 2. As in (i), it is not too difficult to prove that (2.3), with $f = f_\lambda$, cannot be solved identically. But now we have an assertion stronger than before about the 'exceptional' values l: though there are some, like $l = 0$ (one has the 'Pellian' identity $(2z^5 + 1)^2 - f_0(z)(2z^2)^2 = 1$)[10], we proved with

[7] Of course, we must work with *finitely presented* objects, for instance varieties and points defined over $\overline{\mathbb{Q}}$ and given by explicit equations over $\mathbb{Z}$. In this case suitable algorithms come e.g. from reduction modulo two distinct primes. This method is quoted also in [3].

[8] It is easy to see that for $\deg f = 2$ there are always solutions of (2.3) with $x, y \in \mathbb{C}[z]$. This corresponds to a curve H of genus 0, and in fact we have already noted that then any differential can be integrated in finite terms.

[9] This fact, though not difficult, seems not completely obvious to us; a proof comes on using the above described correspondence with torsion on an elliptic curve; see [34], especially p. 92.

[10] The previous values were communicated to me by Masser's student Merkert, whereas this one was found by Masser; both of them used continued fractions.

Masser in [18] that the Pell's equation for $f_l(z)$ is solvable only for finitely many $l \in \mathbb{C}$.[11]

Correspondingly, there are only finitely many $l \in \mathbb{C}$ for which an elementary integral $\int h(z)\mathrm{d}z/\sqrt{f_l(z)}$ exists with $h \in \mathbb{C}[z]$ of degree ≤ 4. One of them, derived from the above identity for $l = 0$, is $\int 5z^2\mathrm{d}z/\sqrt{z^6+z} = \log(1 + 2z^5 + 2z^2\sqrt{z^6+z})$.

In the sequel we shall again consider families depending on a parameter.

3. Differentials and integrals depending on a parameter

In the last Example 2.4, we have met parametric families of Pell's equations in polynomials, illustrating in particular that a nontrivial solution is quite an exceptional phenomenon. The link provided by Proposition 2.3 then shows that an elementary abelian integral as in (2.4), in such parametric families of differentials, is also quite uncommon.

Then, one may ask what happens for parametric families more general than the above ones. From another viewpoint, it is also natural to ask the above questions also for (functions and) differential forms depending on several variables, and to consider the problem of finding an elementary primitive with respect to some derivation of the relevant field.

Here we shall consider such problems, and stick to the case of two variables z, λ, and again to differentiation with respect to z, so that λ may be seen as a parameter for a pencil, not depending on z (we may also think of λ as a point on an algebraic curve).

We shall soon come back to examples of the previous kind, but first let us briefly pause, as before, on some cases with transcendental functions, which seem puzzling. For instance, let us consider $(\log z)^\lambda \mathrm{d}z$, which we may also write as $e^{\lambda \log\log z}\mathrm{d}z$; to integrate it we may also use the substitution $z \mapsto e^z$, which sends it to $z^\lambda e^z \mathrm{d}z$, and then a *definite* integral relates to the Gamma function. We view λ as a constant, and so we start with a ground field $\mathbb{C}(\lambda)$, or even $L := \overline{\mathbb{C}(\lambda)}$, in place of $\mathbb{C}$. Then we may try to apply Theorem 1.1 to $F = L(z, \log z, \psi)$, where ψ is thought of as $(\log z)^\lambda$, and may be given a meaning as an element whose derivative ψ' equals $\lambda\psi(z\log z)^{-1}$.[12] This field is certainly closed under differentiation with respect to z. Taking into account that $z, \log z, \psi$ may be shown to be algebraically independent over L, one may also show that F has L as field of constants for $\mathrm{d}/\mathrm{d}z$, and after some work, we shall find that $\int \psi \cdot \mathrm{d}z$ is not elementary.[13]

In fact, already thinking of special values of λ yields non elementary integrals, as we have seen to happen with $\lambda = -1$, and actually this continues to hold for λ equal to any negative integer, since substitution $z \mapsto e^z$ and integration by parts reduces this to $\lambda = -1$. But for $\lambda \in \mathbb{N}$ we find elementary integrals as is immediately verified by the same steps. This sort of ambiguous behaviour inspires further research.

Actually, this matter of specialization may raise issues already at the moment of giving an exact definition; for instance, for special values of λ we could find poles of some coefficients introduced along the process of constructing an elementary extension. However, if we argue (as above) with algebraic functions of λ, and since in each elementary extension only finitely many functions are involved, these poles may occur only at finitely many points, and we may

[11] As in Example (i), only the algebraic l are of interest for this case.

[12] One has to use here basic notions of differential algebra, as presented for instance in [29].

[13] Here we should add the condition that also the extension field has L as a field of constants, but in fact this may be dispensed with, see Remark 1.2(i) above. Or else, if we want to avoid this point, we can also look at λ as a 'generic' complex number itself and look at F as a field of meromorphic functions in some region, as before, so that $\mathbb{C}$ remains the field of constants.

specialize without ado outside this set. Hence, for instance, we may immediately deduce that if an integral is 'generically' elementary (i.e., if it is elementary over the constant field $\overline{\mathbb{C}(\lambda)}$) then it remains elementary for all but finitely many complex values of λ.[14] It is the converse issue which is (at least for us) much less clear, and much more interesting: *If an integral is not 'generically' elementary, for 'which' values of the parameter can it become elementary?*

Let us call 'exceptional' these values. Of course, by 'which' we may mean several things; for instance we may aim to prove that the exceptional values are in some sense 'sparse', or even that they form a finite set, or at least that their set has infinite complement, depending on the case. Note that the above example shows that in general there is no finiteness assertion for them.[15]

Such a kind of problem was actually raised explicitly by J. Davenport in the already quoted book [6]; he considered throughout especially differentials on an algebraic curve, and in Ch. 3.6 he allowed dependence on a parameter, also varying algebraically. We may view these data as a pencil of (curve+differential)s, or else as a single differential on a curve, both defined over the function field of a(nother) curve. In the sequel we shall tacitly switch between these viewpoints.

Notation. For later convenience, once the differential ω_λ is given (assumed not to have an elementary primitive identically in λ), let us denote by $\mathcal{E}$ the set of the 'exceptional' values of the parameter for which the specialized differential admits an elementary primitive . (It shall be clear from the context whether we consider any complex value or only algebraic ones.)

The said book considers, among others, the problem of proving that *if such a differential does not admit an elementary primitive, then the set of values of the parameter for which the specialized differential admits such a primitive is finite* (i.e., $\mathcal{E}$ is finite). So, for instance we have already noted that this fails for the above example of $(\log z)^\lambda$, but now we are confined to algebraic differentials.

The book contains a Theorem 7 in this direction; however there seems not to be a complete proof of all the stated assertions. In particular, a difficulty which that treatment does not overcome concerns torsion on a Jacobian (similarly to what we have seen in §2 in connection to the Pell's equation). It would be much easier to prove the weaker assertion that *the differential admits an elementary primitive if and only if the specialized one admits such a primitive for all but finitely many values of the parameter*.[16] Such a result may remind of the Hilbert's irreducibility theorem. But here we shall be concerned with the more delicate question of the *finiteness* of the set $\mathcal{E}$ of exceptional values.

---

[14] This yields another proof that $(\log z)^\lambda \mathrm{d}z$ is not elementary, because, as noted above, it is not such for λ equal to any negative integer.

[15] Through Theorem 1.1 it may be proved that we find an elementary integral of $(\log z)^\lambda$ precisely for $\lambda \in \mathbb{N}$.

[16] This amounts to $\mathcal{E}$ having infinite complement. Even this kind of assertion contains different levels of subtlety, depending on whether we allow 'generic' specializations, or merely algebraic ones. For instance, the assertion of the above Example 2.4(ii) may be weakened either to '*there are infinitely many values of* $\lambda \in \mathbb{C}$ *for which the Pell is not solvable for* f_λ' or to the same statement with $\mathbb{C}$ replaced by $\overline{\mathbb{Q}}$ or even $\mathbb{Q}$. Now, we had already noted that the first assertion amounts to the fact that (2.3) is not solvable identically in λ; actually (even limiting to a single transcendental value of λ), this is almost tautological, which is not the case for the second claim. This last, at any rate in the 'non-constant' cases, follows from a general result by Silverman predicting bounded Weil height for the exceptional points, or also from suitable applications of Hilbert's irreducibility Theorem, as in work by Néron.

A link with Unlikely Intersections. In a series of papers in collaboration with Masser [14]–[19] we had considered problems involving torsion on pencils of abelian varieties. These papers actually prove, under suitable conditions, general finiteness theorems for the set of specializations where a non-torsion section for a pencil of abelian surfaces may become torsion.[17] (We analyzed separately the cases when the abelian surface is isogenous to a product of elliptic curves, or is simple, because the arguments present some differences.) Such an issue may be considered a 'relative' case of the celebrated Manin-Mumford conjecture (a theorem of Raynaud since the 80s), and is a special case of a general conjecture of Pink [22]. The general topic is often referred to as 'Unlikely Intersections', because the specializations in question are 'unexpected' (mainly for dimensional reasons) and should be hard to come by. See [34] and [35] for a discussion.

Now, we have already pointed out in §2 that the solvability of the Pell's equation is related to torsion on a suitable Jacobian variety, and this simple observation (well known since long ago) provides the link of the said context with the present one.

Indeed, at some point Masser noticed the book [6] and eventually, especially through the role of the Pell's equation, we realized that the finiteness questions raised therein could have been obtained as a consequence of our methods. As will be seen in some examples below, the present context shall involve other algebraic groups beyond abelian varieties; these arise as generalized Jacobians of suitable curves, and, for the cases of interest here, are extensions of abelian varieties by additive groups $\mathbb{G}_{\mathrm{a}}^{r}$. (See what follows for more; see also [35], §4, for another link of generalized Jacobians with certain 'degenerate' Pell's equations.)

The proof of a full finiteness statement for any pencil of differentials on a curve is in the course of being written down. Here we shall merely give some examples, collected in the next section, illustrating the mentioned connection; the examples should also isolate the main issues which appear in a general finiteness proof. (In a subsequent section we shall say more on this general case.)

Remark 3.1. In this context λ shall indicate a generic point of a given curve over $\overline{\mathbb{Q}}$; the arguments are not affected by the structure of this curve, and hence for simplicity in the sequel we shall work mainly with $\mathbb{P}_1$, i.e. letting λ be an indeterminate.

Also, we shall let throughout $\overline{\mathbb{Q}}$ be a ground field for the coefficients involved in our functions and curves. The case when arbitrary complex coefficients appear may be treated in a completely similar way, but requires a corresponding extension of the said auxiliary results obtained with Masser; such an extension, which seems more involved than may be expected, has been carried out partly in [16] and [17], but the general case is still in progress (in joint work also with Corvaja).

4. Some examples of finiteness of the set of exceptional specializations

To better illustrate the alluded finiteness problem in this article, and keep continuity with the above, we start by analyzing the hyperelliptic differentials already considered in (2.1), actually for the particular pencils which appear in a previous example. For these cases, we shall sketch a deduction of finiteness of the set denoted $\mathcal{E}$ above, relying on the mentioned results with Masser and related ones.

[17] On the contrary, except for some degenerate 'constant' cases which may be classified, there are infinitely many torsion specializations on a pencil of elliptic curves, as happens in connection with Example 2.4(i) above.

Example 4.1. We shall refer to differentials and curves as in (2.1), but depending on a parameter λ, so we take therein $h(z) = h_\lambda(z)$, $f(z) = f_\lambda(z)$, polynomials in $\overline{\mathbb{Q}}(\lambda)[z]$, and we consider a corresponding $\omega_\lambda = h_\lambda/w$. Let us discuss the specialization issue in some detail, using the pencils appearing in Example 2.4. We shall presently reverse the order of the two instances therein, because the second one leads more rapidly to a connection with our context.

(i) Let us choose $f_\lambda(z) = z^6 + z + \lambda$, as in 2.4(ii); now the curve $H = H_\lambda$ is given by $w^2 = z^6 + z + \lambda$, and has (generically) genus 2.

Let us suppose that ω_λ is not identically integrable in finite terms (i.e. considered over the field $\overline{\mathbb{Q}(\lambda)}$) and let us pick $l \in \overline{\mathbb{Q}}$ such that ω_l is integrable in finite terms; as above, we denote by $\mathcal{E}$ the set of these 'exceptional' numbers. (We tacitly disregard the finitely many l such that f_l has a double root, so that H_l shall have also genus 2.)

We shall use Proposition 2.3. Take first the case when Pell's equation (2.3) for f_l does not have a solution. Then, we infer that the differential equation $2\varphi' f_l + \varphi f_l' = 2h_l$ has a polynomial solution. This solvability amounts to the one of a (inhomogeneous) system of $\leq \deg h_l + 1$ linear equations, where the unknowns are the coefficients of φ (which must have degree $\leq \deg h_l - \deg f_l + 1 \leq \deg h_\lambda - 5$) and where the entries of the system are expressed linearly in l and the coefficients of h_l. By Capelli's criterion, the solvability may be controlled through the vanishing of suitable minors of the matrix of the system. Now, we have a similar matrix with l replaced by λ, with a corresponding solvability condition. Each minor which is not identically zero in λ may become zero only for finitely many algebraic numbers l; hence, if l is taken outside this finite set (for any of the minors) we cannot have solvability (because we are assuming that the system is not identically solvable).

On the other hand, we have already remarked that the Pell's equation may have solutions only for finitely many l, as is proved in [18]. (It should be noted that this second finiteness, contrary to the elementary nature of the arguments for the former case, requires much more effort and the use of rather deep tools.)

In conclusion, combination of these remarks yields finiteness for the set $\mathcal{E}$.

(ii) Let us now choose $f_\lambda(z) = z^4 + z + \lambda$, as in 2.4(i); now H_λ has genus 1 (and is isomorphic to the curve with Weierstrass equation $y^2 = x^3 - \lambda x + 1/4$). We again suppose that ω_λ is not identically integrable in finite terms and analyze the set $\mathcal{E}$, assuming by contradiction that it is infinite.

Arguing as in part (i), we obtain finiteness of the set of those $l \in \mathcal{E}$ such that the Pell's equation for f_l is not solvable.

Suppose now that the Pell's equation for f_l has a solution; we have remarked in Example 2.4 that for the present f_λ, of degree 4, this set is infinite (forgetting that $l \in \mathcal{E}$), so we cannot argue as in (i) and have to find supplementary arguments.

Say that $l \in \mathcal{E}$ and that (x_l, y_l) is a Pell's solution with minimal degrees. Now the differential equation of Proposition 2.3 is $2\varphi' f_l + \varphi f_l' = 2h_l - 2cq_l$, where c is any constant and $q_l = x_l'/y_l$ has degree 1. One problem in carrying out the previous analysis is that we have little information on q_l; for instance it shall follow that not even its zero can be expressed algebraically in l (restricting to the numbers $l \in \mathcal{E}$ in question). Still less we know x_l, y_l; note that they are polynomials of minimal degree so that $u_l := x_l + w_l y_l$ has divisor $m_l(\infty_+ - \infty_-)$, where $m_l \neq 0$ is an integer depending on l, and $|m_l| = \deg x_l$. Certainly $|m_l|$ shall tend to infinity (for otherwise, as is very easy to see, the Pell's equation would be identically solvable). All of this suggests that the quantities appearing in the specialized

system are not part of an algebraic (continuous) family.

We may cope with this serious difficulty in some steps, as follows. Since the differential equation $2\varphi' f_l + \varphi f_l' = 2h_l - 2cq_l$ has (for some constant c) a polynomial solution $\varphi = \varphi_l$, say, and since $\deg q_l = 1$, we see that for infinitely many l the polynomial $2h_l$ is in the image, *modulo linear polynomials*, of the differential operator $2f_l \mathrm{d}/\mathrm{d}z + f_l'$. But then, by elementary arguments of linear algebra very similar to those in the first part, we deduce the same for $2h_\lambda$, namely that there exists a polynomial $\varphi_\lambda \in \overline{\mathbb{Q}}(\lambda)[z]$ such that $2\varphi_\lambda' f_\lambda + \varphi_\lambda f_\lambda' - 2h_\lambda$ has degree ≤ 1.

This amounts to the fact that $\omega_\lambda - \mathrm{d}(\varphi_\lambda w)$ is of the shape $Q_\lambda(z)\mathrm{d}z/w$, for a polynomial $Q_\lambda \in \overline{\mathbb{Q}}(\lambda)[z]$, linear in z, and on subtracting $\mathrm{d}(\varphi_\lambda w)$ we may directly assume that $\omega_\lambda = Q_\lambda(z)\mathrm{d}z/w$. In turn, this yields that ω_λ has at most simple poles at ∞_+, ∞_- and no other poles.[18]

Then for infinitely many $l \in \mathcal{E}$ we obtain that $\omega_l = c_l \mathrm{d}u_l/u_l$, for a constant c_l, which we may assume $\neq 0$, for otherwise we fall in the previous case (and ω_λ would be identically integrable in finite terms).

Now, looking at poles shall not yield further information, and then let us look at zeros of these differentials. Let $\zeta_\lambda \in \overline{\mathbb{Q}}(\lambda)$ be the unique zero of Q_λ. Then we may specialize it at almost all $l \in \overline{\mathbb{Q}}$, obtaining a zero denoted ζ_l, and we deduce that both points above $z = \zeta_l$ on H_l, i.e. both points $\xi_l^\pm \in H_l$ such that $z(\xi_l^\pm) = \zeta_l$, are zeros of $\mathrm{d}u_l$.

Hence, for these $l \in \mathcal{E}$ there is a rational function on H_l having divisor of the shape $m_l(\infty_+ - \infty_-)$ and such that its differential vanishes at both points above ζ_l, so vanishes at ξ_l^+. Now, while the first condition says that $\infty_+ - \infty_-$ is torsion on the Jacobian of H_l (which is essentially H_l), adding the second condition says that $\infty_+ - \infty_-$ *is torsion on the generalized Jacobian of* H_l *with respect to the modulus* $2\xi_l^+$.

For notions related to generalized Jacobians we refer to Serre's book [32], and, for the present case, also to the recent paper [5] by Corvaja, Masser and the writer. Here, we merely mention that this generalized Jacobian, denoted here Γ_l, is an extension of the elliptic curve H_l (for instance with ∞_- as origin), by the additive algebraic group $\mathbb{G}_\mathrm{a}$; i.e., there is an exact sequence of algebraic groups $0 \to \mathbb{G}_\mathrm{a} \to \Gamma_l \to H_l \to 0$. It may be described in several ways; one of them, especially relevant here, uses a 'strong' equivalence of divisors on a curve H: two such divisors D, D' are *strongly equivalent* if their difference is in the first place a principal divisor, i.e. $D - D' = \mathrm{div}(u)$ for some rational function u on H (which is usual equivalence); it is then required that the differential $\mathrm{d}u$ vanishes at a prescribed point. So, this is precisely what happens in the above situation, where $H = H_l$ and ξ_l^+ is the point in question.

Naturally, we may consider the 'generic' extension Γ_λ of H_λ by $\mathbb{G}_\mathrm{a}$, corresponding to the modulus $2\xi_\lambda^+$.[19] And we find that for the relevant $l \in \mathcal{E}$, the image in Γ_l of the divisor $\infty_+ - \infty_-$ is torsion. Now, Γ_λ is an algebraic group (e.g., over $\overline{\mathbb{Q}(\lambda)}$) of dimension 2; also, the divisor $\infty_+ - \infty_-$ is not identically torsion therein, for otherwise this would imply in particular that it is torsion on H_λ.

Moreover, it turns out that the extension Γ_λ is *not trivial* (also said *not split*), i.e. not isomorphic to $\mathbb{G}_\mathrm{a} \times H_\lambda$. This fact, proved for instance in [32] (see Ch. VII, Prop. 15), does not appear to be obvious, and is absolutely crucial for the sought finiteness: indeed, for a

[18] We observe that in more general cases these steps may be replaced by a simple use of the Riemann-Roch theorem.

[19] This involves a choice of the sign; indeed, it would be more precise to work with the base curve with function field $\mathbb{C}(\lambda, \sqrt{f(\zeta_\lambda)})$, rather than with the λ-line. However, for brevity we skip such precisions.

split extension, a point could lie in $\{0\} \times H_\lambda$ and could then yield infinitely many torsion specializations (as nothing would change compared to H_λ itself).

At this stage, an analogue of the joint results with Masser, obtained by Masser's student Schmidt [31], asserts that the above torsion-degeneracy in fact may happen only for finitely many values l of λ.

This contradiction concludes the analysis of the present cases.

Note that this implies the previous assertion that the (unique) zero of the polynomial q_l does not vary algebraically in l (for the infinitely many l for which (2.3) is solvable). (Observe also that the leading coefficient of q_l is $\pm m_l$, whose absolute value tends to infinity.)

Remark on effectivity. We note that, though we may calculate the exceptional numbers $l \in \mathcal{E}$ for which the Pell's equation is not solvable, our proof methods do not presently allow to list the finitely many remaining ones. However the mentioned theorem of Silverman giving the boundedness of the Weil height $h(l)$ is effective and can be used to list the relevant l with degree bounded by a prescribed number. Also, there is good hope that the structure of the proofs can be suitably analyzed so to be made completely effective.

Example 4.2. We now present an example, taken from [5], when the curve is 'constant' (i.e. not depending on λ) but instead the differential is variable. Suppose that $f \in \mathbb{C}[z]$ has degree 3 and no multiple roots and consider the projective smooth curve H of genus 1 with affine equation $w^2 = f(z)$. This curve has a unique point O at infinity (i.e. where z has a pole). Take now the differential

$$\omega_\lambda := \frac{\mathrm{d}z}{(z-\lambda)w},$$

where λ is a variable, that is we consider (H and) this differential as defined over $\mathbb{Q}(\lambda)$. This may be easily checked to have a double zero at O and two poles at the points $P_\lambda^\pm$ where $z = \lambda$, with residues $\pm 1/\sqrt{f(\lambda)}$. Theorem 1.1 then easily implies that, if $l \in \mathbb{C}$, $\int \omega_l$ is elementary if and only if there are a nonzero constant c_l and a rational function u_l on H such that $\omega_l = c_l \mathrm{d}u_l/u_l$. Similarly to Example 4.1(ii), this implies that, for some nonzero integer m_l, the class of the divisor $m_l(P_l^+ - P_l^-)$ with respect to the modulus $2O$ is zero; in turn, this yields that $P_l^+ - P_l^-$ is torsion of the corresponding (nontrivial) extension Γ of $J_H(\cong H)$ by $\mathbb{G}_\mathrm{a}$. As in the previous example, though $P_l^+ - P_l^-$ is torsion on J_H for infinitely many l, it may be proved that this stronger torsion condition may be verified only for finitely many $l \in \mathbb{C}$.

In this example actually the double zero of ω at O would imply that the class of $m_l(P_l^+ - P_l^-)$ is zero with respect to the modulus $3O$, which is a condition even stronger than needed. If for instance we replace the present cubic $f(z)$ with, e.g., the quartic z^4+z+1, the corresponding ω_λ has two simple zeros at the two points at infinity. This gives just what is needed, and choosing any of the zeros suffices for the above argument to work.

It is to be noted that this proof can be carried out without appealing to the method used in the above quoted papers with Masser; this is because H is constant, and hence a generalization of Raynaud's theorem obtained by Hindry, or else the method of [5], suffices.[20] See [5] also for references.

[20] In fact, the divisor in question describes, as λ varies, a curve in Γ, and no such curve can contain infinitely many torsion points: otherwise the curve would have to be a translate of an algebraic subgroup, by the cited works. But the only algebraic subgroup of dimension 1 of Γ is $\mathbb{G}_\mathrm{a}$, and each translate of it has at most one torsion point.

5. Some considerations towards a general case

A general criterion. For the analysis of the general case of algebraic differentials over a curve one argues similarly, using a criterion easily derived from Theorem 1.1 and which has been implicit in the above examples. (This appears, e.g., in [26].) Let us describe the essentials.

We let X be an algebraic curve, with function field $k(X)$ (where k is algebraically closed of characteristic zero) and let ω be a (rational) differential on X admitting an elementary integral. Then, by Theorem 1.1 (or at any rate by the mentioned more precise version regarding the constants) ω can be written as in (2.2), with $v, u_1, \ldots, u_n \in k(X)$ and $c_i \in k$, namely $\omega = \mathrm{d}v + \sum_{i=1}^{n} c_i \mathrm{d}u_i/u_i$. As in a previous proof, we can assume that n is minimal, and then the c_i are linearly independent over $\mathbb{Q}$. Any differential $\mathrm{d}u_i/u_i$ has poles of order 1, at the zeros and poles of u_i, with residues given by the respective multiplicities; also, $\mathrm{d}v$ has no residues. Hence, at each point $p \in X$, we have an equation

$$\mathrm{res}_p\, \omega = e_{1p}c_1 + \ldots + e_{np}c_n, \tag{5.1}$$

with integers $e_{ip} = \mathrm{ord}_p(u_i)$. Of course, if we are given in advance only ω, we do not know the possible u_i; however the argument may be partially reversed; let us see how.

In the first place, we may *let $c_1, \ldots, c_n$ be a basis for the vector space spanned over $\mathbb{Q}$ by the residues of ω*, and then for any $p \in X$ we have (uniquely) an equation (5.1), where the e_{ip} are rationals; but in fact (since only finitely many points p are involved) they can be assumed to be integers after rescaling the c_i by a suitable rational. At this stage we have to check whether an equation (2.2) holds, however for some possibly different constants c_i^* (and possibly another n). In this case, by the above remarks, the $\mathbb{Q}$-space spanned by the c_i has to be the same as for the c_i^*. Hence, on expressing the c_i^* linearly in terms of the c_i, with rational coefficients, and using the additive property of the logarithmic derivative, at the cost of multiplying ω by a suitable integer $m \neq 0$, we may assume that $c_i^* = c_i$, and that $m\omega = \mathrm{d}v + \sum c_i \mathrm{d}u_i/u_i$. Hence, we arrive at the previous conclusion, 'only' at the cost of multiplying by m. Necessarily,

$$m \sum_{p \in X} e_{ip} \cdot p = \mathrm{div}(u_i), \qquad i = 1, \ldots, n. \tag{5.2}$$

Note that in this argument the divisors $D_i := \sum_{p \in X} e_{ip} \cdot p \in \mathrm{div}(X)$ are determined in terms only of ω, and have degree zero, because of the $\mathbb{Q}$-independence of the c_i and the sum-formula for residues (see [32], Prop. 6). They may be also shown to be linearly independent over $\mathbb{Z}$. A nonzero integer m as in (5.2) is *a priori* not given (or bounded) in terms of ω (which shall be the main point); in any case, (5.2) yields the following strong necessary condition:

Criterion. *If ω admits an elementary integral, there is a nonzero integer m such that the mD_i are principal divisors, i.e., for $i = 1, \ldots, n$, the class of D_i in the Jacobian J_X of X is torsion.*

So, we see that the torsion condition that we had previously found in connection with the Pell's equation appears indeed generally.

When we have a pencil of differentials, to be specialized, we may compare the 'generic' criterion with the 'special' one, as we have done in the above examples. Through this comparison, in the mentioned joint work with Masser we have very recently obtained the sought

finiteness for the exceptional set $\mathcal{E}$; as in the above simple examples, this heavily relies on the said finiteness results with Masser towards Pink's conjecture (or else on Schmidt's thesis).

However, it is to be remarked that in the course of these applications several new issues appear along the way, apparently with different levels of difficulty. We briefly list some of them, the simplest of which we already met in the examples; we shall not go in detail of how overcoming them (which may be indeed done, and shall appear in a forthcoming paper).

We let X_λ denote a pencil of curves over a(nother) 'base' curve B over $\overline{\mathbb{Q}}$, where λ is a generic point of B (and X_λ is defined over $\overline{\mathbb{Q}}(\lambda)$).[21] We let ω_λ a differential on X_λ and denote by X_l, ω_l, specializations of them, where $l \in B(\overline{\mathbb{Q}})$. We assume that ω_λ is not identically integrable in elementary terms and want to prove the finiteness of the above defined set $\mathcal{E}$.

(a) About the Criterion. We have observed that the condition provided by the 'Criterion' is only necessary for integrability in finite terms, not sufficient. Indeed, in the first place Theorem 1.1 involves also the function v and its differential, not taken into account in the Criterion. However, it is a matter of linear algebra (using the Riemann-Roch theorem) that if $\omega_l - \mathrm{d}v_l$ has only poles of order ≤ 1 for an infinity of special values l and rational functions v_l on X_l, then there exists a function v_λ such that $\omega_\lambda - \mathrm{d}v_\lambda$ has the same property.[22] In this way we may directly assume such a pole structure for ω_λ, and then the same will hold for ω_l, so we may assume $v = 0$ in the application of Theorem 1.1.

However, even with this hypothesis on a differential ω, the condition in the Criterion does not become sufficient. In fact, that condition only ensures that there are rational functions $u_1, \dots, u_n$ such that $m\omega - \sum c_i \mathrm{d}u_i/u_i$ has no poles. But on a curve genus g there are independent g differentials of this sort, so for $g > 0$ we cannot conclude that $m\omega = \sum c_i \mathrm{d}u_i/u_i$. Hence, concerning our purposes, we cannot hope to extract all the needed information from the Criterion. (To supplement this, one has to consider generalized Jacobians.)

(b) About rational dependence of coefficients. Then, there is the question of linear independence over $\mathbb{Q}$ of the c_i. For our purposes we must consider not merely the $c_i = c_i(\lambda)$, which are certain algebraic functions of λ, but also their specializations $c_i(l)$ at exceptional points $l \in \mathcal{E}$, and of course it may happen that the $\mathbb{Q}$-linear independence is destroyed at l. However this may occur at most for a set of algebraic points l of bounded degree over $\mathbb{Q}$ (in fact, when the dimension drops we have a nontrivial equation $\sum \rho_i c_i(l) = 0$ with $\rho_i \in \mathbb{Q}$). Now, the main point here is that, when dealing with torsion, the already mentioned results by Silverman gives boundedness of the height for the special values when a non-torsion section of an abelian family becomes torsion (at least provided the family has no 'constant' factors of dimension > 0); and then Northcott's finiteness theorem may be applied, which yields the sought finiteness. (On the other hand, concerning constant factors, we remark that Silverman's argument still may be partially restored to give sufficient information. See also Masser's Appendix C to [34]; we omit details here.)

Remark 5.1. The book [6] by J. Davenport raises some of these issues, and proposes methods to solve them, which sometimes indeed may work. (See pp. 89–91). However, for instance heights do not appear and a complete treatment is not clarified. In any case, there

[21] For simplicity we had previously considered only the case where $B = \mathbb{P}_1$, and λ is a variable, which causes no real conceptual difference in our arguments.

[22] Here we may have to use also good reduction at the relevant l, which however holds with only finitely many exceptions.

are delicate points still to be mentioned, and we go ahead to illustrate some of them.

(c) About torsion specializations. A major point concerns torsion, and actually this itself turns into several different issues, as we shall now try to illustrate.

The divisors which arise (i.e., the ones called D_i in the Criterion) define points in the Jacobian J_λ of X_λ, or, rather, sections $D_i = D_{i\lambda} : B \to J_\lambda$; indeed, J_λ is a pencil of Jacobians over B. For instance, if all the $D_{i\lambda}$ would be identically torsion, say of order m, then (as in (a) above) we would find suitable functions $u_{i\lambda}$ such that $\delta_\lambda := m\omega_\lambda - \sum c_i \mathrm{d}u_{i\lambda}/u_{i\lambda}$ would be regular. But then we could replace ω_λ with the regular δ_λ and conclude easily; indeed (as in the calculations of §2), $\delta_\lambda \neq 0$ (since ω_λ is not identically integrable in finite terms), hence $\delta_l \neq 0$ for almost all l, finishing the argument because a nonzero regular differential cannot be integrated in finite terms.

We conclude that it suffices to work with the assumption that some $D_{i\lambda}$ is not a torsion section, and one main step in proving finiteness of $\mathcal{E}$ is to show that D_{il} may become a torsion point on J_l only finitely many times.

Actually, such a clear-cut assertion is not always true: if J_λ is an elliptic pencil, then this kind of 'degeneracy' usually happens in an infinite (though sparse) set of $l \in B(\overline{\mathbb{Q}})$; we have found concrete instances of this behaviour in Examples 2.4(i) and 4.1(ii) above. For the assertion to hold, it turns out that we need at least that J_λ has relative (i.e., over B)) dimension ≥ 2, but still this does not suffice. For relative dimension $= 2$, the required finiteness of torsion specializations has been proved in the quoted joint papers with Masser, provided however that the non-torsion section does not map in a proper algebraic subgroup; for instance, this is automatic if J_λ is generically simple. We see that already in this case of surfaces some supplementary hypotheses are needed, and in fact, checking them for the sought application may happen to present difficulties.

Anyway, a first point in this direction has been to generalize [16]–[18] to any relative dimension ≥ 2; this extension is only in part similar to the former case, for instance because the abelian pencils which occur cannot always be represented as Jacobians (so one has to use the Siegel space). Also, one has to look at simple factors of J_λ (after an isogeny), and let A_λ be a typical one; it may happen that A_l becomes non-simple for $l \in \mathcal{E}$, and then to deal with other aspects of the proof (of which we shall say a little more in the next section) one needs for instance results by Masser and Wüstholz (see [13]) to control polarizations.

(d) About splitting of generalized Jacobians. Once such a more general result has been established under the appropriate hypotheses, we have still to deal with the cases when the relevant section does not meet such assumptions, as happens for instance when a nonzero multiple of it lies in some elliptic pencil inside the Jacobian pencil. In these cases, like for Example 4.1(ii), we have to take advantage of a generalized Jacobian, arising from zeros of a suitable differential.

The discussion of all the possibilities which may arise is not brief, and involves further obstacles. Therefore we do not pause further on this, except by pointing out that one such difficulty is related to the possible triviality of the relevant generalized Jacobian (by which we mean that the relevant $\mathbb{G}_{\mathrm{a}}^s$ is a direct factor). For basic cases like in Example 4.1, the triviality never occurs, as we have remarked; but this is not the only possibility. We offer a last example to illustrate this point.

Example 5.2. Let us consider the hyperelliptic curve H_λ of genus 2 with equation $w^2 = z^6 + z^4 + \lambda z^2 + 1$. We have an obvious map $\phi(z, w) = (z^2, w)$ to the curve of genus 1

defined by $v^2 = u^3 + u^2 + \lambda u + 1$. Let us take the point $\xi = \xi_\lambda := (0, 1)$ on H_λ. Then, as in Example 4.1(ii), we may construct the generalized Jacobian $\Gamma_{2\xi}$ of H_λ with respect to the modulus 2ξ; it is an extension of the usual Jacobian $J = J_\lambda$ by $\mathbb{G}_{\mathrm{a}}$:

$$0 \to \mathbb{G}_{\mathrm{a}} \to \Gamma_{2\xi} \xrightarrow{\pi} J \to 0.$$

As before, this extension is not split (as proved in [32]), in the sense that there is no section for π, or, equivalently, $\Gamma_{2\xi}$ is not isomorphic to $\mathbb{G}_{\mathrm{a}} \times J$.

However, it can be shown that there is a section if one restricts above a suitable elliptic curve $\tilde{E} \subset J$, namely there is a map $\eta : \tilde{E} \to \Gamma_{2\xi}$ with $\pi \circ \eta =$ identity of $\tilde{E}$. (One can exhibit $\tilde{E}$; in the first place, there is an isogeny ι from J to $E \times F$, where F is the curve of genus 1 defined by $y^2 = x^3 + \lambda x^2 + x + 1$, and one may put $\tilde{E} = \iota^{-1}(E \times \{0\})$.)

Hence, restricting $\Gamma_{2\xi}$ above $\tilde{E}$ yields a splitting, and this might heavily affect the finiteness proof, for reasons that we have outlined above. However, if we happen to meet such an instance in the course of the above analysis, there is a further information that can be used, and this concerns ramification of the map ϕ at ξ, which may be seen to be indeed related to this splitting. In the present case we find that the ramification index $e_\phi(\xi)$ is 2. But then, it turns out that for our purposes we could consider the extension of J determined by the modulus 3ξ, denoted $\Gamma_{3\xi}$; this is now an extension of J by $\mathbb{G}_{\mathrm{a}}^2$. And it would suffice that this extension, restricted above $E \times \{0\}$, is not totally split (i.e. not isomorphic to $\mathbb{G}_{\mathrm{a}}^2 \times E$). This indeed can be proved, although we omit details here.

6. A few words on the proofs and about further questions

We have illustrated some aspects of the proof of finiteness of the set denoted $\mathcal{E}$ above; and we have remarked (e.g., in (c) above) that they crucially rely on other finiteness proofs obtained with Masser, representing special cases of Pink's conjectures. Hence now we shall only comment on the proofs of these auxiliary ingredients. (See the book [34] for a more complete overview of the topic of Unlikely Intersections, and see the article [22] by Pila in this volume for further and more recent directions and advances in the subject.)

On the proofs of finiteness for torsion specializations. So, we have a pencil of abelian varieties $\mathcal{A} \to B$ over a curve B, and a section $\sigma : B \to \mathcal{A}$, all defined over $\overline{\mathbb{Q}}$.[23] Then, under suitable assumptions (for instance that the image $\sigma(B)$ is not contained in any proper algebraic group subscheme) we aim to prove that $\sigma(l)$ can be torsion only for finitely many $l \in B(\overline{\mathbb{Q}})$. (Below, we also keep the previous conventions, where we had denoted by λ - resp. l - a generic - resp. algebraic - point of B, and by A_λ, A_l the respective fibers.)

A main principle is to view the abelian fibers A_l (also for $l \in \mathbb{C}$) as complex tori (varying analytically) and then the torsion points, read in terms of real coordinates in the lattice bases for the tori, become rational points.

Now, a torsion point $\sigma(l) \in A_l$ yields other ones by conjugation over a number field of definition (on other - conjugate - fibers $A_{l'}$). Here one can get a lower bound for the number of such conjugates through deep results by Masser, David, Masser-Wüstholz; note that by conjugation one jumps around the various fibers, so some uniformity is needed. On the other

[23] We have already remarked that the case when the field of definition is $\mathbb{C}$ requires additional arguments, which in part are still the object of work in progress with Corvaja and Masser.

hand, one can estimate efficiently their number from above, since the corresponding rational points lie on a certain *transcendental* real surface, which is obtained as the inverse image of $\sigma(B)$ by the abelian maps. These last estimates stem from work of Bombieri-Pila for curves, then of Pila for surfaces, and finally of Pila-Wilkie for transcendental varieties of arbitrary dimension.[24] Comparison of estimates yields a contradiction (on the appropriate assumptions) if the torsion order is large enough, concluding the argument.

The question of looking at rational points on transcendental objects (whereas the original motivations mainly concerned algebraic varieties) was raised especially by Sarnak; one should also mention that Sarnak indeed foresaw some principles of this method in the case of tori long ago, in an unpublished paper [30].

Pila and the writer obtained a new proof of the Raynaud's theorem (former Manin-Mumford conjecture) by this method. Masser and the writer first used the method on a 'test' problem of Masser, asking for a proof of finiteness of the set of complex numbers l such that two points with abscissa 2 or 3 on the Legendre curve $E_l : y^2 = x(x-1)(x-l)$ are torsion; later this problem was realized to be a case of the Pink's conjectures. The proof of Masser's expectation was achieved in [14], [15], and the already mentioned further papers extended this systematically. In another direction, Pila applied this kind of method to prove new and considerably general cases of the Andrè-Oort conjecture; moreover he found that the counting of rational points was indeed 'doubly useful' in the whole context. Pila, Tsimerman, Ullmo and others extended this last principle to further issues. (See [34] for more details on this brief history and for some references, and see [22].)

Further (delicate) results needed in this kind of proof are:

- A bound by Silverman (1983) on the height of algebraic points l such that $\sigma(l)$ becomes torsion. (This bound has been already mentioned above; in particular, it is useful to deal with issue (b) in the previous section, but also elsewhere in the proofs. It requires appropriate assumptions on which we do not pause here.)
- An inequality by David (1991) relating the order of a torsion point x of a simple abelian variety $A/\overline{\mathbb{Q}}$ with the degree over $\mathbb{Q}$ of a common field of definition for A, x, and with the Faltings height of A. This is proved by transcendence techniques, and in particular offers explicit dependencies compared to previous bounds of Masser. (See Masser's Appendix D in [34] for a sketch of such a proof in a special case.) It is also to be remarked that our application with Masser of this result of David needs, in the case of 'unexpected' splitting of the abelian variety, other results of Masser-Wüstholz, in order to use this theorem for the simple factors.
- A result by André (1992) asserting algebraic independence, as locally analytic functions on B, of abelian logarithms of the given section and periods for the tori corresponding to the abelian fibers. This is crucial to show that the real surface containing the relevant rational points (to which one has to apply the Pila-Wilkie estimates) is 'highly' transcendental, in the sense that actually it does not contain any arc of a real algebraic curve.

Here we omit references for these auxiliary ingredients, and instead refer to [18], [22] or [34] for that.

[24] Other applications of these results may be found in the already mentioned paper [22] by Pila in the present volume.

Effectivity within this method is expected, but not yet proved. (The mentioned height bounds and lower bounds for degrees are already effective, but the situation is not yet clear for the upper bounds in the Bombieri-Pila-Wilkie work.) This reflects towards an ineffectivity for the finiteness theorems in the quoted papers, and hence also concerning the present context: for instance, at the moment there is no available procedure to exhibit the finite set in Example 2.4(ii).

Analogues and a conjecture of Grothendieck. Indefinite integration of a given differential $h(z)\mathrm{d}z$ corresponds to a linear differential equation $y'(z) := \mathrm{d}y/\mathrm{d}z = h(z)$ for a function $y = y(z)$. (This may be also reduced to the homogeneous form $hy'' = h'y'$.) Needless to say, one can consider more general equations, and especially relevant generalizations are obtained with linear ones of any order, with rational (or algebraic) function coefficients. The study of the fields generated by the solutions originated *differential Galois theory*, and information on the relevant Galois groups often enables one to answer questions like whether the solutions can be expressed in terms of a certain prescribed class of functions. Previously in this paper we have considered the class of elementary functions, but one may change this, and for instance either restrict to algebraic functions, or, conversely, allow more freedom, for instance including in the class all the integrals of algebraic functions, or all solutions of special types of differential equations.[25]

So, we see that there are somewhat different types of questions that may be asked, and it turns out that their study may require different tools. For instance, in the special case of elementary integrals considered above in this paper, the mere differential Galois groups does not provide any useful information. Still, sometimes the issues may be strictly linked; for example, the analysis of §2 led to the equation $c\mathrm{d}y = \omega y$ and in turn to torsion on a Jacobian and the Pell's equation, and this is the same that one finds on asking for algebraic solutions of the differential equation (see also [3], in connection with Remark 2.2 above).

Now, similarly to the problems discussed above, one can consider parametric families of such differential equations, and it makes sense to ask how often the structure of the field generated by the solutions of the specialized equations reflects the generic structure. Here again we may ask several types of questions:

> *For which values of the parameter does the Galois group equal the generic one?*

or

> *For which values of the parameter can the solutions be expressed within some prescribed class, assuming this can't be done for the generic solution?*

And so on. Again, the analysis and the answers may be quite different in the various cases.

For instance, concerning a pencil of differentials ω_λ on X_λ (notation being as above in this section) it would be easy to prove that ω_λ has an *algebraic* (rather than an *elementary*) integral identically in λ if and only if ω_l has an algebraic integral for infinitely many $l \in B(\overline{\mathbb{Q}})$. (A proof follows similarly to the first cases of Example 4.1(i), (ii), in §4 above.) However, even for a differential equation of the first order, we may have algebraic solutions (i.e. finite Galois group) precisely for *rational* values of the parameter, whereas the generic Galois group is $\mathbb{G}_{\mathrm{m}}$: this is what happens for $zy' = \lambda y$, with solutions $c \cdot z^\lambda$.

[25] For examples of this, see the books quoted in §1, and [3]. A relevant category is the class of *liouvillian extensions*, see [24].

In particular, if we formulate our issue in terms of the Galois group, a finiteness assertion for the exceptional specializations does not hold generally even restricting to algebraic function coefficients for our equations.[26] Concerning other classes of functions, here is a (vague) problem in this direction.

Problem. Let us be given a linear differential equation $\sum_{i=0}^{n} a_{i\lambda}(z)y^{(i)} = 0$ with coefficients $a_{i\lambda} \in \overline{\mathbb{Q}}(\lambda)[z]$, and let us assume that it has no nonconstant solution which is an elementary function (in the above sense) over the constant field $\overline{\mathbb{Q}(\lambda)}$. Under what conditions can we conclude that there are only finitely many $l \in \overline{\mathbb{Q}}$ for which the specialized equation $\sum_{i=0}^{n} a_{il}(z)y^{(i)} = 0$ has a nonconstant elementary function solution?

Perhaps the assumptions here could be changed by replacing throughout 'a nonconstant solution' with 'a full system of linearly independent solutions'; and perhaps one should require that all singularities of the operator are regular.[27] Observe that in the above formulation we do not have finiteness in general, as shown by the equation $zy'' - (z+\lambda)y' = 0$, satisfied by $\int z^{\lambda}e^{z}\mathrm{d}z$, which is elementary precisely for $\lambda \in \mathbb{N}$.[28]

In any case, at the moment we have no definite general conjecture in this direction, but we believe it should be interesting to explore the questions, the classes of functions and the conditions under which one can expect finiteness.

Another analogy concerns `reduction modulo a prime` p instead of specialization of a parameter. Namely, we suppose to be given a differential ω on an algebraic curve X defined over $\mathbb{Q}$ (or a number field K), so that we can reduce modulo almost all primes p (or places v of K), to obtain a differential ω_p on the reduced curve X_p, defined over the residue field $\mathbb{F}_p$. This is of course a kind of specialization, and we may ask as before for which primes p is ω_p integrable in finite terms, assuming that ω is not likewise integrable.

Naturally, the fact that the residue field has positive characteristic changes much, if not for the fact that the function field $\mathbb{F}_p(X_p)$ has finite degree over the constant field for a derivation. Indeed, there are several features which distinguish this analysis from the former one. Let us see two simple examples in this direction.

Example 6.1.
(i) Let E be an elliptic curve over $\mathbb{Q}$, with Weierstrass equation $w^2 = f(z)$ ($f \in \mathbb{Q}[z]$ a cubic monic polynomial with simple roots) and let $\omega = \mathrm{d}z/w$ be a regular differential (unique up to constants).

We have seen in §2 that in characteristic zero a regular differential is never integrable in finite terms. On the other hand, concerning the (good) reduction(s) E_p of E, we have:

- *If p is a prime of supersingular reduction for E* [29]*, i.e. if E_p is supersingular, then ω_p is exact, i.e. $\omega_p = \mathrm{d}u_p$ for some $u_p \in \overline{\mathbb{F}_p}(X_p)$.*
- *If E_p is ordinary, then $\omega_p = c_p \cdot \mathrm{d}u_p/u_p$ for some constant c_p and some nonzero $u_p \in \overline{\mathbb{F}_p}(X_p)$.*

See [12], Appendix 2.4, for a proof using the Cartier operator. So, we see that now

[26] In the paper [2], see especially §7, André has studied the context of a conjecture of Grothendieck on reductions modulo p of differential equations (to be mentioned also below) in the case of families in characteristic zero.

[27] This assumption is often relevant in the context, for instance to relate the differential Galois group to monodromy.

[28] Note that this equation has an irregular singularity at infinity.

[29] At least for curves over $\mathbb{Q}$, there are infinitely many such primes after a result of Elkies.

the reduction of ω is *always* integrable in finite terms, and is even exact for infinitely many primes.

This is in marked contrast with our previous conclusions (regarding specializations) in characteristic zero. (On the other hand, an integral of ω_p is algebraic only in the case of supersingular reduction.)

(ii) Let now $r \in \mathbb{Z}$ and, similarly to Example 4.1(ii), set $f(z) = z^4 + z + r$, and consider the curve E defined as a projective smooth model of $w^2 = f(z)$; the function z has two poles denoted $\infty_\pm$ and E becomes an elliptic curve on choosing e.g. ∞_- as an origin.

Consider $\omega := z\mathrm{d}z/w$; as in §2, ω has two simple poles at $\infty_\pm$ and no other poles. As recalled in Example 2.4(i), the Pell's equation (2.3) may be solvable for the present $f(z)$ only for finitely many $r \in \mathbb{Q}$; hence, Proposition 2.3 shows that for almost all $r \in \mathbb{Z}$, as in (i), ω is not integrable in finite terms. Let us then fix one such r, and as above let us p vary.

We haven't mentioned any analogue of Theorem 1.1 in positive characteristic, but we shall only need the sufficiency (for integrability in finite terms) of the condition stated therein. And actually it shall be sufficient to study whether $\omega_p = \mathrm{d}v_p + c_p \cdot \mathrm{d}u_p/u_p$ for rational functions u_p, v_p on E_p and a constant c_p.

Observe, by the way, that after reduction the Pell's equation is always solvable (as in the classical case of $\mathbb{Z}$), because any point on E_p defined over a finite field has finite order; however the analysis parallels only in part the one for Proposition 2.3.

Note that (after a possible finite extension of the constant field) there exists a function u_p whose divisor is of the shape $\infty_+ - \infty_- - pD_p$, where D_p is a divisor of degree 0 on E_p. Indeed, since $J_p := Pic^o(E_p)$ is a divisible group we can find $\gamma \in J_p$ so that $\delta = p\gamma$, where δ is the class of $\infty_+ - \infty_-$. And then, if D_p is a divisor of degree 0 with class γ, we find what is required.

Now, $\mathrm{d}u_p/u_p$ has only simple poles, exactly at $\infty_\pm$, with respective residues ± 1, and adjusting the constant c_p we can ensure that $\omega_p - c_p \cdot \mathrm{d}u_p/u_p$ has no poles, i.e. is a regular differential on E_p (possibly 0): $\omega_p = c_p \cdot \mathrm{d}u_p/u_p + \eta_p$, where η_p is regular on E_p.

But then, by what has been recalled in part (i) of this example, η_p is always either exact or logarithmic up to a constant. We conclude that again ω_p is always integrable in finite terms.

These examples show in particular that in this context things behave rather differently compared to what we have previously seen, since it happens that the reduction is always integrable in finite terms but the differential is not. On the contrary, if we look at an *algebraic* integral, things are different. Let us pause a moment on this interesting issue.

An attractive conjecture attributed to Pólya was as follows:

> *Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be a power series with coefficients $a_n \in \mathbb{Z}$, representing an algebraic function. Suppose that an indefinite integral of $f(z)$ has also integral coefficients (which amounts to $n+1 | a_n$ for all $n \in \mathbb{N}$). Then the integral (that is, the power series $\sum_{n=0}^{\infty} a_n z^{n+1}/(n+1)$) also represents an algebraic function.*

This statement might look perhaps innocuous, however turned out to be deep; a proof for *rational functions* $f(z)$ is not too difficult, and similarly if $(z, f(z))$ describes a rational curve, but the case of positive genus is at a quite different level. A general proof was given by André, partly relying on a method of Chudnovski; see [2], Prop. 6.2.1.

Actually, André worked under weaker assumptions; he looked at the differential equation $y'(z) = f(z)$ and assumed the solvability of the reduction of this equation modulo p, in a power series in $\mathbb{F}_p[[z]]$, for all primes p in a set of density 1 (or even density $> 1/2$), to obtain the same conclusion that the equation has algebraic function solutions. André proved a similar theorem for the equation $y' = f(z)y$.

These theorems of André represent cases of a well-known conjecture of Grothendieck, which (roughly speaking) predicts *a complete system of algebraic function solutions for a linear differential equation over $\mathbb{Q}(z)$ whose reduction modulo p has a complete system of power series solutions over a finite field, for almost all primes p.* Other important cases of this deep conjecture, still open in general, were proved by Katz [9].

Naturally, one may replace $\mathbb{Q}$ by a number field.

We note that for the conclusion it does not suffice that the hypothesis holds for *infinitely many primes p*: indeed, as remarked in Example 6.1(i), a nonzero regular differential on an elliptic curve over $\mathbb{Q}$ becomes exact for the infinitely many primes of supersingular reduction (and if the curve is CM this set of primes has even density $1/2$).

We refer to [2], to §8 of [3] and to Katz's paper [9] for extensive discussions and results towards the Grothendieck conjecture and related topics. (As mentioned above, in [2] an analogue of the conjecture is studied in generality for fields of characteristic zero, including for instance pencils of differential equations over a curve.)

To conclude, we remark it would be interesting to formulate a general statement embracing the examples we have met so far.

Acknowledgements. The author is grateful to the ERC for providing support during the writing of this paper. We thank David Masser, with whom the whole research at the basis of this article has been carried out; he also provided many helpful illustrations and comments. We further thank Yves André and Jonathan Pila for other comments and references.

References

[1] N.H. Abel, *Über die Integration der Differential-Formel $\rho dx/\sqrt{R}$, wenn R und ρ ganze Funktionen sind*, J. für Math. (Crelle) **1** (1826), 185–221.

[2] Y. André, Sur la conjecture des p-courbures de Grothendieck-Katz et un problème de Dwork, Geometric aspects of Dwork theory. Vol. I, II, 55–112, Walter de Gruyter GmbH & Co. KG, Berlin, 2004.

[3] F. Baldassarri and B. Dwork, *On Second Order Linear Differential Equations with Algebraic Solutions*, Amer. J. Math. **101** (1979), 42–76.

[4] E. Bombieri and W. Gubler, *Heights in Diophantine Geometry, New Math. monographs*, vol. 4, Cambidge Univ. Press, 2006.

[5] P. Corvaja, D. Masser, and U. Zannier, *Sharpening Manin-Mumford for certain algebraic groups of dimension* 2, L'Enseign. Math. (with a letter of Serre to Masser as an appendix), to appear.

[6] J. Davenport, *On the integration of algebraic functions*, Lecture Notes in Computer Science, **102**, Springer-Verlag, 1981.

[7] G.H. Hardy, *The Integration of Functions of a Single Variable*, Cambridge Univ. Press, 1916.

[8] I. Kaplanski, *An introduction to differential algebra*, 2nd ed., Hermann, Paris, 1976.

[9] N.M. Katz, *Algebraic Solutions of Differential Equations (p-Curvature and the Hodge Filtration)*, Inventiones Math. **18** (1972), 1–118.

[10] E.R. Kolchin, *Differential Algebra and Algebraic Groups*, Academic Press, 1973.

[11] S. Lang, *Introduction to Algebraic and Abelian Functions*, II ed., Springer-Verlag, 1982.

[12] ———, *Elliptic Functions*, Addison Wesley, 1973.

[13] D. Masser and G. Wüstholz, *Periods and minimal abelian varieties*, Annals of Math. **137** (1993), 407–458.

[14] D. Masser and U. Zannier, *Torsion anomalous points and families of elliptic curves*, C. Rendus Acad. Sci. Paris, **346** (2008), 491–494.

[15] ———, *Torsion anomalous points and families of elliptic curves*, Amer. J. Math. **132** (2010), 1677–1691.

[16] ———, *Torsion points on families of squares of elliptic curves*, Math. Annalen **352** (2012), 453-484.

[17] ———, *Torsion points on families of products of elliptic curves*, preprint 2012, submitted for publication.

[18] ———, *Torsion points on families of simple abelian surfaces and Pell's equation over polynomial rings*, submitted for publication.

[19] ———, *Torsion points on families of abelian varieties, Pell's equation, and integration in elementary terms*, in progress, 2014.

[20] J. Pila, *O-minimality and Diophantine Geometry*, this volume.

[21] J. Pila and U. Zannier, *Rational points in periodic analytic sets and the Manin-Mumford conjecture*, Rend. Mat. Lincei, 2008, 1–14.

[22] R. Pink, *A combination of the conjectures of Mordell-Lang and André-Oort, Geometric Methods in Algebra and Number Theory*, F. Bogomolov and Yu. Tschinkel Edts., 251–282, vol. 253, Birkhauser, Boston, 2005.

[23] A.J. van der Poorten and X.C. Tran, *Quasi-elliptic integrals and periodic continued fractions*, Monatsh. Math. **131** (2000), 155–169.

[24] M. van der Put and M.F. Singer, *Galois Theory of Linear Differential Equations*, 2002.

[25] R.H. Risch, *The problem of integration in finite terms*, Trans. Amer. Math. Soc. **139** (1969), 167–189.

[26] _____, *The solution of the problem of integration in finite terms*, Bull. Amer. Math. Soc. **76** (1970), 605–608.

[27] J.F. Ritt, *Integration in finite terms*, Columbia Univ. Press, New York, 1948.

[28] M. Rosenlicht, *Liouville's theorem on functions with elementary integral*, Pacific J. Math. Vol. 24 (1968), 153–161.

[29] _____, *Integration in finite terms*, Amer. Math. Monthly, 72 (1979), 963–972.

[30] P. Sarnak, *Torsion points on varieties and homology of abelian covers*, manuscript, 1989.

[31] H. Schmidt, PhD thesis in progress, Basel.

[32] J-P. Serre, *Algebraic Groups and Class Fields*, Springer-Verlag GTM 117, 1988.

[33] C.L. Siegel, *Topics in complex function theory*, Vol. I, Wiley, 1969.

[34] U. Zannier, *Some Problems of Unlikely Intersections in Arithmetic and Geometry* (with Appendixes by D. Masser), Annals of Math. Studies, n. 181, Princeton Univ. Pres, 2012.

[35] _____, *Unlikely Intersections and Pell's Equation in Polynomials*, to appear, Springer Verlag.

Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa - ITALY
E-mail: u.zannier@sns.it

Small gaps between primes and primes in arithmetic progressions to large moduli

Yitang Zhang

Abstract. Let p_n denote the n-th prime. We describe the proof of the recent result

$$\liminf_{n\to\infty}(p_{n+1}-p_n)<\infty,$$

which is closely related to the distribution of primes in arithmetic progressions to large moduli. A major ingredient of the argument is a stronger version of the Bombieri-Vinogradov theorem which is applicable when the moduli are free from large prime factors.

Mathematics Subject Classification (2010). 11N05, 11N13.

Keywords. Gaps between primes, primes in arithmetic progressions, Bombieri-Vinogradov theorem, Kloostermann sums.

1. Introduction

Let p_n denote the n-th prime and write $d_n=p_{n+1}-p_n$. The prime number theorem implies that the average value of d_n is $\sim \log p_n$. The study of the upper and lower bounds for d_n is one of the central subjects in analytic number theory.

Currently, the best result on the upper bound for d_n is due to Baker, Harman and Pintz [1] that gives

$$d_n \ll p_n^{21/40}.$$

The lower bound for d_n has attracted great interest. For eighty years, the work on this problem had concentrated on estimating the quantity

$$\Delta=\liminf_{n\to\infty}\frac{d_n}{\log p_n}.$$

The inequality $\Delta\le 1$ is a trivial consequence of the prime number theorem. In 1926 Hardy and Littlewood [13] obtained $\Delta\le 2/3$ on assuming GRH. In 1940 Erdös [8] obtained $\Delta<1-c$ unconditionally with a small unspecified computable constant $c>0$. In 1965 Bombieri and Davenport [2] obtained $\Delta\le 0.4665....$ In 1988 Maier [16] obtained $\Delta\le 0.2484....$ In 2005 Goldston, Pintz and Yildirim [11] eventually proved

$$\Delta=0.$$

In 2013 Zhang [18] proved the following

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Theorem 1.1.

$$\liminf_{n\to\infty} d_n < \infty. \tag{1.1}$$

This result may be regarded as a weak form of the twin prime conjecture

$$\liminf_{n\to\infty} d_n = 2. \tag{1.2}$$

In fact, Zhang proves (1.1) with the right side replaced by 7×10^7. This specific number has been considerably reduced.

A major ingredient of the proof is a stronger version of the Bombieri-Vinogradov theorem which is not implied by the Generalized Riemann Hypothesis directly, but it relies on some deep results in algebraic geometry.

Recently Maynard [15] obtained some stronger results via different methods. In particular, he proved

$$\liminf_{n\to\infty}(p_{n+k} - p_n) < \infty$$

for any fixed $k > 0$.

2. Admissible sets

Let

$$\mathcal{H} = \{h_1, h_2, ..., h_k\}$$

be a set composed of distinct non-negative integers. We say that $\mathcal{H}$ is admissible if $\nu_p(\mathcal{H}) < p$ for every prime p, where $\nu_p(\mathcal{H})$ denotes the number of distinct residue classes modulo p occupied by the h_j.

The reason for introducing admissible sets may be described as follows. Assume that $\mathcal{H}$ is not admissible. Then there is a prime p such that, for any positive integer n, the tuple

$$(n + h_1, n + h_2, ..., n + h_k) \tag{2.1}$$

contains at least one multiple of p. Thus we conclude

Suppose that there are infinitely many positive integers n such that each element in the tuple (2.1) is a prime. Then $\mathcal{H}$ must be admissible.

A conjecture of Hardy and Littlewood asserts the converse. Here we state a weak form of their conjecture only.

Conjecture 2.1. *Suppose that $\mathcal{H}$ is admissible. There are infinitely many positive integers n such that each element in the tuple* (2.1) *is a prime.*

Note that the twin prime conjecture is a special case of this conjecture with $\mathcal{H} = \{0, 2\}$.

It is easy to see that Theorem 1.1 can be deduced from the following

Theorem 2.2. *Suppose that $\mathcal{H}$ is admissible with k sufficiently large. There are infinitely many positive integers n such that the tuple* (2.1) *contains at least two primes.*

Theorem 2.2 actually gives

$$\liminf_{n\to\infty} d_n \le \max_{i\neq j} |h_i - h_j|. \tag{2.2}$$

The initial step in the proof of Theorem 2.2 consists in transferring the problem to evaluate and compare certain arithmetic sums. Write

$$\theta(n) = \begin{cases} \log n & \text{if } n \text{ is prime,} \\ 0 & \text{otherwise.} \end{cases}$$

Let x denote a large number, and let f be defined on $(x, 2x) \cap \mathbf{Z}$ such that

$$f(n) \geq 0. \tag{2.3}$$

We introduce

$$S_1(x, f) = \sum_{x<n<2x} f(n) \tag{2.4}$$

and

$$S_2(x, f; \mathcal{H}) = \sum_{x<n<2x} \left(\sum_{j=1}^{k} \theta(n + h_j) \right) f(n). \tag{2.5}$$

The key point is to prove that, for any sufficiently large x, there is a function f satisfying (2.3), such that

$$\frac{S_2(x, f; \mathcal{H})}{S_1(x, f)} > \log 3x. \tag{2.6}$$

This implies that there is an integer $n \in (x, 2x)$ such that the tuple (2.1) contains at least two primes.

3. The choice of f by Goldsdon-Pintz-Yildirim

From now on, the set $\mathcal{H}$ is assumed to be admissible and fixed.

To choose f and show that (2.6) holds, the following conditions are necessary.

(i) f should be non-negative.

(ii) One should be able to evaluate the sums $S_1(x, f)$ and $S_2(x, f; \mathcal{H})$ efficiently.

(iii) The sum $S_2(x, f; \mathcal{H})$ should be "large" in comparison with $S_1(x, f)$.

Let $\Lambda(n)$ denote the von Mangoldt function. To satisfy the condition (iii), one may choose

$$f(n) = \sum_{j=1}^{k} \Lambda(n + h_j)$$

(the difference $\theta - \Lambda$ is negligible). However, this is not valid in practice since one is unable to evaluate $S_2(x, f; \mathcal{H})$.

It is known that

$$\Lambda(n) = \sum_{d|n} \mu(d) \log \frac{n}{d}, \tag{3.1}$$

where $\mu(d)$ denotes the Möbius function. This relation may not be applicable directly in

many sieve problems, since one is unable to handle the contribution from the terms on the right side with d large. In practice, $\Lambda(n)$ is often approximated (replaced) by

$$\Lambda_D(n) = \sum_{\substack{d|n \\ d<D}} \mu(d) \log \frac{D}{d} \tag{3.2}$$

for $n < 2x$, where D is a parameter $< x$. Usually, D is of the form

$$D = x^b, \qquad 0 < b < \frac{1}{2}. \tag{3.3}$$

If one chooses

$$f(n) = \sum_{j=1}^{k} \Lambda_D(n + h_j),$$

then, with D given by (3.3), $S_1(x, f)$ and $S_2(x, f; \mathcal{H})$ can be evaluated. However, the condition (i) may not hold.

In order to satisfy the condition (i), a simple idea which comes from Selberg's sieve is that f takes the form

$$f(n) = \lambda(n)^2, \qquad \lambda(n) \in \mathbf{R}. \tag{3.4}$$

From now on, we assume that f is of the form (3.4) and write $D = x^b$. The problem is reduced to choosing λ.

It was Goldston, Pintz and Yildirim who found a form of $\lambda(n)$ which proves very efficient. Let

$$P_{\mathcal{H}}(n) = \prod_{j=1}^{k} (n + h_j), \tag{3.5}$$

and let l be a fixed positive integer. They choose

$$\lambda(n) = \frac{1}{(k+l)!} \sum_{\substack{d|P_{\mathcal{H}}(n) \\ d<D}} \mu(d) \bigg(\log \frac{D}{d} \bigg)^{k+l}. \tag{3.6}$$

One may also consider the general form

$$\lambda(n) = \sum_{\substack{d|P_{\mathcal{H}}(n) \\ d<D}} \mu(d) h\bigg(\frac{\log(D/d)}{\log D} \bigg), \tag{3.7}$$

where h is a real polynomial satisfying

$$h^{(i)}(0) = 0, \qquad 0 \le i \le k.$$

The form (3.7) and its variants also find application to other problems in analytic number theory, which is usually called the GPY sieve.

From now on, we assume that $b < 1/2$ and write ν_p, $P(n)$, S_1 and S_2 for $\nu_p(\mathcal{H})$, $P_{\mathcal{H}}(n)$, $S_1(x, f)$ and $S_2(x, f; \mathcal{H})$ respectively; similar abbreviations will apply in the sequel. Let $\lambda(n)$ be as in (3.6). Write

$$g(d) = \frac{1}{(k+l)!} \bigg(\log \frac{D}{d} \bigg)^{k+l} \quad \text{if} \quad d < D \tag{3.8}$$

and

$$g(d) = 0 \quad \text{if} \quad d \geq D, \tag{3.9}$$

so that

$$\lambda(n) = \sum_{d|P(n)} \mu(d)g(d). \tag{3.10}$$

The evaluation of S_1 is not difficult. Squaring out $\lambda(n)^2$ and changing the order of summation we obtain

$$S_1 = \sum_{d_1} \sum_{d_2} \mu(d_1)\mu(d_2)g(d_1)g(d_2) \sum_{\substack{x<n<2x \\ P(n)\equiv 0([d_1,d_2])}} 1, \tag{3.11}$$

where $[d_1, d_2]$ denotes the l.c.m. of d_1 and d_2. Note that $[d_1, d_2]$ is square-free and

$$[d_1, d_2] < D^2 = x^{2b}$$

if $\mu(d_1)\mu(d_2)g(d_1)g(d_2) \neq 0$. Let

$$\varrho_1(d) = |\mu(d)| \prod_{p|d} \nu_p. \tag{3.12}$$

For any square-free d, $d|P(n)$ if and only if n lies in exactly $\varrho_1(d)$ distinct residue classes $(\mathrm{mod}\ d)$. It follows that the innermost sum in (3.11) is equal to

$$\frac{\varrho_1([d_1, d_2])}{[d_1, d_2]} x + O(\varrho_1([d_1, d_2])).$$

This yields

$$S_1 = xT_1 + O(x^{2b+\varepsilon}), \tag{3.13}$$

where

$$T_1 = \sum_{d_1} \sum_{d_2} \frac{\mu(d_1)\mu(d_2)g(d_1)g(d_2)\varrho_1([d_1, d_2])}{[d_1, d_2]}. \tag{3.14}$$

The error term in (3.13) is acceptable since $b < 1/2$.

We now turn to S_2. In a way similar to the proof of (3.11), we deduce that

$$S_2 = \sum_{d_1} \sum_{d_2} \mu(d_1)\mu(d_2)g(d_1)g(d_2) \sum_{j=1}^{k} \sum_{\substack{x<n<2x \\ P(n)\equiv 0([d_1,d_2])}} \theta(n + h_j). \tag{3.15}$$

If d is square-free, $d < D^2$, $n \in (x, 2x)$ and $n + h_j$ is a prime, then $d|P(n)$ if and only if

$$d \equiv c(\mathrm{mod}\ d) \quad \text{for some} \quad c \in \mathcal{C}_j(d),$$

where

$$\mathcal{C}_j(d) = \{c : 1 \leq c \leq d,\ (c, d) = 1,\ P(c - h_j) \equiv 0(\mathrm{mod}\ d)\}. \tag{3.16}$$

Thus the innermost sum in (3.15) is equal to

$$\sum_{c\in\mathcal{C}_j([d_1,d_2])} \sum_{\substack{x<n<2x \\ n\equiv c([d_1,d_2])}} \theta(n + h_j). \tag{3.17}$$

Let

$$\varrho_2(d) = |\mu(d)| \prod_{p|d} (\nu_p - 1). \tag{3.18}$$

For square-free d it can be shown that

$$|\mathcal{C}_j(d)| = \varrho_2(d).$$

It follows from (3.17) and the prime number theorem that the innermost sum in (3.15) is equal to

$$\frac{\varrho_2(d)}{\varphi(d)} x + O\Big(\sum_{c \in \mathcal{C}_j(d)} \mathcal{R}(x; c, d) \Big) + O(x^{\varepsilon}), \tag{3.19}$$

where $\varphi(d)$ denotes the Euler function, and where

$$\mathcal{R}(x; d, c) = \Big| \sum_{\substack{x<n<2x \\ n \equiv c(d)}} \theta(n) - \frac{x}{\varphi(d)} \Big|. \tag{3.20}$$

Inserting (3.19) into (3.15) we obtain

$$S_2 = xT_2 + O((\log x)^{2k+2l} E) + O(x^{2b+\varepsilon}), \tag{3.21}$$

where

$$T_2 = k \sum_{d_1} \sum_{d_2} \frac{\mu(d_1)\mu(d_2) g(d_1) g(d_2) \varrho_2([d_1, d_2])}{\varphi([d_1, d_2])} \tag{3.22}$$

and

$$E = \sum_{d < D^2} |\mu(d)| \tau_3(d) \varrho_2(d) \sum_{j=1}^{k} \sum_{c \in \mathcal{C}_j(d)} \mathcal{R}(x; c, d).$$

Here $\tau_3(d)$ denotes the 3-fold divisor function.

The problem is now reduced to evaluation T_1 and T_2 and estimating E.

4. The main term

By (3.13) and (3.21), we need only prove the following which lead to (2.6).

(a) With k sufficiently large and with l appropriately chosen,

$$\frac{T_2}{T_1} > (1 + c) \log x \tag{4.1}$$

for some positive constant c.

(b) The sum E can be efficiently bounded, namely,

$$E \ll x(\log x)^{-A} \tag{4.2}$$

for any large constant A.

Note that ϱ_1 is supported on square-free integers. Substituting $d_0 = (d_1, d_2)$ and rewriting d_1 and d_2 for d_1/d_0 and d_2/d_0 respectively, we deduce that

$$T_1 = \sum_{d_0} \sum_{d_1} \sum_{d_2} \frac{\mu(d_1 d_2)\varrho_1(d_0 d_1 d_2)}{d_0 d_1 d_2} g(d_0 d_1) g(d_0 d_2). \tag{4.3}$$

Similarly,

$$T_2 = k \sum_{d_0} \sum_{d_1} \sum_{d_2} \frac{\mu(d_1 d_2)\varrho_2(d_0 d_1 d_2)}{\varphi(d_0 d_1 d_2)} g(d_0 d_1) g(d_0 d_2). \tag{4.4}$$

The right sides of (4.3) and (4.4) can be evaluated by standard methods. It is shown that

$$T_1 \sim \frac{1}{(k+2l)!} \binom{2l}{l} \mathfrak{S} (\log D)^{k+2l}$$

and

$$T_2 \sim \frac{k}{(k+2l+1)!} \binom{2l+2}{l+1} \mathfrak{S} (\log D)^{k+2l+1},$$

where

$$\mathfrak{S} = \prod_p \left(1 - \frac{\nu_p}{p}\right) \left(1 - \frac{1}{p}\right)^{-k}.$$

(Note that $\mathfrak{S} = 0$ if $\mathcal{H}$ is not admissible.) It follows that

$$\frac{T_2}{T_1} \sim \frac{bk}{k+2l+1} \frac{(2l+2)(2l+1)}{(l+1)^2} \log x. \tag{4.5}$$

If $b > 1/4$, then, with k is sufficiently large in terms of b and with l appropriately chosen, we have

$$\frac{bk}{k+2l+1} \frac{(2l+2)(2l+1)}{(l+1)^2} > 1. \tag{4.6}$$

This leads to (4.1). On the other hand, the relation (4.6) is not valid for $b \le 1/4$.

Thus we conclude

Theorem 2.2 will be valid if (4.2) *holds for some* $b > 1/4$.

5. The error term

We say that the primes have level of distribution ϑ if, for any large constant A and for any small positive constant ε,

$$\sum_{d<Q} \max_{y \le x} \max_{(c,d)=1} \left| \sum_{\substack{n \le y \\ n \equiv c(d)}} \Lambda(n) - \frac{y}{\varphi(d)} \right| \ll x (\log x)^{-A} \tag{5.1}$$

with $Q = x^{\vartheta - \varepsilon}$. Since the estimate

$$\sum_{\substack{n \le y \\ n \equiv c(d)}} \Lambda(n) - \frac{y}{\varphi(d)} \ll y^{1/2+\varepsilon}$$

is implied by the GRH, the primes have level of distribution $1/2$ if GRH is true. This result was unconditionally proved by Bombieri and Vinogradov in 1965.

Theorem 5.1. *The primes have level of distribution* $1/2$.

Elliott and Halberstam [7] conjectured that the primes have level of distribution 1. By Cauchy's inequality, we see that (4.2) will be valid if the primes have level of distribution $\vartheta = 2b + \varepsilon$. Thus, by the discussion in Section 4, we conclude

> *Theorem 2.2 will be valid if the primes have level of distribution* ϑ *for some* $\vartheta > 1/2$.

6. Refinement of the GPY method

One is unable to prove that the primes have level of distribution ϑ for some $\vartheta > 1/2$, even if the GRH is assumed. We introduce a refinement of the GPY method which applies to the main and error terms equally.

Let $\varpi > 0$ be a small constant, and let $b = 1/4 + \varpi$. Our first observation is that, on the right sides of (4.3) and (4.4), the terms with $d_0d_1d_2$ having a large prime factor p, say $p \geq x^{\varpi}$, make minor contribution. Thus, if we modify the sum in (3.6) by imposing the constraint that $p|d$ implies $p < x^{\varpi}$, the resulting main terms in S_1 and S_2 will have minor changes only. This was independently observed by Motohashi and Pintz. Our second observation, which is the most novel part of the proof, is that with such a constraint imposed in (3.6), the resulting error in S_2 can be efficiently bounded.

From now on, let S_1, S_2 and $f(n)$ be as in (2.4), (2.5) and (3.4) respectively, but let $\lambda(n)$ be redefined by

$$\lambda(n) = \sum_{d|(P(n),\mathcal{P})} \mu(d)g(d)$$

with $g(d)$ given by (3.8) and (3.9), where

$$\mathcal{P} = \prod_{p<x^{\varpi}} p.$$

Repeating the arguments in Section 4 we obtain

$$S_1 = xT_1^* + O(x^{2b+\varepsilon})$$

and

$$S_2 = xT_2^* + O((\log x)^{2k+2l}E^*) + O(x^{2b+\varepsilon}),$$

where

$$T_1^* = \sum_{d_1|\mathcal{P}} \sum_{d_2|\mathcal{P}} \frac{\mu(d_1)\mu(d_2)g(d_1)g(d_2)\varrho_1([d_1,d_2])}{[d_1,d_2]},$$

$$T_2^* = k \sum_{d_1|\mathcal{P}} \sum_{d_2|\mathcal{P}} \frac{\mu(d_1)\mu(d_2)g(d_1)g(d_2)\varrho_2([d_1,d_2])}{\varphi([d_1,d_2])}$$

and

$$E^* = \sum_{\substack{d<D^2\\ d|\mathcal{P}}} |\mu(d)|\tau_3(d)\varrho_2(d)\sum_{j=1}^{k}\sum_{c\in\mathcal{C}_j(d)}\mathcal{R}(x;c,d).$$

By routine estimations of the differences $T_1 - T_1^*$ and $T_2 - T_2^*$ and (4.5), it can be shown that

$$\frac{T_1^*}{T_2^*} > (1+c)\log x$$

for some constant $c > 0$, if k is sufficiently large and if l is appropriately chosen. The proof of Theorem 2.2 is therefore reduced to proving

Theorem 6.1. *If ϖ is sufficiently small, then*

$$E^* \ll x(\log x)^{-A}$$

for any large constant A.

7. Sketch of the proof of Theorem 6.1

Note that $D^2 = x^{1/2+2\varpi}$. By Theorem 5.1 and Cauchy's inequality, the proof of Theorem 6.1 is reduced to showing that

$$\sum_{\substack{x^{1/2-\varepsilon}<d<D^2\\ d|\mathcal{P}}}\sum_{c\in\mathcal{C}_j(d)}\mathcal{R}(x;c,d) \ll x(\log x)^{-2A} \qquad 1\le j\le k. \tag{7.1}$$

The constraint $d|\mathcal{P}$ is crucial in the proof of (7.1). This is originally due to the simple fact that if $1 < R < d$, $d > x^{1/2-\varepsilon}$ and $d|\mathcal{P}$, then d can be factored as

$$d = rq \qquad \text{with} \qquad R < r < x^{\varpi}R. \tag{7.2}$$

We can replace θ by Λ in the expression for $\mathcal{R}(x;c,d)$. The proof of (7.1) is described as follows. For any function γ supported on $(x, 2x)\cap\mathbf{Z}$ we define

$$\Delta(\gamma;c,d) = \Big|\sum_{n\equiv c(d)}\gamma(n) - \frac{1}{\varphi(d)}\sum_{(n,d)=1}\gamma(n)\Big|, \qquad (c,d)=1.$$

Using a combinatorial identity for Λ due to Heath-Brown [14], the proof of (7.1) is reduced to estimating the sum of $\Delta(\gamma;c,d)$ for certain Dirichlet convolutions γ. There are three types of the convolutions involved in the proof. Write

$$\eta = 1 + (\log x)^{-2A}, \qquad x_1 = x^{3/8+8\varpi}, \qquad x_2 = x^{1/2-4\varpi}.$$

In the first two types the function γ is of the form $\gamma = \alpha * \beta$ such that

(i) $\alpha = (\alpha(m))$ is supported on $[M,\ \eta^{19}M)$,

(ii) $\beta = (\beta(n))$ is supported on $[N,\ \eta^{19}N)$,

(iii) $MN \in (x, 2x)$.

(There are some other conditions on $\alpha(m)$ and $\beta(n)$.) We say that γ is of Type I if $x_1 < N \le x_2$; we say that γ is of Type II if $x_2 < N < 2x^{1/2}$; we say that γ is of Type III if it is of the form $\gamma = \alpha * \varkappa_{N_1} * \varkappa_{N_2} * \varkappa_{N_3}$ where α is as in (i) and $\varkappa_N$ is the characteristic function of $[N,\, \eta N) \cap \mathbf{Z}$, such that

(iv) $N_3 \le N_2 \le N_1, \qquad MN_1 \le x_1$,

(v) $MN_1N_2N_3 \in (x, 2x)$.

It should be stressed that, without the constraint $d|\mathcal{P}$, we are unable to efficiently bound the sum

$$\sum_{x^{1/2-\varepsilon}<d<D^2} \sum_{c\in\mathcal{C}_j(d)} \Delta(\gamma; c, d)$$

if γ is of Type I or II. However, by (7.2), the Type I and II estimates are reduced to bounding sums of the type

$$\sum_{R<r<2R} \sum_{Q<q<2Q} \sum_{c\in\mathcal{C}_j(rq)} \Delta(\gamma; c, rq) \tag{7.3}$$

with R appropriately chosen such that it is close to N in the logarithmic scale. Thus, using the dispersion method due to Bombieri, Fouvry, Friedlander and Iwaniec [3–5, 9], and Weil's bound for Kloosterman sums, we can prove that the sum (7.3) is

$$\ll x(\log x)^{-43A}. \tag{7.4}$$

The Type III estimate essentially relies on the Birch-Bombieri result in the appendix to Friedlander and Iwaniec [10]. This result in turn relies on Deligne's proof of the Riemann Hypothesis for varieties over finite fields (the Weil Conjecture)[6]. Assuming γ is of Type III, we estimate each $\Delta(\gamma; c, d)$ directly. However, without the constraint $d|\mathcal{P}$, efficient bounds for $\Delta(\gamma; c, d)$ can not be obtained unless we relax the condition $MN_1 \le x_1$. Our argument is carried out by combining the method in [10] and the factorization (7.2) (here r is chosen to be relatively small), the latter will allow us to save a factor $r^{1/2}$. Thus we obtain

$$\Delta(\gamma; c, d) \ll \frac{x^{1-\varepsilon}}{d} \tag{7.5}$$

for $x^{1/2-\varepsilon} < d < D^2$, $d|\mathcal{P}$ and $(c, d) = 1$.

The estimate (7.1) follows from (7.4) and (7.5).

References

[1] R. C. Baker, G. Harman, and J. Pintz, *The difference between consecutive primes II*, Proc. London Math. Soc. (3)**83** (2001), 532–562.

[2] E. Bombieri and H. Davenport, *Small differences between prime numbers*, Proc. Roy. Soc. Ser. A **293** (1966), 1–18.

[3] E. Bombieri, J. B. Friedlander, and H. Iwaniec, *Primes in arithmetic progressions to large moduli*, Acta Math. **156** (1986), 203–251.

[4] E. Bombieri, J. B. Friedlander, and H. Iwaniec, *Primes in arithmetic progressions to large moduli II*, Math. Ann. **277** (1987), 361–393.

[5] ———, *Primes in arithmetic progressions to large moduli III*, J. Am. Math. Soc. **2** (1989), 215–224.

[6] P. Deligne, *La conjecture de Weil I*, Publ. Math. IHES. **43** (1974), 273–307.

[7] P. D. T. A. Elliott and H. Halberstam, *A conjecture in prime number theory*, in Symposia Mathematica, Vol. IV (INDAM, Roma, 1968/1969), Academic Press, London, 1970, pp 59–72.

[8] P. Erdös, *The difference of consecutive primes*, Duke Math. J. **6** (1940), 438–441.

[9] J. B. Fouvry and H. Iwaniec, *Primes in arithmetic progressions*, Acta Arith. **42** (1983), 197–218.

[10] J. B. Friedlander and H. Iwaniec, *Incomplete Kloosterman sums and a divisor problem*, Ann. Math. **121** (1985), 319–350.

[11] D. A. Goldston, J. Pintz, and C. Y. Yildirim, *Primes in Tuples I*, Ann. Math. **170** (2009), 819–862.

[12] ———, *Primes in Tuples II*, Acta Math. **204** (2010), 1–47.

[13] G. H. Hardy and J. E. Littlewood, *unpublished manuscript.*

[14] D. R. Heath-Brown, *Prime numbers in short intervals and a generalized Vaughan identity*, Canad. J. Math. **34** (1982), 1365–1377.

[15] J. Maynard, *Small gaps between primes*, preprint, arXiv:1311.4600.

[16] H. Maier, *Small differences between prime numbers*, Michigan Math. J. **35** (1988), 323–344.

[17] K. Soundararajan, *Small gaps between prime numbers: the work of Goldston-Pintz-Yildirim*, Bull. Amer. Math. Soc. **44** (2007), 1–18.

[18] Y. Zhang, *Bounded gaps between primes*, Ann.of Math. **179** (2014), 1121–1174.

Department of Mathematics and Statistics, University of New Hampshire, Durham, NH 03824
E-mail: yitangz@unh.edu

Linear equations in primes and dynamics of nilmanifolds

Tamar Ziegler

Abstract. In this paper we survey some of the ideas behind the recent developments in additive number theory, combinatorics and ergodic theory leading to the proof of Hardy-Littlewood type estimates for the number of prime solutions to systems of linear equations of finite complexity.

Mathematics Subject Classification (2010). Primary 11B30, 37A30 ; Secondary 11B25, 37A45.

Keywords. Multiple recurrence, arithmetic progressions, Szemerédi's Theorem, Gowers norms, Hardy-Littlewood conjectures.

1. Introduction

A famous conjecture of Hardy and Littlewood [29] predicts that given a k-tuple of integers $\mathcal{H} = \{h_1, \dots, h_k\}$, there are infinitely many k-tuples

$$x + h_1, \dots, x + h_k,$$

such that all elements are simultaneously prime unless there is an obvious divisibility obstruction. Denote by $\nu_{\mathcal{H}}(p)$ the number of congruence classes modulo p that $\mathcal{H}$ occupies, and call a k-tuple of integers admissible if $\nu_{\mathcal{H}}(p) < p$ for all primes p. Then the Hardy-Littlewood conjecture amounts to the statement that $x + h_1, \dots, x + h_k$ are simultaneously prime infinitely often if and only if $\mathcal{H}$ is admissible. Moreover, they conjectured a precise formula for the asymptotic number of k-tuples for an admissible $\mathcal{H}$: Let $\mathbb{P}$ denote the set of primes, then

$$|\{x \in [1, N], \{x + h_1, \dots, x + h_k\} \subset \mathbb{P}\}| \sim \mathfrak{S}(\mathcal{H}) \frac{N}{(\log N)^k}.$$

The constant $\mathfrak{S}(\mathcal{H})$ is an Euler product and is called the singular series.[1] While there have recently been extraordinary developments towards our understanding of gaps between primes and prime tuples [17, 36, 37, 51], some of them presented at the current ICM, we are still far from proving this conjecture.

One can relax the conjecture by looking for prime points in higher rank affine sublattices of $\mathbb{Z}^k$. In a series of papers by Green-Tao [21, 22], Green-Tao-Z [24] we prove:

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Theorem 1.1 (Green-Tao-Z (2012)). *Let $\{\psi_i(\vec{x})\}_{i=1}^k$ be a collection of k affine linear forms in m variables with integer coefficients, $\psi_i(\vec{x}) = \sum_{j=1}^m a_{ij}x_j + b_i$. Suppose no two forms are affinely dependent*[2]*. Then*

$$|\{\vec{x} \in [0,N]^m, \{\psi_1(\vec{x}), \dots, \psi_k(\vec{x})\} \subset \mathbb{P}\}| \sim \mathfrak{S}(\vec{\psi})\frac{N^m}{(\log N)^k}$$

where $\mathfrak{S}(\vec{\psi})$ is an explicit Euler product (analogous to $\mathfrak{S}(H)$).

As a special case of this theorem we obtain the asymptotic number of k-term arithmetic progressions of primes. The reader will observe that the condition that no two forms are affinely dependent rules out the important case of twin primes, or more generally any k-tuple with bounded gaps as described above, however its non-homogeneous nature allows one to use it in various applications that were previously conditional on the Hardy-Littlewood conjectures (see for example [8, 30]). Theorem 1.1 may be viewed as a vast generalization of Vinogradov's 3-prime theorem [49]: any large enough odd number is a sum of three primes. We remark that very recently Vinogradov's result has been extended to include all odd numbers greater that 5 [31], thus verifying the weak Goldbach conjecture.

In this paper we give an outline of intertwining developments in ergodic theory, combinatorics and additive number theory leading to Theorem 1.1.

2. Arithmetic progressions in sets of positive density

Our starting point on the combinatorial front is the following result of K. Roth [39]. Let $E \subset \mathbb{N}$. The *upper density* of E is defined to be

$$\bar{d}(E) = \limsup_{N\to\infty} \frac{|E \cap [1,N]|}{N}.$$

Theorem 2.1 (Roth 1953). *Let $E \subset \mathbb{N}$ be a set of positive upper density, then E contains a non trivial 3-term arithmetic progression.*

Roth's proof plays an important role in later developments - we outline the idea below. Let $\delta > 0$, and suppose E has density δ in an arithmetic progression P of size N, namely $E \subset P$ and $|E| = \delta N$. We first observe that if each element in P were to be chosen independently at random to be in E with probability δ then E would typically contain many 3-term progressions - approximately $\delta^3 N^2$. In view of this, Roth's argument is based on the following:

- either E has at least $\frac{\delta^3 N^2}{2}$ 3-term progressions, or

[1]The singular series $\mathfrak{S}(\mathcal{H})$ is given by the Euler product

$$\mathfrak{S}(\mathcal{H}) = \prod_p \left(1 - \frac{\nu_{\mathcal{H}}(p)}{p}\right)\left(1 - \frac{1}{p}\right)^{-k}.$$

We refer the reader to [42] for an excellent exposition of the heuristics leading to the conjecture above. We write $a(N) \sim b(N)$ if $a(N) = b(N)(1 + o(1))$.

[2]Affine linear forms are affinely dependent is their linear parts are linearly dependent; e.g. the forms x and $x+2$ are affinely dependent. A collection of k affine linear forms no two forms are affinely dependent is said to be of *finite complexity* [21].

- E has density at least $\delta + c(\delta)$ on a sub-progression $Q \subset P$ of size $N^{\frac{1}{3}}$, where c is a decreasing positive function.

Our starting point is a subset $E \subset P = [1, N]$, of density δ. After running the above argument at most $s = 1/c(\delta)$ times we obtain a subset $E' \subset E$ which is of density (exactly) 1 in a subprogression $P' \subset [1, N]$ of size at least $N^{\frac{1}{3^s}}$. Namely, either at some point we have many 3-term arithmetic progressions, or after finitely many steps we find an arithmetic progression of size $N^{\frac{1}{3^s}}$ in E; if N is sufficiently large then $N^{\frac{1}{3^s}} \geq 3$.

We remark that a more careful analysis allows one to have the density δ depend on N in the form $\delta = 1/(\log\log N)^t$ [3].

The main issue is, of course, the second step in this argument - namely, obtaining increased density on a large subprogression. This can be achieved via discrete Fourier analysis - one considers E as a subset of $\mathbb{Z}_N = \mathbb{Z}/N\mathbb{Z}$. Denoting 1_E the characteristic function of E, one shows that if E does not contain roughly the expected number of 3-term progressions, then the function $1_E - \delta$ has a large non trivial Fourier coefficient, namely, there exist an integer r such that

$$\left| \frac{1}{N} \sum_{x \in \mathbb{Z}_N} (1_E - \delta)(x) e^{2\pi i x \frac{r}{N}} \right| \geq c(\delta).$$

Using equidistribution properties of the sequence $\{x\frac{r}{N}\}$ mod 1, one finds a large subprogression Q - of size $N^{\frac{1}{3}}$ - on which $x\frac{r}{N}$ is roughly constant. This in turn can be translated into an increased density of at least $\delta + c(\delta)$ on (many) translates of Q. This type of argument is referred to nowadays as a *density increment argument*.

Generalizing Roth's theorem to k-term progressions for $k > 3$ turned out to be very difficult, and was shown by Szemerédi in his famous theorem [44]:

Theorem 2.2 (Szemerédi 1975). *Let E be a set of positive upper density, then E contains a non trivial k-term arithmetic progression.*

By now there are many proofs of Szemerédi's theorem. In this paper we will focus on two of them: Furstenberg's ergodic theoretic proof, which marked the beginning of the ergodic theoretic side of our story, and Gowers's proof, which pioneered the application of tools from additive combinatorics to the study of arithmetic progressions.

3. Furstenberg's proof of Szemerédi's theorem

Shortly after Szemerédi proved the theorem on arithmetic progressions in sets of positive upper density in the integers, Furstenberg gave an ergodic theoretic proof of Szemerédi's theorem [15]. The ideas behind this proof initiated a new field in ergodic theory, referred to as *ergodic Ramsey theory*, and are the foundation of all subsequent ergodic theoretic developments on which the story in our paper is based.

Furstenberg first observed that one can translate questions about patterns in subsets of positive density in the integers to return time questions for sets of positive measure in a measure preserving system. More precisely:

[3] The state of the art in the question of 3-term progressions is the recent result of T. Sanders stating that one can have the density as small as $\delta = 1/\log N^{1-o(1)}$ [41].

Theorem 3.1 (Furstenberg correspondence principle). *Let $\delta > 0$, and let $E \subset \mathbb{N}$ be a set with positive upper density*[4]*. There exists a probability measure preserving system*[5] *$(X, \mathcal{B}, \mu, T)$, and a measurable set A with $\mu(A) > 0$, such that the following holds: if for some integers $n_1, \ldots, n_k$*

$$\mu(A \cap T^{-n_1} A \cap \ldots \cap T^{-n_k} A) > 0,$$

then

$$\bar{d}(E \cap (E - n_1) \cap \ldots \cap (E - n_k)) > 0.$$

In particular, there exists an integer x such that $x, x + n_1, \ldots, x + n_k \in E$.

It follows that if we seek a $k+1$ term arithmetic progression in E, it suffices to show that for any probability measure preserving system $(X, \mathcal{B}, \mu, T)$, and any A with $\mu(A) > 0$, there is a positive integer n with $\mu(A \cap T^{-n} A \cap \ldots \cap T^{-kn} A) > 0$. Observe that the case $k = 1$ is the famous Poincaré recurrence theorem. Indeed, Furstenberg proves the following theorem:

Theorem 3.2 (Furstenberg multiple recurrence theorem). *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system, and let A be with $\mu(A) > 0$. Then for any $k > 0$*

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{n \leq N} \mu(A \cap T^{-n} A \cap \ldots \cap T^{-kn} A) > 0. \tag{3.1}$$

On first impression, it might seem that in replacing the arbitrary set E of positive density with an arbitrary set A of positive measure, our situation is not much improved. However, in the ergodic theoretic context one might hope to prove and apply useful structure theorems. In the case at hand - the averages (3.1) are studied via morphism to more structured measure preserving systems, as we will try to demonstrate below.

We will henceforth assume that the system $\mathbf{X}$ is *ergodic*, namely any T-invariant set is of measure either 0 or 1. Any system can be decomposed to its ergodic components, thus we lose no generality in Theorem 3.2 by making this assumption.

We first briefly discuss Furstenberg's ergodic theoretic proof of Roth's theorem on 3-term progressions. We wish to evaluate the average

$$\frac{1}{N} \sum_{n \leq N} \mu(A \cap T^{-n} A \cap T^{-2n} A) = \frac{1}{N} \sum_{n \leq N} \int 1_A(x) 1_A(T^n x) 1_A(T^{2n} x) d\mu$$

where $1_A(x)$ is the characteristic function of A. Furstenberg proves that there exists a measure preserving system $\mathbf{Z} = (Z, \mathcal{B}_Z, \mu_Z, T_Z)$ that is a *Kronecker system*[6], and a morphism[7]

[4]Furstenberg's correspondence principle as well as his multiple recurrence theorem hold in the more general context when one considers the *upper Banach density* of the set E,

$$d^*(E) = \limsup_{N - M \to \infty} \frac{|E \cap [M, N-1]|}{N - M}.$$

We will keep to the upper density for simplicity.

[5]A probability *measure preserving system* $\mathbf{X} = (X, \mathcal{B}, \mu, T)$ consists of a probability space $(X, \mathcal{B}, \mu)$ and an invertible measurable map $T : X \to X$ with $T_* \mu = \mu$.

[6]A Kronecker system $\mathbf{Z} = (Z, \mathcal{B}_Z, \mu_Z, T_Z)$ is a system where Z is a compact Abelian group, $\mathcal{B}_Z$ the Borel σ-algebra, μ_Z the Haar measure, and T_Z is a rotation $T_Z(x) = x + \alpha$ for some $\alpha \in Z$

[7]A morphism between measure preserving systems $\mathbf{X}, \mathbf{Y}$ is a measure preserving map between the corresponding measure spaces that intertwines the actions of T_X, T_Y. In this case $\mathbf{Y}$ is called a *factor* of $\mathbf{X}$.

$\pi : \mathbf{X} \to \mathbf{Z}$ such that for any $f_0, f_1, f_2 \in L^\infty(X)$,

$$\frac{1}{N}\sum_{n\leq N}\int f_0(x)f_1(T^n x)f_2(T^{2n}x)d\mu$$

is asymptotically the same as

$$\frac{1}{N}\sum_{n\leq N}\int \pi_* f_0(z)\pi_* f_1(T_Z^n z)\pi_* f_2(T_Z^{2n}z)d\mu_Z.$$

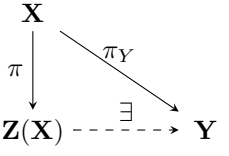
That is, rather than trying to evaluate the average in an arbitrary (ergodic) system, we need to evaluate it in a very special system - a compact abelian group rotation system: we are left with evaluating

$$\lim \frac{1}{N}\sum_{n\leq N}\int \pi_* 1_A(z)\pi_* 1_A(z+n\alpha)\pi_* 1_A(z+2n\alpha)d\mu_Z.$$

Via Fourier analysis the above limit is easily seen to equal

$$\int \pi_* 1_A(z)\pi_* 1_A(z+b)\pi_* 1_A(z+2b)d\mu_Z(z)d\mu_Z(b).$$

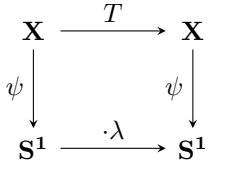
Now the projection π_* is a positive operator, namely if $f \geq 0$ then $\pi_* f \geq 0$. It follows that $\pi_* 1_A \geq 0$, and since $\int \pi_* 1_A d\mu_Z = \int 1_A d\mu = \mu(A) > 0$, the above average is clearly positive. The system $\mathbf{Z} = \mathbf{Z}(\mathbf{X})$ is called the *Kronecker factor* of $\mathbf{X}$ and satisfies the following universal property. If $\mathbf{Y}$ is a Kronecker system that is a factor of X and $\pi_Y : \mathbf{X} \to \mathbf{Y}$ the factor map, then π_Y factors through $\mathbf{Z}(\mathbf{X})$ as demonstrated in the diagram below:



The factor $\mathbf{Z}(\mathbf{X})$ is constructed via the eigenfunctions of $\mathbf{X}$. Let us demonstrate why a non trivial eigenfunction implies the existence of a non-trivial circle rotation factor. Let ψ be an eigenfunction of $\mathbf{X}$,

$$\psi(Tx) = \lambda\psi(x).$$

The function $|\psi|$ is a T-invariant function, and by ergodicity $|\psi|$ is constant a.e. Thus we can normalize ψ to take values in the unit circle. Any normalized eigenfunction gives rise to a morphism to a circle rotation system $\psi : \mathbf{X} \to (S^1, \mathrm{Borel}, \mathrm{Haar}, \cdot\lambda)$:



The factor $\mathbf{Z}(\mathbf{X})$ would then be the image of the map $(\psi_i) : X \to (S^1)^{\mathbb{N}}$ given by $x \to (\psi_i(x))$, where $\{\psi_i\}$ is the collection of normalized eigenfunctions[8] of $\mathbf{X}$.

If $\mathbf{X}$ has no non-trivial eigenfunctions, then $\mathbf{Z}(\mathbf{X})$ is trivial (a point system), and thus $\pi_* f = \int f \, d\mu$. In this case $\mathbf{X}$ is called *weakly mixing*. We then have

$$\frac{1}{N}\sum_{n=1}^{N}\int f(x)f(T^n x)f(T^{2n}x)\,d\mu \to \left(\int f\,d\mu\right)^3,$$

and we can thus think of the points $x, T^n x, T^{2n}x$ as asymptotically independent on average. The content of Furstenberg's argument is then that if $x, T^n x, T^{2n}x$ are *not* asymptotically independent on average, then the obstruction lies in an Abelian group rotation factor. We remark that it is clear that an Abelian group rotation factor is an obstruction as in Abelian groups $z + 2n\alpha$ is determined by $z, z + n\alpha$.

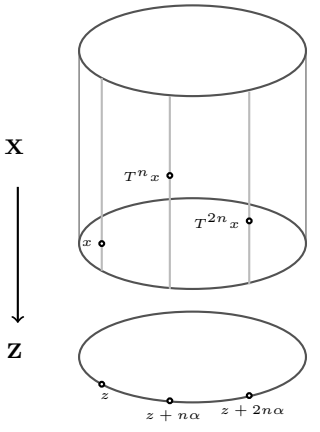


Figure 3.1. The points $x, T^n x, T^{2n}x$ are independent (asymptotically on average) in the fibers over the maximal Abelian group rotation factor.

To summarize, Furstenberg's proof of Roth's Theorem is based on the following dichotomy:

- either $\mathbf{X}$ is weakly mixing, or
- there is a morphism from $\mathbf{X}$ to a non trivial group rotation system.

The above argument motivates the following definition ([16]):

Definition 3.3 (k-characteristic factor). Let $\mathbf{Y}$ be a factor of $\mathbf{X}$, and let $\pi : \mathbf{X} \to \mathbf{Y}$ be the factor map. We say that $\mathbf{Y}$ is *k-characteristic* if

$$\frac{1}{N}\sum_{n=1}^{N}\int f_0(x)f_1(T_{\mathbf{X}}^n x)\dots f_k(T_{\mathbf{X}}^{kn}x)d\mu_{\mathbf{X}}$$

[8]We implicitly assume that the system $\mathbf{X}$ is separable and thus has at most countably many normalized eigenfunctions.

is asymptotically the same as

$$\frac{1}{N}\sum_{n=1}^{N}\int \pi_* f_0(y)\pi_* f_1(T_{\mathbf{Y}}^n y)\dots\pi_* f_k(T_{\mathbf{Y}}^{kn} y)d\mu_{\mathbf{Y}}.$$

We make the following observations:

- The system $\mathbf{X}$ itself is k-characteristic for all k.
- The trivial system is 1-characteristic. In this case $\pi_* f(x) = \int f(x)d\mu_{\mathbf{X}}$, and by the *mean ergodic theorem*

$$\frac{1}{N}\sum_{n=1}^{N}\int f(x)f(T_{\mathbf{X}}^n x)d\mu_{\mathbf{X}} \sim \left(\int f d\mu_{\mathbf{X}}\right)^2.$$

- The Kronecker factor $\mathbf{Z}(\mathbf{X})$ is 2-characteristic (Furstenberg [15]).

The Furstenberg-Zimmer structure theorem [15, 55] relativizes the dichotomy between weak mixing and an abelian rotation factor. One can show, using spectral theory, that $\mathbf{X}$ being weakly mixing is equivalent to the product system with the diagonal action $\mathbf{X}\times\mathbf{X}$ being ergodic. One can relativize this notion as follows. Let $\mathbf{X}\times_{\mathbf{Y}}\mathbf{X}$ be the fiber product over $\mathbf{Y}$. Say that $\pi:\mathbf{X}\to\mathbf{Y}$ is a *relatively weak mixing extension* if the map $\pi\times_{\mathbf{Y}}\pi:\mathbf{X}\times_{\mathbf{Y}}\mathbf{X}\to\mathbf{Y}$ is relatively ergodic, namely any $T\times T$ invariant subset in $\mathbf{X}\times_{\mathbf{Y}}\mathbf{X}$ is lifted from $\mathbf{Y}$ via the map $\pi\times_{\mathbf{Y}}\pi$. The role of the compact abelian group rotation is replaced by the notion of an isometric extension. Say that $\pi:\mathbf{X}\to\mathbf{Y}$ is an *isometric extension* if $\mathbf{X}=\mathbf{Y}\times_\sigma\mathbf{M}$ where $\mathbf{M}=(M,\mathcal{B}_{\mathbf{M}},\mu_{\mathbf{M}})$ with M a compact metric space, $\mathcal{B}_{\mathbf{M}}$ the Borel σ-algebra and μ_M the probability measure invariant under the the action of the isometry group of M, $T_{\mathbf{X}}(y,m)=(T_{\mathbf{Y}}y,\sigma(y)m)$, where σ is a (measurable) map from Y to the isometry group of M, and $\mu_{\mathbf{X}}=\mu_{\mathbf{Y}}\times\mu_M$.

Theorem 3.4 (Furstenberg-Zimmer structure theorem [15, 55]). *There exists a sequence of factors*

$$\mathbf{X}\to\ldots\to\mathbf{Z}_k(\mathbf{X})\to\mathbf{Z}_{k-1}(\mathbf{X})\to\ldots\to\mathbf{Z}_1(\mathbf{X})\to\star$$

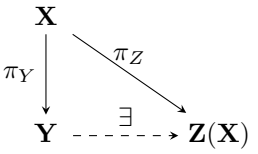
such that for each k, either $\mathbf{X}\to\mathbf{Z}_k(\mathbf{X})$ is relatively weakly mixing, or there is a morphism from $\mathbf{X}$ to a non trivial isometric extension of $\mathbf{Z}_k(\mathbf{X})$.

Theorem 3.5 (Furstenberg [15]). *The factors $\mathbf{Z}_k(\mathbf{X})$ are $(k+1)$-characteristic.*

Observe that the factor $\mathbf{Z}_0(\mathbf{X})$ is the trivial factor and the factor $\mathbf{Z}_1(\mathbf{X})$ is the Kronecker factor. With the above structure theorem at hand it then suffices to prove the multiple recurrence theorem for systems which are towers of isometric extensions. Furstenberg utilizes this structure to show multiple recurrence - the idea being that if the multiple recurrence property holds for any k for a system $\mathbf{Y}$, and $\mathbf{X}$ is an isometric extension of $\mathbf{Y}$, then multiple recurrence holds for $\mathbf{X}$ as well.

4. Obstructions to 4-term progressions

The Kronecker factor $\mathbf{Z}_1(\mathbf{X})=\mathbf{Z}(\mathbf{X})$ is also a *universal 2-characteristic factor* : it satisfies the property that if $\mathbf{Y}$ is any 2-characteristic factor and $\pi_Y:\mathbf{X}\to\mathbf{Y}$ the factor map, then the factor map $\pi_Z:\mathbf{X}\to\mathbf{Z}(\mathbf{X})$ factors through $\mathbf{Y}$ as demonstrated in the diagram below:



The factors $\mathbf{Z}_k(\mathbf{X})$ that were constructed by Furstenberg are *not* universal $(k+1)$-characteristic for $k > 1$. This raises the following natural problem: classify the universal $(k+1)$-characteristic factors $Z_k(X)$. In other words, we try to understand the exact obstructions on the points $x, T^n x, \ldots, T^{(k+1)n} x$ preventing them from moving about freely in X.

For the case $k = 1$, the upshot of the discussion regarding Furstenberg's proof of Roth's theorem on 3-term progressions in the previous section was that the only obstructions to the independence (asymptotically on average) of $x, T^n x, T^{2n} x$ come from a compact abelian group rotation factor, associated to the non trivial eigenfunctions of $\mathbf{X}$. Already in the case $k = 2$ (corresponding to 4-term progressions) we have new obstructions. Consider for example the system

$$\mathbf{Y} = (\mathbb{T} \times \mathbb{T}, \text{Borel}, \text{Haar}, T_{\mathbf{Y}})$$

where

$$T_{\mathbf{Y}} y = T_{\mathbf{Y}}(z, w) = (z + \alpha, w + 2z + \alpha),$$

where α is irrational. Iterating S we obtain

$$T^n_{\mathbf{Y}} y = T^n_{\mathbf{Y}}(z, w) = (z + n\alpha, w + 2nz + n^2\alpha).$$

We now observe that

$$y = 3T^n_{\mathbf{Y}} y - 3T^{2n}_{\mathbf{Y}} y + T^{3n}_{\mathbf{Y}} y$$

Namely, the point y is determined by the three points $T^n_{\mathbf{Y}} y, T^{2n}_{\mathbf{Y}} y, T^{3n}_{\mathbf{Y}} y$.

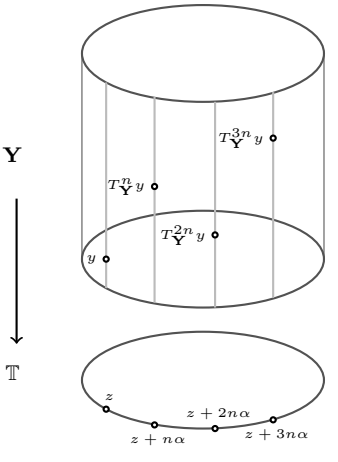


Figure 4.1. The points $y, T^n_{\mathbf{Y}} y, T^{2n}_{\mathbf{Y}} y, T^{3n}_{\mathbf{Y}} y$ are not independent in the fibers over $\mathbb{T}$.

If there is a morphism $\mathbf{X} \to \mathbf{Y}$, these new obstructions to the (asymptotic on average) independence of the points $x, T_{\mathbf{X}}^n x, T_{\mathbf{X}}^{2n} x, T_{\mathbf{X}}^{3n} x$ will surface. Another way to see the obstructions coming from the system $\mathbf{Y}$ is by observing that the system $\mathbf{Y}$ exhibits *second order eigenfunctions*, namely functions ϕ satisfying $\phi(T_{\mathbf{Y}} y) = \psi(y)\phi(y)$ where ψ is an ordinary (first order) eigenfunction; for example the function $\phi(y) = \phi(z, w) = e^{2\pi i w}$ is a second order eigenfunction. Any second order eigenfunction satisfies

$$\phi(y) = \phi^3(T_{\mathbf{Y}}^n y)\phi^{-3}(T_{\mathbf{Y}}^{2n} y)\phi(T_{\mathbf{Y}}^{3n} y)$$

Thus choosing $f_0 = \phi^{-1}$, $f_1 = \phi^3$, $f_2 = \phi^{-3}$, and $f_3 = \phi$ we see that

$$\begin{aligned} 1 &= \int f_0(x) f_1(T_{\mathbf{Y}}^n y) f_2(T_{\mathbf{Y}}^{2n} y) f_3(T_{\mathbf{Y}}^{3n} x) dm \\ &= \frac{1}{N} \sum_{n \le N} \int f_0(x) f_1(T_{\mathbf{Y}}^n y) f_2(T_{\mathbf{Y}}^{2n} y) f_3(T_{\mathbf{Y}}^{3n} x) dm. \end{aligned}$$

On the other hand one can verify that a (non trivial) 2nd order eigenfunction ϕ (and its powers) is orthogonal to ordinary eigenfunctions, thus for any $i = 0, 1, 2, 3$ the projection of the function f_i on the Kronecker factor is 0.

It turns out, however, that second order eigenfunctions are not the only obstructions. Consider the Heisenberg nilsystem: the phase space Y is the Heisenberg nilmanifold

$$Y = N/\Gamma = \begin{pmatrix} 1 & \mathbb{R} & \mathbb{R} \\ 0 & 1 & \mathbb{R} \\ 0 & 0 & 1 \end{pmatrix} \Big/ \begin{pmatrix} 1 & \mathbb{Z} & \mathbb{Z} \\ 0 & 1 & \mathbb{Z} \\ 0 & 0 & 1 \end{pmatrix}$$

equipped with the Borel σ-algebra and the Haar measure, and the transformation $T_{\mathbf{Y}}$ given by $T_{\mathbf{Y}} g\Gamma = ag\Gamma$, where

$$a = \begin{pmatrix} 1 & \alpha & 0 \\ 0 & 1 & \beta \\ 0 & 0 & 1 \end{pmatrix}.$$

Topologically Y is a circle bundle over a two dimensional torus. This system shares with the system in the above example the property that the point $g\Gamma$ is determined by $a^n g\Gamma, a^{2n} g\Gamma$, $a^{3n} g\Gamma$. However this dependence can not be described by a simple equation as in the previous example. Moreover, $\mathbf{Y}$ has *no* non-trivial second order eigenfunctions[9].

The Heisenberg nilsystem is a special case of the following system:

$$\mathbf{Y} = (N/\Gamma, \text{Borel}, \text{Haar}, T_{\mathbf{Y}}),$$

where N/Γ a 2-step nilmanifold, and

$$T_{\mathbf{Y}} : g\Gamma \to ag\Gamma \qquad a \in N.$$

The system $\mathbf{Y}$ is called a 2-step nilsystem. It turns out that we need not look for further obstructions in the case $k = 2$ - all obstructions to 4-term progressions come from 2-step pro-nilsystems - inverse limits of 2-step nilsystems [10–12, 16]:

Theorem 4.1 (Conze-Lesigne, Furstenberg-Weiss). *Let* $\mathbf{X}$ *be an ergodic measure preserving system. There exists a* 2*-step pro-nilsystem* $\mathbf{Y}$ *and a morphism* $\pi : \mathbf{X} \to \mathbf{Y}$ *such that* $\mathbf{Y}$ *is*

[9]The easiest way to see this is via equidistribution properties of polynomial orbits on nilmanifolds [34].

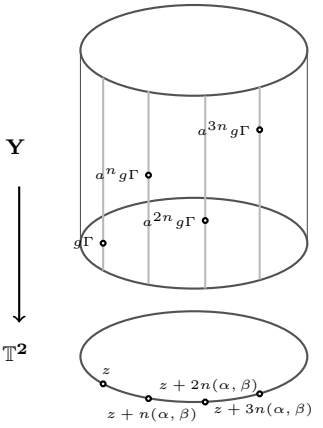


Figure 4.2. The points $g\Gamma, a^n g\Gamma, a^{2n} g\Gamma, a^{3n} g\Gamma$ are not independent in the fibers over the two dimensional torus.

the universal 3*-characteristic factor of* $\mathbf{X}$*, namely*

$$\frac{1}{N}\sum_{n=1}^{N}\int f(x)f(T_{\mathbf{X}}^{n}x)f(T_{\mathbf{X}}^{2n}x)f(T_{\mathbf{X}}^{3n}x)d\mu_{\mathbf{X}}$$

is asymptotically the same as

$$\frac{1}{N}\sum_{n=1}^{N}\int \pi_* f(y)\pi_* f(T_{\mathbf{Y}}^{n}y)\pi_* f(T_{\mathbf{Y}}^{2n}y)\pi_* f(T_{\mathbf{Y}}^{3n}y)d\mu_{\mathbf{Y}}.$$

We can now prove Szemerédi's theorem for 4-term progressions by verifying that in a 2-step nilsystem the above limit is positive (when $f = 1_A$).

Let us say a few words about the proof. By Furstenberg's structure theorem it is sufficient to study systems $\mathbf{X}$ of the form $\mathbf{Z} \times_\sigma \mathbf{M}$ where $\mathbf{Z} = \mathbf{Z}_1(\mathbf{X})$ is the Kronecker factor, $T_{\mathbf{Z}}(z) = z + \alpha$, and M is a compact metric space and $\sigma : Z \to \mathrm{ISO}(M)$. It is then shown that one can further reduce to the case where M is a compact abelian group and $\sigma : Z \to M$ satisfies a functional equation now called the Conze-Lesigne equation: for all b, a.e z

$$\sigma(z+b) - \sigma(z) = c(b) + F_b(z+\alpha) - F_b(z). \tag{4.1}$$

Describing how one can solve the above equation is beyond the scope of this paper, but let us hint how this equation is related to nilpotency. Consider the group

$$G = \{(b, f) : b \in Z, f : Z \to M \text{ measurable}\}$$

with the action

$$(b, f) * (c, g) = (b + c, f^c \cdot g)$$

where $f^c(z) = f(z+c)$. Then condition (4.1) can be interpreted as the fact that $[(\alpha, \sigma), (b, F_b)]$ is in the center of G, which hints at 2-step nilpotent behavior.

We mention another observation regarding equation (4.1). Upon examination one sees that

$$c(b_1 + b_2) - c(b_1) - c(b_2)$$

is an eigenvalue of $T_{\mathbf{Z}}$, and using the fact that there are only countably many of those, one can modify $c(b), F_b(z)$ so that $c(b)$ is linear in b in a neighborhood of zero in Z. A similar feature will surface in the combinatorial analysis described in section 8 below, devoted to the Inverse Theorem for the Gowers norms, which is why we mention it here.

5. Gowers proof of Szemerédi's Theorem

The next advancement (chronologically) was in the combinatorial front. Gowers gave a new proof for Szemerédi's theorem [27]. His proof is a generalization of Roth's argument to arbitrarily long arithmetic progressions using an ingenious combination of discrete Fourier analysis and additive combinatorics; in particular Gowers obtains a Roth type bound for the density of the form $1/(\log\log N)^{c(k)}$ for some constant depending on k - the length of the progression.

We first fix some notation. We denote by $[N]$ the interval $[1, N]$. For a finite set E we denote by $\mathbb{E}_{x\in E} f(x)$ the average $\frac{1}{|E|}\sum_{x\in E} f(x)$. For two functions $f, g : [N] \to \mathbb{C}$ we write $f(x) \ll g(x)$ if $|f(x)| \leq Cg(x)$ for some constant C independent of N, and we write $f(x) \ll_A g(x)$ if $|f(x)| \leq C(A)g(x)$ for some constant $C(A)$ independent of N.

In the course of the proof Gowers defines the following norms which play a very important role in further developments.

Definition 5.1 (Gowers norms). Let $f : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$. For $h \in \mathbb{Z}/N\mathbb{Z}$ define the discrete derivative in direction h

$$\Delta_h f(x) = f(x + h)\overline{f(x)}$$

We define the k-th Gowers uniformity norms U_k on $\mathbb{C}^N$ by

$$\|f\|^{2^k}_{U^k[N]} = \mathbb{E}_{x,h_1,\dots h_k\in[N]}\Delta_{h_1}\dots\Delta_{h_k} f(x)$$

Remark 5.2. One can define the Gowers norms on any abelian group; of special interest is the group $\mathbb{F}_2^n$ where the Gowers norms are intimately related to polynomial testing.

We make a few initial observations. For 1-bounded functions f ($\|f\|_\infty \leq 1$)

- $\|f\|_{U^k[N]} = 1$ if an only if $f(x) = e^{2\pi i q(x)}$ where q is a polynomial of degree $< k$.
- By repeated application of the Cauchy-Schwarz inequality, if f correlates with $e^{2\pi i q(x)}$ where q is a polynomial of degree $< k$ then f has large Gowers norms; namely

$$|\mathbb{E}_{x\in[N]} f(x)e^{-2\pi i q(x)}| > \delta \implies \|f\|_{U^k[N]} \gg_\delta 1.$$

- If f is a random function taking the values ± 1 with probability $1/2$ for any $x \in [N]$, then by the law of large numbers, $\|f\|_{U^k[N]} = o(1)$.

The Gowers uniformity norms play an important role in the study of arithmetic progressions. If f and g are close in the U_k norm, i.e $\|f-g\|_{U^k[N]}$ is small, then they have approximately the same number of $k+1$ term progressions. Denote by $AP_k(f)$ the number of $(k+1)$-term progressions in f: denote

$$AP_k(f) = \mathbb{E}_{x,d\in[N]} f(x)f(x+d)\dots f(x+kd).$$

Then

$$|AP_k(f) - AP_k(g)| \ll_k \|f-g\|_{U^k[N]}. \tag{5.1}$$

In fact a more general statement regarding linear forms is true:

Proposition 5.3. *Let $f_1,\dots,f_k$ be 1-bounded functions. Let $L_1(\vec{x}),\dots,L_m(\vec{x})$ be k affine linear forms in d variables with integer coefficients: $L_i(\vec{x}) = \sum_{j=1}^{d} l_{ij}x_j + b_i$, no two of which are affinely dependent. Then there exists $k>0$ such that*

$$|\mathbb{E}_{\vec{x}\in[N]^d} f_1(L_1(\vec{x}))\cdots f_k(L_m(\vec{x}))| \ll \min_j \|f_j\|_{U^k[N]}.$$

The proposition is proved via repeated applications of the Cauchy-Schwarz inequality, and this is where the Gowers norms enter the picture in the proof of Theorem 1.1; it is the source of the condition that no two forms are affinely dependent.

The strategy of Gowers is similar in spirit to that of Roth. The idea is as follows. Let $E\subset[N]$ be with $|E|=\eta N$. Then

- either the number of $(k+1)$-term progressions is more than half of that expected in random set, namely $\geq \eta^{k+1}N^2/2$, or
- $\|1_E-\eta\|_{U^k[N]} \gg_\eta 1$.

In order to proceed one needs to understand the condition $\|1_E-\eta\|_{U^k[N]} \gg_\eta 1$. For $k=2$ we observe that

$$\|f\|_{U^2[N]}^4 = \|\hat{f}\|_4^4 \leq \|\hat{f}\|_2^2\|\hat{f}\|_\infty^2.$$

Thus if $\|f\|_2\leq 1$ then we find that $\|f\|_{U^2[N]}\geq\eta$ implies $\|\hat{f}\|_\infty\geq\eta^2$. This implies that f has a large Fourier coefficient, namely

$$|\mathbb{E}_{x\in[N]} f(x)e(x\alpha)| \geq \eta^2.$$

For larger k the situation is much more complicated. Gowers proves the following local inverse theorem for higher Gowers norms.

Theorem 5.4 (Local inverse theorem for Gowers norms). *Let $f:\mathbb{Z}/N\mathbb{Z}\to\mathbb{C}$ be with $|f|\leq 1$. Then*

$$\|f\|_{U^k[N]}\geq\delta \implies |\mathbb{E}_{x\in P} f(x)e^{2\pi i q(x)}| \gg_\delta 1,$$

where P is a progression of length at least N^t, $q(x)$ is a polynomial of degree $k-1$, and t depends[10] on k,δ.

[10] In fact Gowers shows that one can find many such progressions: one can partition $\mathbb{Z}/N\mathbb{Z}$ into progressions $P_1,\dots,P_M$ of average length greater than N^t, such that $\sum_{i=1}^{M}|\sum_{x\in P_i} f(x)e^{2\pi i q(x)}| \gg_\delta N$.

The word 'local' in this context refers to the fact that the correlation in the above theorem is obtained not on the full interval $[N]$ but rather on a short progression of length at least N^t with $t < 1$ (for $k > 2$). This theorem provides sufficient structure to obtain increased density on a subprogression of length at least N^s: we apply Theorem 5.4 to the function $1_E - \eta$, and use the equidistribution properties of the sequence $\{q(x)\}$ mod 1 to find an arithmetic progression of length at least N^s ($s < t$) on which $\{q(x)\}$mod 1 is roughly constant.

6. Classification of universal k-characteristic factors

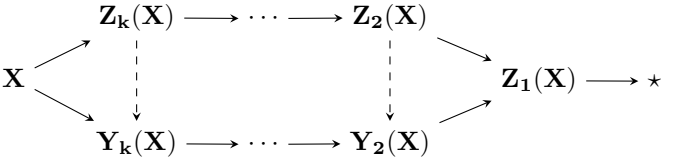
We return now to the question of classifying k-characteristic factors. Recall that we are interested in the averages

$$\frac{1}{N}\sum_{n\leq N}\int f(x)f(T^n x)f(T^{2n}x)\dots f(T^{kn}x)d\mu. \tag{6.1}$$

The universal 4-characteristic factors were classified by Host and Kra [33], and independently in the author's PhD thesis [52], and were shown to be 3-step pro-nilsystems. Both methods were extended to work for general k - by Host and Kra in [32], and by the author in [54].

Theorem 6.1 (Host-Kra (05), Z (07)). *Let* $\mathbf{X}$ *be an ergodic measure preserving system. The universal* k*-characteristic factor* $\mathbf{Y_k}(\mathbf{X})$ *is a* $(k-1)$*-step pro-nilsystem.*

We have the following diagram displaying the relation between the factors $\mathbf{Z_k}(\mathbf{X})$ defined by Furstenberg in his proof of Szemerédi's theorem and the pro-nilfactors $\mathbf{Y_k}(\mathbf{X})$ which are the universal characteristic factors:



As a corollary of this structure theorem one can calculate the asymptotic formula for the averages in (6.1) via a limit formula for the corresponding averages on nilsystems [53].

Theorem 6.2 (Z (05)). *Let* $\mathbf{X}$ *be a* $(k-1)$*-step nilsystem. Then*

$$\lim \frac{1}{N}\sum_{n\leq N}\int f_0(x)f_1(T^n x)\dots f_k(T^{kn}x)d\mu = \int f_0(x_0)f_1(x_1)\dots f_k(x_k)dm_H$$

where m_H *is the Haar measure on the subnilmanifold* $H\Gamma^{k+1}/\Gamma^{k+1} \subset X^{k+1} = (G/\Gamma)^{k+1}$*, where* H *is the subgroup*

$$\left\{\left(g_0, g_0g_1, g_0g_1^2g_2, g_0g_1^3g_2^3g_3, \dots, g_0g_1^k g_2^{\binom{k}{2}}\dots g_{k-2}^{\binom{k}{k-2}}\right) : g_i \in G_i\right\}$$

where $\{1\}=G_{k-1}\subset G_{k-2}\subset\dots\subset G_1=G_0=G$ *is the derived series, i.e.* $G_{i+1} = [G_i, G]$.

One can now prove Szemerédi's theorem by showing that the above limit is positive if $f_i = 1_A$ for $i = 0, \dots, k$. This approach to proving Szemerédi's theorem (and various generalizations) was taken in [6].

The proof in [54] generalizes the methods in [10–12]. Inductively, one is led to the problem of solving a functional equation similar in nature to equation (4.1), only the extension cocycles are now defined on a (pro)-nilmanifolds (rather than a compact abelian group). Such cocycles are in general much more difficult to handle, but one can still use the fact that orbits on products of nilmanifolds are well understood and have a nice algebraic nature (as one can see in Theorem 6.2 above).

The proof in [32] introduces seminorms, which are similar, at least semantically, to the Gowers uniformity norms[11]

Definition 6.3 (Host-Kra-Gowers semi-norms).

$$\|f\|_{U^k(\mathbf{X})}^{2^k} := \lim_{N\to\infty} \mathbb{E}_{h_1,\dots,h_k\in[N]} \int \Delta_{h_1} \dots \Delta_{h_k} f(x) d\mu(x).$$

It is then proved that characteristic factors for averages associated with the ergodic U_k semi-norms defined above are also pro-nilsystems. Or, in a different formulation:

$$\|f\|_{U^{k+1}(\mathbf{X})} > 0 \implies \pi : \mathbf{X} \to k\text{-step nilsystem}, \quad \pi_* f \neq 0.$$

This suggests a far reaching generalization of the Gowers local inverse theorem, which we will discuss in section 8 below.

7. The Green-Tao theorem from a characteristic factor point of view

In their famous paper, Green and Tao prove a Szemerédi's type theorem in the prime numbers [19]:

Theorem 7.1 (Green-Tao (05)). *Let $E \subset \mathbb{P}$ of positive relative density, then E contains long arithmetic progressions.*

We present the idea of the proof from a characteristic factor point of view. Our starting point will be the following version of Szemerédi's theorem: Let $f : [N] \to [0,1]$ be a function with $|\mathbb{E}_{n\in[N]} f(x)| > \delta$. Then for any integer $k > 0$

$$AP_k(f) = \mathbb{E}_{x,d\in[N]} f(x) f(x+d) \dots f(x+kd) \geq c(\delta) + o(1) \tag{7.1}$$

where $c(\delta) > 0$, and is independent of N.

If we naively try to apply this theorem for a subset E of the prime numbers of relative density δ, we run into an obvious problem that $\mathbb{E}_{x\in[N]} 1_E(x) = o(1)$. We can try to fix this problem by putting a weight on each prime - we consider the von-Mangoldt function $\Lambda(x)$ which takes the value $\log p$ if x is a positive power of p and 0 otherwise. In this case we will have

$$\mathbb{E}_{x\in[N]} \Lambda(x) 1_E(x) = \delta + o(1),$$

[11] Such averages as the one below were studied in the case $k = 2$ already by Bergelson in [4].

for some constant $\delta > 0$ (independent of N). But now we face the problem that the function $\tilde{1}_E(x) = \Lambda(x)1_E(x)$ does not take values in $[0,1]$; in fact the function $\tilde{1}_E(x)$ is unbounded. Green and Tao show that for a certain class of unbounded functions (functions bounded by a k-psuedorandom function) one can find a "k-characteristic factor" for the average (7.1) generated by *bounded* functions ! We can summarize the procedure as follows:

- Introduce combinatorial notions of (approximate) *factor* and *projection* onto a factor.
- Find a convenient combinatorial "k-characteristic factor" for averages associated with the U_k norms, in this case a factor of functions bounded by a constant $C(k)$ (depending only on k).
- Let $\pi_*(\tilde{1}_E)$ be the (approximate) projection on the factor. Then $0 \leq \pi_*(\tilde{1}_E) \leq C(k)$, the average of the function $\pi_*\tilde{1}_E$ is approximately the same as that of $\tilde{1}_E$, namely approximately δ, and $\|\tilde{1}_E - \pi_*(\tilde{1}_E)\|_{U^k[N]}$[12] is small. A version of the Gowers-Cauchy-Schwarz inequality (for functions bounded by k-pseudoradnom functions) gives then, as in (5.1), that

$$|AP_k(\tilde{1}_E) - AP_k(\pi_*\tilde{1}_E)| \ll \|\tilde{1}_E - \pi_*\tilde{1}_E\|_{U^k[N]} \tag{7.2}$$

- Apply Szemerédi's Theorem to the $C(k)$-bounded function $\pi_*\tilde{1}_E$, to obtain $AP_k(\pi_*\tilde{1}_E) \gg_\delta 1$, and thus $AP_k(\tilde{1}_E) \gg_\delta 1$

A different way to say this is that given $\epsilon > 0$ we can decompose

$$\tilde{1}_E = g + h \tag{7.3}$$

where g is a $C(k)$-bounded function, and h is a function with $\|h\|_{U^k[N]} < \epsilon$. This type of theorem is now referred to as a *decomposition* theorem. There is a very nice modern and more abstract treatment of general decomposition theorems in [28], and [38] using the Hahn-Banach theorem. We remark that Theorem 7.1 has since been extended to include polynomial configurations [46], and multidimensional configurations [13, 14, 48].

8. The Inverse Theorem for the Gowers Norms

The argument in the Green-Tao theorem is based on Szemerédi's theorem which is valid for *any* subset of positive density in the integers. This has two major caveats. The first is that it can not lead to an asymptotic formula for the number of arithmetic progressions, only a lower bound. The second is that it can not be used to study non homogeneous linear configurations, since there are counter examples within periodic sets of positive density. How then can we hope to get an asymptotic formula as in Theorem 1.1 ?

[12]We defined the Gowers norms for functions f on the group $\mathbb{Z}/N\mathbb{Z}$. We can define Gowers norms for functions $f : [N] \to \mathbb{C}$, setting $G := \mathbb{Z}/\tilde{N}\mathbb{Z}$ for some integer $\tilde{N} \geq 2^d N$, and defining a function $\tilde{f} : G \to \mathbb{C}$ by $\tilde{f}(x) = f(x)$ for $x = 1, \ldots, N$ and $\tilde{f}(x) = 0$ otherwise. We then set

$$\|f\|_{U^d[N]} := \|\tilde{f}\|_{U^d(G)} / \|1_{[N]}\|_{U^d(G)},$$

where $1_{[N]}$ is the indicator function of $[N]$. It is easy to see that this definition is independent of the choice of $\tilde{N}$.

We recall now that - in the ergodic theoretic context - to get a limit formula we needed to identify the universal characteristic factors. Motivated by theorem 6.1, Green and Tao conjectured in 2006 that the combinatorial "universal characteristic factors" for the U_k norm come from *nilsequences* - sequences arising in a natural way from nilsystems.

Conjecture 8.1 (Inverse conjecture for the Gowers norms ($\mathrm{GI}(s)$)). *Let $s \geq 0$ be an integer, and let $0 < \delta \leq 1$. Then there exists a finite collection $\mathcal{M}_{s,\delta}$ of s-step nilmanifolds G/Γ, each equipped with some smooth Riemannian metric $d_{G/\Gamma}$ as well as constants $C(s,\delta), c(s,\delta) > 0$ with the following property. Whenever $N \geq 1$ and $f : [N] \to \mathbb{C}$ is a 1-bounded function such that $\|f\|_{U^{s+1}[N]} \geq \delta$, there exists a nilmanifold $G/\Gamma \in \mathcal{M}_{s,\delta}$, some $g \in G$ and a function $F : G/\Gamma \to \mathbb{C}$ bounded in magnitude by 1 and with Lipschitz constant at most $C(s,\delta)$ with respect to the metric $d_{G/\Gamma}$, such that*

$$|\mathbb{E}_{n\in[N]} f(n)\overline{F(g^n x)}| \geq c(s,\delta). \tag{8.1}$$

That is, the global obstruction (scale N) to Gowers uniformity come from sequences arising from nilsystems. Recall that the local theorem for the Gowers norms shows that local obstructions (at scale N^t) to Gowers U^{s+1} uniformity norms come from phase polynomials of degree s. We remark that the converse to Conjecture 8.1 is true and relatively easy to prove via repeated applications of the Cauchy-Schwarz inequality. Namely, if (8.1) holds then $\|f\|_{U^{s+1}[N]} \gg_\delta 1$. We also mention that if δ is sufficiently close to 1 then the conjecture is true; moreover, f is close (in L^1) to a genuine (unique) phase polynomial [1], and thus correlates with a (unique) phase polynomial[13]; uniqueness allows one to try an intelligent guess. In the realm when $\delta > 0$, we cannot expect uniqueness, and as it turns out, we also can't expect correlation with a genuine phase polynomial.

One can ask a similar question in the context of finite field geometry. Given a function $f : \mathbb{F}_p^n \to \mathcal{D}$ with large Gowers norm (fixing p and letting n approach ∞), what can be said about f? It was conjectured that such functions would correlate with polynomial phase functions. More precisely:

Conjecture 8.2 (Inverse conjecture for the Gowers norms in finite fields). *Let p be a prime and let $f : \mathbb{F}_p^n \to \mathbb{C}$ be 1-bounded, with $\|f\|_{U^{s+1}[\mathbb{F}_p^n]} \geq \delta$. Then there exists a polynomial $P : \mathbb{F}_p^n \to \mathbb{F}_p$ of degree $\leq k$ such that*

$$|\mathbb{E}_{x\in\mathbb{F}_p^n} f(x) e^{2\pi i P(x)/p}| \geq c(s,\delta).$$

The case $s = 1$ of both conjectures follows from a short Fourier-analytic computation. The case $s = 2$ of Conjecture 8.1 was proved in [20]. The case $s = 2$ of Conjecture 8.2 was proved in [20] for odd p and for $p = 2$ in [40]. Surprisingly, Conjecture 8.2 turned out to be *false*; a counter example for the $U^4[\mathbb{F}_2^n]$ was constructed independently in [18, 35]. However, it turned out that with a small modification ofConjecture 8.2 is actually true [7, 45, 47]. Call $P : \mathbb{F}_p^n \to \mathbb{C}$ a non-standard polynomial of degree $< k$ if for all $h_1, \ldots, h_s \in \mathbb{F}_p^n$ we have

$$\Delta_{h_1} \ldots \Delta_{h_s} P \equiv 1$$

If char $\mathbb{F} \geq s$, then a non-standard polynomial is a standard phase polynomial, i.e $e^{2\pi i P(x)/p}$ where $P : \mathbb{F}_p^n \to \mathbb{F}_p$ a polynomial of degree $< s$, but otherwise the class of non-standard polynomials is larger.

[13] One can exhibit a polynomial phase function $e^{p(x)}$ as a nilsequence see e.g. [24]

Theorem 8.3 (Bergelson-Tao-Z (10), Tao-Z (10,12)). *Let p be a prime and let $f : \mathbb{F}_p^n \to \mathbb{C}$ be 1-bounded, with $\|f\|_{U^{s+1}[\mathbb{F}_p^n]} \geq \delta$. Then there exists a non-standard polynomial P of degree $\leq s$, and a constant $c(s,\delta) > 0$ such that*

$$|\mathbb{E}_{x \leq \mathbb{F}_p^n} f(x) e^{P(x)}| \gg c(s,\delta).$$

The proof of theorem 8.3 is via an ergodic theoretic structure theorem, similar in nature to Theorem 6.1, and a correspondence theorem - translating the finitary question to a question about limiting behavior of multiple averages for an $\oplus \mathbb{F}_p$ ergodic action.

Finally Conjecture 8.1 was proved [24]:

Theorem 8.4 (Green-Tao-Z (12)). *The inverse conjecture for the Gowers norms $GI(s)$ norms is true.*

The proof of Theorem 8.4 is long an complicated and is carried out in [24]. For a more gentle introduction to the proof we refer the reader to either [26], where the case $k = 3$ (the U^4 norm) is handled, or to the announcement in [25]. We now try to give the flavor of the proof. Suppose $\|f\|_{U^{s+1}[N]} \geq \delta$, then by definition

$$\mathbb{E}_{h \in N} \|\Delta_h f(n)\|_{U^s[N]}^{2^s} \gg_\delta 1.$$

It follows that for all h in a set H of size $\gg_\delta N$ we have $\|\Delta_h f(n)\|_{U^s[N]} \gg_\delta 1$. Now, inductively we know that, for $h \in H$, $\Delta_h f(n)$ correlates with an $s-1$-step nilsequence $F_h(g_h^n x_h \Gamma)$ (of complexity $\ll_\delta 1$), namely

$$|\mathbb{E}_{h \in N} \Delta_h f(n) F_h(g_h^n x_h \Gamma)| \gg_\delta 1$$

In the case $GI(2)$, this 1-step nilsequence can be taken to be $e^{2\pi i \lambda_h n}$, but in general we can't hope for anything as simple. The key difficulty now is to try to find some extra structure relating the nilsequences $F_h(g_h^n x_h \Gamma)$ for different values of h. This is already quite difficult in the $GI(2)$ case. In this case, an ingenious argument of Gowers involving tools from additive combinatorics, coupled with some geometry of numbers allows one to linearize λ_h on a nice set - a generalized arithmetic progression (GAP). This argument is then combined with a symmetry argument to construct a 2-step nilsequence $g(h)$ with $\Delta_h g(n) = e^{2\pi i \lambda_h n}$ for many values of h [20]). For general s, we follow the same strategy, however it turns out to be rather difficult to extract some algebraic structure relating the various nilsequences $F_h(g_h^n x_h \Gamma)$. An alternate approach to the inverse theorem was subsequently developed by Szegedy [9, 43]. We remark that both Theorems 8.3, 8.4 are qualitative; it is a major open question to find quantitative proofs for them.

How can one apply Theorem 8.4 to obtain Theorem 1.1? We give a very rough sketch. One needs to calculate the projection of the function $\tilde{1}_{\mathbb{P}}(n) = (\log n) 1_{\mathbb{P}}(n)$ onto the combinatorial nil-factor. It turns out that the projection essentially lies in the much smaller factor of periodic functions (with bounded period). More precisely, one first performs pre-sieving to eliminate the periodic contributions. Let $W = \prod_{p<w} p$ for w a slowly increasing function of N. For $(b,W) = 1$ consider $\tilde{1}_{W,b,\mathbb{P}}(n) = \tilde{1}_{\mathbb{P}}(Wn+b)$. The projection of this function on the combinatorial nil-factor should be constant, and since its average is 1 - this constant should be 1; namely one must show that

$$\|\tilde{1}_{W,b,\mathbb{P}}(n) - 1\|_{U^k[N]} = o(1).$$

Suppose $\|\tilde{1}_{W,b,\mathbb{P}}(n)-1\|_{U^k[N]}>\delta$. Fix $\varepsilon>0$, and decompose as in (7.3)

$$\tilde{1}_{W,b,\mathbb{P}}(x)-1=f+g$$

where f is bounded $f\ll_k 1$, and $\|g\|_{U^k[N]}<\varepsilon$. Then since U_k is a norm we get that $\|f\|_{U^k[N]}>\delta/2$. By Theorem 8.4 there is a nilsequence $F(g^nx\Gamma)$ of complexity $\ll_\delta 1$ (i.e. all parameters associated with the nilsequence such as the dimension of the nilmanifold are bounded in terms of δ), such that $|\mathbb{E}f(x)F(g^nx\Gamma)|\gg_\delta 1$. From the easy direction of Theorem 8.4 (which is valid for non bounded functions as well, via repeated applications of the Cauchy-Schwarz inequality) we have $|\mathbb{E}g(x)F(g^nx\Gamma)|<c(\varepsilon)$ (with c a decreasing function). In [22] it is shown that for any bounded complexity nilsequence we have $\mathbb{E}(\tilde{1}_{W,b,\mathbb{P}}(n)-1)F(g^nx\Gamma)=o(1)$. Choosing ε sufficiently small in the decomposition (7.3), we get a contradiction.

Acknowledgements. The author is supported by ISF grant 407/12. I thank H. Furstenberg for introducing me to ergodic theory and to the rich subject of multiple recurrence. I thank V. Bergelson and T. Tao for their valuable comments on an earlier version of this paper.

References

[1] N. Alon, T. Kaufman, M. Krivelevich, S. Litsyn, and D. Ron, *Testing low-degree polynomials over GF(2)*, Approximation, randomization, and combinatorial optimization, 188-199, Lecture Notes in Comput. Sci., **2764**, Springer, Berlin, 2003.

[2] T. Austin, *On the norm convergence of nonconventional ergodic averages*, Ergodic Theory Dynam. Systems **30** (2010), no. 2, 321–338.

[3] V. Bergelson, B. Host, and B. Kra, *Multiple recurrence and nilsequences. With an appendix by Imre Ruzsa*, Invent. Math. **160** (2005), no. 2, 261–303.

[4] V. Bergelson, *The multifarious Poincaré recurrence theorem. Descriptive set theory and dynamical systems* (Marseille-Luminy, 1996), 31–57, London Math. Soc. Lecture Note Ser., **277**, Cambridge Univ. Press, Cambridge, 2000.

[5] V. Bergelson and A. Leibman, *Polynomial extensions of van der Waerden's and Szemerédi's theorems*. J. Amer. Math. Soc. **9** (1996), no. 3, 725–753.

[6] V. Bergelson, A. Leibman, and E. Lesigne, *Intersective polynomials and the polynomial Szemerédi theorem*, Adv. Math. **219** (2008), no. 1, 369–388.

[7] V. Bergelson, T. Tao and T. Ziegler, *An inverse theorem for the uniformity seminorms associated with the action of* $\mathbb{F}_p^\infty$, Geom. Funct. Anal. **19** (2010), No. 6, 1539–1596.

[8] T. Browning and L. Matthiesen, *Norm forms for arbitrary number fields as products of linear polynomials*, arXiv:1307.7641.

[9] O. A. Camarena and B. Szegedy *Nilspaces, nilmanifolds and their morphisms*. arXiv: 1009.3825.

[10] J.P. Conze and E. Lesigne, *Théorèmes ergodique por les mesures diagonales*, Bull. Soc. Math. France **112** (1984), 143–175.

[11] ———, *Sur un théorème ergodique pour des mesures diagonales*, Probabilities, 1–31, Publ. Inst. Rech. Math. Rennes, 1987-1, Univ. Rennes I, Rennes, 1988.

[12] ———, *Sur un théorème ergodique pour des mesures diagonales*, C. R. Acad. Sci. Paris, Série I, **306** (1988), 491–493.

[13] B. Cook, A. Magyar, and T. Titichetrakun, *A Multidimensional Szemerédi Theorem in the primes*, arXiv:1306.3025.

[14] J. Fox and Y. Zhao, *A short proof of the multidimensional Szemerédi theorem in the primes*, arXiv:1307.4679.

[15] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.

[16] H. Furstenberg and B. Weiss, *A mean ergodic theorem for* $1/N\sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$. Convergence in ergodic theory and probability (Columbus, OH, 1993), 193–227, Ohio State Univ. Math. Res. Inst. Publ., 5 de Gruyter, Berlin, 1996.

[17] D. Goldston, J. Pintz, and C. Yıldırım, *Primes in tuples. I*, Ann. of Math. (2) **170** (2009), no. 2, 819–862.

[18] B. Green and T. Tao, *The distribution of polynomials over finite fields, with applications to the Gowers norms*, Contrib. Discrete Math., (2)**4** (2009), 1–36.

[19] ———, *The primes contain arbitrarily long arithmetic progressions*, Annals of Math. **167** (2008), 481-547.

[20] ———, *An inverse theorem for the Gowers* $U^3(G)$ *norm*, Proc. Edin. Math. Soc. **51** (2008), 73–153.

[21] ———, *Linear equations in primes*, Ann. of Math. (2) **171** (2010), no. 3, 1753-1850.

[22] ———, *The Möbius function is strongly orthogonal to nilsequences*, Ann. of Math. (2) **175** (2012), no. 2, 541–566.

[23] ———, *The quantitative behaviour of polynomial orbits on nilmanifolds*, Ann. of Math. (2) **175** (2012), no. 2, 465–540.

[24] B. Green, T. Tao, and T. Ziegler, *An inverse theorem for the Gowers* $U^{s+1}[N]$ *norm*, Ann. Math. (2) **176** (2012), no. 2, 1231–1372.

[25] ———, *An inverse theorem for the Gowers* $U^{s+1}[N]$ *norm*, Electron. Res. Announc. Math. Sci. **18** (2011), 69–90.

[26] ———, *An inverse theorem for the Gowers U4-norm*, Glasg. Math. J. **53** (2011), no. 1, 1–50.

[27] T. Gowers, *A new proof of Szemerédi's theorem*, Geom. Func. Anal., **11** (2001), 465–588.

[28] _____, *Decompositions, approximate structure, transference, and the Hahn-Banach theorem*, Bull. Lond. Math. Soc. **42** (2010), no. 4, 573–606.

[29] G.H. Hardy and J.E. Littlewood, *Some problems of Parititio Numerorum (III): On the expression of a number as a sum of primes*, Acta Math. **44** (1922), 1–70.

[30] Y. Harpaz, A. Skorobogatov, and O.Wittenberg, *The Hardy-Littlewood conjecture and rational points*, arXiv:1304.3333.

[31] H. Helfgott, *The ternary Goldbach conjecture is true*, arXiv:1312.7748.

[32] B. Host and B. Kra, *Nonconventional ergodic averages and nilmanifolds,* Ann. of Math. (2) **161** (2005), no. 1, 397–488.

[33] B. Host and B. Kra, personal communication.

[34] A. Leibman, *Orbits on a nilmanifold under the action of a polynomial sequence of translations*, Ergodic Theory and Dynamical Systems **27** (2007), 1239–1252.

[35] S. Lovett, R. Meshulam, and A. Samorodnitsky, *Inverse conjecture for the Gowers norm is false*, STOC 2008.

[36] J. Maynard, *Small gaps between primes*, arXiv:1311.4600.

[37] D.H.J. Polymath *New equidistribution estimates of Zhang type, and bounded gaps between primes*, arXiv:1402.0811.

[38] O. Reingold, L. Trevisan, M. Tulsiani, and S. Vadhan, *Dense subsets of pseudorandom sets*, Electronic Colloquium on Computational Complexity, Proceedings of 49th IEEE FOCS, 2008.

[39] K. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 104–109.

[40] A. Samorodnitsky, *Low-degree tests at large distances*, STOC 2007.

[41] T. Sanders, *On Roth's theorem on progressions*, Ann. of Math. (2) **174** (2011), no. 1, 619–636.

[42] K. Soundararajan, *Small gaps between prime numbers: the work of Goldston-Pintz-Yildirim*, Bull. Amer. Math. Soc. (N.S.) **44** (2007), no. 1, 1–18.

[43] B. Szegedy, *On higher order Fourier analysis*, arXiv:1203.2260.

[44] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 299–345.

[45] T. Tao, T. Ziegler, *The inverse conjecture for the Gowers norm over finite fields via the correspondence principle*, Analysis & PDE Vol. **3** (2010), No. 1, 1–20.

[46] _____, *The primes contain arbitrarily long polynomial progressions*, Acta Math. **201** (2008) 213–305.

[47] ———, *The inverse conjecture for the Gowers norm over finite fields in low characteristic*, Ann. Comb. **16** (2012), no. 1, 121–188.

[48] ———, *A multi-dimensional Szemerédi theorem for the primes via a correspondence principle*. Israel Journal of Math, to appear.

[49] I. M. Vinogradov, *Elements of number theory*. Translated by S. Kravetz. Dover Publications, Inc., New York, 1954.

[50] M. Walsh, *Norm convergence of nilpotent ergodic averages*, Ann. of Math. (2) **175** (2012), no. 3, 1667–1688.

[51] Y. Zhang, *Bounded gaps between primes*, Annals of Math, to appear.

[52] T. Ziegler, *Non conventional ergodic averages*, PhD Thesis, Hebrew University 2003.

[53] ———, *A Non Conventional Ergodic Theorem for a Nil-System*, Ergodic Theory and Dynamical Systems **25** (2005) no. 4, 1357–1370.

[54] ———, *Universal characteristic factors and Furstenberg averages*, J. Amer. Math. Soc. **20** (2007), 53–97.

[55] R. Zimmer, *Ergodic actions with generalized discrete spectrum*, Illinois J. Math. **20** (1976), no. 4, 555–588.

Einstein Institute of Mathematics, Edmond J. Safra Campus, Givat Ram The Hebrew University of Jerusalem, Jerusalem, 91904, Israel; Mathematics Department, Technion - Israel Institute of Technology Haifa, 32000, Israel.

E-mail: tamarz@math.huji.ac.il

4. Algebraic and Complex Geometry

On the virtual fundamental class

Kai Behrend

Abstract. We make a few general remarks about derived schemes, and explain the formalism of the virtual fundamental class. We put particular emphasis on the case of symmetric obstruction theories, and explain why the associated intersection numbers and enumerative invariants (such as those of Donaldson-Thomas) exhibit motivic behaviour. Motivated by this, we raise the question of categorification, and explain why this leads into derived symplectic geometry.

Mathematics Subject Classification (2010). Primary 14N35; Secondary 14D20.

Keywords. Virtual fundamental class, symmetric obstruction theory, motivic invariants, derived geometry.

1. Introduction

One of the most remarkable success stories in algebraic geometry in the last 20 years has been the advances made in enumerative geometry by the introduction of new invariants such as those of Gromov-Witten [17], or Donaldson-Thomas [26], [20]. A key concept in these developments is the virtual fundamental class [4], [19]. It provides a substitute for smoothness in the case that moduli spaces are far from non-singular. Before it was even constructed, it was clear [16] that the virtual fundamental class should be some kind of 'classical shadow' of a 'derived moduli scheme'.

In differential geometry, such issues are usually dealt with by moving objects into general position (for example, replacing holomorphic curves by pseudo-holomorphic curves), but in algebraic geometry, this is often not possible, and a more intrinsic approach is required. Thus, one seeks to construct the virtual fundamental class from data intrinsic to the moduli problem in question: this is the derived geometry of the moduli problem. The virtual fundamental class provides a cycle against which one can integrate natural cohomology classes to obtain enumerative invariants.

There is a fundamental difference between the enumerative theories à la Gromov-Witten, and those à la Donaldson-Thomas. The latter belong to derived symplectic geometry, and therefore display motivic behaviour (which the former do not). That these invariants might satisfy some kind of motivic behaviour was realized right away, because it is apparent when the moduli spaces are non-singular. In the general case, the proof of this fact was then furnished in [2]. The key realization was, that the correct way to count the contribution of a point of the moduli space to the virtual count was a generalization of the Milnor number of a critical point. This opened the door to motivic generalizations of Donaldson-Thomas theory [18], and derived symplectic geometry [24].

Proceedings of the International Congress of Mathematicians, Seoul, 2014

This article is divided into two parts.

In the first, we give a brief idea of some properties of derived schemes in general, and delve into the theory of the virtual fundamental class in more detail. We will not define derived schemes, and will completely ignore many important issues, such as, for example, all higher categorical phenomena.

In the second part we consider Lagrangian intersections, and explain how this leads to the motivic nature of Donaldson-Thomas invariants and the discovery of derived symplectic geometry.

2. The virtual fundamental class

The virtual fundamental class is a *classical shadow* of a *derived space*. We start by explaining what this means.

We will generally put ourselves in the context of algebraic geometry over the field of complex numbers. By a *classical scheme* we mean a 'usual scheme' as discussed in [12], for example. Let us agree that all our schemes are of finite type over $\mathbb{C}$.

2.1. Derived schemes. We will not define what a derived scheme is, but rather list the most important of its 'classical shadows'. Suppose $\mathfrak{X}$ is a derived scheme.

2.1.1. Classical locus. There is a classical scheme $X \subset \mathfrak{X}$, contained as a closed subscheme inside $\mathfrak{X}$. Usually the classical scheme X is highly singular, even if $\mathfrak{X}$ is not. It may happen that $\mathfrak{X} = X$.

2.1.2. Amplitude. Derived schemes have an 'amplitude of smoothness'. This is an integer $n \geq 0$. If $n = 0$, then $\mathfrak{X} = X$ and X is non-singular. (In fact, every smooth scheme can be considered as a derived scheme of amplitude 0.)

If the amplitude satisfies $n \leq 1$, the derived scheme $\mathfrak{X}$ is often called *quasi-smooth*. If $\mathfrak{X}$ has amplitude 1, and $X = \mathfrak{X}$, then X is necessarily a local complete intersection scheme. (Conversely, every local complete intersection scheme is a quasi-smooth derived scheme.)

The smoothness amplitude may be ∞, although that case should be considered pathological. (Classical schemes which are not local complete intersections are always of infinite amplitude when considered as derived schemes.)

2.1.3. Virtual dimension. If the amplitude is finite, the derived scheme $\mathfrak{X}$ has a virtual dimension. This agrees with the usual notion of dimension in the case of amplitude zero (and also in the case of amplitude 1, if $\mathfrak{X} = X$).

2.1.4. Virtual fundamental class. This exists only in the case that the smoothness amplitude is at most 1, i.e., the quasi-smooth case.

If the amplitude is 0, the virtual fundamental class is the usual fundamental class $[X]$ of X, as an oriented real manifold. If X is compact, it is an element of $H_{2d}(X, \mathbb{Z})$, where d is the complex dimension of X, in the general case it is an element of Borel-Moore homology, i.e., homology with locally finite supports. In algebraic geometry, we usually prefer to use Chow homology, then the fundamental class is in $A_d(X)$, the group of cycles of dimension d modulo rational equivalence. We will use the Chow homology version.

2.1.5. Tangent complex. There is a derived category object $\mathbb{T}_{\mathfrak{X}} \in D^b_{\mathrm{coh}}(\mathscr{O}_X)$ in the derived category of the classical locus called the *tangent complex*. In the amplitude zero case, $\mathbb{T}_{\mathfrak{X}}$ is the tangent bundle of the classical scheme $X = \mathfrak{X}$. More generally, in the case $X = \mathfrak{X}$, the tangent complex $\mathbb{T}_{\mathfrak{X}}$ is the dual of the more fundamental cotangent complex of X.

The tangent complex determines the amplitude: if the tangent complex $\mathbb{T}_{\mathfrak{X}}$ is a complex of vector bundles $E^0 \to E^1 \to \ldots \to E^n$ over X, then the derived scheme $\mathfrak{X}$ has amplitude in $[0, n]$. More precisely, $\mathfrak{X}$ has amplitude in $[0, n]$, if and only if, *locally in* X, the tangent complex is quasi-isomorphic to a complex $E^0 \to \ldots \to E^n$ of vector bundles over X.

The tangent complex also determines the virtual dimension. In fact, the virtual dimension of $\mathfrak{X}$ is the rank of $\mathbb{T}_{\mathfrak{X}}$, which is equal to $\sum_{i=0}^{\text{amplitude}} (-1)^i \dim E^i$, for a (local) presentation $\mathbb{T}_{\mathfrak{X}} = E^\bullet$.

Let us denote the cohomology sheaves of $\mathbb{T}_{\mathfrak{X}}$ by $h^i(\mathbb{T}_{\mathfrak{X}})$. There are coherent sheaves on X. We always have that $h^0(\mathbb{T}_{\mathfrak{X}}) = T_X = \mathscr{D}er(\mathscr{O}_X, \mathscr{O}_X)$ is the Zariski tangent sheaf of X.

In the quasi-smooth case, we denote $h^1(\mathbb{T}_{\mathfrak{X}})$ by $ob_{\mathfrak{X}}$, and call it the *obstruction sheaf*. In this latter case, we may think of $\mathbb{T}_{\mathfrak{X}}$ as a 2-extension of $ob_{\mathfrak{X}}$ by T_X, i.e., an element of $\mathrm{Ext}^2_{\mathscr{O}_X}(ob_{\mathfrak{X}}, T_X)$:

$$0 \longrightarrow T_X \longrightarrow \mathbb{T}^0_{\mathfrak{X}} \longrightarrow \mathbb{T}^1_{\mathfrak{X}} \longrightarrow ob_{\mathfrak{X}} \longrightarrow 0$$

In case X is non-singular as a scheme, $ob_{\mathfrak{X}}$ is a vector bundle, and so the 2-extension $\mathbb{T}_{\mathfrak{X}}$ is a cohomology class in $H^2\big(X, ob^\vee_{\mathfrak{X}} \otimes T_X\big)$.

2.1.6. Canonical bundle. In the finite amplitude case, the determinant of $\mathbb{T}_{\mathfrak{X}}$ is a well-defined line bundle on X. Its dual is denoted by $K_{\mathfrak{X}}$ and called the *virtual canonical bundle* of $\mathfrak{X}$.

2.1.7. Higher structure sheaves. There is a graded sheaf of $\mathscr{O}_X$-algebras on the classical locus X, denoted $\pi_*(\mathscr{O}_{\mathfrak{X}})$. It satisfies $\pi_0(\mathscr{O}_{\mathfrak{X}}) = \mathscr{O}_X$ and $\pi_i(\mathscr{O}_{\mathfrak{X}}) = 0$, for all $i < 0$. Every $\pi_i(\mathscr{O}_{\mathfrak{X}})$ is a coherent sheaf of $\mathscr{O}_X$-modules. Moreover, $\pi_*(\mathscr{O}_{\mathfrak{X}})$ is a *graded commutative* sheaf of $\mathscr{O}_X$-algebras, so there are operations $\pi_i(\mathscr{O}_{\mathfrak{X}}) \otimes_{\mathscr{O}_X} \pi_j(\mathscr{O}_{\mathfrak{X}}) \to \pi_{i+j}(\mathscr{O}_{\mathfrak{X}})$. In the quasi-smooth case, $\pi_*(\mathscr{O}_{\mathfrak{X}})$ is bounded above. In general (even in the finite amplitude case) it will be unbounded.

2.2. Examples. Here are a few typical sources of derived schemes.

2.2.1. Affine variety cut out by a set of equations. Any collection of r polynomials $f_1, \ldots, f_r \in \mathbb{C}[x_1, \ldots, x_n]$ in n variables defines a derived scheme $\mathfrak{X}$ of amplitude contained in $[0, 1]$, i.e. a quasi-smooth derived scheme. Its virtual dimension is equal to $n - r$, the 'expected' dimension of the common zero locus of r equations in n variables. The underlying classical scheme is $X = \mathrm{Spec}\, \mathbb{C}[x_1, \ldots, x_n]/(f_1, \ldots, f_r)$. Every irreducible component of X has dimension $n - r$ or larger. If X is of pure dimension $n - r$, then it is a local complete intersection scheme, and $\mathfrak{X} = X$. All derived phenomena are due to excess dimension in the intersection of the r hypersurfaces $Z(f_1), \ldots, Z(f_r)$.

2.2.2. Toy model of quasi-smooth derived schemes. Slightly more generally, suppose that M is a non-singular $\mathbb{C}$-variety, and $E \to M$ an algebraic vector bundle over M, with a global section $s : M \to E$. In this case the scheme-theoretic zero locus $X \subset M$ of s is naturally

endowed with a derived scheme structure $X \subset \mathfrak{X} \subset M$. The smoothness amplitude of $\mathfrak{X}$ is contained in $[0,1]$, and the virtual dimension is $\dim M - \operatorname{rk} E$.

Differentiating the section $s : M \to E$ gives rise to a homomorphism of vector bundles $T_M|_X \to E|_X$ over X, this represents the tangent complex $\mathbb{T}_{\mathfrak{X}} \in D^b_{\mathrm{coh}}(\mathscr{O}_X)$, if we put $T_M|_X$ in degree 0 and $E|_X$ in degree 1. The kernel of $T_M|_X \to E|_X$, i.e., $h^0(\mathbb{T}_{\mathfrak{X}})$, is the Zariski tangent sheaf of X, denoted T_X. It is isomorphic to the sheaf of derivations $\mathscr{O}_X \to \mathscr{O}_X$, and does not depend on $\mathfrak{X}$, only on X. The obstruction sheaf $ob_{\mathfrak{X}} = h^1(\mathbb{T}_{\mathfrak{X}})$, is the cokernel of $T_M|_X \to E|_X$. It depends on $\mathfrak{X}$ in an essential way. There is an exact sequence of coherent sheaves on X

$$0 \longrightarrow T_X \longrightarrow T_M|_X \longrightarrow E|_X \longrightarrow ob_{\mathfrak{X}} \longrightarrow 0 \ .$$

(To get the signs right, let us remark that, strictly speaking, the complex $\mathbb{T}_{\mathfrak{X}}$ is given by the canonical homomorphism $T_M|_X \to E|_X$, *multiplied by* -1, as it is the dual of the cotangent complex of $\mathfrak{X}$, which is considered more fundamental.)

The name *obstruction sheaf* comes from deformation theory. The obstruction sheaf contains all obstructions to the smoothness of X, although it can be much bigger. (Moreover, the obstructions to smoothness of X are, of course, intrinsic to X, whereas $ob_{\mathfrak{X}}$ is not.)

The virtual fundamental class of $\mathfrak{X}$ is the *localized top Chern class* of E, which is an element of $H_{2\dim M - 2\operatorname{rk} E}(X, \mathbb{Z})$, or $A_{\dim M - \operatorname{rk} E}(X)$, respectively. Let us review its construction via *deformation to the normal cone*, [11].

The graph of s is a subvariety Γ_s of the total space of E, isomorphic to the base M via the projection $\pi : E \to M$. We multiply Γ_s by a scalar $\lambda \in \mathbb{C}^*$ using the vector bundle structure on E, and let $\lambda \to \infty$. The limit maps to $X \subset M$ via the projection π, and it is invariant under the $\mathbb{C}^*$-action on E. Therefore, it is is a cone C_X inside the restriction $E|_X$ of the bundle E to X. The cone C_X is known as the *normal cone* of X in M. As an abstract scheme, $C_X \to X$ depends only on the embedding $X \hookrightarrow M$, in fact, $C_X = \operatorname{Spec}_X \bigoplus_{i=0}^{\infty} I^i/I^{i+1}$, where I is the ideal sheaf of X in M. (The embedding $C_X \hookrightarrow E|_X$ comes from writing X as the zero locus of the section $s : M \to E$ of π.) The cone $C_{X/M}$ is of pure dimension $\dim M$.

We are not interested in this cone as a scheme, but only in its *fundamental cycle*. If $C_1, \ldots, C_s$ are the irreducible components of $C_{X/M}$ and $r_1, \ldots, r_s$ their multiplicities in the scheme $C_{X/M}$, then the fundamental cycle is $[C_{X/M}] = r_1[C_1] + \ldots + r_s[C_s] \in A_{\dim M}(E|_X)$. The images of the components C_i under π are subvarieties of X, the *distinguished subvarieties*. They will play a role in 3.2.5, below.

Pulling back cycles via the projection $\pi : E|_X \to X$ induces an isomorphism on Chow groups $A(X) \to A(E|_X)$, by homotopy invariance of Chow homology. The inverse of this isomorphism is the *Gysin map* $0^!_E : A(E|_X) \to A(X)$. It lowers degrees by $\operatorname{rk} E$, because pulling back increases degrees by $\operatorname{rk} E$. The localized top Chern class is $c_{\mathrm{top}}(E) = 0^!_E[C_{X/M}]$. As mentioned, this is the virtual fundamental class of the derived scheme $\mathfrak{X}$.

$$[\mathfrak{X}]^{\mathrm{vir}} = c_{\mathrm{top}}(E) = 0^!_E[C_{X/M}] \, .$$

The image of this class in $A(M)$ is the usual top Chern class of E as a bundle over M.

If all irreducible components of X have the expected dimension $\dim M - \operatorname{rk} E$, we are in the case where $X = \mathfrak{X}$ and X is a local complete intersection. Then $[\mathfrak{X}]^{\mathrm{vir}} = [X]$.

In the case where X is non-singular (as a scheme), the obstruction sheaf $ob_{\mathfrak{X}}$ is a vector bundle. Its rank is the excess dimension determined by the equation

$$\dim X = \dim M - \operatorname{rk} E + \operatorname{rk} ob_{\mathfrak{X}} \ .$$

Its top Chern class is the virtual fundamental class:

$$[\mathfrak{X}]^{\mathrm{vir}} = c_{\mathrm{top}}(ob_{\mathfrak{X}}) \cap [X] \,.$$

2.2.3. Intersection theory. Still more generally, consider two smooth subvarieties M, N inside a smooth ambient variety V. The classical intersection scheme is the fibered product $X = M \times_V N$. If X has the expected dimension, its scheme structure carries all relevant intersection theory phenomena, in particular the intersection multiplicities. In the case of excess intersection dimension, there is a refined fibered product $\mathfrak{X}$ in the category of derived schemes, whose underlying classical scheme is X. The virtual dimension of $\mathfrak{X}$ is $\dim V - \dim M - \dim N$.

Again, the amplitude of $\mathfrak{X}$ is contained in $[0, 1]$. It is equal to 0 if and only if the intersection of M and N inside V is *transverse*, which means that at every point P in the intersection X, we have $T_M|_P + T_N|_P = T_V|_P$.

Again, $X = \mathfrak{X}$ if and only if X has dimension $\dim V - \dim M - \dim N$ at every one of its points, i.e., if and only if the intersection is *proper*.

The tangent complex is $\mathbb{T}_{\mathfrak{X}} = [T_M|_X \oplus T_N|_X \to T_V|_X]$. The obstruction sheaf fits into the exact sequence

$$0 \longrightarrow T_X \longrightarrow T_M|_X \oplus T_N|_X \longrightarrow T_V|_X \longrightarrow ob_{\mathfrak{X}} \longrightarrow 0 \ .$$

So if X is non-singular, then $ob_{\mathfrak{X}}$ is equal to the *excess bundle* of the intersection.

The virtual fundamental class is the *refined intersection product* $[\mathfrak{X}]^{\mathrm{vir}} = [M] \cdot [N] \in A(X)$, which maps to the (global) intersection product in $A(V)$. (This is the top Chern class of the excess bundle in the case that X is non-singular.)

The higher structure sheaves are derived tor-sheaves:

$$\pi_i(\mathscr{O}_{\mathfrak{X}}) = \mathscr{T}or_i^{\mathscr{O}_V}(\mathscr{O}_M, \mathscr{O}_N) \,.$$

(This is the origin of the word *derived scheme*.)

If M and N have complementary dimension, then the virtual dimension of $\mathfrak{X}$ is 0, and hence $[\mathfrak{X}]^{\mathrm{vir}}$ is a 0-cycle on X. If, moreover, X is proper (the algebraic analogue of compact), then $[\mathfrak{X}]^{\mathrm{vir}}$ has a well-defined degree. This is the *intersection number* of M and N in V, which we think of as the virtual number of points of $\mathfrak{X}$:

$$\#^{\mathrm{vir}}(\mathfrak{X}) = \deg[\mathfrak{X}]^{\mathrm{vir}} = \int_{[\mathfrak{X}]^{\mathrm{vir}}} 1 \in \mathbb{Z} \,.$$

If the intersection is transverse, $\#^{\mathrm{vir}}(\mathfrak{X}) = \#(X)$. If it is proper, $\#^{\mathrm{vir}}(\mathfrak{X})$ is equal to the length of the 0-dimensional scheme X.

2.2.4. Gauge theory. Suppose that $L = L^{\geq 0}$ is a *differential graded Lie algebra*. As a standard example, let Y be a compact C^∞-manifold, and let

$$L^k = C^\infty(Y, \Omega_Y^k \otimes_{\mathbb{C}} M_{n\times n})$$

be the $\mathbb{C}$-vector space of matrix-valued C^∞-forms of degree k on Y. The differential is the de Rham differential, the Lie bracket is the anti-commutator of the natural associative product on L given by combining wedge product of forms with matrix multiplication in the natural way.

The *curvature map* is the quadratic map $F : L^1 \to L^2$, defined by

$$x \mapsto dx + \tfrac{1}{2}[x, x]\,.$$

The vanishing locus of the curvature in L^1 is the *Maurer-Cartan locus* of L. In our example, L^1 is the set of all connections on the trivial vector bundle of rank n on Y, and the Maurer-Cartan locus is the set of flat connections.

Let G be a *gauge group* for L. This is a complex Lie group whose Lie algebra is L^0, and which acts on L by automorphism of the differential graded Lie algebra structure, in such a way that the derivative of the action of G on L^k is the Lie algebra action of L^0 and L^k given by the graded Lie algebra structure of L. In our example, $G = C^\infty(Y, GL_n)$ is the group of invertible sections in $L^0 = C^\infty(Y, M_{n\times n})$.

Moreover, assume given a *gauge cocycle*, i.e., a map $\gamma : G \to L^1$, satisfying

1. $\gamma(gh) = {}^g\gamma(h) + \gamma(g)$,
2. $d\gamma(g) + \frac{1}{2}[\gamma(g), \gamma(g)] = 0$,
3. ${}^g\big(d({}^{g^{-1}}x)\big) = dx + [\gamma(g), x]$,

then we define the *gauge action* of G on L^1 by $g * x = {}^g x + \gamma(g)$. In our example, $\gamma(g) = -(dg)g^{-1}$.

The *moduli space* of (L, G, γ) is the quotient of the Maurer-Cartan locus by the gauge action. In our example, this is the set of flat connections on the trivial bundle modulo gauge equivalence, or, equivalently, the set of isomorphism classes of flat vector bundles of rank n, whose underlying C^∞-bundle is trivial. In other words, it is the moduli space of (topologically trivial) local systems on Y.

For the quotient by G to be well-behaved, we have to restrict to an open subset of L^1, where G has trivial, or almost trivial stabilizers. In our example, this leads to the moduli space of *simple* local systems. Unfortunately, it is not compact.

Strictly speaking (in our example), each L^k, as well as G, is infinite-dimensional, so the above constructions do not make sense within algebraic geometry, but the ideas are nevertheless very important. On the other hand, there are many examples where L and G are finite-dimensional, so we are within algebraic geometry, and the quotient by the gauge group can be treated with geometric invariant theory, and the well-behaved subspace of L^1 is the *stable* locus with respect to a linearization of the gauge group action.

For example, Kapranov[13] has shown how our example can be made finite-dimensional by triangulating the manifold Y, and considering simplicial matrix valued cochains.

Anyway, let us call the moduli space $X = MC(L)/G$. It is the classical locus of a derived scheme $\mathfrak{X}$. To construct the tangent complex $\mathbb{T}_{\mathfrak{X}}$, we start with the trivial graded vector bundle with fibre L over $MC(L)$, and endow it with the differential which is given by d^x in the fibre over the Maurer-Cartan element x. Here d^x is the differential d twisted by $\mathrm{ad}(x)$, i.e., $d^x(y) = dy + [x, y]$. The gauge groups acts on the complex of vector bundles (L, d^x) over X, on the fibres via the linear action, on the base via the gauge action. The complex (L, d^x) therefore descends to a complex of vector bundles on the quotient. This is the shift $\mathbb{T}_{\mathfrak{X}}[-1]$. Therefore, $H^i(\mathbb{T}_{\mathfrak{X}}|_x) = H^{i+1}(L, d^x)$. (If $H^0(L, d^x) \neq 0$, then the stabilizer of G at x, whose Lie algebra is $H^0(L, d^x)$, is positive-dimensional. This stabilizer would have been 'ignored' when taking the G-quotient, and hence in this case $\mathbb{T}_{\mathfrak{X}}[-1]$ is obtained as the truncation $\tau_{\geq 1}(L, d^x)$.)

In our example, (L, d^x) is the de Rham complex of Y with values in the endomorphism bundle of the flat bundle E defined by x. Therefore its cohomology groups are $H^i(L, d^x) = H^i\big(Y, \mathrm{End}(E)\big)$, and hence $H^i(\mathbb{T}_{\mathfrak{X}}|_x) = H^{i+1}\big(Y, \mathrm{End}(E)\big)$. The virtual dimension of this derived scheme is therefore

$$\sum_{i=0}^{\dim Y - 1} (-1)^i H^{i+1}\big(Y, \mathrm{End}(E)\big) = 1 - \chi\big(Y, \mathrm{End}(E)\big)\,,$$

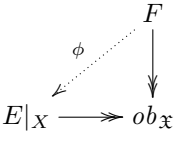
and its amplitude is contained in the interval $[0, \dim Y - 1]$.

For the Zariski tangent space of the point of X defined by the flat bundle E we get $T_X|_E = H^1\big(Y, \mathrm{End}(E)\big)$. This is the space of infinitesimal deformations of E, and this result would have been predicted by deformation theory. The next space $H^2\big(Y, \mathrm{End}(E)\big)$ contains the obstructions to the smoothness of X at the point represented by E. This is another fact proved in deformation theory.

If we pull back the higher structure sheaves form X to $MC(L)$, and take global sections, we get essentially the differential graded Lie algebra cohomology of L (a generalization of ordinary Lie algebra cohomology to the differential graded case), with values in the trivial module $\mathbb{C}$.

So if Y is a surface, then $\mathfrak{X}$ is quasi-smooth, and we should have a virtual fundamental class. But it turns out, at least when Y is orientable, that there is more symmetry (see 3.3.1, below), and therefore X is smooth. The spaces X we get are known as *character varieties*, there is a rich body of literature devoted to their study.

2.3. Construction of the virtual fundamental class. Suppose $\mathfrak{X}$ is a quasi-smooth derived scheme. For simplicity, let us assume that the underlying classical scheme X is quasi-projective. Then it is possible to write the obstruction sheaf $ob_{\mathfrak{X}} = h^1(\mathbb{T}_{\mathfrak{X}})$ as a quotient of a locally free sheaf $F \twoheadrightarrow ob_{\mathfrak{X}}$. There is then an associated scheme of closed subcones $C \hookrightarrow F$. It is characterized by the property that whenever $\mathfrak{X}$ is (locally!) written as the derived scheme associated to a toy model $M \xrightarrow{s} E$, then for any lift ϕ as in the diagram



we have $C = \phi^{-1}(C_{X/M})$.

It can be shown that C is the preimage under the epimorphism $F \twoheadrightarrow ob_{\mathfrak{X}}$ of the cone $cv_{\mathfrak{X}} \subset ob_{\mathfrak{X}}$ of *small curvi-linear obstructions*.

The virtual fundamental class of $\mathfrak{X}$ is then

$$[\mathfrak{X}]^{\mathrm{vir}} = 0^!_F[C] \in A(X)\,.$$

It does not depend on the choice of F nor on the choice of the epimorphism $F \twoheadrightarrow ob_{\mathfrak{X}}$.

As this construction shows, the virtual fundamental class of $\mathfrak{X}$ depends only on the tangent complex $\mathbb{T}_{\mathfrak{X}}$, and its properties. Abstracting these properties, leads to the notion of *perfect obstruction theory*. We explain this formalism next.

2.3.1. The truncated tangent complex of a scheme. Given a scheme X, which is embedded as a closed subscheme into a non-singular scheme M, there is a canonical homomor-

phism of coherent sheaves on X

$$T_M|_X \longrightarrow \mathscr{N}_{X/M} \ .$$

Here, $\mathscr{N}_{X/M}$ is the *normal sheaf* of X in M, it is the dual of the coherent $\mathscr{O}_X$-module I/I^2, where I is the ideal sheaf of X in M. A standard argument shows that the associated object of $D^b_{\mathrm{coh}}(\mathscr{O}_X)$, obtained by putting $T_M|_X$ in degree 0 and $\mathscr{N}_{X/M}$ in degree 1 (and multiplying the map by -1), does not depend on the embedding $X \hookrightarrow M$ chosen, and exists naturally even when no such embedding exists globally. This derived category object is the *truncated tangent complex* of X, notation $\mathbb{T}_X$. (It is the dual of $\tau_{\geq -1}L_X$, where L_X is the cotangent complex of the scheme X.)

As $\mathscr{N}_{X/M}$ is the dual of a coherent sheaf, it is, in fact, a scheme over X. The normal cone of the embedding $X \subset M$ is naturally a $\mathbb{C}^*$-invariant closed subscheme $C_{X/M} \hookrightarrow \mathscr{N}_{X/M}$.

2.3.2. Perfect obstruction theories. A *perfect obstruction theory* on the scheme X is an object of the derived category $F \in D^b_{\mathrm{coh}}(\mathscr{O}_X)$, together with a morphism $\phi : \mathbb{T}_X \to F$, such that

1. F is of perfect amplitude contained in $[0,1]$, in other words, X being quasi-projective, F is given by a two term complex of vector bundles $F^0 \to F^1$ over X,
2. whenever we represent ϕ (locally over X) by a homomorphism of complexes

$$\begin{array}{ccc} T_M|_X & \longrightarrow & \mathscr{N}_{X/M} \\ {\scriptstyle \phi^0}\downarrow & & \downarrow{\scriptstyle \phi^1} \\ F^0 & \longrightarrow & F^1 \end{array} \tag{2.1}$$

the square we obtain is a pullback diagram of schemes over X.

Given a perfect obstruction theory $\mathbb{T}_X \to F$ on X, we define the corresponding obstruction sheaf to be $ob_F = h^1(F)$. Any way of representing F as $F^0 \to F^1$ gives an epimorphism $F^1 \to ob_F$ from a vector bundle, and hence, as explained above, a cone $C \hookrightarrow F^1$, and the virtual fundamental class

$$[X,F]^{\mathrm{vir}} = 0^!_{F^1}[C] \in A(X) \, .$$

which depends only on the scheme X, and the perfect obstruction theory $\mathbb{T}_X \to F$. (For any diagram (2.1), the cone $C \hookrightarrow F^1$ pulls back to the normal cone $C_{X/M} \subset \mathscr{N}_{X/M}$ under ϕ^1.)

Whenever $\mathfrak{X}$ is a quasi-smooth derived scheme, there is a natural homomorphism $\mathbb{T}_X \to \mathbb{T}_{\mathfrak{X}}$ in $D^b_{\mathrm{coh}}(\mathscr{O}_X)$, which is, in fact, a perfect obstruction theory. The virtual fundamental class of $\mathfrak{X}$ is then equal to the virtual fundamental class of the underlying classical scheme X with respect to the obstruction theory $\mathbb{T}_{\mathfrak{X}}$.

In general, perfect obstruction theories are much easier to construct than derived schemes. Usually, they come directly from deformation theory, and the fact that they are perfect also follows from standard facts in deformation theory (if it is true).

2.3.3. K-theory fundamental class. The original suggestion of Kontsevich to construct the virtual fundamental class was via the higher structure sheaves. In fact, for any quasi-smooth derived scheme $\mathfrak{X}$, the following is true [7]:

$$[\mathfrak{X}]^{\mathrm{vir}} = \tau_X[\pi_*(\mathscr{O}_{\mathfrak{X}})] \cdot \mathrm{Td}^{-1}(\mathbb{T}_{\mathfrak{X}}) \, .$$

Here τ_X is the homological Chern character of Baum-Fulton-MacPherson, applied to the class of the alternating sum of the higher structure sheaves in K-theory of coherent sheaves on X.

2.4. Applications. When the scheme X is endowed with a virtual fundamental class $[X,F]^{\mathrm{vir}} \in A_d(X)$, defined by a perfect obstruction theory F (which may or may not come from a quasi-smooth derived scheme $\mathfrak{X}$ via $F = \mathbb{T}_{\mathfrak{X}}$), and X is *proper*, then the virtual fundamental class defines a homomorphism

$$A^d(X) \longrightarrow \mathbb{Z}\,, \qquad \psi \longmapsto \int_{[X]^{\mathrm{vir}}} \psi\,.$$

Here d is the virtual dimension of (X,F). The cohomology classes ψ of degree d which are integrated against the virtual fundamental class to obtain integers are often called *insertions*. Thus combining the virtual fundamental class with suitable insertions gives integers, which are often referred to as *invariants*, because being constructed by means of intersection theory, they are usually invariant under deformations of the underlying geometry.

2.4.1. Donaldson Invariants. As an example, consider the algebraic geometry analogue of our gauge theory example, above. Instead of flat bundles on a compact C^∞-manifold, we consider holomorphic bundles on a compact complex manifold, or algebraic vector bundles on a non-singular projective variety, denoted Y. For a vector bundle E over Y, its infinitesimal deformations are given by the vector space $H^1\big(Y,\mathrm{End}(E)\big)$, and obstructions are contained in the vector space $H^2\big(Y,\mathrm{End}(E)\big)$. This suggests, by analogy with the above gauge theory example, that the derived moduli scheme $\mathfrak{X}$ of bundles on Y should have a tangent complex $\mathbb{T}_{\mathfrak{X}}$ with the property that $H^i(\mathbb{T}_{\mathfrak{X}}|_E) = H^{i+1}\big(Y,\mathrm{End}(E)\big)$ at every point E of the classical moduli scheme X of bundles on Y. This leads us to expect that the derived category object $\big(\tau_{\geq 1} R\pi_*\,\mathscr{H}om(\mathscr{E},\mathscr{E})\big)[1]$ on X might serve as a perfect obstruction theory, and define a virtual fundamental class on X. (Here $\mathscr{E}$ is the universal bundle over $X\times Y$, and $\pi : X\times Y \to X$ is the projection.) As one of the requirements on a perfect obstruction theory is that it has to exist in the interval $[0,1]$, we will need to require that $H^i\big(Y,\mathrm{End}(E)\big) = 0$, for all $i > 2$. The best way to assure this in general, is to assume that Y is of dimension 2, i.e., an algebraic surface. We then do, indeed, get a perfect obstruction theory, and a virtual fundamental class for X. To compactify X, we pass to stable sheaves. In good cases, i.e., where there are no strictly semi-stable sheaves, the moduli space will be compact, and with suitable insertions, we get numerical invariants. These are, essentially, *Donaldson invariants* [10]. See [22] for a treatment using virtual fundamental classes.

2.4.2. Gromov-Witten invariants. An example for which the theory of perfect obstruction theories and virtual fundamental classes was very successful is that of Gromov-Witten invariants. The relevant moduli space is essentially a moduli space of morphisms $X = \mathrm{Mor}(C,V)$, where C is a fixed algebraic curve (with at most nodal singularities) and V a fixed non-singular projective variety. Deformation theory tells us that infinitesimal deformations of $f : C \to V$ are given by $H^0(C, f^*T_V)$, i.e., by vector fields tangent to V, supported by C. Moreover, obstructions are contained in $H^1(C, f^*T_V)$. So $R\pi_* f^*T_V$ is a perfect obstruction theory for X. (Here $f : X\times C \to V$ is the universal map, and $\pi : X\times C \to X$ the projection.) The key construction that makes this useful, is the compactification of $X = \mathrm{Mor}(C,V)$ using *stable maps* due to Kontsevich. In the compact-

ification the curve C exhibits 'bubbling phenomena', i.e. it sprouts off trees of projective lines (or Riemann spheres), connected to each other by nodal singularities. No matter how bad the bubbling, we always have $H^i(C, f^*T_V) = 0$, for $i > 1$, because C always stays one-dimensional.

For the theory of Gromov-Witten invariants one makes this construction relative to the moduli space of all stable curves. So one keeps V fixed, but varies C. The additional complication that arises, is that the automorphism groups of curves, although finite, may jump in families. This means that the moduli space will be a *stack of Deligne-Mumford type* rather than a scheme. The only difference this makes is that the integrals of the insertions against the virtual fundamental class, and hence the Gromov-Witten invariants, will take values in rational numbers, rather than integers.

3. The virtual fundamental class in symplectic geometry

We continue in the context of algebraic geometry over $\mathbb{C}$. Thus, a *symplectic manifold* is a non-singular scheme over $\mathbb{C}$, endowed with an everywhere non-degenerate closed algebraic (hence holomorphic) 2-form. A *Lagrangian submanifold* is an algebraic (in particular holomorphic) submanifold of half the dimension of the ambient symplectic manifold, whose tangent space is everywhere isotropic for the symplectic form.

3.1. Lagrangian intersections. In symplectic geometry it is natural to consider the intersection of two Lagrangian submanifolds inside a symplectic manifold. First of all, we notice that the expected dimension of such an intersection is always 0, because two Lagrangian submanifolds always have complementary dimensions. Thus, if the intersection is compact, intersection theory gives rise to intersection numbers, without the need for cohomological insertions.

3.1.1. Motivic nature of intersection. Now something quite unexpected happens: the intersection number is *motivic*, i.e., it behaves like an Euler characteristic. To explain what we mean by this, suppose that S is a symplectic manifold of dimension $2n$, and let L, M be Lagrangian submanifolds, with intersection $X = L \cap M$.

1. the intersection number $\#^{\mathrm{vir}}(X)$ makes sense whether or not the intersection X is compact,
2. the intersection number is additive over open covers: if $S = U \cup V$, for Zariski open subsets U, V of S, then

$$\#^{\mathrm{vir}}(X) + \#^{\mathrm{vir}}(X \cap U \cap V) = \#^{\mathrm{vir}}(X \cap U) + \#^{\mathrm{vir}}(X \cap V)\,.$$

One of our goals is to explain why this is so.

3.1.2. Punctual contributions to the intersection number. The two motivic properties imply that we can make sense of the contribution to the intersection number of every single point $P \in X$. Simply set

$$\nu_X(P) = \#^{\mathrm{vir}}(X) - \#^{\mathrm{vir}}(X \setminus P)\,.$$

It turns out that the function

$$\nu_X : X \longrightarrow \mathbb{Z}$$

is *constructible* (i.e. constant along the strata of a suitable stratification of X), and that

$$\#^{\mathrm{vir}}(X) = \sum_{n\in\mathbb{Z}} n\,\chi^{\mathrm{top}}\{x \in X \colon \nu_X = n\}$$

is the weighted Euler characteristic with respect to the constructible function ν_X.

Another surprising fact is that $\nu_X(P)$ depends only on the scheme structure of X near P, i.e., the singularity of X at P. Thus the constructible function ν_X is *intrinsic* to X.

3.1.3. Obstruction sheaf. Another unexpected fact is that also the obstruction sheaf is intrinsic to the classical intersection scheme X. To determine the obstruction sheaf, recall the tangent complex $\mathbb{T}_{\mathfrak{X}} = [T_L \oplus T_M \to T_S]|_X$. We take its dual, and shift it back into the interval $[0,1]$: $\mathbb{T}^\vee_{\mathfrak{X}}[-1] = [\Omega_S \to \Omega_L \oplus \Omega_M]|_X$. Now the symplectic form defines an isomorphism $\sigma : T_S \to \Omega_S$, such that $\sigma^\vee = -\sigma$. We use σ to construct the homomorphism $\theta = \frac{1}{2}(\sigma \oplus \sigma) : T_L|_X \oplus T_M|_X \to \Omega_S|_X$ giving rise to a commutative diagram

$$\begin{array}{ccc} \mathbb{T}_{\mathfrak{X}} & & T_L|_X \oplus T_M|_X \xrightarrow{1\oplus -1} T_S|_X \\ {\scriptstyle(\theta,-\theta^\vee)}\downarrow & & {\scriptstyle\theta}\downarrow \qquad\qquad \downarrow{\scriptstyle -\theta^\vee} \\ \mathbb{T}^\vee_{\mathfrak{X}}[-1] & & \Omega_S|_X \xrightarrow{-1\oplus 1} \Omega_L|_X \oplus \Omega_M|_X \end{array}$$

Since the kernel of $\Omega_S|_X \to \Omega_L|_X$ is equal to $T_L|_X$, by the Lagrangian nature of L, and a similar fact holds for M, we see that the kernel of $[\Omega_S \to \Omega_L \oplus \Omega_M]|_X$ is equal to T_X, and that θ induces the identity on T_X. Similar reasoning proves that the cokernel of $[T_L \oplus T_M \to T_S]|_X$ is equal to Ω_X, and that $-\theta^\vee = \theta$ induces the identity on Ω_X. We deduce that $(\theta, -\theta^\vee) : \mathbb{T}_{\mathfrak{X}} \to \mathbb{T}^\vee_{\mathfrak{X}}[-1]$ is a canonical quasi-isomorphism, hence an isomorphism in the derived category. We conclude that

$$ob_{\mathfrak{X}} = h^1(\mathbb{T}_{\mathfrak{X}}) = h^1(\mathbb{T}^\vee_{\mathfrak{X}}[-1]) = \Omega_X\,.$$

From this it also follows that

$$T_X = h^0(\mathbb{T}_{\mathfrak{X}}) = ob^\vee_{\mathfrak{X}}\,,$$

in other words, *deformations are dual to obstructions*, and *the obstruction sheaf is the sheaf of differentials*.

3.1.4. Smooth intersection. Let us explain the origin of the motivic nature of Lagrangian intersection numbers in the case where the intersection X is smooth (as a scheme). (Sometimes, this condition is referred to as *clean* intersection.) It means that the dimension of the intersection $T_L|_x \cap T_M|_x$ inside $T_S|_x$ is a locally constant function of $x \in X$. The obstruction sheaf is locally free, and called the excess bundle. We see that the excess bundle of the intersection is always the cotangent bundle of the intersection variety. Its top Chern class is therefore equal to $c_{\mathrm{top}}(\Omega_X) = (-1)^{\dim X} c_{\mathrm{top}}(T_X)$, i.e., up to sign, equal to the Euler class of X. By the Gauß-Bonnet formula we conclude, if X is proper:

$$\#^{\mathrm{vir}}(X) = \int_{[X]} c_{\mathrm{top}}(ob_{\mathfrak{X}}) = (-1)^{\dim X} \int_{[X]} c_{\mathrm{top}}(T_X) = (-1)^{\dim X} \chi^{\mathrm{top}}(X)\,, \qquad (3.1)$$

where χ^{top} denotes the topological Euler characteristic. (Apologies for using 'top' in two different senses.) Unlike the intersection number, the topological Euler characteristic is well-defined also for non-proper X. Moreover, it is additive over open covers.

We see that for the case of smooth intersection, the weight function is constant (and intrinsic to X):

$$\nu_X \equiv (-1)^{\dim X}\,.$$

Every point $P \in X$ contributes $(-1)^{\dim X}$ to the intersection number.

3.1.5. Toy model: Critical locus of a regular function. The toy model for a Lagrangian intersection is the critical locus of a regular function. Suppose that M is a manifold and $f : M \to \mathbb{C}$ a regular function. The graph of the exact differential form df is a Lagrangian submanifold inside the cotangent bundle Ω_M of M with its tautological symplectic structure. The zero section of the cotangent bundle is another Lagrangian, and the intersection of these two Lagrangians is the critical locus of f in M.

$$X = \mathrm{Crit}(f)\,.$$

For simplicity of exposition, assume that X is contained (at least set-theoretically) in the fibre of f over 0. As f is necessarily constant on every component of its critical set, this is not a serious restriction. Then X is a natural scheme structure on the set of singularities of the fibre $f^{-1}(0) \subset M$.

Let $P \in X$ be a point. The *Milnor fibre* of f at P is the intersection of a nearby fibre of $f : M \to \mathbb{C}$ with a small ball around $P \in M$:

$$F_f(P) = \{x \in M\colon ||x|| \leq \epsilon \text{ and } f(x) = \eta\} \tag{3.2}$$

for sufficiently small $\epsilon \gg \eta > 0$. The Milnor fibre $F_f(P)$ is a manifold with boundary (the boundary is the *link* of the singularity of f at P), which is independent of the choice of metric on M used to define the ball of radius ϵ, and of ϵ and η, as long as ϵ is sufficiently small, and η sufficiently small with respect to ϵ. It is called the Milnor fibre, because it is diffeomorphic to the fibre of the fibration

$$B_\epsilon(P) \setminus f^{-1}(0) \longrightarrow S^1\,, \qquad x \longmapsto \frac{f(x)}{||f(x)||}\,.$$

studied by Milnor [21].

In his well-known textbook [ibid.], Milnor studied mainly the case were P is an isolated singularity of $f^{-1}(0)$, i.e., an isolated point of X. In this case he proved that $F_f(P)$ has the homotopy type of a bouquet of spheres of real dimension $\dim_{\mathbb{C}} M - 1$. The number of spheres, which can be expressed as

$$\mu_f(P) = (-1)^{\dim M}\big(1 - \chi^{\mathrm{top}} F_f(P)\big)\,, \tag{3.3}$$

is called the *Milnor number*, and Milnor proved that it is equal to the dimension of $\mathscr{O}_{X,P}$, i.e., the multiplicity of P as a point on the scheme X. In local coordinates $x_1, \ldots, x_n$ on M near P, this is the dimension of

$$\mathbb{C}[[x_1, \ldots, x_n]]/(\partial_1 f, \ldots, \partial_n f)\,,$$

which is finite, as f has an isolated critical point at $x = 0$.

When X has positive dimension, so that P is not an isolated critical point of f, the expression (3.3) is still a well-defined integer. Just as in the case of an isolated singularity, the number $\mu_f(P)$ depends only on the scheme structure of X near P (not on the function f whose critical scheme X is), and it varies in a constructible fashion over X. It is a consequence of the *microlocal index theorem* of Kashiwara and MacPherson (and the fact that the characteristic variety of the perverse sheaf of vanishing cycles is equal to the normal cone of the critical set) that, if $\mathrm{Crit}\, f$ is proper,

$$\#^{\mathrm{vir}}(\mathrm{Crit}\, f) = \chi^{\mathrm{top}}(\mathrm{Crit}\, f, \mu_f)\,.$$

This is an analogue of formula (3.1), and proves the motivic nature of $\#^{\mathrm{vir}}(\mathrm{Crit}\, f)$, in the case $X = \mathrm{Crit}\, f$.

We see that in this case $\nu_X(P) = \mu_f(P)$, so the contribution of the point P to the virtual count is, up to sign, the reduced Euler characteristic of the Milnor fibre.

3.1.6. Homogeneous case. We explain one case where the generalized Milnor number $\nu_X(P) = \mu_f(P)$ of (3.3) is easy to determine. Consider a set of non-zero integers $r_1, \ldots, r_n \in \mathbb{Z}$, called *weights*, and assume we have a polynomial $f \in \mathbb{C}[x_1, \ldots, x_n]$ which is weighted homogeneous of degree 0, if we assign to x_i the weight r_i. Let $X \subset \mathbb{A}^n$ be the critical scheme of f, i.e., the affine scheme with affine coordinate ring $\mathbb{C}[x_1, \ldots, x_n]/(\partial_1 f, \ldots, \partial_n f)$. Let us further assume that $f \in (x_1, \ldots, x_n)^3$. This is not a serious restriction, if we are interested in the critical locus of f. It ensures that at the origin $P \in \mathbb{A}^n$, we have $T_X|_P = T_{\mathbb{A}^n}|_P$, and hence that $\dim T_X|_P = n$.

In this case the generalized Milnor number (3.3) is

$$\mu_X(P) = (-1)^n\,.$$

This is easy to see: consider the circle group $S^1 \subset \mathbb{C}^*$ and its action on $\mathbb{A}^n = \mathbb{C}^n$ given by $\theta \cdot (x_1, \ldots, x_n) = (\theta^{r_1} x_1, \ldots, \theta^{r_n} x_n)$. The Milnor fibre (3.2) is invariant under this S^1-action. Also, this S^1-action on the Milnor fibre is fixed point free. Therefore, the Euler characteristic of the Milnor fibre vanishes.

This calculation has non-trivial applications, for example, it implies that for the Hilbert scheme of n points on a smooth scheme of dimension 3 (proper or not), the weighted Euler characteristic is, up to sign, equal to the topological Euler characteristic:

$$\chi^{\mathrm{top}}(\mathrm{Hilb}^n Y, \nu_{\mathrm{Hilb}^n Y}) = (-1)^n \chi^{\mathrm{top}}(\mathrm{Hilb}^n Y)\,. \tag{3.4}$$

3.2. Symmetric obstruction theories.

3.2.1. Lagrangian intersection case. In the Lagrangian intersection case, the tangent complex is endowed with extra structure, namely an *odd pairing*. Above, we constructed the quasi-isomorphism $\Theta = (\theta, -\theta^\vee) : \mathbb{T}_{\mathfrak{X}} \to \mathbb{T}_{\mathfrak{X}}^\vee[-1]$. It has the property that $\Theta^\vee[-1] = -\Theta$. This means that it can be equivalently thought of as an alternating pairing of degree -1

$$\mathbb{T}_{\mathfrak{X}} \otimes \mathbb{T}_{\mathfrak{X}} \longrightarrow \mathscr{O}_X[-1]\,,$$

or a global section of $(\Lambda^2 \mathbb{T}_{\mathfrak{X}}^\vee)[-1]$. (This latter class is the classical shadow a (-1)-shifted symplectic structure on the derived scheme $\mathfrak{X}$ underlying this Lagrangian intersection. The closedness is not seen at the classical level.) The second exterior power of Θ is an isomorphism

$$\Lambda^2\Theta : \Lambda^2 \mathbb{T}_{\mathfrak{X}} \longrightarrow \Lambda^2(\mathbb{T}_{\mathfrak{X}}^\vee[-1]) = \mathrm{Sym}^2\, \mathbb{T}_{\mathfrak{X}}^\vee[-2]\,,$$

via which we can turn the pairing $\Lambda^2\mathbb{T}_{\mathfrak{X}} \to \mathscr{O}_X[-1]$ defined by Θ into a pairing $\mathrm{Sym}^2\,\mathbb{T}_{\mathfrak{X}}^\vee \to \mathscr{O}_X[1]$. This symmetric pairing is called the *symmetric obstruction theory* defined by the Lagrangian intersection $X = L \cap M$. (In the literature, it is usually $\mathbb{T}_{\mathfrak{X}}^\vee$ which is called the obstruction theory, not $\mathbb{T}_{\mathfrak{X}}$.)

3.2.2. General case. This leads us to the following definition. A perfect obstruction theory $\mathbb{T}_X \to F$ is *symmetric*, if F is endowed with a non-degenerate alternating pairing of degree -1. (Equivalently, $F^\vee$ is endowed with a non-degenerate symmetric pairing of degree $+1$, and this is where the name comes from.) This means that we are given an isomorphism $\Theta : F \to F^\vee[-1]$, such that $\Theta^\vee[-1] = -\Theta$.

In the toy model case $X = \mathrm{Crit}\, f$, the perfect obstruction theory is given by the *Hessian* of f, this is a self dual map $H(f) : T_M|_X \to \Omega_M|X$. In fact, the obstruction theory is

$$F = [T_M|_X \xrightarrow{-H(f)} \Omega_M|_X] \ .$$

Its shifted dual is

$$F^\vee[-1] = [T_M|_X \xrightarrow{H(f)} \Omega_M|_X] \ ,$$

and so $\Theta = (\mathrm{id}, -\mathrm{id}) : F \to F^\vee[-1]$ defines the required pairing.

In the general case, just like in the Lagrangian intersection case, the obstruction sheaf is always canonically isomorphic to the sheaf of differentials: $h^1(F) = \Omega_X$.

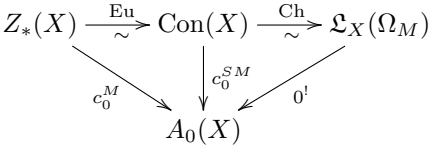
We will now explain what the presence of this additional structure says about the virtual fundamental class.

3.2.3. Almost closed 1-forms. In the non-symplectic case, there is no difference between the toy model for an intersection, and the toy model for a perfect obstruction theory. Every intersection of smooth varieties (étale or analytically) locally looks like the zero set of a section of a vector bundle on a smooth variety, and every perfect obstruction theory is locally isomorphic to the perfect obstruction theory given by a such a toy model.

In the present case this is no longer true. Every Lagrangian intersection looks locally like the critical scheme of a holomorphic function (by the holomorphic Darboux theorem), but this is not true for symmetric obstruction theories.

The local model for a scheme with symmetric obstruction theory is a non-singular variety with an *almost closed* 1-form on it. A 1-form on a non-singular scheme M is *almost closed*, if $d\omega$ vanishes in $\Omega^2_M|_X$, where X is the zero locus of ω. In other words, the equations $\partial_i f_j = \partial_j f_i$, saying that the 1-form $\omega = \sum f_i dx_i$ is closed, have to be satisfied only modulo the ideal (f_i). Almost closed 1-forms give rise to symmetric obstruction theories on their vanishing loci, and every scheme with symmetric obstruction theory locally comes from an almost closed 1-form. There are examples of almost closed 1-forms, whose associated symmetric obstruction theory does not admit a description as the symmetric obstruction theory of a critical set [23].

3.2.4. Microlocal geometry. Suppose X is a scheme embedded as a closed subscheme in the smooth scheme M. Microlocal geometry provides us with a commutative diagram



Here $Z_*(X)$ is the group of cycles on X, it is the free abelian group generated by the irreducible closed subvarieties (prime cycles) of X. The group $\mathrm{Con}(X)$ is the abelian group of all $\mathbb{Z}$-valued constructible functions on X. On the right, $\mathfrak{L}_X(\Omega_M)$ is the group of conic Lagrangian cycles on the cotangent bundle of M, which lie over X. The homomorphism Eu is MacPherson's local Euler obstruction, and Ch is the characteristic cycle map. The downward maps are the degree 0 Chern-Mather class, the degree 0 Schwartz-MacPherson Chern class, and the intersection with the zero section (Gysin map), respectively. The left hand triangle does not depend on the embedding into M.

The easiest of the horizontal maps to describe explicitly is the composition $L = \mathrm{Ch} \circ \mathrm{Eu}$. It maps a prime cycle V in X to the closure of the conormal bundle inside Ω_M of any non-singular open subset of V, multiplied by $(-1)^{\dim V}$.

The next easiest is the inverse of Ch. Let P be a point in X. Choose a Euclidean distance function from P in the complex manifold M, denote its square by ρ, and let Δ be the graph of $d\rho$ inside Ω_M. Then the value of the constructible function $\mathrm{Ch}^{-1}([C])$ at P is

$$\mathrm{Ch}^{-1}([C])(P) = I_{\{P\}}([C], [\Delta])\,,$$

the topological intersection number of the conic Lagrangian cycle $[C]$ with $[\Delta]$ at P. This is well-defined, because P is an isolated point of the intersection $C \cap \Delta$.

Thus, by composition, we have also described $\mathrm{Eu} = \mathrm{Ch}^{-1} \circ L$.

3.2.5. The distinguished cycle and its Euler obstruction. Let $C_{X/M}$ be the normal cone of X in M, and $\{C_i\}$ its irreducible components with their multiplicities r_i. Let $\mathfrak{c}_i$ be the prime cycle in X obtained as the image of C_i under the projection $C_{X/M} \to X$. Form the cycle

$$\mathfrak{c}_X = \sum (-1)^{\dim \mathfrak{c}_i} r_i \mathfrak{c}_i$$

on X. It is called the *distinguished* cycle of X. It is a standard fact that $\mathfrak{c}_X$ is intrinsic to X: it does not depend on the chosen embedding $X \hookrightarrow M$.

Define[1] $\nu_X = \mathrm{Eu}(\mathfrak{c}_X)$. This is a constructible function on X, which is intrinsic to X. The value $\nu_X(P)$ of ν_X at the point $P \in X$ only depends on an analytic neighbourhood of P in X, it is an invariant of the singularity of X at the point P. When not mentioned otherwise, the *weighted Euler characteristic* of a scheme is the weighted Euler characteristic with respect to the weight function ν_X.

If X is smooth near P, then $\nu_X(P) = (-1)^{\dim X}$. If X is the scheme-theoretic critical locus of an analytic function f on M near P, then $\nu_X(P) = \mu_f(P)$.

(Even in the presence of a symmetric obstruction theory, it is not true that $\nu_X(P)$ is always a generalized Milnor number for a function f, such that $X = \mathrm{Crit}(f)$ near P, by the above mentioned [23]. For such cases $\mu_f(P)$ is not defined, so in [2], we prove a

[1]This function ν_X is sometimes referred to as the *Behrend function* in the literature on Donaldson-Thomas theory.

formula expressing $\nu_X(P)$ as a linking number, which is similar in spirit to the Milnor fibre definition, and always applies.)

3.2.6. The main theorem. [2] Now suppose that X is endowed with a symmetric obstruction theory. The embedding $X \hookrightarrow M$ provides us with an epimorphism $\Omega_M|_X \twoheadrightarrow \Omega_X$ from a vector bundle to the obstruction sheaf. The obstruction theory then gives us, as explained in 2.3, an obstruction cone $C \subset \Omega_M|_X \subset \Omega_M$, and the virtual fundamental class is $[X]^{\mathrm{vir}} = 0^![C]$.

The obstruction cone is locally (as a cone scheme over X) isomorphic to the normal cone $C_{X/M}$. The local isomorphisms are given by locally existing almost closed 1-forms cutting out X in M.

The key fact is that the conic cycle $[C]$ in Ω_M is Lagrangian. This is a local calculation, so we can reduce to the case where X is the zero locus of an almost closed 1-form ω on M, and the symmetric obstruction theory is given by ω, too. The graph of ω is not Lagrangian in Ω_M, as ω is not closed, but multiplying the graph by a scalar, end letting the scalar go to infinity, we obtain C, which does turn out to be Lagrangian, as a conic cycle.

Therefore $[C]$ is obtained via $\mathrm{Ch} \circ \mathrm{Eu}$ from unique cycle in X. Because C is locally isomorphic to $C_{X/M}$, this cycle in X can only be the distinguished cycle $\mathfrak{c}_X$. We conclude that $[C] = \mathrm{Ch}(\nu_X)$, and

$$[X]^{\mathrm{vir}} = 0^![C] = c_0^{SM}(\nu_X) = c_0^M(\mathfrak{c}_X)\,.$$

So, just as the obstruction sheaf, the virtual fundamental class is intrinsic to X.

Now MacPherson's theorem, which generalizes the Gauß-Bonnet theorem to singular schemes (and is also equivalent to Kashiwara's microlocal index theorem), says that, if X is proper

$$\chi^{\mathrm{top}}(X, \nu) = \deg(c_0^{SM}\nu)\,,$$

for any constructible function ν on X. Applying this to our distinguished function ν_X, we get, if X is proper

$$\#^{\mathrm{vir}}(X) = \deg[X]^{\mathrm{vir}} = \deg(c_0^{SM}\nu_X) = \chi^{\mathrm{top}}(X, \nu_X)\,.$$

We conclude that the virtual count is equal to the weighted Euler characteristic, and hence satisfies the two motivic properties.

3.3. Discussion. Let us start by discussing a prototypical source of symmetric obstruction theories, which is not *a priori* an algebraic Lagrangian intersection.

3.3.1. Gauge theory. Let us continue with our above discussion of gauge theory. The additional ingredient we need, to move it into derived symplectic geometry is a symmetric bilinear pairing $\kappa : L \otimes L \to \mathbb{C}[-\ell]$ of degree $-\ell$, on our differential graded Lie algebra L. (The most important case is $\ell = 3$.) This pairing needs to be *cyclic*, which means, besides being symmetric, that

1. $\kappa(dx, y) + (-1)^{\deg x}\kappa(x, dy) = 0$,
2. $\kappa([x, y], z) = \kappa(x, [y, z])$.

Finally, κ needs to be non-degenerate. For the purposes of this superficial exposition, let us agree that this means that κ sets up a perfect pairing between L^i and $L^{\ell-i}$, for all i. It has

as immediate consequence that L exists entirely within the interval $[0, \ell]$. (In practice, this condition is too strong, it is almost never satisfied. Instead, one has to deal with κ which only induce perfect pairings on cohomology. This is, in fact, the source of many subtleties of derived symplectic geometry.)

In our standard example, where $L = C^\infty(Y, \Omega_Y^\bullet \otimes M_{n\times n})$, we need to assume that Y is oriented, so that it has a fundamental class $[Y]$. The pairing κ is then given by

$$\kappa(x, y) = \int_{[Y]} \mathrm{tr}(x \circ y)\,.$$

(The circle denotes the associative product on L.) We will ignore issues of topology on the infinite-dimensional L, and instead pretend that L is finite-dimensional.

The isomorphism of complexes $\kappa : (L, d^x) \to (L, d^x)^\vee[-\ell]$ over $MC(L)$ is invariant under the gauge action, and so descends to the moduli space X, and provides us with an isomorphism $\mathbb{T}_{\mathfrak{X}}[-1] \xrightarrow{\sim} \mathbb{T}_{\mathfrak{X}}[-1]^\vee[-\ell]$, hence $\mathbb{T}_{\mathfrak{X}} \xrightarrow{\sim} \mathbb{T}^\vee_{\mathfrak{X}}[2-\ell]$. As before, we need to truncate (L, d^x) into the interval $[1, \ell-1]$, if G acts with positive-dimensional stabilizer. If $\ell = 3$, we obtain $\mathbb{T}_{\mathfrak{X}}$ in the interval $[0, 1]$, together with an isomorphism $\mathbb{T}_{\mathfrak{X}} \xrightarrow{\sim} \mathbb{T}^\vee_{\mathfrak{X}}[-1]$, i.e., a symmetric obstruction theory.

In our example, the case $\ell = 3$ is the case where Y is a 3-manifold. A rigorous treatment would identify the corresponding virtual counts as *Casson invariants*. (Although, because of the technical difficulties of the gauge theoretic approach, the Casson invariant is usually treated differently.)

Let us continue discussing the case $\ell = 3$. The curvature map $F : L^1 \to L^2$ is now a map $F : L^1 \to (L^1)^\vee$, and can therefore be thought of as an algebraic differential form on the linear space L^1. The cyclicity of κ implies that this form is closed. It is therefore exact, and an antiderivative is easily written down:

$$f(x) = \tfrac{1}{2}\kappa(x, dx) + \tfrac{1}{6}\kappa(x, [x, x])\,.$$

This cubic function $f : L^1 \to \mathbb{C}$ is known as the *Chern-Simons* function. It satisfies $df = F$, and so the Maurer-Cartan locus $MC(L) = Z(F) = \mathrm{Crit}(f)$ is the critical locus of the Chern-Simons function.

It is almost true that $f : L^1 \to \mathbb{C}$ is invariant under the gauge action and induces a function $\tilde{f} : L^1/G \to \mathbb{C}$ of which the moduli space $X = MC(L)/G$ is the critical set. In fact, $\tilde{f}$ has values in $\mathbb{C}/\mathbb{Z}$, but this is sufficient to make sense of its critical locus.

Unfortunately, it is not straightforward to follow these arguments through in Kapranov's finite-dimensional model for moduli of local systems, because the cup product (unlike the wedge product) is not commutative on the level of cochains.

3.3.2. Donaldson-Thomas theory. Considering the algebraic geometry analogue of this gauge theory example leads to a holomorphic analogue of the Casson invariant, which is known as the *Donaldson-Thomas* invariant. If we are only interested in virtual counts, we can avoid all gauge theoretic complications, and directly construct a symmetric obstruction theory on the relevant moduli space of stable coherent sheaves on a fixed Calabi-Yau threefold Y. In fact, the perfect obstruction theory is essentially the same we mentioned above in 2.4.1, namely $\big(\tau_{[1,2]} R\pi_* \mathscr{H}om(\mathscr{E}, \mathscr{E})\big)[1]$. The symmetric structure is given by Serre duality, which implies that deformations, given by $\mathrm{Ext}^1(E, E)$ are dual to obstructions, which are given by $\mathrm{Ext}^2(E, E)$. Because Y is Calabi-Yau, its canonical sheaf, which would usually feature in Serre duality, is trivial.

As the virtual fundamental class comes from a symmetric obstruction theory, the associated virtual counts, i.e., the Donaldson-Thomas invariants, are of a motivic nature. This sets them apart from many other counting invariants for Calabi-Yau threefolds, and is the reason why they have been so intensely studied in recent years.

The simplest non-trivial example of a Donaldson-Thomas moduli space is the Hilbert scheme $\mathrm{Hilb}^n Y$. Using the above mentioned result (3.4), one can prove that for every projective Calabi-Yau 3-fold Y, the virtual count is, up to sign, equal to the Euler characteristic:

$$\#^{\mathrm{vir}}(\mathrm{Hilb}^n Y) = (-1)^n \chi^{\mathrm{top}}(\mathrm{Hilb}^n Y)\,.$$

This formula was first conjectured in [20].

3.3.3. Motivic invariants. Let $K(\mathrm{Var})$ be the Grothendieck group of varieties. It is generated as an abelian group by symbols $[X]$, for all finite type $\mathbb{C}$-schemes X, subject to the *scissor relations*, namely that, whenever $Z \hookrightarrow X$ is a closed immersion of schemes, we have $[X] = [Z] + [X/Z]$. The group $K(\mathrm{Var})$ is called the group of *motivic weights*, or sometimes simply the group of *motives*.

Mapping a scheme or variety X to $[X]$ defines the universal Euler characteristic. Every map from the category of schemes to an abelian group satisfying the scissor relations factors uniquely through $K(\mathrm{Var})$. So, since the virtual counts in Donaldson-Thomas theory behave somewhat like an Euler characteristic, it is tempting to try to construct, for every moduli space of sheaves X (compact or not) on a Calabi-Yau threefold Y, an element $DT(X)$ in $K(\mathrm{Var})$, such that $\chi^{\mathrm{top}}(DT(X)) = \chi^{\mathrm{top}}(X, \nu_X)$.

The main challenge is to replace the integer weights, given by the constructible function ν_X, by motivic weights. For this purpose, it turns out the symmetric obstruction theory is not sufficient. One has to construct locally Chern-Simons type functions and find motivic versions of μ_f, rather than ν_X. It is then natural to use motivic vanishing cycles [9] as motivic weights. This programme has been carried out by Kontsevich-Soibelman [18], in an even more ambitious context where sheaves on a Calabi-Yau threefold are replaced by objects of more general Calabi-Yau-three categories.

The simplest case is the Hilbert scheme. In [3] we write down the generating series

$$Z_Y(t) = \sum_{n=0}^{\infty} [\mathrm{Hilb}^n Y]_{\mathrm{mot}}\, t^n\,,$$

where Y is a Calabi-Yau threefold (compact or not), and $[\mathrm{Hilb}^n Y]_{\mathrm{mot}}$ denotes the motivic virtual count associated to the Hilbert scheme of n points on Y. For the case $Y = \mathbb{A}^3$, we have

$$Z_{\mathbb{A}^3}(t) = \prod_{m=1}^{\infty} \prod_{k=0}^{m-1} (1 - \mathbb{L}^{k+2-\frac{m}{2}} t^m)^{-1} = \mathrm{Exp}\Big(\frac{-\mathbb{L}^{\frac{3}{2}} t}{(1+\mathbb{L}^{\frac{1}{2}} t)(1+\mathbb{L}^{-\frac{1}{2}} t)}\Big)\,.$$

And for general Y

$$Z_Y(t) = \mathrm{Exp}\Big([Y] \frac{-\mathbb{L}^{-\frac{3}{2}} t}{(1+\mathbb{L}^{\frac{1}{2}} t)(1+\mathbb{L}^{-\frac{1}{2}} t)}\Big)\,.$$

(Here Exp is the motivic exponential.)

3.3.4. Categorification. The Euler characteristic of an algebraic variety X over $\mathbb{C}$ is given by

$$\chi^{\mathrm{top}}(X) = \sum_i (-1)^i \dim H^i(X^{\mathrm{top}}, \mathbb{C}) ,$$

the alternating sum of the dimensions of the cohomology groups of the topological space X^{top} (analytic topology) associated to X. We say that the cohomology spaces *categorify* the Euler characteristic. The terminology is justified by the fact that the cohomology lies in the *category* of vector spaces, rather the *set* of numbers. This is an important step, because it allows the full machinery of homological algebra to be applied to the calculation of Euler characteristics, and it gives much deeper information then the Euler characteristic alone.

The question arises, if one can write weighted Euler characteristics in a similar fashion. In the toy model case where $X = \mathrm{Crit}\, f$, for a function $f : M \to \mathbb{C}$ on a smooth scheme X, we have the perverse sheaf of vanishing cycles Φ_f on X, and

$$\chi^{\mathrm{top}}(X, \nu_X) = \chi^{\mathrm{top}}(X, \mu_f) = \sum_i (-1)^i \dim H^i(X^{\mathrm{top}}, \Phi_f) .$$

So we say that the perverse sheaf of vanishing cycles categorifies the virtual number of critical points of f.

Recently [6, 15], global versions of Φ_f on moduli spaces of sheaves have been constructed, thus categorifying Donaldson-Thomas theory in this sense. Gluing the locally defined sheaves of vanishing cycles requires an *orientation* on the moduli space.

3.3.5. Orientation. Even for the case of Lagrangian intersections, the question of orientation is non-trivial. For the derived scheme $\mathfrak{X}$, defined by the Lagrangian intersection $L \cap M$ inside S we have

$$\det \mathbb{T}_{\mathfrak{X}} = \det T_L|_X \otimes \det T_M|_X \otimes (\det T_S|_X)^{-1} = \det T_L|_X \otimes \det T_M|_X ,$$

because $\det \Omega_S = \Lambda^{2n}\Omega_S$ is trivial (the n-th power of the symplectic form trivializes it). So for the canonical line bundle, we have

$$K_{\mathfrak{X}} = K_L|_X \otimes K_M|_X .$$

A line bundle Υ on X, such that $\Upsilon^{\otimes 2} = K_{\mathfrak{X}}$ is called an *orientation* of the Lagrangian intersection. For example, if $K_L|_X = K_M|_X$, then the intersection is canonically oriented.

The systematic study of motivic invariants and categorification requires a thorough understanding of orientations.

3.4. Derived symplectic geometry. We will say a few words about the classical shadows of a derived symplectic scheme. Then we will discuss the toy model in more depth.

3.4.1. Tangent complex. On the level of tangent complexes, a symplectic structure on a derived scheme $\mathfrak{X}$ will induce an isomorphism

$$\theta : \mathbb{T}_{\mathfrak{X}} \longrightarrow \mathbb{T}^{\vee}_{\mathfrak{X}}[-n] \tag{3.5}$$

of some degree, which we have denoted by $-n$. As $\mathbb{T}_{\mathfrak{X}}$ is in degrees ≥ 0 and $\mathbb{T}^{\vee}_{\mathfrak{X}}$ in degrees ≤ 0, it follows that n is equal to the amplitude of $\mathfrak{X}$.

Moreover, θ is antisymmetric, i.e., satisfies $\theta = -\theta^\vee[-n]$.

For example, if the degree of θ is 0, then the complex $\mathbb{T}_\mathfrak{X}$ in the interval $[0, \ell]$ is quasi-isomorphic to the complex $\mathbb{T}_\mathfrak{X}^\vee$ in the interval $[-\ell, 0]$, which means that $\mathbb{T}_\mathfrak{X}$ is in fact concentrated in degree 0, and is simply a vector bundle. So if the symplectic structure θ is of degree $n = 0$, the derived scheme $\mathfrak{X}$ is a classical scheme X endowed with a classical symplectic structure.

If the degree of θ is -1, then the complex $\mathbb{T}_\mathfrak{X}$ in the interval $[0, \ell]$ is quasi-isomorphic to the complex $\mathbb{T}_\mathfrak{X}^\vee[-1]$ in the interval $[1-\ell, 1]$, which forces $\ell \leq 1$, so that $\mathfrak{X}$ is quasi-smooth, and θ is nothing but a symmetric obstruction theory on the underlying classical scheme X, as discussed above.

3.4.2. Higher structure sheaves. On the level of higher structure sheaves, a symplectic structure on $\mathfrak{X}$ will induce an analogue of the Poisson bracket. This will be a bracket $\{\,,\} : \pi_k(\mathscr{O}_\mathfrak{X}) \otimes_\mathbb{C} \pi_\ell(\mathscr{O}_\mathfrak{X}) \to \pi_{k+\ell-n}(\mathscr{O}_\mathfrak{X})$, which is a derivation with respect to the commutative product in each argument, and satisfies the graded Jacobi identity.

In the case of Lagrangian intersections, this bracket was constructed in [5]. It was then discovered [1], that this bracket comes naturally out of considering *deformation quantization up to first order*.

3.4.3. Quantization. To explain the approach to categorification via quantization, we cannot get by with our classical shadows any longer. So let us finish by discussing a toy model for a derived symplectic scheme. We restrict to the *odd* case.

We take the point of view that a derived scheme is a graded manifold with a homological vector field on it (in other words, a differential graded manifold). Thus, let $V = V^0 \oplus \ldots \oplus V^n$ be a finite-dimensional graded vector space which we think of as a linear graded manifold (so that the graded tangent space at every point is equal to V). The algebra of functions on V is $\operatorname{Sym} V^\vee$, the graded symmetric algebra generated by the dual $V^\vee$ of V. On $\operatorname{Sym} V^\vee$ we have a derivation $Q : \operatorname{Sym} V^\vee \to \operatorname{Sym} V^\vee$, of degree $+1$, which satisfies $[Q, Q] = 0$, or equivalently $Q \circ Q = 0$. The derived scheme is $\mathfrak{X} = (V, Q)$.

For example, if we consider the gauge theory context, and discard L^0 (and the gauge group), as well as L^ℓ (where ℓ is the top degree), we may take $V = L^{[1,\ell-1]}[1]$. The derivation Q corresponds to the vector field given by the algebraic map $x \mapsto dx + \frac{1}{2}[x, x]$, from V to V. (In general, Q may have higher order terms: this makes L an L_∞-algebra, instead of a differential graded Lie algebra.)

The classical shadows of $\mathfrak{X}$ are as following: the underlying classical scheme is $\operatorname{Spec} h^0(\operatorname{Sym} V^\vee, Q)$. The amplitude is n. The virtual dimension is $\sum_i (-1)^i \dim V^i$. The tangent complex is the trivial graded vector bundle with fibre V, endowed with the differential which is given by the derivative of Q, thought of as an algebraic map $Q : V \to V$. The higher structure sheaves are $\pi_i(\mathscr{O}_\mathfrak{X}) = h^{-i}(\operatorname{Sym} V^\vee, Q)$.

Given a non-degenerate alternating pairing $\kappa \in \Lambda^2 V^\vee\,[-n]$ of degree $-n$, we have an induced isomorphism $\kappa : V \xrightarrow{\sim} V^\vee\,[-n]$, such that $\kappa^\vee[-n] = -\kappa$. Let us assume that it commutes with Q. Then we have a symplectic structure on (V, Q). (The 2-form of degree $-n$ given by κ is closed, because it is constant.) In the gauge theory context, compatibility with Q is equivalent to cyclicity as defined in 3.3.1, above.

Let us now assume that n is odd. The second symmetric power of κ induces an isomor-

phism

$$(\mathrm{Sym}^2\,\kappa)[n] : \mathrm{Sym}^2\,V\,[n] \xrightarrow{\sim} \mathrm{Sym}^2(V^\vee[-n])\,[n] = \Lambda^2 V^\vee\,[-n]\ .$$

Taking the preimage of κ itself under this isomorphism, we obtain $\Delta \in \mathrm{Sym}^2\,V\,[n]$. We may interpret this as a map $\mathrm{Sym}^2\,V^\vee \to \mathbb{C}\,[n]$, which extends uniquely to a second order differential operator $\Delta : \mathrm{Sym}\,V^\vee \to \mathrm{Sym}\,V^\vee\,[n]$ of degree n, vanishing on $V^\vee$, called the *Batalin-Vilkovisky* operator.

We can also consider $\Delta : \mathrm{Sym}^2\,V^\vee \to \mathbb{C}[n]$ as a symmetric pairing, which extends uniquely to a symmetric biderivation $\{\,,\,\} : \mathrm{Sym}\,V^\vee \otimes \mathrm{Sym}\,V^\vee \to \mathbb{C}[n]$. This is the *Poisson bracket* of the symplectic structure κ. (It also exists in the case that n is even, although it is anti-symmetric in this case.) The BV operator Δ generates the Poisson bracket in the sense that

$$\Delta(xy) - (-1)^x x\,\Delta(y) - \Delta(x)\,y = \{x, y\}\,.$$

Let us now restrict to the case $n = 1$, so that $V = V^0 \oplus V^1$, and κ identifies V^1 with $(V^0)^\vee$. Then $\mathrm{Sym}\,V^\vee$ is identified with the algebra of polyvector fields on V^0, the Poisson bracket with the Schouten bracket, and Δ with the divergence operator associated to the linear structure on V. (Via the identification of polyvector fields with differentials given by a volume form compatible with the linear structure, Δ corresponds to the de Rham differential.)

As in the above gauge theory example, there exists a Chern-Simons map $f \in \mathrm{Sym}\,V^\vee$ of degree 0, such that $Q = \{f, \cdot\}$. The only difference is that f may have higher order terms. We may think of f as a polynomial function $f : V^0 \to \mathbb{C}$.

As Δ satisfies $\Delta^2 = 0$, it defines a differential on $\mathrm{Sym}\,V^\vee$, which commutes with Q. To obtain a double complex from the two commuting differential Q and Δ, we we introduce a formal variable $\hbar$, and pass to $\mathrm{Sym}\,V^\vee(\!(\hbar)\!)$, with differential $\Delta + \hbar Q$. Then it follows from a theorem recently proved by Sabbah [25], that for the generalized Milnor number of f at the origin $0 \in V^0$, we have

$$\mu_f(0) = \sum (-1)^k \dim_{\mathbb{C}(\!(\hbar)\!)} H^k\big(\mathrm{Sym}\,V^\vee(\!(\hbar)\!), Q + \hbar\Delta\big)$$

Thus the $\mathbb{C}(\!(\hbar)\!)$-vector spaces $H^k\big(\mathrm{Sym}\,V^\vee(\!(\hbar)\!), Q + \hbar\Delta\big)$ categorify the Donaldson-Thomas type virtual count.

Currently there is ongoing research to generalize this toy model to actual moduli spaces. As mentioned, the main difficulty is that many properties of κ, such as the non-degeneracy, or the compatibility with Q, are only satisfied up to homotopy (unless one is willing to work in infinite dimensions, which has its own problems). See [24].

3.4.4. Lagrangian intersections. For the case of Lagrangian intersections, the categorification via quantization was achieved by Kashiwara-Schapira [14], see also [8].

References

[1] V. Baranovsky and V. Ginzburg, *Gerstenhaber-Batalin-Vilkovisky structures on coisotropic intersections*, Math. Res. Lett., **17**(2) (2010), 211–229.

[2] K. Behrend, *Donaldson-Thomas type invariants via microlocal geometry*, Ann. of Math. (2), **170**(3) (2009), 1307–1338.

[3] K. Behrend, J. Bryan, and B. Szendrői, *Motivic degree zero Donaldson-Thomas invariants*, Invent. Math., **192**(1) (2013), 111–160.

[4] K. Behrend and B. Fantechi, *The intrinsic normal cone*, Invent. Math., **128**(1) (1997), 45–88.

[5] ———, *Gerstenhaber and Batalin-Vilkovisky structures on Lagrangian intersections*, In Algebra, arithmetic, and geometry: in honor of Yu. I. Manin. Vol. I, volume 269 of Progr. Math., pp. 1–47. Birkhäuser Boston, Inc., Boston, MA, 2009.

[6] V. Bussi, D Joyce, and S. Meinhardt, *On motivic vanishing cycles of critical loci*, arXiv: 1305.6428 [math.AG].

[7] I. Ciocan-Fontanine and M. Kapranov, *Virtual fundamental classes via dg-manifolds*, Geom. Topol., **13**(3) (2009), 1779–1804.

[8] A. D'Agnolo and M. Kashiwara, *On quantization of complex symplectic manifolds*, Comm. Math. Phys., **308**(1) (2011), 81–113.

[9] J. Denef and F. Loeser, *Geometry on arc spaces of algebraic varieties*, In European Congress of Mathematics, Vol. I (Barcelona, 2000), volume 201 of Progr. Math., pp. 327–348. Birkhäuser, Basel, 2001.

[10] S. K. Donaldson, *Polynomial invariants for smooth four-manifolds*, Topology, **29**(3) (1990), 257–315.

[11] W. Fulton, *Intersection Theory*, volume 2 of Ergebnisse der Mathematik und ihrer Grenzgebiete (3), Springer-Verlag, Berlin, 1984.

[12] R. Hartshorne, *Algebraic Geometry*, Graduate Texts in Mathematics, No. 52. Springer-Verlag, New York, 1977.

[13] M. Kapranov, *Injective resolutions of BG and derived moduli spaces of local systems*, J. Pure Appl. Algebra, **155**(2-3) (2001), 167–179.

[14] M. Kashiwara and P. Schapira, *Constructibility and duality for simple holonomic modules on complex symplectic manifolds*, Amer. J. Math., **130**(1) (2008), 207–237.

[15] Y.-H. Kiem and J. Li, *Categorification of Donaldson-Thomas invariants via perverse sheaves*, arXiv:1305.6428 [math.AG].

[16] M. Kontsevich, *Enumeration of rational curves via torus actions*, In The moduli space of curves (Texel Island, 1994), volume 129 of Progr. Math., pp. 335–368. Birkhäuser, Boston, 1995.

[17] M. Kontsevich and Yu. Manin, *Gromov-Witten classes, quantum cohomology, and enumerative geometry*, Comm. Math. Phys., **164** (1994), 525–562.

[18] M. Kontsevich and Y. Soibelman, *Motivic Donaldson-Thomas invariants: summary of results*, In Mirror symmetry and tropical geometry, volume 527 of Contemp. Math., pp. 55–89. Amer. Math. Soc., Providence, RI, 2010.

[19] J. Li and G. Tian, *Virtual moduli cycles and Gromov-Witten invariants of algebraic varieties*, J. Amer. Math. Soc., **11**(1) (1998), 119–174.

[20] D. Maulik, N. Nekrasov, A. Okounkov, and R. Pandharipande, *Gromov-Witten theory and Donaldson-Thomas theory. I*, Compos. Math., **142**(5) (2006), 1263–1285.

[21] J. Milnor, *Singular points of complex hypersurfaces*, Annals of Mathematics Studies, No. 61. Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1968.

[22] T. Mochizuki, *Donaldson type invariants for algebraic surfaces*, volume 1972 of Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2009. Transition of moduli stacks.

[23] R. Pandharipande and R. P. Thomas, *Almost closed 1-forms*, Glasg. Math. J., **56**(1) (2014), 169–182.

[24] T. Pantev, B. Toën, M. Vaquié, and G. Vezzosi, *Shifted symplectic structures*, arXiv: 1111.3209 [math.AG], 2011.

[25] C. Sabbah, *On a twisted de Rham complex II*, arXiv:1012.3818 [math.AG].

[26] R. P. Thomas, *A holomorphic Casson invariant for Calabi-Yau 3-folds, and bundles on $K3$ fibrations*, J. Differential Geom., **54**(2) (2000), 367–438.

1984 Mathematics Road, Vancouver B.C., Canada V6T 1Z2
E-mail: behrend@math.ubc.ca

Quasimap theory

Ionuţ Ciocan-Fontanine and Bumsig Kim

Abstract. We provide a short introduction to the theory of ε-stable quasimaps and its applications via wall-crossing to Gromov-Witten theory of GIT targets.

Mathematics Subject Classification (2010). Primary 14D20, 14D23, 14N35.

Keywords. GIT quotients, Quasimaps, Gromov-Witten Theory, Mirror Symmetry, Cohomological Field Theory, Gauged Linear σ-models.

1. Introduction

This note is intended as a brief survey of the theory of quasimaps from curves to a certain (large) class of GIT quotients, and of its applications to Gromov-Witten theory, as developed in the papers [6, 10–14]. The theory may be viewed as an algebro-geometric realization of Witten's Gauged Linear σ-model (GLSM) [52] in the geometric phases. The study of GLSM and of its relation to Mirror Symmetry has been a very active area in String Theory, see [1, 22, 31, 32, 34, 42] for a (very incomplete) sampling of developments.

When such a geometric phase (a target with a GIT presentation) is fixed, there is a family of quasimap theories indexed by a stability parameter $\varepsilon \in \mathbb{Q}_{>0}$. When $\varepsilon > 1$ one recovers the "nonlinear σ-model", i.e., the Gromov-Witten theory of the target. There is a wall-and-chamber structure on $\mathbb{Q}_{>0}$, with walls at $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{d}, \ldots$, such that the theory stays unchanged in the chamber $(\frac{1}{d+1}, \frac{1}{d}]$. Wall-crossing formulas relating the invariants in different chambers of nonsingular targets are conjectured (and are established in many cases) in [11] for genus zero, and in [12] for all genera; the genus zero case is extend to orbifolds in [6]. These results are described in §4-5 of the paper. As explained there, the wall-crossing formulas may be viewed as generalizations in many directions of Givental's Mirror Theorems [26] for (complete intersections in) toric manifolds with semi-positive anti-canonical class. In addition, the *mirror map* is given a geometric interpretation as the generating series of primary quasimap invariants with a fundamental class insertion.

There is also an extension of the theory in a different direction, allowing the domain curves of quasimaps to carry weighted markings. When (some) markings are given infinitesimally small weights, this produces for many targets a *closed form* expression of a "big I-function" defined on the entire parameter space $H^*(X, \mathbb{Q})$ associated to the GIT target X. By a result in [13] the big J-function of the Gromov-Witten theory of X is obtained from this new big I-function via the "Birkhoff factorization" procedure of [18]. As a result, one obtains an explicit determination of all the genus zero Gromov-Witten invariants of X.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. Maps from curves to quotient targets

2.1. A class of GIT quotients. Let $W = \mathrm{Spec}A$ be an affine algebraic variety over $\mathbb{C}$ and let $\mathbf{G}$ be a reductive algebraic group over $\mathbb{C}$, acting upon W from the right. Choose a character of $\mathbf{G}$, $\theta \in \mathrm{Hom}(\mathbf{G}, \mathbb{C}^*)$. Denote by $\mathbb{C}_\theta$ the associated 1-dimensional $\mathbf{G}$-representation space. This determines a $\mathbf{G}$-equivariant line bundle $L_\theta := W \times \mathbb{C}_\theta$ on W.

There are four quotients as follows.

(i) The *affine quotient* $W/\!\!/_{\mathrm{aff}}\mathbf{G} := \mathrm{Spec}(A^{\mathbf{G}})$, which is of finite type over $\mathbb{C}$ by Hilbert's Theorem.

(ii) The *stack quotient* $[W/\mathbf{G}]$ (see [20, 40]). One departure when working with algebraic stacks versus working with schemes is that algebraic stacks are groupoid-valued functors from the category of schemes, while schemes are set-valued functors. By the Yoneda lemma, the category $(\mathrm{Sch}/\mathbb{C})$ of $\mathbb{C}$-schemes is embedded into the category of functors of points. For schemes $X, Y \in (\mathrm{Sch}/\mathbb{C})$, the set of Y-points of X is $\mathrm{Hom}_{(\mathrm{Sch}/\mathbb{C})}(Y, X)$, i.e., the set of all morphisms from Y to X over $\mathbb{C}$. The stack $[W/\mathbf{G}]$ can be considered as a functor from $(\mathrm{Sch}/\mathbb{C})$ to the category of groupoids, defined as follows. A morphism from Y to $[W/\mathbf{G}]$ is by definition a triple $(Y, P, \tilde{f})$, where P is a principal $\mathbf{G}$-bundle on Y (which is trivializable in the étale topology of Y) and $\tilde{f} : P \to W$ is a $\mathbf{G}$-equivariant morphism. Equivalently, it is a triple (Y, P, f), with f a section of the induced fiber bundle $P \times_{\mathbf{G}} W \to Y$ with fiber W. An isomorphism from (Y, P, f) to (Y, P', f') is a $\mathbf{G}$-bundle homomorphism $\varphi : P \to P'$ such that $f' \circ \varphi = f$. Suppose that $Y = \mathrm{Spec}\mathbb{C}$; then the $\mathbb{C}$-points of $[W/\mathbf{G}]$ form a groupoid, the collection of orbits with the isomorphisms described above. A $\mathbb{C}$-point has non-trivial automorphisms if and only if the corresponding $\mathbf{G}$-orbit in Y has non-trivial stabilizer group.

Consider the trivial $\mathbf{G}$-bundle $W \times \mathbf{G}$ on W. It comes with the $\mathbf{G}$-equivariant map $W \times \mathbf{G} \to W$ given by the action. This gives a canonical morphism from W to $[W/\mathbf{G}]$, fitting in the cartesian diagram

$$\begin{array}{ccc} P & \xrightarrow{f} & W \\ \downarrow & & \downarrow \\ Y & \longrightarrow & [W/\mathbf{G}]. \end{array}$$

The geometry of $[W/\mathbf{G}]$ is encoded by the "atlas" $W \to [W/\mathbf{G}]$.

(iii) The *GIT quotient* $W/\!\!/\mathbf{G} := W/\!\!/_\theta\mathbf{G} := \mathrm{Proj}(\oplus_{n\geq 0}\Gamma(W, L_\theta^{\otimes n})^{\mathbf{G}})$. This is a quasi-projective scheme, equipped with a canonical *projective* morphism to $W/\!\!/_{\mathrm{aff}}\mathbf{G}$. It is called the *Proj quotient in direction* θ in [43, §6.13].

(iv) The *GIT stack quotient* $[W^{ss}/\mathbf{G}]$. This is an open substack of $[W/\mathbf{G}]$ since $W^{ss} := \{p \in W : s(p) \neq 0 \text{ for some } n > 0, s \in \Gamma(W, L_\theta^n)^{\mathbf{G}}\}$ is a $\mathbf{G}$-invariant open subset of W.

Assumption 2.1. $\mathbf{G}$ *acts on* W^{ss} *with at most finite stabilizers.*

Assumption 2.1 is equivalent to requiring that there are no strictly semi-stable points. Under this assumption, the GIT stack quotient $[W^{ss}/\mathbf{G}]$ is a Deligne-Mumford (DM) stack.

It follows that there is a natural commuting diagram of morphisms:

$$\begin{array}{ccc} [W^{ss}/\mathbf{G}] & \longrightarrow & [W/\mathbf{G}] \\ {\scriptstyle \text{proper}}\downarrow & & \downarrow \\ W/\!\!/\mathbf{G} & \xrightarrow[\text{projective}]{} & W/_{\text{aff}}\mathbf{G}. \end{array}$$

The left vertical morphism is proper, see e.g. [35].

2.2. Examples.

1. *Projective Spaces.* Let $\mathbf{G} = \mathbb{C}^*$ diagonally act on $V = \mathbb{C}^{n+1}$ and let $\theta = \mathrm{id}_{\mathbb{C}^*}$. Then
$$V/\!\!/\mathbf{G} = \mathbb{P}^n \subset [V/\mathbf{G}] = [\mathbb{C}^{n+1}/\mathbb{C}^*]$$
and L_θ restricted to $V/\!\!/\mathbf{G}$ is $\mathcal{O}(1)$.

   Note that the set of $\mathbb{C}$-points of $[V/G]$ contains one more element $[0, ..., 0]$ other than those in the projective space $\mathbb{P}^n$. This point has nontrivial automorphisms and is called a stacky point. Even when $n = 0$, the stack $[\mathbb{C}/\mathbb{C}^*]$ is interesting. This stack parameterizes pairs (L, s) with L a line bundle and s a section of L.
2. *Grassmannians.* Let $V = \mathrm{Hom}(\mathbb{C}^r, \mathbb{C}^n)$, $\mathbf{G} = GL(r, \mathbb{C})$ and $\theta = \det$. Then $V/\!\!/\mathbf{G} = Gr(r, n)$, the Grassmannian of r-planes in $\mathbb{C}^n$. A similar description works for a type A flag variety, see e.g. [5].
3. *Toric DM-stacks.* $V = \mathbb{C}^N$ and $\mathbf{G} = (\mathbb{C}^*)^r$; there are many choices of θ and the GIT quotient $[V^{ss}/\mathbf{G}]$ is a toric DM-stack.
4. *Complete Intersections.* Any projective variety $X \subset \mathbb{P}^{n-1}$ is a GIT quotient: $X = W/\!\!/\mathbb{C}^*$, with $W = C(X) \subset \mathbb{C}^n$, the affine cone over X, but only *complete intersections* lead to good theories (see Remark 3.3 below).
5. *Zero locus of regular sections of homogenous vector bundles.* Let $V, \mathbf{G}, \theta$ define a GIT quotient as in §2.1 and let E be a $\mathbf{G}$-representation with induced vector bundle $\mathcal{E} = V^{ss} \times_{\mathbf{G}} E$ on $V/\!\!/\mathbf{G}$. Let $s \in \Gamma(V, V \times E)^{\mathbf{G}}$ be regular with induced $\bar{s} \in \Gamma(V/\!\!/\mathbf{G}, \mathcal{E})$. If we set $W := Z(s) \subset V$ (note that W is lci), then $W/\!\!/\mathbf{G} = Z(\bar{s}) \subset V/\!\!/\mathbf{G}$. For example, complete intersections in toric varieties are obtained in this way, but there are many more non-abelian examples with indecomposable bundles $\mathcal{E}$ which are *not* complete intersection.

   According to Coates, Corti, Galkin, and Kasprzyk [17] who rework the Mori-Mukai classification of Fano 3-folds, every smooth Fano 3-fold can be realized as an example of this type. We remark that the Rødland's Pfaffian Calabi-Yau 3-fold and the determinantal Gulliksen-Negård Calabi-Yau 3-fold are also of this type (see [50, §2], [34, §5] respectively).
6. *Nakajima Quiver Varieties.* Nakajima quiver varieties ([47]) give a large class of typically quasi-projective only GIT quotients of the kind we are interested in, see [14, Example 6.3.2]. Particularly interesting such examples are certain Hilbert schemes of

points on non-compact surfaces. For example, let

$$V = \mathrm{Hom}(\mathbb{C}^n, \mathbb{C}^n)^{\oplus 2} \oplus \mathrm{Hom}(\mathbb{C}, \mathbb{C}^n) \oplus \mathrm{Hom}(\mathbb{C}^n, \mathbb{C}),$$
$$W := \{(A, B, i, j) \in V \mid [A, B] + ij = 0\},$$
$$\mathbf{G} = GL(n, \mathbb{C}), \text{ and } \theta = \det .$$

Then $W/\!\!/\mathbf{G} = \mathrm{Hilb}_n(\mathbb{C}^2)$ and $W/_{\mathrm{aff}}\mathbf{G} = \mathrm{Sym}^n(\mathbb{C}^2)$. This is the well-known ADHM presentation of the Hilbert scheme of points in the plane.

More generally, let $\Gamma \subset SL(2, \mathbb{C})$ be a finite subgroup. Let

$$X := \Gamma\text{-Hilb}(\mathbb{C}^2) := \{Z \subset \mathbb{C}^2 : \mathcal{O}_Z \cong \mathbb{C} \cdot \Gamma\}.$$

It is the crepant resolution of $\mathbb{C}^2/\Gamma$. Using an appropriate Fourier-Mukai functor $\Phi : D(X) \to D^\Gamma(\mathbb{C}^2)$, the Hilbert scheme $\mathrm{Hilb}_n(X)$ can be realized as the Nakajima quiver variety associated to the framed affine Dynkin diagram with a certain King's stability condition, see [37], [48].

7. *Local Targets.* Let $V, \mathbf{G}, \theta$ define a projective GIT quotient and let E be a $\mathbf{G}$-representation space, with an induced vector bundle $\mathcal{E} = V^{ss} \times_{\mathbf{G}} E$ on $V/\!\!/\mathbf{G}$. Assume E is a sum of $\mathbb{C}_{k_i\theta}$ for some negative integers $k_1, \ldots, k_r$. If $W := V \times E$, then θ gives a linearization and $W/\!\!/\mathbf{G}$ is the total space of $\mathcal{E}$ over $V/\!\!/\mathbf{G}$. Again, it is only quasi-projective. These are usually called *local targets* in Gromov-Witten theory.
8. $SU_C(2, L)$. Let C be a nonsingular projective curve. Then the moduli space of rank 2 stable vector bundles on C with an odd determinant L, $\deg L \geq 4g(C) - 1$ is realized as the GIT quotient of an affine variety by a general linear group (see [43, Theorem 10.1]).

2.3. Moduli of maps to the stack quotient. To keep the presentation simple, from now on we assume that *the* $\mathbf{G}$*-action on* W^{ss} *is free.* The general case is referred to [6].

Let $(C, p_1, \cdots, p_k)$ be a pointed, genus g prestable curve, i.e., C is a connected projective curve at worst with nodal singularities, p_i are ordered nonsingular points of C, and the arithmetic genus of C is g.

As explained, a map $C \xrightarrow{[u]} [W/\mathbf{G}]$ is described by the data

$$\big((C, \{p_i\}), P, u\big)$$

with P a principal $\mathbf{G}$-bundle on C and u a section of the induced W-bundle $P \times_{\mathbf{G}} W \xrightarrow{\rho} C$. Any such $[u] : C \to [W/\mathbf{G}]$ defines

$$\beta \in \mathrm{Pic}([W/\mathbf{G}])^\vee = \mathrm{Hom}_{\mathbb{Z}}(\mathrm{Pic}^{\mathbf{G}}(W), \mathbb{Z}), \quad \beta(L) = \deg_C \mathcal{L},$$

where $\mathcal{L} := u^*(P \times_{\mathbf{G}} L)$ (a line bundle on C). This β is called the *numerical class* of the triple $\big((C, \{p_i\}), P, u\big)$.

Consider the moduli stack $\mathfrak{M}_{g,k}([W/\mathbf{G}], \beta)$ parametrizing all $\big((C, \{p_i\}), P, u\big)$ as above. It is a non-separated Artin stack of infinite type. We describe here its obstruction theory. Consider the natural morphisms:

$$\begin{array}{ccc} \mathfrak{M}_{g,k}([W/\mathbf{G}], \beta) & \xrightarrow{\mu} & \mathfrak{B}un_{\mathbf{G}}^{g,k} \\ & \searrow^{\nu} & \downarrow \phi \\ & & \mathfrak{M}_{g,k} \end{array}$$

where

- $\mathfrak{M}_{g,k}$ is the moduli stack of prestable k-pointed curves of genus g;
- $\phi : \mathfrak{B}un_{\mathbf{G}}^{g,k} \longrightarrow \mathfrak{M}_{g,k}$ is the relative moduli stack of principal $\mathbf{G}$-bundles on the fibers of the universal curve $\mathfrak{C}_{g,k} \longrightarrow \mathfrak{M}_{g,k}$;
- μ and ν are the natural forgetful morphisms.

Both $\mathfrak{M}_{g,k}$ and $\mathfrak{B}un_{\mathbf{G}}^{g,k}$ are smooth Artin stacks and ϕ is a smooth morphism. It follows that the natural obstruction theory to consider is the μ-relative obstruction theory governing deformations of sections u. Over the open substack $M_{g,k}(W/\!\!/\mathbf{G}, \beta)$, this induces the usual *absolute* obstruction theory of maps to the GIT quotient. The stack $M_{g,k}(W/\!\!/\mathbf{G}, \beta)$ parameterizes the triples $\big((C, \{p_i\}), P, u\big)$ with C irreducible and the image of u contained in $P \times_{\mathbf{G}} W^{ss}$.

The above discussion suggests several natural questions to address:

1. The Kontsevich compactification $\overline{M}_{g,k}(W/\!\!/\mathbf{G}, \beta)$ is an open and closed substack of $\mathfrak{M}_{g,k}([W/\mathbf{G}], \beta)$. Using the linearization θ, can we impose *stability conditions* to single out other *Deligne-Mumford* open and closed substacks containing $M_{g,k}(W/\!\!/\mathbf{G}, \beta)$, and which are *proper* (over $W/\!\!/_{\mathrm{aff}}\mathbf{G}$)?
2. If in addition the restriction of obstruction theory is perfect, these substacks will have a virtual class, hence we get "numerical invariants" associated to the triple $(W, \mathbf{G}, \theta)$. When is the obstruction theory perfect?
3. Assuming the first two questions have been answered satisfactorily, how do the invariants change when varying the stability condition? Can one obtain explicit "wall-crossing" formulas?

In the rest of the paper we explain how quasimap theory provides some answers to the above questions. The first two questions are discussed in §3, while §4 and §5 deal with the wall-crossing phenomenon and its relation to Mirror Symmetry.

3. Quasimaps and ε-stability

3.1. Stable quasimaps.

Definition 3.1.

(i) $\big((C, \{p_i\}), P, u\big)$ is called a θ-**quasimap** (or simply quasimap) to $W/\!\!/\mathbf{G}$ if

$$\#\{u(C) \cap W^{us}\} < \infty,$$

where $W^{us} := W \setminus W^{ss}$. Hence, $C \overset{[u]}{\dashrightarrow} W/\!\!/\mathbf{G}$ is a rational map with finitely many base points.

(ii) A θ-quasimap is called **prestable**, if the base points are away from the nodes and markings of C.

For such a prestable quasimap and $x \in C$, define

$$\ell(x) := \mathrm{length}(\mathcal{O}_{x,C}/[u]^\sharp I_{[W^{us}/\mathbf{G}]}\mathcal{O}_{x,C}) \in \mathbb{Z}_{\geq 0}.$$

(iii) Fix $\varepsilon \in \mathbb{Q}_{>0}$. A prestable quasimap is called ε-**stable**, if

1. the line bundle $\omega_C(\sum p_i) \otimes \mathcal{L}_\theta^\varepsilon$ on C is ample, where $\mathcal{L}_\theta = u^*(P \times_{\mathbf{G}} L_\theta) = P \times_{\mathbf{G}} \mathbb{C}_\theta$.
2. $\varepsilon \ell(x) \leq 1$ for every nonsingular point $x \in C$.

There is also an "asymptotic" stability condition, obtained by requiring only the ampleness condition, but for *every* $\varepsilon \in \mathbb{Q}_{>0}$. We denote it by $\varepsilon = 0+$.

Denote by $\mathrm{Q}_{g,k}^\varepsilon(W/\!\!/\mathbf{G}, \beta)$ the moduli stack parameterizing ε-stable quasimaps of type (g, k, β).

Theorem 3.2 ([14]). *For every $\varepsilon \geq 0+$, $\mathrm{Q}_{g,k}^\varepsilon(W/\!\!/\mathbf{G}, \beta)$ is a DM stack with a natural proper morphism to the affine quotient. (In particular, if $W/\!\!/\mathbf{G}$ is projective, then $\mathrm{Q}_{g,k}^\varepsilon(W/\!\!/\mathbf{G}, \beta)$ is proper.)*

If W^{ss} is nonsingular and W has at worst lci singularities (necessarily in W^{us}), then the canonical obstruction theory on $\mathrm{Q}_{g,k}^\varepsilon(W/\!\!/\mathbf{G}, \beta)$ (relative to $\mathfrak{Bun}_{\mathbf{G}}^{g,k}$) is perfect.

From now on we assume the lci condition on W, so that $\mathrm{Q}_{g,k}^\varepsilon(W/\!\!/\mathbf{G}, \beta)$ carries a virtual fundamental class.

Remark 3.3.

1. The theory depends on the triple $(W, \mathbf{G}, \theta)$, not just on the geometric target $W/\!\!/\mathbf{G}$.
2. Assume $(g, k) \neq (0, 0)$.
   - If $\varepsilon > 1$ we get the Kontsevich stable maps to $W/\!\!/\mathbf{G}$; the obstruction theory is then perfect for *all* W (of course W^{ss} is assumed to be nonsingular). However, for $\varepsilon \leq 1$ the lci condition is necessary.
   - If $0 < \varepsilon \leq \frac{1}{\beta(L_\theta)}$, all lengths of base points are allowed and the domain curve has *no rational tails*. The asymptotic stability condition says that we are in this chamber for *all* β.
   - There are finitely many "chambers" $(\frac{1}{d+1}, \frac{1}{d}]$ such that the moduli spaces stay constant for $\varepsilon \in (\frac{1}{d+1}, \frac{1}{d}]$; intuitively, when crossing the wall we trade rational tails of degree d (with respect to $\mathcal{O}(\theta)$) with base points of length d.

3.2. Some history.

- For *fixed* curve with *no markings* and $\varepsilon = 0+$, many earlier compactifications are unified by this construction:
  - Drinfeld's quasimaps to $\mathbb{P}^n$, see [3]. However, note that the moduli of Drinfeld's quasimaps to flag varieties considered in [3] are defined using the Plücker embeddings and therefore fit into the situation described in Example 4 of §2.2. Since under the Plücker embedding the flag varieties are not complete intersections, the canonical obstruction theory of the moduli spaces is not perfect.
  - Gauged linear σ-models for toric targets (Witten [52], Morrison - Plesser [42], Givental [25, 26]).
  - Quot schemes for Grassmannians ((Strømme [49], Bertram [2]); their generalization to type A flag varieties due to Laumon [38, 39] (and rediscovered under the names hyperquot or flag-quot schemes in [9, 36]).

  - ADHM sheaves for $\mathrm{Hilb}_n(\mathbb{C}^2)$ (Diaconescu [21]).

- For the case when the complex structure of the domain curves varies and/or markings are allowed, the starting point was the work by Marian, Oprea, and Pandharipande [41] on moduli of stable quotients (in the terminology introduced above, this corresponds to target a Grassmannian and $\varepsilon = 0+$). Inspired by their paper, the authors developed the toric case and realized that the GIT point of view is the correct generalization of both ([10]). The ε-stability idea appeared first in work by Mustaţă and Mustaţă for target $\mathbb{P}^n$ ([46]). For Grassmannian targets, Toda introduced and studied ε-stable quotients in [51].
- There's a long (ongoing) related story in the symplectic category concerned with the study of (compactifications of) the moduli space of solutions to vortex equations, starting with work of Cieliebak - Gaio - Salamon and of Mundet i Riera, see [7, 8, 29, 30, 44, 45, 54]. An algebraic version of this theory is developed by Woodward in [53].
- Frenkel - Teleman - Tolland are developing a general formalism of a Gromov-Witten type theory of quotient stacks $[Y/G]$, see [24].

3.3. Quasimap Invariants. There are evaluation maps ev_i to $W/\!\!/\mathbf{G}$ (by the prestable condition) and tautological line bundles M_i on $\mathrm{Q}^{\varepsilon}_{g,k}(W/\!\!/\mathbf{G}, \beta)$ with fiber the cotangent line to C at the i^{th} marking:

$$ev_i : \mathrm{Q}^{\varepsilon}_{g,k}(W/\!\!/\mathbf{G}, \beta) \to W/\!\!/\mathbf{G}.$$

As usual, denote $\psi_i := c_1(M_i)$. Given

$$\delta_1, \dots \delta_k \in H^*(W/\!\!/\mathbf{G}, \mathbb{Q})$$

and integers $a_1, \dots, a_k \geq 0$, we define ε-quasimap invariants

$$\langle \delta_1 \psi_1^{a_1}, \dots, \delta_k \psi_k^{a_k} \rangle^{\varepsilon}_{g,k,\beta} := \int_{[\mathrm{Q}^{\varepsilon}_{g,k}(W/\!\!/\mathbf{G},\beta)]^{\mathrm{vir}}} \prod \psi_i^{a_i} \prod ev_i^*(\delta_i)$$

for all $\varepsilon \geq 0+$.

If $\varepsilon > 1$ (we write $\varepsilon = \infty$ for all such stability conditions), these are the descendant Gromov-Witten invariants of $W/\!\!/\mathbf{G}$.

The definition above requires $W/\!\!/\mathbf{G}$ projective; in the quasi-projective case there are equivariant versions available for all interesting targets, e.g., toric varieties, local targets, and Nakajima quiver varieties. Precisely, what is needed in order to have a good theory for non-compact targets $W/\!\!/\mathbf{G}$ is that there is an action on W by an algebraic torus $\mathbf{T} \cong (\mathbb{C}^*)^r$, commuting with the $\mathbf{G}$-action and such that the $\mathbf{T}$-fixed locus on the affine quotient $W/_{\mathrm{aff}}\mathbf{G}$ is proper (and therefore a finite set). To get a unified framework, we will make this assumption from now, allowing the case $r = 0$ of a trivial torus.

The invariants satisfy the "splitting axiom" and in fact they form the degree zero sector of a Cohomological Field Theory (CohFT) on $H^*(W/\!\!/\mathbf{G})$. However, for general targets $W/\!\!/\mathbf{G}$ and $\varepsilon \leq 1$, the *string equation* may fail so the CohFT will not have a flat identity. We refer the reader to [12, §2] for some more details on the quasimap CohFT.

4. Genus zero wall-crossing and mirror maps

It is natural to expect that different stability chambers carry the same information. This will be expressed via wall-crossing formulas for generating functions of the invariants. In this section we explain why the wall-crossing formulas for genus zero invariants are generalizations (in many directions) of Givental's Toric Mirror Theorems.

First we fix some notations:

- $H^*(W/\!\!/\mathbf{G})$ denotes the localized $\mathbf{T}$-equivariant cohomology with $\mathbb{Q}$-coefficients.
- $\langle\ ,\ \rangle$ is the intersection pairing on $H^*(W/\!\!/\mathbf{G})$.
- $\{\gamma_1 = \mathbb{1}, \dots, \gamma_s\}$ and $\{\gamma^1, \dots, \gamma^s\}$ are dual bases of $H^*(W/\!\!/\mathbf{G})$ with respect to $\langle\ ,\ \rangle$. Here $\mathbb{1}$ denotes the cohomology class dual to the fundamental cycle.
- $\mathrm{Eff}(W, \mathbf{G}, \theta)$ denotes the semigroup of numerical classes $\beta \in \mathrm{Pic}([W/\mathbf{G}])^\vee$ represented by θ-quasimaps with possibly disconnected domain. (Note that $\mathrm{Eff}(W, \mathbf{G}, \theta)$ is in general bigger than the cone of effective curves in $W/\!\!/\mathbf{G}$.)
- $\Lambda \cong \mathbb{Q}[[q]]$ denotes the Novikov ring of the theory, that is, the q-adic completion of the semigroup ring $\mathbb{C}[\mathrm{Eff}(W, \mathbf{G}, \theta)]$, $\beta \leftrightarrow q^\beta$.

4.1. *S*-operators. For $\delta_i \in H^*(W/\!\!/\mathbf{G})$ and integers $a_i \geq 0$, put

$$\langle\!\langle \delta_1\psi_1^{a_1}, \dots, \delta_k\psi_k^{a_k} \rangle\!\rangle^\varepsilon_{g,k} = \sum_{\beta \in \mathrm{Eff}(W,\mathbf{G},\theta)} \sum_{m \geq 0} \frac{q^\beta}{m!} \langle \delta_1\psi_1^{a_1}, \dots, \delta_k\psi_k^{a_k}, t, \dots, t) \rangle^\varepsilon_{g,k+m,\beta}.$$

It is a formal function of $t = \sum_{i=1}^s t_i\gamma_i \in H^*(W/\!\!/\mathbf{G})$.

Define, for $\gamma \in H^*(W/\!\!/\mathbf{G}, \Lambda)$ and a formal variable z,

$$S^\varepsilon_t(z)(\gamma) := \sum_{i=1}^s \gamma_i \langle\!\langle \frac{\gamma^i}{z - \psi}, \gamma \rangle\!\rangle^\varepsilon_{0,2}(t).$$

Here $\psi = \psi_1$ and the right-hand side is interpreted as usual by expanding $1/(z - \psi)$ as a geometric series in ψ/z. By convention, $\langle \frac{\gamma^i}{z-\psi}, \gamma \rangle^\varepsilon_{0,2,0} = \langle \gamma^i, \gamma \rangle$. We think of S^ε_t as a family (parametrized by t) of operators on $H^*(W/\!\!/\mathbf{G}, \Lambda)$. When the variable z is understood we drop it from the notation.

In Gromov-Witten theory, the operator S^∞_t is well-known. Its matrix is the (inverse of) a fundamental solution for the quantum differential equation. Furthermore, by the string equation for Gromov-Witten invariants, $S^\infty_t(\mathbb{1})$ coincides with Givental's (big) J-function of $W/\!\!/\mathbf{G}$ (we will come back to J-functions in the next subsection). The operator S^∞_t determines the entire genus zero sector of the Gromov-Witten theory of $W/\!\!/\mathbf{G}$ by a standard reconstruction procedure, essentially due to Dubrovin [23]. As shown in [12], the same reconstruction works for ε-quasimap invariants for all $\varepsilon \geq 0+$. The key point where a new idea is needed is the proof of the following result, which reconstructs invariants with *two* descendant insertions.

Theorem 4.1. *Let z, w be formal variables and define*

$$V^\varepsilon_t(z, w) := \sum_{i,j=1}^s \gamma_i \otimes \gamma_j \langle\!\langle \frac{\gamma^i}{z - \psi}, \frac{\gamma^j}{w - \psi} \rangle\!\rangle^\varepsilon_{0,2}(t),$$

where the convention

$$\sum_{i,j=1}^{s} \gamma_i \otimes \gamma_j \langle \frac{\gamma^i}{z-\psi}, \frac{\gamma^j}{w-\psi} \rangle_{0,2,0}^{\varepsilon} = \frac{[\Delta]}{z+w}$$

is made for the unstable term in the double bracket, with $[\Delta]$ *the diagonal class. Then*

$$V_t^{\varepsilon} = \frac{S_t^{\varepsilon}(z) \otimes S_t^{\varepsilon}(w)([\Delta])}{z+w}.$$

The usual - and very easy - argument that proves the above theorem in Gromov-Witten theory (see [27, item (4) on p.117]) requires the string equation and therefore does not extend to stability parameters $0+ \leq \varepsilon \leq 1$. The new proof from [12] is uniform for all values of ε.

4.2. Wall-crossing for S-operators. The most general wall-crossing formula in genus zero applies to the operators S_t^{ε}, see [11, Theorem 7.3.1]. We state here a slightly more special case.

Theorem 4.2. *Assume that there is an action by a torus* $\mathbf{T}$ *on* W*, commuting with the action of* $\mathbf{G}$*, and such that the induced* $\mathbf{T}$*-action on* $W/\!\!/\mathbf{G}$ *has isolated fixed points. For every* $\varepsilon \geq 0+$

$$S_t^{\varepsilon}(\mathbb{1}) = S^{\infty}_{\tau^{\varepsilon}(t)}(\mathbb{1}),$$

where the (invertible) transformation $\tau^{\varepsilon}(t)$ *is the series of primary* ε*-quasimap invariants*

$$\begin{aligned} \tau^{\varepsilon}(t) &= \sum_{i=1}^{s} \gamma_i \langle\!\langle \gamma^i, \mathbb{1} \rangle\!\rangle_{0,2}^{\varepsilon}(t) - \mathbb{1} \\ &= t + \sum_{i=1}^{s} \gamma_i \sum_{\beta \neq 0} \sum_{m \geq 0} \frac{q^{\beta}}{m!} \langle \gamma^i, \mathbb{1}, t, \dots, t \rangle_{0,2+m,\beta}^{\varepsilon}. \end{aligned}$$

Moreover, the same statement holds for E*-twisted theories, where* E *is any convex* $\mathbf{G}$*-representation.*

A $\mathbf{G}$-representation is called convex if the $\mathbf{G}$-equivariant bundle $W \times E$ on W is generated by $\mathbf{G}$-equivariant sections. By twisted theories in the last statement we mean that the twisting is by the top Chern class, in the sense of Coates - Givental [18], as extended for quasimap invariants in [14, §6.2]. The twisting vector bundle $\mathcal{E}$ on $W/\!\!/\mathbf{G}$ is descended from the representation E.

Note that no positivity assumptions are made in Theorem 4.2 on $(W, \mathbf{G}, \theta)$, or on $(W, E, \mathbf{G}, \theta)$ in the twisted case, and also that no assumption is made on the 1-dimensional orbits of the $\mathbf{T}$ action on $W/\!\!/\mathbf{G}$.

Theorem 4.2 applies to essentially every example listed earlier: toric manifolds, flag manifolds, some (but not all) Nakajima quiver varieties, and local targets over them all admit torus actions with the required property. Of course, the statement is conjectured to hold without the existence of a torus action with isolated fixed points, see [11, Conjecture 6.1.1]. In fact, the part of the Theorem involving twisted theories already covers such targets. This is because the E-twisted quasimap invariants give (almost all of) the genus zero quasimap invariants of the zero-locus of a regular section of the bundle $\mathcal{E} = W^{ss} \times_{\mathbf{G}} E$ and this zero-locus generally is not $\mathbf{T}$-invariant.

4.3. J^ε-functions and Birkhoff factorization. Recall first the big J-function of Gromov-Witten theory:

$$J^\infty(q,t,z) = \mathbb{1} + \frac{t}{z} + \sum_i \gamma_i \langle\!\langle \frac{\gamma^i}{z(z-\psi)} \rangle\!\rangle_{0,1}^\infty(t)$$
$$= \mathbb{1} + \frac{t}{z} + \sum_{\beta,k} \frac{q^\beta}{k!} (ev_1)_* \frac{[\overline{M}_{0,1+k}(W/\!\!/\mathbf{G},\beta)]^{\mathrm{vir}} \cap \prod_{j=2}^{1+k} ev_j^* t}{z(z-\psi)}$$

(the last sum is over $(\beta,k) \neq (0,0),(0,1)$). We want to extend it to all $\varepsilon \geq 0+$. The problem is that the spaces $\mathrm{Q}^\varepsilon_{0,1}(W/\!\!/\mathbf{G},\beta)$ do not exist for $\varepsilon \leq \frac{1}{\beta(L_\theta)}$. To resolve it we use the interpretation of the J-function as a sum of certain virtual localization residues for the natural $\mathbb{C}^*$-action on the Gromov-Witten *graph spaces* $\overline{M}_{0,k}(W/\!\!/\mathbf{G} \times \mathbb{P}^1, (\beta,1))$.

Specifically, for *all* $0+ \leq \varepsilon$, $k \geq 0$, we have the *quasimap graph space*

$$QG^\varepsilon_{0,k,\beta}(W/\!\!/\mathbf{G}) = \{\big((C,\{p_i\}),P,u,\varphi\big) \mid \varphi : C \to \mathbb{P}^1, \varphi_*[C] = [\mathbb{P}^1]\}.$$

This is the moduli space of (genus zero, k-pointed) ε-stable quasimaps whose domain curve contains a component C_0 which is a parametrized $\mathbb{P}^1$. The ampleness part of the ε-stability condition involves only $\overline{C \setminus C_0}$, while the length condition remains the same. These spaces are defined for all $k \geq 0$ and the analogue of Theorem 3.2 holds for them. For toric targets and $\varepsilon = 0+$ they were introduced in [10], the general case is in [14].

The $\mathbb{C}^*$-action on $\mathbb{P}^1$ induces an action on $QG^\varepsilon_{0,k,\beta}(W/\!\!/\mathbf{G})$. Let z denote the equivariant parameter.

Consider the fixed locus F_0 of quasimaps for which all markings and the entire degree β are over $0 \in \mathbb{P}^1 \cong C_0$. There are two cases:

- $k \geq 1$, or $\varepsilon > \frac{1}{\beta(L_\theta)}$. Then $F_0 \cong \mathrm{Q}^\varepsilon_{0,1+k}(W/\!\!/\mathbf{G},\beta)$ with its canonical virtual class and $\mathrm{e}_{\mathbb{C}^*}(N^{\mathrm{vir}}) := \mathrm{e}_{\mathbb{C}^*}(N^{\mathrm{vir}}_{F_0/QG^\varepsilon_{0,k,\beta}(W/\!\!/\mathbf{G})}) = z(z-\psi)$. We also have the evaluation map $ev = ev_1 : F_0 \to W/\!\!/\mathbf{G}$.
- $k = 0$ and $\varepsilon \leq \frac{1}{\beta(L_\theta)}$. Then $F_0 = \{(\mathbb{P}^1,P,u)\}$, with u having a base point of (maximal) length $\beta(L_\theta)$ at $0 \in \mathbb{P}^1$. We define $ev : F_0 \to W/\!\!/\mathbf{G}$ by taking evaluation at the generic point of $\mathbb{P}^1$. In this case $\mathrm{e}_{\mathbb{C}^*}(N^{\mathrm{vir}})$ changes with β.

Now for each $\varepsilon \geq 0+$ we define the big J^ε-function by

$$J^\varepsilon(q,t,z) := \sum_{\beta,k\geq 0} \frac{q^\beta}{k!} ev_* \mathrm{Res}_{F_0}\big([QG^\varepsilon_{0,k,\beta}(W/\!\!/\mathbf{G})]^{\mathrm{vir}} \cap \prod_{j=1}^k ev_j^* t\big)$$
$$= \mathbb{1} + \frac{t}{z} + \sum_{0<\beta(L_\theta)\leq 1/\varepsilon} q^\beta ev_* \frac{[F_0]}{\mathrm{e}_{\mathbb{C}^*}(N^{\mathrm{vir}})}$$
$$+ \sum_{\beta,k} \frac{q^\beta}{k!} (ev_1)_* \frac{[\mathrm{Q}^\varepsilon_{0,1+k}(W/\!\!/\mathbf{G},\beta)]^{\mathrm{vir}} \cap \prod_{j=2}^{1+k} ev_j^* t}{z(z-\psi)}.$$

The *small* J^ε-function is by definition the specialization at $t=0$,

$$J^\varepsilon_{sm}(q,z) := J^\varepsilon(q,0,z).$$

For the asymptotic stability $\varepsilon = 0+$ we have the *small I-function*

$$I_{sm}(q,z) = J^{0+}(q,0,z) = \mathbb{1} + \sum_{\beta\neq 0} q^\beta ev_* \frac{[F_0]}{\mathrm{e}_{\mathbb{C}^*}(N^{\mathrm{vir}})}.$$

When $W/\!\!/\mathbf{G}$ is a nonsingular toric variety, or a complete intersections in a toric variety, the small I-function is (essentially up to an exponential factor) the cohomology valued hypergeometric q-series introduced by Givental, see [26].

Closed expressions for I_{sm} are known also for many non-abelian GIT quotients: flag manifolds of classical type, zero loci of sections of homogeneous bundles in them, local targets over them, the Hilbert scheme of points in $\mathbb{C}^2$ ([4, 5, 15, 16]).

In general, the big J^ε-function and the operator S^ε_t are related by "Birkhoff factorization". This is the content of the following Theorem.

Theorem 4.3 ([11]). *For any GIT target and any $\varepsilon \geq 0+$*

$$J^\varepsilon(q,t,z) = S^\varepsilon_t(P^\varepsilon(q,t,z))$$

where $P^\varepsilon(q,t,z)$ is a power series *in z. (In fact, P^ε is naturally a generating function of $\mathbb{C}^*$-equivariant graph space integrals, see [11, §5.4].)*

4.4. The case of semi-positive targets. The triple $(W,\mathbf{G},\theta)$ is called *semi-positive* if

$$\beta(\det(T_W)) \geq 0$$

for every $\beta \in \mathrm{Eff}(W,\mathbf{G},\theta)$. Here T_W is the (virtual) tangent bundle of the lci $\mathbf{G}$-variety W, viewed as an element in the equivariant K-group $K^0_{\mathbf{G}}(W)$. We note that semi-positivity implies that the anti-canonical class of a projective $W/\!\!/\mathbf{G}$ is nef, but the converse need not be true.

The Birkhoff Factorization in Theorem 4.3 simplifies drastically for semi-positive targets. If $(W,\mathbf{G},\theta)$ is semi-positive, easy dimension counting arguments show that for every $\varepsilon \geq 0+$ the function J^ε contains no positive powers of z. Hence we have the asymptotic expansions

$$J^\varepsilon_{sm}(q,z) = J^\varepsilon_0(q)\mathbb{1} + \frac{J^\varepsilon_1(q)}{z} + O(1/z^2),$$

$$J^\varepsilon(q,t,z) = J^\varepsilon_0(q)\mathbb{1} + \frac{t + J^\varepsilon_1(q)}{z} + O(1/z^2).$$

In particular, we have

$$I_{sm}(q,z) = I_0(q)\mathbb{1} + I_1(q)\frac{1}{z} + O(1/z^2),$$

defining the q-series $I_0(q)$ and $I_1(q)$. They satisfy $I_0(q) = 1 + O(q) \in \Lambda$ and $I_1 \in qH^{\leq 2}(W/\!\!/\mathbf{G},\Lambda)$. For $\varepsilon > 0$, the coefficients $J^\varepsilon_0(q)$ and $J^\varepsilon_1(q)$ are polynomial truncations of the series I_0 and I_1. Note that since there are explicit closed formulas for I_{sm} in almost all examples, the series $I_0(q)$ and $I_1(q)$ are also explicit.

It follows that Theorem 4.3 specializes to the following Corollary (a very special case of this result is due to [19], by different methods).

Corollary 4.4 ([11]). *Let $(W, \mathbf{G}, \theta)$ be semi-positive and let $\varepsilon \geq 0+$ be arbitrary. Then the J-function and the S-operator are related by*

$$S_t^\varepsilon(\mathbb{1}) = \frac{J^\varepsilon(q,t,z)}{J_0^\varepsilon(q)}.$$

The transformation $\tau_\varepsilon(t) = \sum_{i=1}^s \gamma_i \langle\!\langle \gamma^i, \mathbb{1} \rangle\!\rangle_{0,2}^\varepsilon(t) - \mathbb{1}$ satisfies

$$\tau_\varepsilon(t) = \frac{t + J_1^\varepsilon(q)}{J_0^\varepsilon(q)},$$

and in particular

$$\sum_{i=1}^s \gamma_i \sum_{\beta \neq 0} q^\beta \langle \gamma^i, \mathbb{1} \rangle_{0,2,\beta}^\varepsilon = \frac{J_1^\varepsilon(q)}{J_0^\varepsilon(q)}.$$

Combining Theorem 4.2 with Corollary 4.4 gives the following Corollary.

Corollary 4.5 ([11]). *Assume $(W, \mathbf{G}, \theta)$ is semi-positive and there is a $\mathbf{T}$-action on $W/\!\!/\mathbf{G}$ with isolated fixed points. Then*

$$J^\infty\left(q, \frac{t + J_1^\varepsilon(q)}{J_0^\varepsilon(q)}, z\right) = \frac{J^\varepsilon(q,t,z)}{J_0^\varepsilon(q)}.$$

The same is true for E-twisted theories on $W/\!\!/\mathbf{G}$, where E is a convex $\mathbf{G}$-representation such that, for all θ-effective β,

$$\beta(\det(T_W)) - \beta(W \times \det(E)) \geq 0.$$

Let $\varepsilon = 0+$. After making $t = 0$ in the last Corollary and applying the string and divisor equations in the GW side, we obtain the usual formulation of the genus zero Mirror Theorem for the small J-function of Gromov-Witten theory

$$e^{\frac{1}{z}\frac{I_1(q)}{I_0(q)}} J_{sm}^\infty(Q, z) = I_{sm}(q, z)$$

after the change of variable $Q^\beta = q^\beta e^{\int_\beta \frac{I_1(q)}{I_0(q)}}$. For Calabi-Yau complete intersections in toric varieties, this change of variables is precisely the mirror map obtained from the solutions to the Picard-Fuchs equations associated to the mirror manifolds. Note that by the last equation in Corollary 4.4 the mirror map acquires a geometric interpretation in terms of two-point primary $(0+)$-quasimap invariants with a fundamental class insertion, as suggested by Jinzenji [33].

Hence the genus zero wall-crossing formula in Theorem 4.2 generalizes the mirror theorems as follows:

- from abelian to non-abelian quotients
- from the small to the big phase space
- from $\varepsilon = 0+$ to all ε
- from semi-positive GIT triples to all such triples.

4.5. Wall-crossing for J^ε-functions in the general case. Without the semi-positivity assumption the relation between $J^\varepsilon(q,t,z)$ and $J^\infty(q,t,z)$ is more complicated than the one given by Corollary 4.5. The most concise formulation is given by the following Conjecture [11, Conjecture 6.4.2].

Conjecture 4.6. *For all GIT triples $(W,\mathbf{G},\theta)$ and all stability parameters $\varepsilon \geq 0+$ the function $J^\varepsilon(q,t,z)$ is on the Lagrangian cone of the Gromov-Witten theory of $W/\!\!/\mathbf{G}$.*

Recall that for a general target X Givental introduced a formalism which encodes the genus zero sector of the Gromov-Witten theory of X via an overruled Lagrangian cone in an appropriate infinite-dimensional symplectic vector space, see [18, 28]. The Lagrangian cone is generated by the big J-function (this statement is another formulation of the Dubrovin reconstruction mentioned earlier). The conjecture then implies that $J^\infty(q,\tau_{\infty,\varepsilon}(q,t),z)$ is a linear combination of the derivatives $\partial_{t_i}J^\varepsilon(q,t,z)$ with uniquely determined coefficients (depending on q,t, and z) and unique change of variables $t\mapsto\tau_{\infty,\varepsilon}(q,t)$.

Theorem 4.7 ([11]). *Assume there is a $\mathbf{T}$-action on W such that the induced action on $W/\!\!/\mathbf{G}$ has isolated fixed points and isolated 1-dimensional orbits. Then Conjecture 4.6 holds true.*

4.6. Big $\mathbb{I}$-functions. The results described so far in this section elucidate the relationship between quasimap and GW invariants of $W/\!\!/\mathbf{G}$ in genus zero. However, if one is primarily interested in calculating GW invariants, the applicability of these results is restricted only to invariants with (descendant) insertions at one marking. This is because only for the small I-function one can write down explicit closed formulas. In general, quasimap invariants with two or more insertions are equally difficult to determine for all values of the stability parameter ε. To improve the situation the authors have introduced in [13] a new version of big $\mathbb{I}$-function of $(W,\mathbf{G},\theta)$ by considering a theory of $(0+)$-stable quasimaps with infinitesimally weighted markings. We conjectured that this function lies on the Lagrangian cone of the Gromov-Witten theory of $W/\!\!/\mathbf{G}$. Arguments parallel to the ones in the unweighted case are used to prove the following Theorem.

Theorem 4.8 ([13]). *Let $(W,\mathbf{G},\theta)$ be a GIT triple. Assume there is a $\mathbf{T}$-action on W such that the induced action on $W/\!\!/\mathbf{G}$ has isolated fixed points and isolated 1-dimensional orbits. Then the big $\mathbb{I}$-function associated to $(W,\mathbf{G},\theta)$ is on the Lagrangian cone of the Gromov-Witten theory of $W/\!\!/\mathbf{G}$. Furthermore, if E is a convex $\mathbf{G}$-representation, then the E-twisted $\mathbb{I}$ is on the E-twisted Lagrangian cone of $W/\!\!/\mathbf{G}$.*

As a consequence, the big J function of the (E-twisted) Gromov-Witten theory of $W/\!\!/\mathbf{G}$ is obtained from $\mathbb{I}$ via the Birkhoff factorization procedure of Coates and Givental [18]. The advantage is that one can calculate again explicit closed formulas for this new big $\mathbb{I}$-function in many cases. In [13] it is explained how to do so for toric varieties and for complete intersections in them. For example, if $\mathbb{C}^{n+1}/\!\!/_{\mathrm{id}}\mathbb{C}^* = \mathbb{P}^n$ is the standard GIT presentation of the projective space and $E=\mathbb{C}_{l(\mathrm{id})}$ is the 1-dimensional $\mathbb{C}^*$-representation with weight $l\in\mathbb{Z}_{>0}$, then one finds

$$\mathbb{I}^E_{\mathbb{C}^{n+1}/\!\!/\mathbb{C}^*}(t)=\sum_{d=0}^{\infty}q^d\frac{\exp(\sum_{i=0}^n t_i(H+dz)^i/z)}{\prod_{k=1}^d(H+kz)^{n+1}}\prod_{k=0}^{ld}(lH+kz),$$

where H is the hyperplane class and $t=\sum_{i=0}^n t_iH^i$ is the general element of $H^*(\mathbb{P}^n,\mathbb{Q})$.

Observe that if we denote by $t_{sm} = t_0 \mathbb{1} + t_1 H$ the restriction of t to the small parameter space $H^0(\mathbb{P}^n, \mathbb{Q}) \oplus H^2(\mathbb{P}^n, \mathbb{Q})$, then

$$\mathbb{I}^E_{\mathbb{C}^{n+1}/\!\!/\mathbb{C}^*}(t_{sm}) = \exp(\frac{t_0 \mathbb{1} + t_1 H}{z}) \sum_{d=0}^{\infty} q^d \exp(dt_1) \frac{\prod_{k=0}^{ld}(lH + kz)}{\prod_{k=1}^{d}(H + kz)^{n+1}},$$

which is precisely Givental's small I-function of a hypersurface of degree l in $\mathbb{P}^n$ (and differs from the function I_{sm} from §4.3 by the overall exponential factor $\exp(\frac{t_0 \mathbb{1} + t_1 H}{z})$ and the change $q \mapsto q \exp(t_1)$).

Remark 4.9. The results described in §4.2 - §4.6 above have been extended in [6] to the case of "orbifold GIT targets", that is, to the case when $[W^{ss}/\mathbf{G}]$ is a nonsingular Deligne-Mumford stack. A result related to Theorem 4.8 has been obtained earlier by Woodward, [53, Theorem 1.6].

5. Higher genus wall-crossing for semi-positive targets

In this section we discuss the wall-crossing formulas for higher genus ε-quasimap descendant invariants in the case of semi-positive triples $(W, \mathbf{G}, \theta)$.

Let

$$\mathbf{t}(\psi) := t_0 + t_1 \psi + t_2 \psi^2 + t_3 \psi^3 + \dots,$$

with $t_j = \sum_i t_{ji} \gamma_i \in H^*(W/\!\!/\mathbf{G}, \mathbb{Q})$ general cohomology classes.

By definition, the genus g, ε-descendant potential of $(W, \mathbf{G}, \theta)$ is

$$F^{\varepsilon}_g(\mathbf{t}) := \sum_{\beta \in \mathrm{Eff}(W,\mathbf{G},\theta)} \sum_{m \geq 0} \frac{q^\beta}{m!} \langle \mathbf{t}(\psi_1), \mathbf{t}(\psi_2), \dots \mathbf{t}(\psi_m) \rangle^{\varepsilon}_{g,m,\beta}.$$

As usual, we omit from the sum the unstable terms corresponding to $(g, m, \beta, \varepsilon)$ for which the moduli spaces are not defined.

Conjecture 5.1 ([12]). *For a semi-positive triple $(W, \mathbf{G}, \theta)$, and every $\varepsilon \geq 0+$*

$$(J^{\varepsilon}_0(q))^{2g-2} F^{\varepsilon}_g(\mathbf{t}(\psi)) = F^{\infty}_g \left(\frac{\mathbf{t}(\psi) + J^{\varepsilon}_1(q)}{J^{\varepsilon}_0(q)} \right). \tag{5.1}$$

Further, for every $\varepsilon_1 \neq \varepsilon_2$

$$(J^{\varepsilon_1}_0(q))^{2g-2} F^{\varepsilon_1}_g (J^{\varepsilon_1}_0(q) \mathbf{t}(\psi) - J^{\varepsilon_1}_1(q)) = (J^{\varepsilon_2}_0(q))^{2g-2} F^{\varepsilon_2}_g \left(J^{\varepsilon_2}_0(q) \mathbf{t}(\psi) - J^{\varepsilon_2}_1(q) \right). \tag{5.2}$$

To be precise, in the case $g = 0$ the equalities are conjectured to hold modulo terms of degree ≤ 1 in the coordinates t_{ji} (but see [12, Remark 3.1.3] for an explanation on how to extend the statement to an equality up to constants).

Note that the (a priori stronger) wall-crossing formula (5.2) follows from (5.1).

Considering the Taylor coefficients on both sides gives the following equivalent formulation of (5.1): If $2g - 2 + k \geq 0$, then for arbitrary $\delta_1, \dots \delta_k \in H^*(W/\!\!/\mathbf{G}, \mathbb{Q})$ and integers $a_1, \dots, a_k \geq 0$,

$$(J_0^\varepsilon(q))^{2g-2+k}\sum_\beta q^\beta\langle\delta_1\psi_1^{a_1},\dots,\delta_n\psi_k^{a_k}\rangle_{g,k,\beta}^\varepsilon=$$

$$\sum_\beta q^\beta\sum_{m=0}^\infty\frac{1}{m!}\left\langle\delta_1\psi_1^{a_1},\dots,\delta_k\psi_k^{a_k},\frac{J_1^\varepsilon(q)}{J_0^\varepsilon(q)},\dots,\frac{J_1^\varepsilon(q)}{J_0^\varepsilon(q)}\right\rangle_{g,k+m,\beta}^\infty.$$

Combining Corollary 4.5 with reconstruction for ε-quasimap invariants proves the Conjecture in genus zero.

Theorem 5.2 ([12]). *Let $(W,\mathbf{G},\theta)$ be semi-positive. Assume there is an action by a torus $\mathbf{T}$, such that the fixed points of the induced $\mathbf{T}$-action on $W/\!\!/\mathbf{G}$ are isolated. Then the $g=0$ case of Conjecture 5.1 holds. Moreover, if E is a convex $\mathbf{G}$-representation such that $\beta(\det(T_W))-\beta(W\times\det(E))\geq 0$ for all θ-effective β, then the conjecture also holds at $g=0$ for the E-twisted ε-quasimap theories of $W/\!\!/\mathbf{G}$.*

A more convincing piece of evidence for the validity of the Conjecture is provided by the following result:

Theorem 5.3 ([12]). *Let X be a nonsingular quasi-projective toric variety of dimension n, obtained as the GIT quotient of a semi-positive triple $(\mathbb{C}^{n+r},(\mathbb{C}^*)^r,\theta)$. Then Conjecture 5.1 holds for X.*

It is easy to see that toric varieties (in any semi-positive GIT presentation) have $I_0(q)=1$ (and hence $J_0^\varepsilon=1$ for all ε). When X is a nonsingular and projective toric Fano and we take its "standard" GIT presentation (as considered e.g. in [10]), then $I_1(q)=0$ as well. Hence we obtain the following

Corollary 5.4. *If X is a nonsingular projective Fano toric variety, then its quasimap invariants (for the standard GIT presentation) are independent on ε:*

$$F_g^\varepsilon(\mathbf{t}(\psi))=F_g^\infty(\mathbf{t}(\psi)),\ \ \forall\varepsilon\geq 0+.$$

The first statement of the kind in the Corollary was established by Marian - Oprea - Pandharipande [41] for $W/\!\!/\mathbf{G}$ a Grassmannian and for $\varepsilon=0+$. Their result was extended to all ε in [51] by Toda.

Remark 5.5.

1. The most interesting case covered by Theorem 5.3 is that of toric Calabi-Yau targets. For 3-folds, our theorem says that $F_g^{0+}|_{\mathbf{t}(\psi)=0}$ is *equal* to the B-model genus g prepotential, expanded near a large complex structure point for the mirror of X.
2. The arguments proving Theorem 5.3 also apply to show that the higher genus wall-crossing of Conjecture 5.1 holds for some non-abelian local Calabi-Yau targets, namely local Grassmannians, and in fact local type A flag manifolds, see [12, Theorem 1.3.4].
3. The remaining challenge is to prove Conjecture 5.1 for compact Calabi-Yau targets at $g\geq 1$.

Acknowledgements. The authors thank Daewoong Cheong, Duiliu-Emanuel Diaconescu, and Davesh Maulik for collaboration on parts of this quasimap project. The first named author was partially supported by NSF DMS-1305004 and the second named author was partially supported by KRF 2007-0093859. In addition, the second named author thanks University of Minnesota for hospitality during the writing of the paper.

References

[1] Benini, F. and Cremonesi, S., *Partition functions of $N = (2, 2)$ gauge theories on S^2 and vortices*, arXiv:1206.2356.

[2] Bertram, A., *Quantum Schubert calculus*, Adv. Math. **128** (1997), no. 2, 289–305.

[3] Braverman, A., *Spaces of quasi-maps into the flag varieties and their applications*, International Congress of Mathematicians. Vol. II, 1145–1170, Eur. Math. Soc., Zürich, 2006.

[4] Bertram, A., Ciocan-Fontanine, I., and Kim, B., *Two proofs of a conjecture of Hori and Vafa*, Duke Math. J. **126** (2005), no. 1, 101–136.

[5] ———, *Gromov-Witten invariants for abelian and nonabelian quotients*, J. Algebraic Geom. **17**(2) (2008), 275–294.

[6] Cheong, D., Ciocan-Fontanine, I., and Kim, B., *Orbifold quasimap theory*, In preparation.

[7] Cieliebak, K., Gaio, S.R., and Salamon, D.A., *J-holomorphic curves, moment maps, and invariants of Hamiltonian group actions*, Internat. Math. Res. Notices **16** (2000), 831–882.

[8] Cieliebak, K., Gaio, A.R., Mundet i Riera, I., and Salamon, D.A., *The symplectic vortex equations and invariants of Hamiltonian group actions*, J. Symplectic Geom. **1**(3) (2002), 543–645.

[9] Ciocan-Fontanine, I., *The quantum cohomology ring of flag varieties*, Trans. Amer. Math. Soc. **351** (1999), no. 7, 2695–2729.

[10] Ciocan-Fontanine, I. and Kim, B., *Moduli stacks of stable toric quasimaps*, Adv. in Math. **225** (2010), 3022–3051.

[11] ———, *Wall-crossing in genus zero quasimap theory and mirror maps*, arXiv:1304.7056. To appear in Algebraic Geometry.

[12] ———, *Higher genus quasimap wall-crossing for semi-positive targets*, arXiv:1308.6377.

[13] ———, *Big I-functions*, arXiv:1401.7417.

[14] Ciocan-Fontanine, I., Kim, B., and Maulik, D., *Stable quasimaps to GIT quotients*, J. Geom. Phys. **75** (2014), 17–47.

[15] Ciocan-Fontanine, I., Kim, B., Sabbah, C., *The abelian/nonabelian correspondence and Frobenius manifolds*, Invent. Math. **171** (2008), 301–343. and

[16] Ciocan-Fontanine, I., Konvalinka, M., and Pak, I., *Quantum cohomology of $Hilb_n(\mathbb{C}^2)$ and the weighted hook walk on Young diagrams*, Journal of Algebra **349** (2012), 268–283.

[17] Coates, A., Corti, A., Galkin, S., and Kasprzyk, A., *Quantum periods for 3-dimensional Fano manifolds*, arXiv:1303.3288.

[18] Coates T. and Givental, A., *Quantum Riemann-Roch, Lefschetz, and Serre*, Ann. Math. (2) **165**(1), (2007), 15–53.

[19] Cooper Y. and Zinger, A., *Mirror symmetry for stable quotients invariants*, arXiv:1201.6350.

[20] Deligne, P. and Mumford, D., *The irreducibility of the space of curves of given genus*, Publ. Math. IHES **36** (1969), 75–110.

[21] Diaconescu, D.-E., *Moduli of ADHM sheaves and local Donaldson-Thomas theory*, J. Geom. Phys. **62** (2012), no. 4, 763–799.

[22] Doroud, N., Gomis, J., Le Floch, B., and Lee, S., *Exact results in $D = 2$ supersymmetric gauge theories*, J. High Energy Phys. 2013, no. 5, 093, front matter+69 pp.

[23] Dubrovin, B., *Geometry of 2D topological field theories*, Integrable systems and quantum groups (Montecatini Terme, 1993), 120–348, Lecture Notes in Math., 1620, Springer, Berlin, 1996.

[24] Frenkel, E., Teleman, C., and Tolland, A.J., *Gromov-Witten gauge theory I*, arXiv:0904.4834.

[25] Givental, A., *Equivariant Gromov-Witten invariants*, Internat. Math. Res. Notices (1996), no. 13, 613–663.

[26] Givental, A., *A mirror theorem for toric complete intersections*, Topological field theory, primitive forms and related topics (Kyoto, 1996), 141–175, Progr. Math., 160, Birkhäuser Boston, Boston, MA, 1998.

[27] ———, *Elliptic Gromov-Witten invariants and the generalized mirror conjecture*, Integrable systems and algebraic geometry (Kobe/Kyoto, 1997), 107–155, World Sci. Publ., River Edge, NJ, 1998.

[28] ———, *Symplectic geometry of Frobenius structures*, Frobenius manifolds, 91–112, Aspects Math., E36, Friedr. Vieweg, Wiesbaden, 2004.

[29] Gonzalez, E. and Woodward, C., *Area dependence in gauged Gromov-Witten theory*, arXiv:0811.3358.

[30] ———, *Gauged Gromov-Witten theory for small spheres*, Math. Z. **273** (2013), no. 1-2, 485–514.

[31] Herbst, M., Hori K., and Page, D., *Phases of $N = 2$ theories in $1 + 1$ dimensions with boundary*, arXiv:0803.2045.

[32] Hori, K. and Vafa, C., *Mirror symmetry*, arXiv:hep-th/0002222.

[33] Jinzenji, M., *Mirror map as generating function of intersection numbers: toric manifolds with two Kähler forms*, Comm. Math. Phys. **323** (2013), no. 2, 747–811.

[34] Jockers, H., Kumar, V., Lapan, J., Morrison, D., and Romo, M., *Two-Sphere Partition Functions and Gromov-Witten Invariants*, Comm. Math. Phys. **325** (2014), no. 3, 1139–1170.

[35] Keel, S. and Mori, S., *Quotients by groupoids*, Ann. of Math. (2) **145** (1997), no. 1, 193–213.

[36] Kim, B., *Gromov-Witten invariants for flag manifolds*, Thesis (Ph.D.) University of California, Berkeley. 1996.

[37] Kuznetsov, A., *Quiver varieties and Hilbert schemes*, Mosc. Math. J. **7** (2007), no. 4, 673–697.

[38] Laumon, G., *Un anlogue global du cône nilpotent*, Duke Math. J. **57** (1988), 647–671.

[39] _____, *Faisceaux automorphes liés aux sèries d'Eisenstein*, Automorphic forms, Shimura varieties and L-functions, Vol. I (Ann Arbor, MI, 1988), 227–281, Perspect. Math., 10, Academic Press, Boston, MA, 1990.

[40] Laumon, G. and Moret-Bailly, L., *Champs algébriques*, Ergegnisse der Math. und ihrer Grenzgebiete **3**. Folge, 39 (Springer Verlag) (2000).

[41] Marian, A., Oprea, D., and Pandharipande, R., *The moduli space of stable quotients*, Geometry & Topology **15** (2011), 1651–1706.

[42] Morrison, D. and Plesser, R.R., *Summing the instantons: quantum cohomology and mirror symmetry in toric varieties*, Nuclear Phys. B 440 (1–2) (1995), 279–354.

[43] Mukai, S., *An introduction to invariants and moduli*, Translated from the 1998 and 2000 Japanese editions by W. M. Oxbury. Cambridge Studies in Advanced Mathematics, 81. Cambridge University Press, Cambridge, 2003. xx+503 pp.

[44] Mundet i Riera, I., *Hamiltonian Gromov-Witten invariants*, Topology **42**(3) (2003), 525–553.

[45] Mundet i Riera, I. and Tian, G., *A compactification of the moduli space of twisted holomorphic maps*, Adv. Math. **222**(4) (2009), 1117–1196.

[46] Mustaţă A., *Intermediate moduli spaces of stable maps*, Invent. Math. **167**(1) (2007), 47–90.

[47] Nakajima, H., *Instantons on ALE spaces, quiver varieties, and Kac-Moody algebras*, Duke Math. J. **76** (1994), no. 2, 365–416.

[48] Nakajima, H., *Sheaves on ALE spaces and quiver varieties*, Moscow Math. J. **7** (4) (2007), 699–722.

[49] Strømme, S., *On parametrized rational curves in Grassmann varieties*, Space curves (Rocca di Papa, 1985), 251–272, Lecture Notes in Math., 1266, Springer, Berlin, 1987.

[50] Tjøtta, E., *Quantum cohomology of a Pfaffian Calabi-Yau variety: verifying mirror symmetry predictions*, Compositio Mathematica **126** (2001), 79–89.

[51] Toda, Y., *Moduli spaces of stable quotients and wall-crossing phenomena*, Compositio Math. **147** (2011), 1479–1518.

[52] Witten, E., *Phases of $N = 2$ theories in two dimensions*, Nuclear Phys. B **403** (1993), no. 1-2, 159–222.

[53] Woodward, C., *Quantum Kirwan morphism and Gromov-Witten invariants of quotients*, arXiv:1204.1765.

[54] Ziltner, F., *A quantum Kirwan map: bubbling and Fredholm theory for symplectic vortices over the plane*, arXiv:1209.5866. To appear in Mem. Amer. Math. Soc.

School of Mathematics, University of Minnesota, 206 Church St. SE, Minneapolis MN, 55455, USA; and School of Mathematics, Korea Institute for Advanced Study, 85 Hoegiro, Dongdaemun-gu, Seoul, 130-722, Republic of Korea

E-mail: ciocan@math.umn.edu

School of Mathematics, Korea Institute for Advanced Study, 85 Hoegiro, Dongdaemun-gu, Seoul, 130-722, Republic of Korea

E-mail: bumsig@kias.re.kr

Semiorthogonal decompositions in algebraic geometry

Alexander Kuznetsov

Abstract. In this review we discuss what is known about semiorthogonal decompositions of derived categories of algebraic varieties. We review existing constructions, especially the homological projective duality approach, and discuss some related issues such as categorical resolutions of singularities.

Mathematics Subject Classification (2010). Primary 18E30; Secondary 14F05.

Keywords. Semiorthogonal decompositions, exceptional collections, Lefschetz decompositions, homological projective duality, categorical resolutions of singularities, Fano varieties.

1. Introduction

In recent years an extensive investigation of semiorthogonal decompositions of derived categories of coherent sheaves on algebraic varieties has been done, and now we know quite a lot of examples and some general constructions. With time it is becoming more and more clear that semiorthogonal components of derived categories can be thought of as the main objects in noncommutative algebraic geometry. In this paper I will try to review what is known in this direction — how one can construct semiorthogonal decompositions and how one can use them.

In section 2 we will recall the basic notions, discuss the most frequently used semiorthogonal decompositions, and state the base change formula. In section 3 we review the theory of homological projective duality which up to now is the most powerful method to construct semiorthogonal decompositions. In section 4 we discuss categorical resolutions of singularities, a subject interesting by itself, and at the same time inseparable from homological projective duality. In section 5 examples of homologically projectively dual varieties are listed. Finally, in section 6 we discuss semiorthogonal decompositions of varieties of small dimension.

I should stress that in the area of algebraic geometry described in this paper there are more questions than answers, but it really looks very promising. Also, due to volume constraints I had to leave out many interesting topics closely related to the main subject, such as the categorical Griffiths component, Hochschild homology and cohomology, and many others.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. Semiorthogonal decompositions

This paper can be considered as a continuation and a development of the ICM 2002 talk [7] of Alexei Bondal and Dmitri Orlov. So I will freely use results and definitions from [7] and restrict myself to a very short reminder of the most basic notion. In particular, the reader is referred to [7] for the definition of a Serre functor, Fourier–Mukai transform, etc.

2.1. A short reminder. Recall that a semiorthogonal decomposition of a triangulated category $\mathcal{T}$ is a collection $\mathcal{A}_1, \dots, \mathcal{A}_n$ of full triangulated subcategories such that:

(a) for all $1 \le j < i \le n$ and any objects $A_i \in \mathcal{A}_i$, $A_j \in \mathcal{A}_j$ one has $\mathsf{Hom}_{\mathcal{T}}(A_i, A_j) = 0$;

(b) the smallest triangulated subcategory of $\mathcal{T}$ containing $\mathcal{A}_1, \dots, \mathcal{A}_n$ coincides with $\mathcal{T}$.

We will use the notation $\mathcal{T} = \langle \mathcal{A}_1, \dots, \mathcal{A}_n \rangle$ for a semiorthogonal decomposition of $\mathcal{T}$ with components $\mathcal{A}_1, \dots, \mathcal{A}_n$.

We will be mostly interested in semiorthogonal decompositions of $\mathbf{D}^b(\mathsf{coh}(X))$, the *bounded derived category of coherent sheaves* on an algebraic variety X which in most cases will be assumed to be smooth and projective over a base field k.

Recall that a full triangulated subcategory $\mathcal{A} \subset \mathcal{T}$ is admissible if its embedding functor $i : \mathcal{A} \to \mathcal{T}$ has both left and right adjoint functors $i^*, i^! : \mathcal{T} \to \mathcal{A}$. An admissible subcategory $\mathcal{A} \subset \mathcal{T}$ gives rise to a pair of semiorthogonal decompositions

$$\mathcal{T} = \langle \mathcal{A}, {}^{\perp}\mathcal{A} \rangle \qquad \text{and} \qquad \mathcal{T} = \langle \mathcal{A}^{\perp}, \mathcal{A} \rangle, \tag{2.1}$$

where

$${}^{\perp}\mathcal{A} := \{T \in \mathcal{T} \mid \mathsf{Hom}(T, A[t]) = 0 \text{ for all } A \in \mathcal{A}, t \in \mathbb{Z}\}, \tag{2.2}$$

$$\mathcal{A}^{\perp} := \{T \in \mathcal{T} \mid \mathsf{Hom}(A[t], T) = 0 \text{ for all } A \in \mathcal{A}, t \in \mathbb{Z}\}, \tag{2.3}$$

are the left and the right orthogonals to $\mathcal{A}$ in $\mathcal{T}$. More generally, if $\mathcal{A}_1, \dots, \mathcal{A}_m$ is a semiorthogonal collection of admissible subcategories in $\mathcal{T}$, then for each $0 \le k \le m$ there is a semiorthogonal decomposition

$$\mathcal{T} = \langle \mathcal{A}_1, \dots, \mathcal{A}_k, {}^{\perp}\langle \mathcal{A}_1, \dots, \mathcal{A}_k \rangle \cap \langle \mathcal{A}_{k+1}, \dots, \mathcal{A}_m \rangle^{\perp}, \mathcal{A}_{k+1}, \dots, \mathcal{A}_m \rangle. \tag{2.4}$$

The simplest example of an admissible subcategory is the one generated by an exceptional object. Recall that an object E is exceptional if one has $\mathsf{Hom}(E, E) = \mathsf{k}$ and $\mathsf{Hom}(E, E[t]) = 0$ for $t \ne 0$. An exceptional collection is a collection of exceptional objects $E_1, E_2, \dots, E_m$ such that $\mathsf{Hom}(E_i, E_j[t]) = 0$ for all $i > j$ and all $t \in \mathbb{Z}$. An exceptional collection in $\mathcal{T}$ gives rise to a semiorthogonal decomposition

$$\mathcal{T} = \langle \mathcal{A}, E_1, \dots, E_m \rangle \quad \text{with} \quad \mathcal{A} = \langle E_1, \dots, E_m \rangle^{\perp}. \tag{2.5}$$

Here E_i denotes the subcategory generated by the same named exceptional object. If the category $\mathcal{A}$ in (2.5) is zero the exceptional collection is called full.

2.2. Full exceptional collections. There are several well-known and quite useful semiorthogonal decompositions. The simplest example is the following

Theorem 2.1 (Beilinson's collection). *There is a full exceptional collection*

$$\mathbf{D}^b(\mathsf{coh}(\mathbb{P}^n)) = \langle \mathcal{O}_{\mathbb{P}^n}, \mathcal{O}_{\mathbb{P}^n}(1), \dots, \mathcal{O}_{\mathbb{P}^n}(n) \rangle. \tag{2.6}$$

Of course, twisting by $\mathcal{O}_{\mathbb{P}^n}(t)$ we get $\langle \mathcal{O}_{\mathbb{P}^n}(t), \mathcal{O}_{\mathbb{P}^n}(t+1), \dots, \mathcal{O}_{\mathbb{P}^n}(t+n)\rangle$ which is also a full exceptional collection for each $t \in \mathbb{Z}$.

A bit more general is the Grassmannian variety:

Theorem 2.2 (Kapranov's collection, [19]). *Let $\mathsf{Gr}(k,n)$ be the Grassmannian of k-dimensional subspaces in a vector space of dimension n. Let $\mathcal{U}$ be the tautological subbundle of rank k. If $\mathsf{char}\,\mathsf{k} = 0$ then there is a semiorthogonal decomposition*

$$\mathbf{D}^b(\mathsf{coh}(\mathsf{Gr}(k,n))) = \langle\, \Sigma^\alpha \mathcal{U}^\vee \,\rangle_{\alpha \subset R(k,n-k)}, \tag{2.7}$$

where $R(k,n-k)$ is the $k\times(n-k)$ rectangle, α is a Young diagram, and Σ^α is the associated Schur functor.

When $\mathsf{char}\,\mathsf{k} > 0$ there is an exceptional collection as well, but it is a bit more complicated, see [9].

Another interesting case is the case of a smooth quadric $Q^n \subset \mathbb{P}^{n+1}$.

Theorem 2.3 (Kapranov's collection, [19]). *When $\mathsf{char}\,\mathsf{k} \neq 2$ and k is algebraically closed, there is a full exceptional collection*

$$\mathbf{D}^b(\mathsf{coh}(Q^n)) = \begin{cases} \langle \mathsf{S}, \mathcal{O}_{Q^n}, \mathcal{O}_{Q^n}(1), \dots, \mathcal{O}_{Q^n}(n-1)\rangle, & \text{if } n \text{ is odd} \\ \langle \mathsf{S}^-, \mathsf{S}^+, \mathcal{O}_{Q^n}, \mathcal{O}_{Q^n}(1), \dots, \mathcal{O}_{Q^n}(n-1)\rangle, & \text{if } n \text{ is even} \end{cases} \tag{2.8}$$

where S and $\mathsf{S}^\pm$ are the spinor bundles.

Many exceptional collections have been constructed on other rational homogeneous spaces, see e.g. [48], [32], [47], [42], [11], and [41]. Full exceptional collections on smooth toric varieties (and stacks) were constructed by Kawamata [20]. Also exceptional collections were constructed on del Pezzo surfaces [46], some Fano threefolds [24, 45] and many other varieties.

2.3. Relative versions. Let S be a scheme and E a vector bundle of rank r on it. Let $\mathbb{P}_S(E)$ be its projectivization, $f : \mathbb{P}_S(E) \to S$ the projection, and $\mathcal{O}_{\mathbb{P}_S(E)/S}(1)$ the Grothendieck line bundle on $\mathbb{P}_S(E)$.

Theorem 2.4 ([46]). *For each $i \in \mathbb{Z}$ the functor*

$$\Phi_i : \mathbf{D}^b(\mathsf{coh}(S)) \to \mathbf{D}^b(\mathsf{coh}(\mathbb{P}_S(E))), \qquad F \mapsto Lf^*(F) \overset{\mathbb{L}}{\otimes} \mathcal{O}_{\mathbb{P}_S(E)/S}(i) \tag{2.9}$$

is fully faithful, and there is a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(\mathbb{P}_S(E))) = \langle \Phi_0(\mathbf{D}^b(\mathsf{coh}(S))), \dots, \Phi_{r-1}(\mathbf{D}^b(\mathsf{coh}(S)))\rangle. \tag{2.10}$$

Of course, analogously to the case of a projective space, one can replace the sequence of functors $\Phi_0, \dots, \Phi_{r-1}$ by $\Phi_t, \dots, \Phi_{t+r-1}$ for any $t \in \mathbb{Z}$.

An interesting new feature appears for Severi–Brauer varieties. Recall that a Severi–Brauer variety over S is a morphism $f : X \to S$ which étale locally is isomorphic to a projectivization of a vector bundle. A Severi–Brauer variety X can be constructed from a torsion element in the Brauer group $\mathsf{Br}(S)$ of S.

Theorem 2.5 (Bernardara's decomposition, [3]). *Let $f : X \to S$ be a Severi–Brauer variety of relative dimension n and $\beta \in \mathsf{Br}(S)$ its Brauer class. Then for each $i \in \mathbb{Z}$ there is a fully faithful functor $\Phi_i : \mathbf{D}^b(\mathsf{coh}(S, \beta^i)) \to \mathbf{D}^b(\mathsf{coh}(X))$ and a semiorthogonal decomposition*

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \Phi_0(\mathbf{D}^b(\mathsf{coh}(S))), \Phi_1(\mathbf{D}^b(\mathsf{coh}(S, \beta))), \dots, \Phi_n(\mathbf{D}^b(\mathsf{coh}(S, \beta^n)))\rangle. \quad (2.11)$$

Here $\mathsf{coh}(S, \beta^i)$ is the category of β^i-twisted coherent sheaves on S and the functor Φ_i is given by $F \mapsto Lf^*(F) \overset{\mathbb{L}}{\otimes} \mathcal{O}_{X/S}(i)$, where the sheaf $\mathcal{O}_{X/S}(i)$ is well defined as a $f^*\beta^{-i}$-twisted sheaf.

Another important semiorthogonal decomposition can be constructed for a smooth blowup. Let $Y \subset X$ be a locally complete intersection subscheme of codimension c and let $\tilde{X}$ be the blowup of X with center in Y. Let $f : \tilde{X} \to X$ be the blowup morphism and $D \subset \tilde{X}$ the exceptional divisor of the blowup. Let $i : D \to \tilde{X}$ be the embedding and $p : D \to Y$ the natural projection (the restriction of f to D). Note that $D \cong \mathbb{P}_Y(\mathcal{N}_{Y/X})$ is the projectivization of the normal bundle.

Theorem 2.6 (Blowup formula, [46]). *The functor $Lf^* : \mathbf{D}^b(\mathsf{coh}(X)) \to \mathbf{D}^b(\mathsf{coh}(\tilde{X}))$ as well as the functors*

$$\Psi_k : \mathbf{D}^b(\mathsf{coh}(Y)) \to \mathbf{D}^b(\mathsf{coh}(\tilde{X})), \qquad F \mapsto Ri_*(Lp^*(F) \overset{\mathbb{L}}{\otimes} \mathcal{O}_{D/Y}(k)),$$

are fully faithful for all $k \in \mathbb{Z}$, and there is a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(\tilde{X})) = \langle Lf^*(\mathbf{D}^b(\mathsf{coh}(X))), \Psi_0(\mathbf{D}^b(\mathsf{coh}(Y))), \dots, \Psi_{c-2}(\mathbf{D}^b(\mathsf{coh}(Y)))\rangle. \quad (2.12)$$

Finally, consider a flat fibration in quadrics $f : X \to S$. In other words, assume that $X \subset \mathbb{P}_S(E)$ is a divisor of relative degree 2 in a projectivization of a vector bundle E of rank $n + 2$ on a scheme S corresponding to a line subbundle $\mathscr{L} \subset S^2E^\vee$.

Theorem 2.7 (Quadratic fibration formula, [31]). *For each $i \in \mathbb{Z}$ there is a fully faithful functor*

$$\Phi_i : \mathbf{D}^b(\mathsf{coh}(S)) \to \mathbf{D}^b(\mathsf{coh}(X)), \qquad F \mapsto Lf^*(F) \overset{\mathbb{L}}{\otimes} \mathcal{O}_{X/S}(i)$$

and a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathbf{D}^b(\mathsf{coh}(S, \mathcal{C}\ell_0)), \Phi_0(\mathbf{D}^b(\mathsf{coh}(S))), \dots, \Phi_{n-1}(\mathbf{D}^b(\mathsf{coh}(S)))\rangle, \quad (2.13)$$

where $\mathcal{C}\ell_0$ is the sheaf of even parts of Clifford algebras on S associated with the quadric fibration $X \to S$.

The sheaf $\mathcal{C}\ell_0$ is a sheaf of $\mathcal{O}_S$-algebras which as an $\mathcal{O}_S$-module is isomorphic to

$$\mathcal{C}\ell_0 \cong \mathcal{O}_S \oplus (\Lambda^2 E \otimes \mathscr{L}) \oplus (\Lambda^4 E \otimes \mathscr{L}^2) \oplus \dots$$

and equipped with an algebra structure via the Clifford multiplication. If the dimension n of fibers of $X \to S$ is odd, then $\mathcal{C}\ell_0$ is a sheaf of Azumaya algebras on the open subset of S corresponding to nondegenerate quadrics (which of course may be empty). On the other hand, if n is even then $\mathcal{O}_S \oplus \Lambda^n E \otimes \mathscr{L}^{n/2}$ is a central subalgebra in $\mathcal{C}\ell_0$, so the latter gives a sheaf $\tilde{\mathcal{C}\ell}_0$ of algebras on the twofold covering

$$\tilde{S} := \mathsf{Spec}_S(\mathcal{O}_S \oplus \Lambda^n E \otimes \mathscr{L}^{n/2}) \quad (2.14)$$

of S, and $\tilde{\mathcal{C}\ell}_0$ is a sheaf of Azumaya algebras on the preimage of the open subset of S corresponding to nondegenerate quadrics.

2.4. Base change. A triangulated category $\mathcal{T}$ is S-linear if it is equipped with a module structure over the tensor triangulated category $\mathbf{D}^b(\mathsf{coh}(S))$. In particular, if X is a scheme over S and $f : X \to S$ is the structure morphism then a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{A}_1, \dots, \mathscr{A}_m \rangle \tag{2.15}$$

is S-linear if each of subcategories $\mathscr{A}_k$ is closed under tensoring with an object of $\mathbf{D}^b(\mathsf{coh}(S))$, i.e. for $A \in \mathscr{A}_k$ and $F \in \mathbf{D}^b(\mathsf{coh}(S))$ one has $A \overset{\mathbb{L}}{\otimes} Lf^*(F) \in \mathscr{A}_k$.

The semiorthogonal decompositions of Theorems 2.4, 2.5 and 2.7 are S-linear, and the blowup formula of Theorem 2.6 is X-linear. The advantage of linear semiorthogonal decompositions lies in the fact that they obey a base change result. For a base change $T \to S$ denote by $\pi : X \times_S T \to X$ the induced projection.

Theorem 2.8 ([37]). *If X is an algebraic variety over S and* (2.15) *is an S-linear semiorthogonal decomposition then for a change of base morphism $T \to S$ there is, under a certain technical condition, a T-linear semiorthogonal decomposition*

$$\mathbf{D}^b(\mathsf{coh}(X \times_S T)) = \langle \mathscr{A}_{1T}, \dots, \mathscr{A}_{mT} \rangle$$

such that $\pi^*(A) \subset \mathscr{A}_{iT}$ for any $A \in \mathscr{A}_i$ and $\pi_*(A') \subset \mathscr{A}_i$ for any $A' \in \mathscr{A}_{iT}$ which has proper support over X.

2.5. Important questions. There are several questions which might be crucial for further investigations.

Question 2.9. *Find a good condition for an exceptional collection to be full.*

One might hope that if the collection generates the Grothendieck group (or the Hochschild homology) of the category then it is full. However, recent examples of quasiphantom and phantom categories (see section 6.2) show that this is not the case. Still we may hope that in the categories generated by exceptional collections there are no phantoms. In other words one could hope that the following is true.

Conjecture 2.10. *Let $\mathcal{T} = \langle E_1, \dots, E_n \rangle$ be a triangulated category generated by an exceptional collection. Then any exceptional collection of length n in $\mathcal{T}$ is full.*

If there is an action of a group G on an algebraic variety X, one can consider the equivariant derived category $\mathbf{D}^b(\mathsf{coh}^G(X))$ along with the usual derived category. In many interesting cases (flag varieties, toric varieties, GIT quotients) it is quite easy to construct a full exceptional collection in the equivariant category. It would be extremely useful to find a way to transform it into a full exceptional collection in the usual category. In some sense the results of [41] can be considered as an example of such an approach.

Another very important question is to find possible restrictions for existence of semiorthogonal decompositions. Up to now there are only several cases when we can *prove* indecomposability of a category. The first is the derived category of a curve of positive genus. The proof (see e.g. [44]) is based on special properties of categories of homological dimension 1. Another is the derived category of a Calabi–Yau variety (smooth connected variety with trivial canonical class). Its indecomposability is proved by a surprisingly simple argument due to Bridgeland [8]. This was further generalized in [21] to varieties with globally generated canonical class. On the other hand, the original argument of Bridgeland generalizes to any connected Calabi–Yau category (i.e. with the Serre functor isomorphic to a shift and Hochschild cohomology in degree zero isomorphic to k).

3. Homological projective duality

The starting point of a homological projective duality (HP duality for short) is a smooth projective variety X with a morphism into a projective space and a semiorthogonal decomposition of $\mathbf{D}^b(\mathsf{coh}(X))$ of a very special type.

3.1. Lefschetz decompositions. Let X be an algebraic variety and $\mathscr{L}$ a line bundle on X.

Definition 3.1. A right Lefschetz decomposition of $\mathbf{D}^b(\mathsf{coh}(X))$ with respect to $\mathscr{L}$ is a semiorthogonal decomposition of form

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{A}_0, \mathscr{A}_1 \otimes \mathscr{L}, \dots, \mathscr{A}_{m-1} \otimes \mathscr{L}^{m-1} \rangle \tag{3.1}$$

with $0 \subset \mathscr{A}_{m-1} \subset \cdots \subset \mathscr{A}_1 \subset \mathscr{A}_0$. In other words, each component of the decomposition is a subcategory of the previous component twisted by $\mathscr{L}$.

Analogously, a left Lefschetz decomposition of $\mathbf{D}^b(\mathsf{coh}(X))$ with respect to $\mathscr{L}$ is a semiorthogonal decomposition of form

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{B}_{m-1} \otimes \mathscr{L}^{1-m}, \dots, \mathscr{B}_1 \otimes \mathscr{L}^{-1}, \mathscr{B}_0 \rangle \tag{3.2}$$

with $0 \subset \mathscr{B}_{m-1} \subset \cdots \subset \mathscr{B}_1 \subset \mathscr{B}_0$.

The subcategories $\mathscr{A}_i$ (resp. $\mathscr{B}_i$) forming a Lefschetz decomposition will be called blocks, the largest will be called the first block. Usually we will consider right Lefschetz decompositions. So, we will call them simply Lefschetz decompositions.

Beilinson's collection on $\mathbb{P}^n$ is an example of a Lefschetz decomposition with $\mathscr{A}_0 = \mathscr{A}_1 = \cdots = \mathscr{A}_n = \langle \mathcal{O}_{\mathbb{P}^n} \rangle$. Kapranov's collection on the Grassmannian $\mathsf{Gr}(k, n)$ also has a Lefschetz structure with the category $\mathscr{A}_i$ generated by $\Sigma^\alpha \mathcal{U}^\vee$ for $\alpha \subset R(k-1, n-k-i)$.

Note that in Definition 3.1 one can replace the twist by a line bundle with any other autoequivalence of $\mathbf{D}^b(\mathsf{coh}(X))$ and get the notion of a Lefschetz decomposition with respect to an autoequivalence. This may be especially useful when dealing with arbitrary triangulated categories.

It is also useful to know that for a given line bundle $\mathscr{L}$ a Lefschetz decomposition is completely determined by its first block. Moreover, an admissible subcategory extends to a right Lefschetz decomposition if and only if it extends to a left Lefschetz decomposition. The simplest example of an admissible subcategory which does not extend to a Lefschetz decomposition is the subcategory $\langle \mathcal{O}_{\mathbb{P}^2}, \mathcal{O}_{\mathbb{P}^2}(2) \rangle \subset \mathbf{D}^b(\mathsf{coh}(\mathbb{P}^2))$.

Question 3.2. *Find a good sufficient condition for a Lefschetz extendability of an admissible subcategory $\mathscr{A}_0 \subset \mathbf{D}^b(\mathsf{coh}(X))$.*

One can define a partial ordering on the set of all Lefschetz decompositions of $\mathbf{D}^b(\mathsf{coh}(X))$ by inclusions of their first blocks. As we will see soon, the most interesting and strong results are obtained by using minimal Lefschetz decompositions.

3.2. Hyperplane sections. Let X be a smooth projective variety with a morphism into a projective space $f : X \to \mathbb{P}(V)$ (not necessarily an embedding). Put $\mathcal{O}_X(1) := f^*\mathcal{O}_{\mathbb{P}(V)}(1)$ and assume that a right Lefschetz decomposition with respect to $\mathcal{O}_X(1)$

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{A}_0, \mathscr{A}_1(1), \dots, \mathscr{A}_{m-1}(m-1) \rangle \tag{3.3}$$

is given (we abbreviate $\mathscr{A}_i(i) := \mathscr{A}_i \otimes \mathcal{O}_X(i)$). Consider the dual projective space $\mathbb{P}(V^\vee)$. By the base change (Theorem 2.8) the product $X \times \mathbb{P}(V^\vee)$ inherits a $\mathbb{P}(V^\vee)$-linear semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(X \times \mathbb{P}(V^\vee))) = \langle \mathscr{A}_{0\mathbb{P}(V^\vee)}, \mathscr{A}_1(1)_{\mathbb{P}(V^\vee)}, \dots, \mathscr{A}_{m-1}(m-1)_{\mathbb{P}(V^\vee)} \rangle$$

Consider the universal hyperplane section of X, $\mathscr{X} := X \times_{\mathbb{P}(V)} Q \subset X \times \mathbb{P}(V^\vee)$, where $Q \subset \mathbb{P}(V) \times \mathbb{P}(V^\vee)$ is the incidence quadric and denote by $\alpha : \mathscr{X} \to X \times \mathbb{P}(V^\vee)$ the natural embedding.

Lemma 3.3. *The functor $L\alpha^* : \mathbf{D}^b(\mathsf{coh}(X \times \mathbb{P}(V^\vee))) \to \mathbf{D}^b(\mathsf{coh}(\mathscr{X}))$ is fully faithful on each of the subcategories $\mathscr{A}_1(1)_{\mathbb{P}(V^\vee)}, \dots, \mathscr{A}_{m-1}(m-1)_{\mathbb{P}(V^\vee)}$ and induces a $\mathbb{P}(V^\vee)$-linear semiorthogonal decomposition*

$$\mathbf{D}^b(\mathsf{coh}(\mathscr{X})) = \langle \mathscr{C}, \mathscr{A}_1(1)_{\mathbb{P}(V^\vee)}, \dots, \mathscr{A}_{m-1}(m-1)_{\mathbb{P}(V^\vee)} \rangle. \tag{3.4}$$

The first component $\mathscr{C}$ of this decomposition is called the HP dual category of X. It is a very interesting category, especially if it can be identified with the derived category of some algebraic variety Y. In this case this variety is called the HP dual variety of X.

Definition 3.4. An algebraic variety Y equipped with a morphism $g : Y \to \mathbb{P}(V^\vee)$ is called homologically projectively dual to $f : X \to \mathbb{P}(V)$ with respect to a given Lefschetz decomposition (3.3) if there is given an object $\mathcal{E} \in \mathbf{D}^b(\mathsf{coh}(Q(X,Y)))$ such that the Fourier–Mukai functor $\Phi_{\mathcal{E}} : \mathbf{D}^b(\mathsf{coh}(Y)) \to \mathbf{D}^b(\mathsf{coh}(\mathscr{X}))$ is an equivalence onto the HP dual subcategory $\mathscr{C} \subset \mathbf{D}^b(\mathsf{coh}(\mathscr{X}))$ of (3.4).

Here $Q(X,Y) = (X \times Y) \times_{\mathbb{P}(V)\times\mathbb{P}(V^\vee)} Q = \mathscr{X} \times_{\mathbb{P}(V^\vee)} Y$. If a homological projective duality between varieties X and Y is established then there is an interesting relation between derived categories of their linear sections.

3.3. Homologically projectively duality statement. For each linear subspace $L \subset V^\vee$ denote its orthogonal complement in V by $L^\perp := \mathsf{Ker}(V \to L^\vee)$. Further denote

$$X_L := X \times_{\mathbb{P}(V)} \mathbb{P}(L^\perp), \qquad Y_L := Y \times_{\mathbb{P}(V^\vee)} \mathbb{P}(L). \tag{3.5}$$

Varieties defined in this way are called mutually orthogonal linear sections of X and Y. We will say that X_L and Y_L have expected dimensions if

$$\dim X_L = \dim X - r \qquad \text{and} \qquad \dim Y_L = \dim Y - (N - r),$$

where $N = \dim V$ and $r = \dim L$ (so that $N - r = \dim L^\perp$).

Theorem 3.5 (Homological projective duality, [29]). *Let (Y, g) be an HP dual variety for (X, f) with respect to* (3.3)*. Then*

(1) *Y is smooth and $\mathbf{D}^b(\mathsf{coh}(Y))$ has an admissible subcategory $\mathscr{B}_0$ equivalent to $\mathscr{A}_0$ and extending to a left Lefschetz decomposition*

$$\mathbf{D}^b(\mathsf{coh}(Y)) = \langle \mathscr{B}_{n-1}(1-n), \dots, \mathscr{B}_1(-1), \mathscr{B}_0 \rangle, \quad \mathscr{B}_{n-1} \subset \dots \subset \mathscr{B}_1 \subset \mathscr{B}_0. \tag{3.6}$$

(2) *(X, f) is HP dual to (Y, g) with respect to* (3.6)*.*

(3) *The set of critical values of g is the classical projective dual of X.*

(4) *For any subspace $L \subset V^\vee$ if X_L and Y_L have expected dimensions then there are semiorthogonal decompositions*

$$\mathbf{D}^b(\mathsf{coh}(X_L)) = \langle \mathscr{C}_L, \mathscr{A}_r(r), \dots, \mathscr{A}_{m-1}(m-1) \rangle, \tag{3.7}$$

$$\mathbf{D}^b(\mathsf{coh}(Y_L)) = \langle \mathscr{B}_{n-1}(1-n), \dots, \mathscr{B}_{N-r}(r-N), \mathscr{C}_L \rangle \tag{3.8}$$

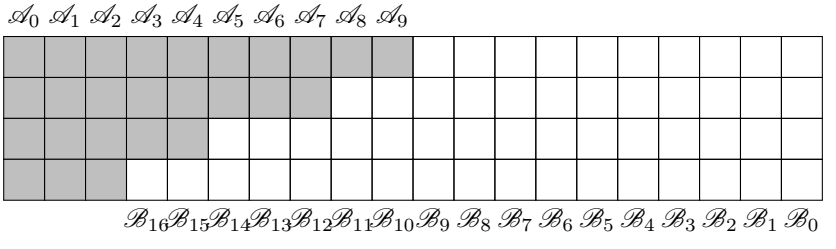
with the same triangulated category $\mathscr{C}_L$ appearing in the RHS.

The decomposition (3.6) of $\mathbf{D}^b(\mathsf{coh}(Y))$ will be referred to as the HP dual Lefschetz decomposition. The common component $\mathscr{C}_L$ of decompositions (3.7) and (3.8) will be referred to as the nontrivial part of the derived categories of X_L and Y_L, while the subcategories $\mathscr{A}_i(i)$ and $\mathscr{B}_j(-j)$ (one checks that the pullback functors for the embeddings $X_L \to X$ and $Y_L \to Y$ are fully faithful on the subcategories $\mathscr{A}_i$ and $\mathscr{B}_j$ for $i \geq r$ and $j \geq N - r$) are considered as trivial (in the sense that they come from the ambient varieties).

The first two statements of this Theorem show that the relation we are dealing with is indeed a duality, the third statement shows the relation to the classical projective duality (and so justifies the word "projective" in the name), and the last statement is the real result. We will soon see how powerful it is.

Note also that in the statement of the Theorem the linear sections X_L and Y_L need not be smooth. In fact, one can show that for HP dual varieties X and Y a section X_L is smooth if and only if its orthogonal section Y_L is smooth, but no matter whether this is the case or not, the decompositions (3.7) and (3.8) hold true.

Now let us say some words about the relations of the Lefschetz decompositions (3.3) and (3.6) for HP dual varieties. As it was already mentioned, the largest components of those are just equivalent $\mathscr{B}_0 \cong \mathscr{A}_0$. Further, the component $\mathscr{B}_i$ is very close to the orthogonal complement of $\mathscr{A}_{N-1-i}$ in $\mathscr{A}_0$. More precisely, these two categories have semiorthogonal decompositions with the same components but with in general different gluing functors. This can be visualized by a picture.



The gray part of the picture corresponds to the initial Lefschetz decomposition, the columns correspond to its blocks, while the white part corresponds to the dual decomposition, the complementary columns correspond to the complementary subcategories of the dual Lefschetz decomposition. The number of rows is equal to the number of different components in the initial (and the dual) Lefschetz decomposition. In this example picture $\mathscr{A}_0 = \mathscr{A}_1 = \mathscr{A}_2 \neq \mathscr{A}_3 = \mathscr{A}_4 \neq \mathscr{A}_5 = \mathscr{A}_6 = \mathscr{A}_7 \neq \mathscr{A}_8 = \mathscr{A}_9$, and so one can say that the rows correspond to the "primitive parts" $(\mathscr{A}_3)^\perp_{\mathscr{A}_0}$, $(\mathscr{A}_5)^\perp_{\mathscr{A}_3}$, $(\mathscr{A}_8)^\perp_{\mathscr{A}_5}$, and $\mathscr{A}_8$ of all the categories in the picture, the length of the initial decomposition is $m = 10$, the length of the dual decomposition is $n = 17$, while the dimension of the ambient space is $N = 20$.

Note that $\mathscr{B}_i = 0$ if and only if $\mathscr{A}_{N-1-i} = \mathscr{A}_0$, so the number n of components in (3.6) equals N minus the number of components in (3.3) equal to $\mathscr{A}_0$.

In fact, the best (in many aspects) situation is when in the original Lefschetz decomposition (3.3) all components coincide $\mathscr{A}_0 = \mathscr{A}_1 = \cdots = \mathscr{A}_{m-1}$ (such Lefschetz decompositions are called rectangular). Then the HP dual Lefschetz decomposition is also rectangular, has the same components $\mathscr{B}_0 = \mathscr{B}_1 = \cdots = \mathscr{B}_{n-1} \cong \mathscr{A}_0$ and

$$n = N - m$$

(in particular in a picture analogous to the above the gray and the white parts are rectangles, which explains the name "rectangular"). Moreover, in this case for any $0 < r < N$ one has either $r \geq m$ or $N - r \geq n$, hence in decompositions (3.7) and (3.8) either the first or the second category has only one component $\mathscr{C}_L$ and nothing else. Then the other decomposition shows that the nontrivial component of the derived category of a linear section is equivalent to the derived category of the orthogonal linear section of the dual variety.

3.4. HP duality and noncommutative varieties. In general, the homologically projectively dual category $\mathscr{C} \subset \mathbf{D}^b(\mathsf{coh}(\mathscr{X}))$ defined by (3.4) need not be equivalent to $\mathbf{D}^b(\mathsf{coh}(Y))$ for an algebraic variety Y. In fact, only a few such cases are known — the linear duality, the duality for quadrics, the duality for Grassmannians $\mathsf{Gr}(2,4)$ and $\mathsf{Gr}(2,5)$, and the spinor variety $\mathbb{S}_5$ (see section 5).

One can get many additional interesting examples by allowing Y to be a noncommutative variety. Here a noncommutative variety can be understood in different ways. If one uses the most general sense — as a semiorthogonal component of the derived category of an algebraic variety — then tautologically the HP dual category $\mathscr{C}$ itself will provide a noncommutative HP dual variety. In fact, one can develop a theory of HP duality for noncommutative varieties in this sense and prove the same results (see [30]). However, in this most general form the semiorthogonal decompositions provided by the HP duality Theorem will not have an apparent geometric interpretation.

In fact, an interesting geometry arises in HP duality if the dual variety Y is close to a commutative variety. For example, it often happens that there is a (commutative) algebraic variety Y_0 with a map $g_0 : Y_0 \to \mathbb{P}(V^\vee)$, a sheaf of finite $\mathcal{O}_{Y_0}$-algebras $\mathcal{R}$ on Y_0 whose bounded derived category $\mathbf{D}^b(\mathsf{coh}(Y_0, \mathcal{R}))$ of coherent $\mathcal{R}$-modules on Y_0 is equivalent to the HP dual category $\mathscr{C}$ of X and such that the equivalence $\mathscr{C} \cong \mathbf{D}^b(\mathsf{coh}(Y_0, \mathcal{R}))$ is given by an appropriate object $\mathcal{E} \in \mathbf{D}^b(\mathsf{coh}(Q(X, Y_0), \mathcal{R}))$. Of course, one can easily allow here the variety X also to be noncommutative in the same sense. It is easy to modify all the definitions accordingly.

In section 5 we discuss examples showing that this generalization is meaningful. Among such examples are the Veronese–Clifford duality, the Grassmannian–Pfaffian duality, and their generalizations.

In fact, in some of these examples, the HP duality Theorem 3.5 still gives semiorthogonal decompositions for usual commutative varieties (even though the dual variety is noncommutative). Indeed, the sheaf of algebras $\mathcal{R}$ on Y_0 is frequently isomorphic to a matrix algebra on an open subset of Y_0, typically, on its smooth locus — in fact, in these cases the noncommutative variety $(Y_0, \mathcal{R})$ can be thought of as a categorical resolution of singularities of Y. In this situation, taking a subspace $L \subset V^\vee$ such that Y_{0L} is contained in that open subset, one gets $\mathbf{D}^b(\mathsf{coh}(Y_L)) = \mathbf{D}^b(\mathsf{coh}(Y_{0L}, \mathcal{R})) \cong \mathbf{D}^b(\mathsf{coh}(Y_{0L}))$.

4. Categorical resolutions of singularities

As it was explained above (and we will see in some of the examples below) in many cases the HP dual variety looks as a noncommutative (or categorical) resolution of singularities of a singular variety. So, a good notion of a categorical resolution is necessary for the theory.

4.1. The definition. If $\pi : \tilde{Y} \to Y$ is a resolution of singularities (in the usual sense), we have an adjoint pair of triangulated functors

$$R\pi_* : \mathbf{D}^b(\mathsf{coh}(\tilde{Y})) \to \mathbf{D}^b(\mathsf{coh}(Y))$$

and

$$L\pi^* : \mathbf{D}^{\mathsf{perf}}(Y) \to \mathbf{D}^b(\mathsf{coh}(\tilde{Y}))$$

(here $\mathbf{D}^{\mathsf{perf}}(Y)$ stands for the category of perfect complexes on Y). We axiomatize this situation in the following

Definition 4.1 (cf. [33, 39]). A categorical resolution of singularities of a scheme Y is a smooth triangulated category $\mathcal{T}$ and an adjoint pair of triangulated functors

$$\pi_* : \mathcal{T} \to \mathbf{D}^b(\mathsf{coh}(Y))$$

and

$$\pi^* : \mathbf{D}^{\mathsf{perf}}(Y) \to \mathcal{T}$$

such that $\pi_* \circ \pi^* \cong \mathsf{id}_{\mathbf{D}^{\mathsf{perf}}(Y)}$. In particular, the functor π^* is fully faithful.

We will not discuss the notion of smoothness for a triangulated category. In fact, for our purposes it is always enough to assume that $\mathcal{T}$ is an admissible Y-linear subcategory of $\mathbf{D}^b(\mathsf{coh}(\tilde{Y}))$ for a geometric resolution $\tilde{Y} \to Y$.

Let $(\mathcal{T}, \pi_*, \pi^*)$ and $(\mathcal{T}', \pi'_*, \pi'^*)$ be two categorical resolutions of Y. We will say that $\mathcal{T}$ dominates $\mathcal{T}'$ if there is a fully faithful functor $\epsilon : \mathcal{T}' \to \mathcal{T}$ such that $\pi'_* = \pi_* \circ \epsilon$. Clearly, this is compatible with the usual dominance relation between geometric resolutions — if a resolution $\pi : \tilde{Y} \to Y$ factors as $\tilde{Y} \xrightarrow{f} \tilde{Y}' \xrightarrow{\pi'} Y$ then the pullback functor $\epsilon := Lf^* : \mathbf{D}^b(\mathsf{coh}(\tilde{Y}')) \to \mathbf{D}^b(\mathsf{coh}(\tilde{Y}))$ is fully faithful and

$$R\pi_* \circ Lf^* = R\pi'_* \circ Rf_* \circ Lf^* \cong R\pi'_*.$$

Categorical resolutions have two advantages in comparison with geometric ones. First, if Y has irrational singularities the pullback functor for a geometric resolution is never fully faithful and so its derived category is not a categorical resolution in sense of Definition 4.1. However, it was shown in [39] that any separated scheme of finite type (even nonreduced) over a field of zero characteristic admits a categorical resolution.

The second advantage is that the dominance order for categorical resolutions is more flexible. For example, in many examples one can find a categorical resolution which is much smaller than any geometric resolution. There are strong indications that the Minimal Model Program on the categorical level may be much simpler than the classical one. In particular, we expect the following.

Conjecture 4.2 (cf. [7]). *For any quasiprojective scheme Y there exists a categorical resolution which is minimal with respect to the dominance order.*

4.2. Examples of categorical resolutions. As an evidence for the conjecture we will construct categorical resolutions which are presumably minimal.

Theorem 4.3 ([33]). *Let $f : \tilde{Y} \to Y$ be a resolution of singularities and let E be the exceptional divisor with $i : E \to \tilde{Y}$ being the embedding. Assume that the derived category $\mathbf{D}^b(\mathsf{coh}(E))$ has a left Lefschetz decomposition with respect to the conormal bundle $\mathcal{O}_E(-E)$:*

$$\mathbf{D}^b(\mathsf{coh}(E)) = \langle \mathscr{C}_{m-1}((m-1)E), \dots, \mathscr{C}_1(E), \mathscr{C}_0 \rangle, \tag{4.1}$$

which is Y-linear and has $Li^*(Lf^*(\mathbf{D}^{\mathsf{perf}}(Y))) \subset \mathscr{C}_0$. Then the functor Ri_* is fully faithful on subcategories $\mathscr{C}_k \subset \mathbf{D}^b(\mathsf{coh}(E))$ for $k > 0$, the subcategory

$$\tilde{\mathscr{C}} := \{F \in \mathbf{D}^b(\mathsf{coh}(\tilde{Y})) \mid Li^*(F) \in \mathscr{C}_0\} \tag{4.2}$$

is admissible in $\mathbf{D}^b(\mathsf{coh}(\tilde{Y}))$, and there is a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(\tilde{Y})) = \langle Ri_*(\mathscr{C}_{m-1}((m-1)E)), \dots, Ri_*(\mathscr{C}_1(E)), \tilde{\mathscr{C}} \rangle. \tag{4.3}$$

Moreover, the functor $Lf^* : \mathbf{D}^{\mathsf{perf}}(Y) \to \mathbf{D}^b(\mathsf{coh}(\tilde{Y}))$ factors as a composition of a fully faithful functor $\pi^* : \mathbf{D}^{\mathsf{perf}}(Y) \to \tilde{\mathscr{C}}$ with the embedding $\gamma : \tilde{\mathscr{C}} \to \mathbf{D}^b(\mathsf{coh}(\tilde{Y}))$, and the functors $\pi_* := Rf_* \circ \gamma$ and π^* give $\tilde{\mathscr{C}}$ a structure of a categorical resolution of singularities of Y.

If $\mathscr{C}'_0 \subset \mathscr{C}_0 \subset \mathbf{D}^b(\mathsf{coh}(E))$ are two admissible Lefschetz extendable subcategories (with respect to the conormal bundle) then clearly by (4.2) the categorical resolution $\tilde{\mathscr{C}}'$ constructed from $\mathscr{C}'_0$ is a subcategory in the categorical resolution $\tilde{\mathscr{C}}$ constructed from $\mathscr{C}_0$. Moreover, if $\epsilon : \tilde{\mathscr{C}}' \to \tilde{\mathscr{C}}$ is the embedding functor then $\pi'_* = \pi_* \circ \epsilon$, so $\tilde{\mathscr{C}}$ dominates $\tilde{\mathscr{C}}'$. This shows that minimal categorical resolutions are related to minimal Lefschetz decompositions.

As an example of the application of the above Theorem consider the cone Y over a smooth projective variety $X \subset \mathbb{P}(V)$. Then $\tilde{Y} = \mathsf{Tot}_X(\mathcal{O}_X(-1))$, the total space of the line bundle $\mathcal{O}_X(-1) = \mathcal{O}_{\mathbb{P}(V)}(-1)_{|X}$, is a geometric resolution of Y. The exceptional divisor of the natural morphism $f : \tilde{Y} \to Y$ then identifies with the zero section of the total space, $E = X$, and the conormal bundle identifies with $\mathcal{O}_X(1)$. So, a left Lefschetz decomposition of $\mathbf{D}^b(\mathsf{coh}(X))$ with respect to $\mathcal{O}_X(1)$ gives a categorical resolution of the cone Y over X.

Example 4.4. Take $X = \mathbb{P}^3$ with the double Veronese embedding $f : \mathbb{P}^3 \to \mathbb{P}^9$, so that $f^*\mathcal{O}_{\mathbb{P}^9}(1) = \mathcal{O}_{\mathbb{P}^3}(2)$, and a left Lefschetz decomposition

$$\mathbf{D}^b(\mathsf{coh}(\mathbb{P}^3)) = \langle \mathscr{C}_1(-2), \mathscr{C}_0 \rangle \quad \text{with } \mathscr{C}_0 = \mathscr{C}_1 = \langle \mathcal{O}_{\mathbb{P}^3}(-1), \mathcal{O}_{\mathbb{P}^3} \rangle.$$

Then the category $\tilde{\mathscr{C}} := \{F \in \mathbf{D}^b(\mathsf{coh}(\mathsf{Tot}_{\mathbb{P}^3}(\mathcal{O}_{\mathbb{P}^3}(-2)))) \mid Li^*F \in \langle \mathcal{O}_{\mathbb{P}^3}(-1), \mathcal{O}_{\mathbb{P}^3} \rangle\}$ is a categorical resolution of the Veronese cone, which is significantly smaller than the usual geometric resolution. It is expected to be minimal.

4.3. Crepancy of categorical resolutions. Crepancy is an important property of a resolution which in the geometric situation ensures its minimality. A resolution $f : \tilde{Y} \to Y$ is crepant if the relative canonical class $K_{\tilde{Y}/Y}$ is trivial. There is an analogue of crepancy for categorical resolutions. In fact, there are two such analogues.

Definition 4.5 ([33]). A categorical resolution $(\mathcal{T}, \pi_*, \pi^*)$ of a scheme Y is weakly crepant if the functor $\pi^* : \mathbf{D}^{\mathsf{perf}}(Y) \to \mathcal{T}$ is simultaneously left and right adjoint to the functor $\pi_* : \mathcal{T} \to \mathbf{D}^b(\mathsf{coh}(Y))$.

By Grothendieck duality, the right adjoint of the derived pushforward functor Rf_* : $\mathbf{D}^b(\mathsf{coh}(\tilde{Y})) \to \mathbf{D}^b(\mathsf{coh}(Y))$ is given by $f^!(F) = Lf^*(F) \otimes \mathcal{O}_{\tilde{Y}}(K_{\tilde{Y}/Y})$, so for a geometric resolution crepancy and weak crepancy are equivalent.

Definition 4.6 ([33]). A categorical resolution $(\mathcal{T}, \pi_*, \pi^*)$ of a scheme Y is strongly crepant if the relative Serre functor of $\mathcal{T}$ over $\mathbf{D}^b(\mathsf{coh}(Y))$ is isomorphic to the identity.

Again, Grothendieck duality implies that for a geometric resolution crepancy and strong crepancy are equivalent. Moreover, it is not so difficult to show that strong crepancy of a categorical resolution implies its weak crepancy, but the converse is not true in general. To see this one can analyze the weak and strong crepancy of categorical resolutions provided by Theorem 4.3.

Proposition 4.7. *In the setup of Theorem* 4.3 *assume that the scheme Y is Gorenstein and $K_{\tilde{Y}/Y} = (m-1)E$, where m is the length of the left Lefschetz decomposition* (4.1)*. The corresponding categorical resolution $\tilde{\mathscr{C}}$ of Y is weakly crepant if and only if*

$$Li^*(Lf^*(\mathbf{D}^{\mathsf{perf}}(Y))) \subset \mathscr{C}_{m-1}. \tag{4.4}$$

Furthermore, $\tilde{\mathscr{C}}$ is strongly crepant if and only if (4.1) *is rectangular, i.e.*

$$\mathscr{C}_{m-1} = \cdots = \mathscr{C}_1 = \mathscr{C}_0. \tag{4.5}$$

So, starting from a nonrectangular Lefschetz decomposition it is easy to produce an example of a weakly crepant categorical resolution which is not strongly crepant.

Example 4.8. Take $X = Q^3 \subset \mathbb{P}^4$ and let Y be the cone over X (i.e. a 4-dimensional quadratic cone). Then the left Lefschetz collection

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{C}_2(-2), \mathscr{C}_1(-1), \mathscr{C}_0 \rangle \qquad \text{with } \mathscr{C}_0 = \langle \mathsf{S}, \mathcal{O}_X \rangle, \mathscr{C}_1 = \mathscr{C}_2 = \langle \mathcal{O}_X \rangle$$

(S is the spinor bundle) gives a weakly crepant categorical resolution $\tilde{\mathscr{C}}$ of Y which is not strongly crepant. In fact, if $q : \mathsf{Tot}_X(\mathcal{O}_X(-1)) \to X$ is the canonical projection, the vector bundle $q^*\mathsf{S}$ is a spherical object in $\tilde{\mathscr{C}}$ and the relative Serre functor is isomorphic to the corresponding spherical twist.

4.4. Further questions. Of course, the central question is Conjecture 4.2. Theorem 4.3 shows that it is closely related to the question of existence of minimal Lefschetz decompositions.

Another interesting question is to find new methods of construction of minimal categorical resolutions. An interesting development in this direction is the work [1] in which a notion of a wonderful resolution of singularities (an analogue of wonderful compactifications) is introduced and it is shown that a wonderful resolution gives rise to a weakly crepant categorical resolution. This can be viewed as an advance on the first part of Proposition 4.7. It would be very interesting to find a generalization of the second part of this Proposition in the context of wonderful resolutions.

Another aspect is to find explicit constructions of minimal resolutions for interesting varieties, such as Pfaffian varieties for example. Some of these arise naturally in the context of HP duality as we will see later.

5. Examples of homological projective duality

If an HP duality for two varieties X and Y is proved, one gets as a consequence an identification of the nontrivial components of the derived categories of linear sections of X and Y. Because of that it is clear that such a result is a very strong statement and is usually not so easy to prove. In this section we list several examples of HP duality. We assume that $\mathsf{char}\,\mathsf{k} = 0$ in this section.

5.1. Linear duality. Let $X = \mathbb{P}_S(E)$ be a projectivization of a vector bundle E on a scheme S and assume that the map $f : X \to \mathbb{P}(V)$ is *linear on fibers* of X over S. In other words, we assume that f is induced by an embedding of vector bundles $E \to V \otimes \mathcal{O}_S$ on S. In this case the line bundle $\mathcal{O}_X(1) = f^*\mathcal{O}_{\mathbb{P}(V)}(1)$ is a Grothendieck line bundle for X over S. By Theorem 2.4 we have a rectangular Lefschetz decomposition of $\mathbf{D}^b(\mathsf{coh}(X))$ of length $m = \mathsf{rk}(E)$ with blocks

$$\mathscr{A}_0 = \mathscr{A}_1 = \cdots = \mathscr{A}_{m-1} = p^*(\mathbf{D}^b(\mathsf{coh}(S))),$$

where $p : X \to S$ is the projection. So, we are in the setup of HP duality and one can ask what the dual variety is?

The answer turns out to be given by a projectivization of another vector bundle over S. Define $E^\perp$ as the kernel of the dual morphism

$$E^\perp := \mathsf{Ker}(V^\vee \otimes \mathcal{O}_S \to E^\vee).$$

The projectivization $\mathbb{P}_S(E^\perp)$ comes with a natural morphism $g : \mathbb{P}_S(E^\perp) \to \mathbb{P}(V^\vee)$ and Theorem 2.4 provides $\mathbb{P}_S(E^\perp)$ with a rectangular Lefschetz decomposition of length $N-m$ with blocks $\mathscr{B}_0 = \mathscr{B}_1 = \cdots = \mathscr{B}_{N-m-1} = q^*(\mathbf{D}^b(\mathsf{coh}(S)))$, where $q : \mathbb{P}_S(E^\perp) \to S$ is the projection.

Theorem 5.1 ([29]). *The projectivizations $X = \mathbb{P}_S(E)$ and $Y = \mathbb{P}_S(E^\perp)$ with their canonical morphisms to $\mathbb{P}(V)$ and $\mathbb{P}(V^\vee)$ and the above Lefschetz decompositions are homologically projectively dual to each other.*

The picture visualizing this duality is very simple:

with m gray boxes and $N - m$ white boxes.

In the very special case of $S = \mathsf{Spec}\,\mathsf{k}$ the bundle E is just a vector space and the variety X is a (linearly embedded) projective subspace $\mathbb{P}(E) \subset \mathbb{P}(V)$. Then the HP-dual variety is the orthogonal subspace $\mathbb{P}(E^\perp) \subset \mathbb{P}(V^\vee)$. In particular, the dual of the space $\mathbb{P}(V)$ itself with respect to its identity map is the empty set.

5.2. Quadrics. There are two ways to construct a smooth quadric: one — as a smooth hypersurface of degree 2 in a projective space, and the other — as a double covering of a projective space ramified in a smooth quadric hypersurface. These representations interchange in a funny way in HP duality.

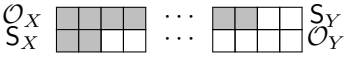
Denote by S the spinor bundle on an odd dimensional quadric or one of the spinor bundles on the even dimensional quadric.

Theorem 5.2.

(1) *If $X = Q^{2m} \subset \mathbb{P}^{2m+1}$ with Lefschetz decomposition given by*

$$\mathscr{A}_0 = \mathscr{A}_1 = \langle \mathsf{S}_X, \mathcal{O}_X \rangle, \qquad \mathscr{A}_2 = \mathscr{A}_3 = \cdots = \mathscr{A}_{2m-1} = \langle \mathcal{O}_X \rangle \tag{5.1}$$

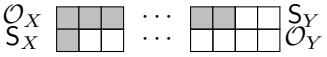
then the HP dual variety is the dual quadric $Y = Q^\vee \subset \check{\mathbb{P}}^{2m+1}$ with the same Lefschetz decomposition.



(2) *If $X = Q^{2m-1} \subset \mathbb{P}^{2m}$ with Lefschetz decomposition given by*

$$\mathscr{A}_0 = \langle \mathsf{S}_X, \mathcal{O}_X \rangle, \qquad \mathscr{A}_1 = \mathscr{A}_2 = \cdots = \mathscr{A}_{2m-2} = \langle \mathcal{O}_X \rangle \tag{5.2}$$

then the HP dual variety is the double covering $Y \to \check{\mathbb{P}}^{2m}$ ramified in the dual quadric $Q^\vee \subset \check{\mathbb{P}}^{2m}$ with Lefschetz decomposition (5.1)



(3) *If $X = Q^{2m-1} \to \mathbb{P}^{2m-1}$ is the double covering ramified in a quadric $\bar{Q} \subset \mathbb{P}^{2m-1}$ with Lefschetz decomposition* (5.2) *then the HP dual variety is the double covering $Y \to \check{\mathbb{P}}^{2m-1}$ ramified in the dual quadric $\bar{Q}^\vee \subset \check{\mathbb{P}}^{2m-1}$ with the same Lefschetz decomposition.*

5.3. Veronese–Clifford duality. Let W be a vector space of dimension n and $V = S^2W$ its symmetric square. We take $X = \mathbb{P}(W)$ and consider its double Veronese embedding $f : \mathbb{P}(W) \to \mathbb{P}(V)$. Then $f^*\mathcal{O}_{\mathbb{P}(V)}(1) = \mathcal{O}_{\mathbb{P}(W)}(2)$. Beilinson's collection (2.6) on $\mathbb{P}(W)$ can be considered as a Lefschetz decomposition (with respect to $\mathcal{O}_{\mathbb{P}(W)}(2)$) of $\mathbf{D}^b(\mathsf{coh}(\mathbb{P}(W)))$ with $\lfloor n/2 \rfloor$ blocks equal to

$$\mathscr{A}_0 = \mathscr{A}_1 = \cdots = \mathscr{A}_{\lfloor n/2 \rfloor - 1} := \langle \mathcal{O}_{\mathbb{P}(W)}, \mathcal{O}_{\mathbb{P}(W)}(1) \rangle,$$

and if n is odd one more block

$$\mathscr{A}_{\lfloor n/2 \rfloor} := \langle \mathcal{O}_{\mathbb{P}(W)} \rangle.$$

The universal hyperplane section $\mathscr{X}$ of X is nothing but the universal quadric in $\mathbb{P}(W)$ over the space $\mathbb{P}(V^\vee) = \mathbb{P}(S^2W^\vee)$ of all quadrics. Then the quadratic fibration formula of Theorem 2.7 gives an equivalence of the HP dual category $\mathscr{C}$ with the derived category $\mathbf{D}^b(\mathsf{coh}(\mathbb{P}(V^\vee), \mathcal{C}\ell_0))$ of coherent sheaves of modules over the even part of the universal Clifford algebra

$$\mathcal{C}\ell_0 = \mathcal{O}_{\mathbb{P}(S^2W^\vee)} \oplus \Lambda^2 W \otimes \mathcal{O}_{\mathbb{P}(S^2W^\vee)}(-1) \oplus \ldots,$$

on the space $\mathbb{P}(S^2W^\vee)$ of quadrics. We will consider the pair $(\mathbb{P}(S^2W^\vee), \mathcal{C}\ell_0)$ as a noncommutative variety and call it the Clifford space.

Theorem 5.3 (Veronese–Clifford duality, [31]). *The homological projective dual of the projective space $X = \mathbb{P}(W)$ in the double Veronese embedding $\mathbb{P}(W) \to \mathbb{P}(S^2W)$ is the Clifford space $Y = (\mathbb{P}(S^2W^\vee), \mathscr{C}\ell_0)$.*

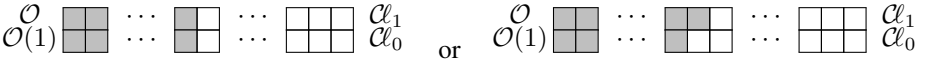
The HP dual Lefschetz decomposition of the Clifford space is given by the full exceptional collection

$$\mathbf{D}^b(\mathsf{coh}(\mathbb{P}(S^2W^\vee, \mathscr{C}\ell_0))) = \langle \mathscr{C}\ell_{1-n^2}, \mathscr{C}\ell_{2-n^2}, \dots, \mathscr{C}\ell_{-1}, \mathscr{C}\ell_0 \rangle,$$

where

$$\mathscr{C}\ell_1 = W \otimes \mathcal{O}_{\mathbb{P}(S^2W^\vee)} \oplus \Lambda^3 W \otimes \mathcal{O}_{\mathbb{P}(S^2W^\vee)}(-1) \oplus \dots$$

is the odd part of the Clifford algebra and $\mathscr{C}\ell_{k-2} = \mathscr{C}\ell_k \otimes \mathcal{O}_{\mathbb{P}(S^2W^\vee)}(-1)$ for each $k \in \mathbb{Z}$. The picture visualizing this duality is:



n even $\qquad$ n odd

For even n it has $n/2$ gray columns and $n^2/2$ white columns, and for odd n it has $(n-1)/2$ gray columns, one mixed column, and $(n^2-1)/2$ white columns.

5.4. Grassmannian–Pfaffian duality. The most interesting series of examples is provided by Grassmannians $\mathsf{Gr}(2, m)$ of two-dimensional subspaces in an m-dimensional vector space.

Let W be a vector space of dimension m and let $V = \Lambda^2 W$, the space of bivectors. The group $\mathsf{GL}(W)$ acts on the projective space $\mathbb{P}(\Lambda^2 W)$ with orbits indexed by the rank of a bivector which is always even and ranges from 2 to $2\lfloor m/2 \rfloor$. We denote by $\mathsf{Pf}(2k, W)$ the closure of the orbit consisting of bivectors of rank $2k$ and call it the k-th Pfaffian variety. Clearly, the smallest orbit $\mathsf{Pf}(2, W)$ is smooth and coincides with the Grassmannian $\mathsf{Gr}(2, W)$ in its Plücker embedding. Another smooth Pfaffian variety is the maximal one — $\mathsf{Pf}(2\lfloor m/2 \rfloor, W) = \mathbb{P}(\Lambda^2 W)$. All the intermediate Pfaffians are singular with $\mathsf{sing}(\mathsf{Pf}(2k, W)) = \mathsf{Pf}(2k-2, W)$. The submaximal Pfaffian variety $\mathsf{Pf}(2\lfloor m/2 \rfloor - 2, W^\vee)$ of the dual space is classically projectively dual to the Grassmannian $\mathsf{Gr}(2, W)$. This suggests a possible HP duality between them.

To make a precise statement we should choose a Lefschetz decomposition of $\mathbf{D}^b(\mathsf{coh}(\mathsf{Gr}(2, W)))$. A naive choice is to take Kapranov's collection (2.7). It can be considered as a Lefschetz decomposition on $X := \mathsf{Gr}(2, m)$ with $m-1$ blocks

$$\mathscr{A}_0 = \langle \mathcal{O}_X, \mathcal{U}_X^\vee, \dots, S^{m-2}\mathcal{U}_X^\vee \rangle,\ \mathscr{A}_1 = \langle \mathcal{O}_X, \mathcal{U}_X^\vee, \dots, S^{m-3}\mathcal{U}_X^\vee \rangle,\ \dots,\ \mathscr{A}_{m-2} = \langle \mathcal{O}_X \rangle.$$

However, it is very far from being minimal. It turns out that a reasonable result can be obtained for another Lefschetz decomposition

$$\mathbf{D}^b(\mathsf{coh}(\mathsf{Gr}(2, m)) = \langle \mathscr{A}_0, \mathscr{A}_1(1), \dots, \mathscr{A}_{m-1}(m-1) \rangle$$

with

$$\mathscr{A}_0 = \dots = \mathscr{A}_{m-1} = \langle \mathcal{O}_X, \mathcal{U}_X^\vee, \dots, S^{(m-1)/2}\mathcal{U}_X^\vee \rangle. \tag{5.3}$$

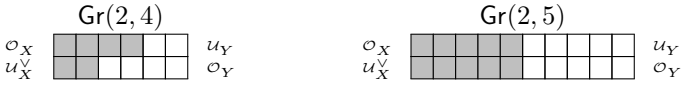
if m is odd, and with

$$\begin{aligned} \mathscr{A}_0 = \dots = \mathscr{A}_{m/2-1} &= \langle \mathcal{O}_X, \mathcal{U}_X^\vee, \dots, S^{m/2-1}\mathcal{U}_X^\vee \rangle, \\ \mathscr{A}_{m/2} = \dots = \mathscr{A}_{m-1} &= \langle \mathcal{O}_X, \mathcal{U}_X^\vee, \dots, S^{m/2-2}\mathcal{U}_X^\vee \rangle, \end{aligned} \tag{5.4}$$

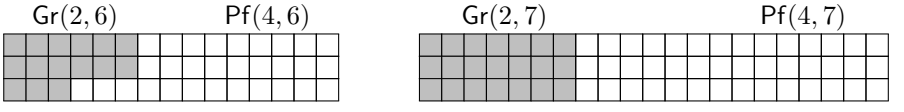
if m is even.

Conjecture 5.4. *The HP dual of the Grassmannian* $\mathsf{Gr}(2, W)$ *with Lefschetz decomposition* (5.4) (*or* (5.3) *depending on the parity of* $m = \dim W$) *is given by a minimal categorical resolution of the submaximal Pfaffian* $\mathsf{Pf}(2\lfloor m/2 \rfloor - 2, W^\vee)$. *When* m *is odd, this resolution is strongly crepant.*

This conjecture is proved for $m \le 7$ in [28]. In fact, for $m = 2$ and $m = 3$ one has $\mathsf{Gr}(2, W) = \mathbb{P}(\Lambda^2 W)$ and linear duality applies. For $m = 4$ and $m = 5$ the submaximal Pfaffian $\mathsf{Pf}(2, W^\vee)$ coincides with the Grassmannian, and the above duality is the duality for Grassmannians:



For $m = 6$ and $m = 7$ the submaximal Pfaffian $Y = \mathsf{Pf}(4, W^\vee)$ is singular, but its appropriate categorical resolutions can be constructed by Theorem 4.3. It turns out that these resolutions indeed are HP dual to the corresponding Grassmannians:



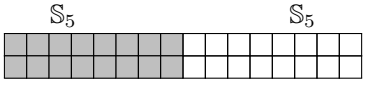
For $m \ge 8$ this construction of a categorical resolution does not work. However it is plausible that the Pfaffians have wonderful resolutions of singularities, so a development of [1] may solve the question.

5.5. The spinor duality. Let W be a vector space of even dimension $2m$ and $q \in S^2 W^\vee$ a nondegenerate quadratic form. The isotropic Grassmannian of m-dimensional subspaces in W has two connected components, abstractly isomorphic to each other and called spinor varieties $\mathbb{S}_m$. These are homogeneous spaces of the spin group $\mathsf{Spin}(W)$ with the embedding into $\mathbb{P}(\Lambda^m W)$ given by the square of the generator of the Picard group, while the generator itself gives an embedding into the projectivization $\mathbb{P}(V)$ of a half-spinor representation V (of dimension 2^{m-1}) of $\mathsf{Spin}(W)$. For small m the spinor varieties are very simple (because the spin-group simplifies): in fact, $\mathbb{S}_1$ is a point, $\mathbb{S}_2 = \mathbb{P}^1$, $\mathbb{S}_3 = \mathbb{P}^3$, and $\mathbb{S}_4 = Q^6$. The first interesting example is $\mathbb{S}_5$.

Theorem 5.5 ([27]). *The spinor variety* $X = \mathbb{S}_5$ *has a Lefschetz decomposition*

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{A}_0, \mathscr{A}_1(1), \dots, \mathscr{A}_7(7) \rangle \quad \textit{with} \quad \mathscr{A}_0 = \dots = \mathscr{A}_7 = \langle \mathcal{O}_X, \mathcal{U}_5^\vee \rangle. \tag{5.5}$$

The HP dual variety is the same spinor variety $Y = \mathbb{S}_5$.



5.6. Incomplete dualities. It is often quite hard to give a full description of the HP dual variety. On the other hand, there is sometimes an open dense subset $U \subset \mathbb{P}(V^\vee)$ for which there is a description of the category $\mathscr{C}_U$ obtained by a base change $U \to \mathbb{P}(V^\vee)$ from the HP dual category $\mathscr{C}$. If $\mathbf{D}^b(\mathsf{coh}(Y_U))$ is a (noncommutative) variety such that $Y_U \cong \mathscr{C}_U$ (a U-linear equivalence), we will say that Y_U is the HP dual of X over U, or an incomplete HP dual variety.

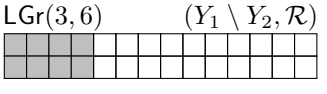
Proposition 5.6. *If Y_U is an HP dual of X over an open subset $U \subset \mathbb{P}(V^\vee)$ then the semiorthogonal decompositions* (3.7) *and* (3.8) *hold for any subspace $L \subset V^\vee$ such that $\mathbb{P}(L) \subset U$ and the varieties X_L and Y_L have expected dimensions.*

There are several examples of HP duality when only an incomplete dual variety is known. Below we discuss two of them. Let W be a vector space of dimension 6 with a symplectic form ω, and $X = \mathsf{LGr}(3, W) \subset \mathbb{P}^{13}$ the corresponding Lagrangian Grassmannian. The classical projectively dual variety $Y_1 := X^\vee \subset \check{\mathbb{P}}^{13}$ is a quartic hypersurface which is singular along a 9-dimensional variety $Y_2 \subset Y_1$. One can check that X has a rectangular Lefschetz decomposition (see [48])

$$\begin{aligned} &\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{A}_0, \mathscr{A}_1(1), \mathscr{A}_2(2), \mathscr{A}_3(3)\rangle \quad \text{with} \\ &\mathscr{A}_0 = \mathscr{A}_1 = \mathscr{A}_2 = \mathscr{A}_3 = \langle \mathcal{O}_X, \mathcal{U}_X^\vee\rangle. \end{aligned} \tag{5.6}$$

The HP dual variety in this case is described only over the open set $U = \check{\mathbb{P}}^{13} \setminus Y_2$, see [27]. For this a morphism $\pi : \tilde{Y} \to Y_1$ which is a nondegenerate conic bundle over $Y_1 \setminus Y_2$ is constructed in *loc. cit.* Let $\mathcal{R}$ be the associated quaternion algebra on $Y_1 \setminus Y_2$.

Theorem 5.7 ([27]). *If $X = \mathsf{LGr}(3, 6) \subset \mathbb{P}^{13}$ with Lefschetz decomposition* (5.6) *then the noncommutative variety $(Y_1 \setminus Y_2, \mathcal{R})$ is the HP dual of X over $\check{\mathbb{P}}^{13} \setminus Y_2$.*



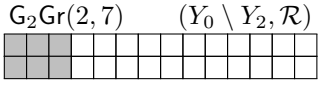
Another example is related to the simple algebraic group G_2. Let X be the orbit of the highest weight vector in the projectivization $\mathbb{P}^{13} = \mathbb{P}(V)$ of the adjoint representation V of G_2. Then X can also be realized as the zero locus of a global section of the vector bundle $\Lambda^4(W/\mathcal{U})$ on the Grassmannian $\mathsf{Gr}(2, W)$ for a 7-dimensional fundamental representation W of G_2 corresponding to a generic 3-form $\lambda \in \Lambda^3 W^\vee$. By that reason we use the notation $\mathsf{G}_2\mathsf{Gr}(2, W)$ for X.

The classical projectively dual variety $Y_1 := X^\vee \subset \check{\mathbb{P}}^{13}$ is a sextic hypersurface which is singular along a 10-dimensional variety $Y_2 \subset Y_1$. One can check that X has a rectangular Lefschetz decomposition

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{A}_0, \mathscr{A}_1(1), \mathscr{A}_2(2)\rangle \quad \text{with} \quad \mathscr{A}_0 = \mathscr{A}_1 = \mathscr{A}_2 = \langle \mathcal{O}_X, \mathcal{U}_X^\vee\rangle. \tag{5.7}$$

As in the previous case, the HP dual variety is described only over the open set $U = \check{\mathbb{P}}^{13} \setminus Y_2$, see [27]. A morphism $\pi : \tilde{Y} \to Y_0$ to the double covering $Y_0 \to \check{\mathbb{P}}^{13}$ ramified in Y_1 which is a Severi–Brauer variety with fiber $\mathbb{P}^2$ over $Y_0 \setminus Y_2$ is constructed in *loc. cit.* Let $\mathcal{R}$ be the associated Azumaya algebra on $Y_0 \setminus Y_2$.

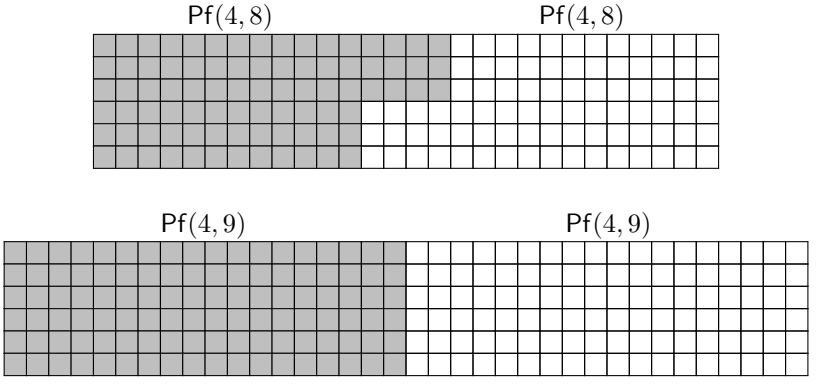
Theorem 5.8 ([27]). *If $X = \mathsf{G}_2\mathsf{Gr}(2, 7)$ with Lefschetz decomposition* (5.7) *then the noncommutative variety $(Y_0 \setminus Y_2, \mathcal{R})$ is the HP dual of X over $\check{\mathbb{P}}^{13} \setminus Y_2$.*



5.7. Conjectures. Note that $\mathsf{Gr}(2, W) = \mathsf{Pf}(2, W)$. This suggests a generalization of the Grassmannian–Pfaffian duality to higher Pfaffians.

Conjecture 5.9. *For any k there is an HP duality between appropriate minimal categorical resolutions of the Pfaffians $\mathsf{Pf}(2k, W)$ and $\mathsf{Pf}(2(\lfloor m/2 \rfloor - k), W^\vee)$. When $m = \dim W$ is odd these resolutions are strongly crepant.*

Below are the expected pictures for the HP duality for $\mathsf{Pf}(4, 8)$ and for $\mathsf{Pf}(4, 9)$:



It is also expected that there is a generalization of the Veronese–Clifford duality. Consider a vector space W of dimension n and its symmetric square $V = S^2W$. The group $\mathsf{GL}(W)$ acts on the projective space $\mathbb{P}(V) = \mathbb{P}(S^2W)$ with orbits indexed by the rank of a tensor. We denote by $\Sigma(k, W) \subset \mathbb{P}(S^2W)$ the closure of the orbit consisting of symmetric tensors of rank k. The smallest orbit $\Sigma(1, W)$ is smooth and coincides with the double Veronese embedding of $\mathbb{P}(W)$. On the other hand, $\Sigma(n, W) = \mathbb{P}(S^2W)$ is also smooth. All the intermediate varieties $\Sigma(k, W)$ are singular with $\mathsf{sing}(\Sigma(k, W)) = \Sigma(k-1, W)$. The classical projective duality acts on these varieties by $\Sigma(k, W)^\vee = \Sigma(n-k, W^\vee)$. However, the HP duality is organized in a much more complicated way.

Besides $\Sigma(k, W)$ itself one can consider its modifications:

- the Clifford modification $(\Sigma(k, W), \mathcal{C}\ell_0)$ for the natural sheaf of even parts of Clifford algebras on it, and
- (for even k) the double covering $\tilde{\Sigma}(k, W)$ of $\Sigma(k, W)$ corresponding to the central subalgebra in $\mathcal{C}\ell_0$ as in (2.14),

and their minimal categorical resolutions. It seems that HP duality interchanges in a complicated way modifications of different type. For example, besides the original Veronese–Clifford duality between $\Sigma(1, n)$ and $(\Sigma(n, n), \mathcal{C}\ell_0)$ there are strong indications that (the minimal resolution of) $\Sigma(2, 4)$ is HP dual to (the minimal resolution of) the double covering $\tilde{\Sigma}(4, 4)$ of $\Sigma(4, 4) = \mathbb{P}^9$ (see [15]), (the minimal resolution of) $\Sigma(2, 5)$ is HP dual to (the minimal resolution of) the double covering $\tilde{\Sigma}(4, 5)$ of $\Sigma(4, 5)$ (see [14]), while (the minimal resolution of) the double covering $\tilde{\Sigma}(2, n)$ of $\Sigma(2, n)$ is HP dual to (the minimal resolution of) $\Sigma(n-1, n)$ for all n (this can be easily deduced from the linear duality).

6. Varieties of small dimension

Let us list what is known about semiorthogonal decompositions of smooth projective varieties by dimension. In this section we assume that $\mathsf{k} = \mathbb{C}$.

6.1. Curves. Curves are known to have no nontrivial semiorthogonal decompositions with the only exception of $\mathbb{P}^1$ (for which every semiorthogonal decomposition coincides with the Beilinson decomposition up to a twist), see [44].

6.2. Surfaces. For surfaces the situation is more complicated. Of course, by the blowup formula any surface has a semiorthogonal decomposition with several exceptional objects and the derived category of a minimal surface as components. In particular, any rational surface has a full exceptional collection. Moreover, for $\mathbb{P}^2$ it is known that any full exceptional collection can be obtained from Beilinson's collection by mutations (which are related to Markov numbers and toric degenerations of $\mathbb{P}^2$). For other del Pezzo surfaces all exceptional objects have been classified [23], and moreover, three-blocks exceptional collections were constructed [18], but the complete picture is not known.

For minimal ruled surfaces, of course there is a semiorthogonal decomposition into two copies of the derived category of the curve which is the base of the ruling.

For surfaces of Kodaira dimension 0 it is well known that there are no nontrivial semiorthogonal decompositions for K3 and abelian surfaces. The same was proved for bielliptic surfaces in [21]. For Enriques surfaces there may be an exceptional collection of line bundles of length up to 10 (see [49]), and for so-called nodal Enriques surfaces the complementary component is related to the Artin–Mumford quartic double solid [15].

For surfaces of Kodaira dimension 1 with $p_q > 0$ there are no semiorthogonal decompositions by [21].

Finally, for surfaces of general type there is an unexpectedly rich theory of semiorthogonal decompositions. In fact, for many surfaces of general type with $p_g = q = 0$ (the classical Godeaux surface, the Beauville surface, the Burniat surfaces, the determinantal Barlow surface, some fake projective planes) exceptional collections of length equal to the rank of the Grothendieck group have been constructed in [2, 4, 5, 10, 12, 13]. The collections, however, are not full. The complementary components have finite (or even zero) Grothendieck group and trivial Hochschild homology and by that reason they are called quasiphantom or phantom categories. The phantoms cannot be detected by additive invariants, but one can use Hochschild cohomology instead, see [38].

An interesting feature here is that the structure of the constructed exceptional collections resembles very much the structure of exceptional collections of del Pezzo surfaces with the same K^2. The only (but a very important) difference is that whenever there is a Hom-space between exceptional bundles on del Pezzo, the corresponding exceptional bundles on the surface of general type have Ext^2-space. This seemingly small difference, however, has a very strong effect on the properties of the category. See more details in *loc. cit.*

6.3. Fano 3-folds. For derived categories of threefolds (and higher dimensional varieties) there are no classification results (as there is no classification of threefolds). Of course, as it already was mentioned for varieties with trivial (or globally generated) canonical class there are no nontrivial decompositions. So, from now on we will discuss Fano varieties.

In dimension 3 all Fano varieties were classified in the works of Fano, Iskovskikh and Mukai. All Fano 3-folds with Picard number greater than 1 are either the blowups of other Fano varieties with centers in points and smooth curves (and then their derived category reduces to the derived category of a Fano 3-fold with smaller Picard number), or conic bundles over rational surfaces (see Tables 12.3–12.6 of [17]). For conic bundles one can use the quadratic bundle formula (Theorem 2.7). It gives a semiorthogonal decomposition with several exceptional objects and the derived category of sheaves of modules over the even part of the Clifford algebra on the base of the bundle.

If the Picard number is 1, the next discrete invariant of a Fano 3-fold to look at is the index, i.e. the maximal integer dividing the canonical class. By Fujita's Theorem the only Fano 3-folds of index greater than 2 are $\mathbb{P}^3$ and Q^3. Their derived categories are well understood, so let us turn to 3-folds of index 2 and 1.

For a Fano 3-fold Y of index 2 the pair of line bundles $(\mathcal{O}_Y, \mathcal{O}_Y(1))$ is exceptional and gives rise to a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(Y)) = \langle \mathscr{B}_Y, \mathcal{O}_Y, \mathcal{O}_Y(1) \rangle. \tag{6.1}$$

The component $\mathscr{B}_Y$ is called the nontrivial component of $\mathbf{D}^b(\mathsf{coh}(Y))$.

A similar decomposition can be found for a Fano 3-fold X of index 1 if its degree $d_X := (-K_X)^3$ is not divisible by 4 (the degree of a 3-fold of index 1 is always even). By a result of Mukai [43] if $d_X > 2$ on such X there is an exceptional vector bundle $\mathcal{E}_X$ of rank 2 with $c_1(\mathcal{E}_X) = K_X$, which is moreover orthogonal to the structure sheaf of X. In other words, $(\mathcal{E}_X, \mathcal{O}_X)$ is an exceptional pair and there is a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{A}_X, \mathcal{E}_X, \mathcal{O}_X \rangle. \tag{6.2}$$

The component $\mathscr{A}_X$ is called the nontrivial component of $\mathbf{D}^b(\mathsf{coh}(X))$.

It is rather unexpected that the nontrivial parts $\mathscr{B}_Y$ and $\mathscr{A}_X$ for a Fano 3-fold Y of index 2 and degree $d_Y := (-K_Y/2)^3$ and for a Fano 3-fold X of index 1 and degree $d_X = 4d_Y + 2$ have the same numerical characteristics, and are, moreover, expected to belong to the same deformation family of categories. In fact, this expectation is supported by the following result. Recall that the degree of a Fano 3-fold of index 2 with Picard number 1 satisfies $1 \le d \le 5$, while the degree of a Fano 3-fold of index 1 with Picard number 1 is even and satisfies $2 \le d \le 22$, $d \ne 20$. So there are actually 5 cases to consider.

Theorem 6.1 ([34]). *For $3 \le d \le 5$ each category $\mathscr{B}_{Y_d}$ is equivalent to some category $\mathscr{A}_{X_{4d+2}}$ and vice versa.*

See *loc. cit.* for a precise statement. In fact, for $d = 5$ the category is rigid and is equivalent to the derived category of representations of the quiver with 2 vertices and 3 arrows from the first vertex to the second (this follows from the construction of explicit exceptional collections in the derived categories of Y_5 and X_{22}, see [45] and [24]). Further, for $d = 4$ each of the categories $\mathscr{B}_{Y_4}$ and $\mathscr{A}_{X_{18}}$ is equivalent to the derived category of a curve of genus 2, and moreover, each smooth curve appears in both pictures. This follows from HP duality for the double Veronese embedding of $\mathbb{P}^5$ and from HP duality for G_2 Grassmannian respectively (Theorem 5.8). Finally, for $d = 3$ no independent description of the category in question is known, but the HP duality for the Grassmannian $\mathsf{Gr}(2, 6)$ gives the desired equivalence (see [25, 28]).

It turns out, however, that already for $d = 2$ the situation is more subtle. It seems that in that case the categories $\mathscr{B}_{Y_2}$ lie at the boundary of the family of categories $\mathscr{A}_{X_{10}}$. And for $d = 1$ the situation is completely unclear.

The situation with Fano 3-folds of index 1 and degree divisible by 4 is somewhat different. For such threefolds it is, in general, not clear how one can construct an exceptional pair. However, for $d_X = 12$ and $d_X = 16$ this is possible. For $d_X = 12$ Mukai has proved [43] that there is an exceptional pair $(\mathcal{E}_5, \mathcal{O}_X)$ where $\mathcal{E}_5$ is a rank 5 exceptional bundle with $c_1(\mathcal{E}_5) = 2K_X$. Using HP duality for the spinor variety $\mathbb{S}_5$ (Theorem 5.5) one can check that this pair extends to a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(X_{12})) = \langle \mathbf{D}^b(\mathsf{coh}(C_7)), \mathcal{E}_5, \mathcal{O}_{X_{12}} \rangle,$$

where C_7 is a smooth curve of genus 7 (see [26, 27]). Analogously, for $d_X = 16$ Mukai has constructed [43] an exceptional bundle $\mathcal{E}_3$ of rank 3 with $c_1(\mathcal{E}_3) = K_X$. Using HP duality for $\mathsf{LGr}(3, 6)$ (Theorem 5.7) one can check that there is a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(X_{16})) = \langle \mathbf{D}^b(\mathsf{coh}(C_3)), \mathcal{E}_3, \mathcal{O}_{X_{16}} \rangle,$$

where C_3 is a smooth curve of genus 3 [27].

6.4. Fourfolds. Of course, for Fano 4-folds we know much less than for 3-folds. So, we will not even try to pursue a classification, but will restrict attention to some very special cases of interest.

Maybe one of the most interesting 4-folds is the cubic 4-fold. One of its salient features is the hyperkähler structure on the Fano scheme of lines, which turns out to be a deformation of the second Hilbert scheme of a K3 surface. This phenomenon has a nice explanation from the derived categories point of view.

Theorem 6.2 ([36]). *Let $Y \subset \mathbb{P}^5$ be a cubic 4-fold. Then there is a semiorthogonal decomposition*

$$\mathbf{D}^b(\mathsf{coh}(Y)) = \langle \mathscr{A}_Y, \mathcal{O}_Y, \mathcal{O}_Y(1), \mathcal{O}_Y(2) \rangle,$$

and its nontrivial component $\mathscr{A}_Y$ is a Calabi–Yau category of dimension 2. Moreover, the category $\mathscr{A}_Y$ is equivalent to the derived category of coherent sheaves on a K3 surface, at least if Y is a Pfaffian cubic 4-fold, or if Y contains a plane Π and a 2-cycle Z such that $\deg Z + Z \cdot \Pi \equiv 1 \bmod 2$.

To establish this result for Pfaffian cubics one can use HP duality for $\mathsf{Gr}(2, 6)$. The associated K3 is then a linear section of this Grassmannian. For cubics with a plane a quadratic bundle formula for the projection of Y from the plane Π gives the result. The K3 surface then is the double covering of $\mathbb{P}^2$ ramified in a sextic curve, and the cycle Z gives a splitting of the requisite Azumaya algebra on this K3.

For generic Y the category $\mathscr{A}_Y$ can be thought of as the derived category of coherent sheaves on a noncommutative K3 surface. Therefore, any smooth moduli space of objects in $\mathscr{A}_Y$ should be hyperkähler, and the Fano scheme of lines can be realized in this way, see [40].

The fact that a cubic 4-fold has something in common with a K3 surface can be easily seen from its Hodge diamond. In fact, the Hodge diamond of Y is

$$\begin{array}{ccccccccccc}
 & & & & 1 & & & & & & \\
 & & & 0 & & 0 & & & & & \\
 & & 0 & & 1 & & 0 & & & & \\
 & 0 & & 0 & & 0 & & 0 & & & \\
0 & & 1 & & 21 & & 1 & & 0 & & \\
 & 0 & & 0 & & 0 & & 0 & & & \\
 & & 0 & & 1 & & 0 & & & & \\
 & & & 0 & & 0 & & & & & \\
 & & & & 1 & & & & & &
\end{array}$$

and one sees immediately the Hodge diamond of a K3 surface in the primitive part of the cohomology of Y. There are some other 4-dimensional Fano varieties with a similar Hodge diamond. The simplest example is the 4-fold of degree 10 in $\mathsf{Gr}(2,5)$ (an intersection of $\mathsf{Gr}(2,5)$ with a hyperplane and a quadric in $\mathbb{P}^9$). Its Hodge diamond is

$$\begin{array}{ccccccccccc}
 & & & & 1 & & & & & & \\
 & & & 0 & & 0 & & & & & \\
 & & 0 & & 1 & & 0 & & & & \\
 & 0 & & 0 & & 0 & & 0 & & & \\
0 & & 1 & & 22 & & 1 & & 0 & & \\
 & 0 & & 0 & & 0 & & 0 & & & \\
 & & 0 & & 1 & & 0 & & & & \\
 & & & 0 & & 0 & & & & & \\
 & & & & 1 & & & & & &
\end{array}$$

and again its primitive part has K3 type. On a categorical level this follows from the following result

Theorem 6.3 ([35]). *Let X be a smooth projective variety of index m with a rectangular Lefschetz decomposition*

$$\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{B}, \mathscr{B}(1), \dots, \mathscr{B}(m-1) \rangle$$

of length m. Let Y_d be the smooth zero locus of a global section of the line bundle $\mathcal{O}_X(d)$ for $1 \le d \le m$. Then there is a semiorthogonal decomposition

$$\mathbf{D}^b(\mathsf{coh}(Y_d)) = \langle \mathscr{A}_{Y_d}, \mathscr{B}, \mathscr{B}(1), \dots, \mathscr{B}(m-d-1) \rangle$$

and moreover, a power of the Serre functor $\mathbf{S}_{\mathscr{A}_{Y_d}}$ is isomorphic to a shift

$$(\mathbf{S}_{\mathscr{A}_{Y_d}})^{d/c} = \left[\frac{d \cdot (\dim X + 1) - 2m}{c} \right], \qquad \text{where } c = \gcd(d, m). \tag{6.3}$$

In particular, if d divides m then the component $\mathscr{A}_{Y_d}$ is a Calabi–Yau category of dimension $\dim X + 1 - 2m/d$.

Remark 6.4. Analogously, one can consider a double covering $Y'_d \to X$ ramified in a zero locus of a global section of the line bundle $\mathcal{O}_X(2d)$ instead. Then there is an analogous semiorthogonal decomposition and the Serre functor has the property

$$(\mathbf{S}_{\mathscr{A}_{Y'_d}})^{d/c} = \tau_*^{(m-d)/c} \circ \left[\frac{d \cdot (\dim X + 1) - m}{c} \right], \tag{6.4}$$

where τ is the involution of the double covering.

Applying this result to a 4-fold Y of degree 10 one constructs a semiorthogonal decomposition $\mathbf{D}^b(\mathsf{coh}(Y)) = \langle \mathscr{A}_Y, \mathcal{O}_Y, \mathcal{U}_Y^\vee, \mathcal{O}_Y(1), \mathcal{U}_Y^\vee(1)\rangle$ with $\mathcal{U}_Y$ being the restriction of the tautological bundle from the Grassmannian $\mathsf{Gr}(2,5)$ and $\mathscr{A}_Y$ a Calabi–Yau category of dimension 2. Again, for some special Y one can check that $\mathscr{A}_Y$ is equivalent to the derived category of a K3 surface and so altogether we get another family of noncommutative K3 categories. Moreover, in this case one can also construct a hyperkähler fourfold from Y. One of the ways is to consider the Fano scheme of conics on Y. It was proved in [16] that it comes with a morphism to $\mathbb{P}^5$ with the image being a singular sextic hypersurface, and the Stein factorization of this map gives a genus zero fibration over the double covering of the sextic, known as a double EPW sextic. This is a hyperkähler variety, deformation equivalent to the second Hilbert square of a K3 surface.

Finally, there is yet another interesting example. Consider a hyperplane section Y of a 5-fold X, which is the zero locus of a global section of the vector bundle $\Lambda^2\mathcal{U}_3^\vee \oplus \Lambda^3(W/\mathcal{U}_3)$ on $\mathsf{Gr}(3,W)$ with W of dimension 7. This variety Y was found by O. Küchle in [22] (variety $c5$ in his table), and its Hodge diamond is as follows

$$\begin{array}{ccccccccc}
 & & & & 1 & & & & \\
 & & & 0 & & 0 & & & \\
 & & 0 & & 1 & & 0 & & \\
 & 0 & & 0 & & 0 & & 0 & \\
0 & & 1 & & 24 & & 1 & & 0 \\
 & 0 & & 0 & & 0 & & 0 & \\
 & & 0 & & 1 & & 0 & & \\
 & & & 0 & & 0 & & & \\
 & & & & 1 & & & &
\end{array}$$

Conjecture 6.5. *The 5-dimensional variety $X \subset \mathsf{Gr}(3,W)$ has a rectangular Lefschetz decomposition $\mathbf{D}^b(\mathsf{coh}(X)) = \langle \mathscr{B}, \mathscr{B}(1)\rangle$ with the category $\mathscr{B}$ generated by 6 exceptional objects. Consequently, its hyperplane section Y has a semiorthogonal decomposition $\mathbf{D}^b(\mathsf{coh}(Y)) = \langle \mathscr{A}_Y, \mathscr{B}\rangle$ with $\mathscr{A}_Y$ being a K3 type category.*

It would be very interesting to understand the geometry of this variety and to find out, whether there is a hyperkähler variety associated to it, analogous to the Fano scheme of lines on a cubic fourfold and the double EPW sextic associated to the 4-fold of degree 10. A natural candidate is the moduli space of twisted cubic curves.

Acknowledgements. In my work I was partially supported by RFFI grant NSh-2998.2014.1, the grant of the Simons foundation, and by AG Laboratory SU-HSE, RF government grant, ag.11.G34.31.0023. I am very grateful to A. Bondal and D. Orlov for their constant help and support. I would like to thank R. Abuaf, T. Bridgeland, T. Pantev and P. Sosna for comments on the preliminary version of this paper.

References

[1] R. Abuaf, *Wonderful resolutions and categorical crepant resolutions of singularities*, preprint, arXiv:1209.1564, to appear in J. Reine Angew. Math.

[2] V. Alexeev and D. Orlov, *Derived categories of Burniat surfaces and exceptional collections*, Math. Ann. **357**, no. 2 (2013), 743–759.

[3] M. Bernardara, *A semiorthogonal decomposition for Brauer–Severi schemes*, Math. Nachr. **282**, no. 10 (2009), 1406–1413.

[4] Ch. Böhning, H.-Ch. Graf von Bothmer, and P. Sosna, *On the derived category of the classical Godeaux surface*, Advances in Math. **243** (2013), 203–231.

[5] Ch. Böhning, H.-Ch. Graf von Bothmer, L. Katzarkov, and P. Sosna, *Determinantal Barlow surfaces and phantom categories*, preprint, arXiv:1210.0343.

[6] A. Bondal and M. Kapranov, *Representable functors, Serre functors, and mutations*, Izv. Akad. Nauk SSSR Ser. Mat. **53** (1989), 1183–1205; English transl., Math. USSR-Izv. **35** (1990), 519–541.

[7] A. Bondal and D. Orlov, *Derived categories of coherent sheaves*, Proc. Internat. Congress of Mathematicians (Beijing, 2002), vol. II, Higher Ed. Press, Beijing (2002), 47–56.

[8] T. Bridgeland, *Equivalences of triangulated categories and Fourier–Mukai transforms*, Bull. London Math. Soc. **31**, no. 1 (1999), 25–34.

[9] R.-O. Buchweitz, G. Leuschke, and M. Van den Bergh, *On the derived category of Grassmannians in arbitrary characteristic*, preprint, arXiv:1006.1633.

[10] N. Fakhruddin, *Exceptional collections on 2-adically uniformised fake projective planes*, preprint, arXiv:1310.3020.

[11] D. Faenzi and L. Manivel, *On the derived category of the Cayley plane II*, preprint, arXiv:1201.6327.

[12] S. Galkin, L. Katzarkov, A. Mellit, and E. Shinder, *Minifolds and phantoms*, preprint, arXiv:1305.4549.

[13] S. Galkin and E. Shinder, *Exceptional collections of line bundles on the Beauville surface*, Advances in Math. **244** (2013), 1033–1050.

[14] Sh. Hosono and H. Takagi, *Duality between* $\mathrm{Chow}^2\mathbb{P}^4$ *and the Double Quintic Symmetroids*, preprint, arXiv:1302.5881.

[15] C. Ingalls and A. Kuznetsov, *On nodal Enriques surfaces and quartic double solids*, preprint, arXiv:1012.3530.

[16] F. Iliev and L, Manivel, *Fano manifolds of degree ten and EPW sextics*, Ann. Sci. Éc. Norm. Supér. (4) **44** (2011), no. 3, 393–426.

[17] V. Iskovskikh and Yu. Prokhorov, *Fano varieties. Algebraic geometry.* V, volume 47 of Encyclopaedia Math. Sci. (1999).

[18] B. Karpov and D. Nogin, *Three-block exceptional collections over Del Pezzo surfaces*, Izv. RAN. Ser. Mat. **62**, no. 3 (1998), 3–38.

[19] M. Kapranov, *On the derived categories of coherent sheaves on some homogeneous spaces*, Invent. Math. **92**, no. 3 (1988), 479–508.

[20] Yu. Kawamata, *Derived categories of toric varieties*, Michigan Math. J. **54**, no. 3 (2006), 517–535.

[21] K. Kawatani and S. Okawa, *Derived categories of smooth proper Deligne-Mumford stacks with* $p_g > 0$, preprint.

[22] O. Küchle, *On Fano 4-folds of index 1 and homogeneous vector bundles over Grassmannians*, Math. Z. **218**, no. 1 (1995), 563–575.

[23] S. Kuleshov and D. Orlov, *Exceptional sheaves on del Pezzo surfaces*, Izv. RAN. Ser. Mat. **58**, no. 3 (1994), 53–87.

[24] A. Kuznetsov, *An exceptional collection of vector bundles on* V_{22} *Fano threefolds*, Vestnik MGU, Ser. **1**, Mat. Mekh. 1996, no. 3, 41–44 (in Russian).

[25] ———, *Derived categories of cubic and* V_{14} *threefolds*, Proc. V.A.Steklov Inst. Math. **246** (2004), 183–207.

[26] ———, *Derived categories of Fano threefolds* V_{12}, Mat. zametki, 78, no. 4 (2005): 579–594, translation in Math. Notes, **78**, no. 4 (2005), 537–550.

[27] ———, *Hyperplane sections and derived categories*, Izvestiya RAN: Ser. Mat. **70**, no. 3 (2006), 23–128 (in Russian); translation in Izvestiya: Mathematics **70**, no. 3 (2006), 447–547.

[28] ———, *Homological projective duality for Grassmannians of lines*, preprint, arXiv: math/0610957.

[29] ———, *Homological projective duality*, Publ. Math. Inst. Hautes Études Sci., **105**, no. 1 (2007), 157–220.

[30] ———, *Homological projective duality* (in Russian), Habilatation thesis, unpublished.

[31] ———, *Derived categories of quadric fibrations and intersections of quadrics*, Advances in Math. **218**, no. 5 (2008), 1340–1369.

[32] ———, *Exceptional collections for Grassmannians of isotropic lines*, Proc. Lond. Math. Soc. **97**, no. 1 (2008), 155–182.

[33] ———, *Lefschetz decompositions and categorical resolutions of singularities*, Selecta Math. **13**, no. 4 (2008), 661–696.

[34] ———, *Derived categories of Fano threefolds*, Proc. V.A.Steklov Inst. Math. **264** (2009), 110–122.

[35] ———, *Calabi–Yau categories*, unpublished.

[36] ———, *Derived categories of cubic fourfolds*, in "Cohomological and Geometric Approaches to Rationality Problems. New Perspectives" Series: Progress in Mathematics, 282 (2010).

[37] ———, *Base change for semiorthogonal decompositions*, Compos. Math. **147**, no. 3, (2011), 852–876.

[38] ———, *Height of exceptional collections and Hochschild cohomology of quasiphantom categories*, preprint, arXiv:1211.4693, to appear in J. Reine Angew. Math.

[39] A. Kuznetsov and V. Lunts, *Categorical resolutions of irrational singularities*, preprint, arXiv:1212.6170.

[40] A. Kuznetsov and D. Markushevich, *Symplectic structures on moduli spaces of sheaves via the Atiyah class*, J. Geom. Phys. **59**, no. 7 (2009), 843–860.

[41] A. Kuznetsov and A. Polishchuk, *Exceptional collections on isotropic Grassmannians*, to appear in IMRN.

[42] L. Manivel, *On the derived category of the Cayley plane*, J. Algebra **330**, no. 1 (2011), 177–187.

[43] S. Mukai, *Fano 3-folds*, in Complex Projective Geometry, Trieste, 1989/Bergen, 1989 (Cambridge Univ. Press, Cambridge, 1992), LMS Lect. Note Ser. **179**, 255–263.

[44] S. Okawa, *Semi-orthogonal decomposability of the derived category of a curve*, Advances in Math. **228**, no. 5 (2011), 2869–2873.

[45] D. Orlov, *Exceptional set of vector bundles on the variety V_5*, Vestnik Moskov. Univ. Ser. I Mat. Mekh. **5** (1991), 69–71.

[46] ———, *Projective bundles, monoidal transformations, and derived categories of coherent sheaves*, Izv. Akad. Nauk SSSR Ser. Mat., 56 (1992): 852–862; English transl., Russian Acad. Sci. Izv. Math. **41** (1993), 133–141.

[47] A. Polishchuk and A. Samokhin, *Full exceptional collections on the Lagrangian Grassmannians $LG(4,8)$ and $LG(5,10)$*, J. Geom. Phys. **61**, no. 10 (2011), 1996–2014.

[48] A. Samokhin, *The derived category of coherent sheaves on $LG_3^{\mathbb{C}}$*, Uspekhi Mat. Nauk 56, no. 3 (2001), 177–178; English transl., Russian Math. Surveys, **56** (2001), 592–594.

[49] S. Zube, *Exceptional vector bundles on Enriques surfaces*, Mat. Zametki **61**, no. 6 (1997), 825–834.

Steklov Mathematical Institute, 8 Gubkin str., Moscow 119991 Russia;
The Poncelet Laboratory, Independent University of Moscow;
Laboratory of Algebraic Geometry, SU-HSE
E-mail: akuznet@mi.ras.ru

K3 surfaces in positive characteristic

Davesh Maulik

Abstract. We describe recent progress in the study of K3 surfaces in characteristic p, focusing on the Tate conjecture and moduli theory. We also discuss some geometric applications and open questions.

Mathematics Subject Classification (2010). Primary 14J28; Secondary 11G25.

Keywords. K3 surfaces, Tate conjecture, moduli spaces.

1. Introduction

K3 surfaces have long been a subject of active study in algebraic geometry, involving a combination of algebraic, arithmetic and complex-geometric perspectives. In the case of K3 surfaces over fields of characteristic $p > 0$, there has been a great deal of progress over the past few years in understanding certain basic questions. In this article, I survey some of these recent advances. In particular, I discuss various proofs of the Tate conjecture over finite fields in odd characteristic and global properties of the moduli space of polarized K3 surfaces, trying to give a sense of the different techniques at our disposal. In the last section, I discuss some geometric applications of this circle of ideas and some open questions.

2. Tate conjecture over finite fields

In this section, we discuss the Tate conjecture for K3 surfaces over finite fields. First, we recall the general statement for divisors on an arbitrary smooth projective variety. Given such a variety X, defined over a finite field $k = \mathbb{F}_q$ of characteristic p, and a prime $\ell \neq p$, the first Chern class in ℓ-adic cohomology defines a map

$$c_1^{\text{ét}} : \mathrm{Pic}(X) \otimes \mathbb{Q}_\ell \to H^2_{\text{ét}}(X, \mathbb{Q}_\ell(1))^{\mathrm{Gal}(\overline{k}/k)},$$

where $\mathrm{Pic}(X)$ denotes the Picard group of X, parametrizing line bundles, and the right-hand side denotes the Galois-invariant part of the (Tate-twisted) second étale cohomology group. The Tate conjecture for X predicts that this map is surjective. It can be viewed as a more difficult analog of the Lefschetz $(1,1)$ theorem for complex varieties, in that it gives a way of understanding the subspace of cohomology generated by divisor classes via linear-algebraic data. It remains completely open for most varieties; in the case of K3 surfaces, however, we have the following:

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Theorem 2.1. *The Tate conjecture holds for K3 surfaces in odd characteristic.*

After a series of earlier results [4, 11, 31, 36, 43], the most general statement given here is due to Madapusi Pera [28], whose proof we discuss in the next section; his result applies not just for finite fields but all finitely generated fields in characteristic p. In this section, we will review a somewhat weaker approach, via work of Nygaard-Ogus [36] in the case of finite height and, in the remaining cases, via work of myself [31] (for primes larger than the polarization degree plus 4) and Charles [11] (for primes at least 5). We then close with some questions regarding the behavior of Picard groups in families of surfaces.

2.1. Earlier results. Before discussing more recent arguments, we mention two significant earlier results (among others)

In the case where the K3 surface admits the structure of an elliptic fibration, the Tate conjecture is equivalent to the statement that the Brauer group $H^2_{\text{ét}}(X, \mathbb{G}_m)$ is finite. In this form, the conjecture was proven in 1974 by Artin and Swinnerton-Dyer [4]. If the group is infinite, they produce twists of the elliptic fibration, trivialized by multisections of increasing degree; a beautiful intersection theory argument shows that eventually these twists must become isomorphic as abstract surfaces which leads to a contradiction. Since K3 surfaces whose Picard groups have rank at least 5 are automatically elliptically fibered, the Tate conjecture follows already in these cases.

A more general breakthrough came via Nygaard [35] and Nygaard-Ogus [36], who handle the case of finite height K3 surfaces. In order to explain this result, we first explain the definition of the height of a K3 surface. Given a K3 surface X, its formal Brauer group $\widehat{\mathrm{Br}}(X)$ of X, defined by Artin and Mazur [3], is a one-dimensional formal group scheme representing the functor

$$T \mapsto [\mathrm{Ker}(\mathrm{Br}(X \times T) \to \mathrm{Br}(X))]\,,$$

on local Artin k-algebras, where $\mathrm{Br}(X \times T)$ is the Brauer group of $X \times T$. The *height* of the K3 surface is simply the height of its formal Brauer group. When finite, it can take values between 1 and 10; otherwise, when $\widehat{\mathrm{Br}} = \widehat{\mathbb{G}_a}$, we say the K3 surface has infinite height or, equivalently, that the K3 surface is *supersingular*.

In [35] and [36], Nygaard and Ogus prove the Tate conjecture for finite height K3 surfaces in characteristic $p \geq 5$; these comprise an open, dense subset in the moduli space of K3 surfaces. Their argument uses the notion of canonical and quasi-canonical lifts of X over the ring of Witt vectors $W(k)$. These are distinguished lifts for which certain properties of X are preserved; in particular, both the cohomological Frobenius action as well as the entire Picard group of X extend to quasi-canonical lifts. Once this is shown, the authors use the Lefschetz $(1,1)$ theorem to construct the expected line bundles in characteristic zero and specialize them back to k.

2.2. Supersingular K3 surfaces. It remains to handle the case supersingular K3 surfaces.

In general, supersingular K3 surfaces exhibit behavior that cannot occur in characteristic zero. For instance, the Tate conjecture is equivalent to a conjecture of Artin which states that, after passage to $\overline{k}$, the geometric Picard rank of X is as large as possible, i.e. $\mathrm{rk}\,\mathrm{Pic}(X_{\overline{k}}) = 22 = b_2(X)$. As an example, if we take $p \equiv 3 \mod 4$, Tate [44] showed that the Fermat quartic

$$\{x^4 + y^4 + z^4 + w^4 = 0\} \subset \mathbb{P}^3_k$$

has geometric Picard rank 22. This phenomenon is quite striking when compared to characteristic zero, where Hodge theory tells us that $\mathrm{rk}\,\mathrm{Pic}(X) \leq 20$.

As another example, it was conjectured by several people that supersingular K3 surfaces are unirational, and can be seen explicitly in the above example. Again, this phenomenon is ruled out in characteristic zero, since the pullback of a non-vanishing two-form to a rational variety would remain non-vanishing while in characteristic p the map can (and must) be inseparable. Quite recently, Liedtke [26] has announced a proof of this conjecture.

Unfortunately, due to this unusual behavior, the strategy of quasi-canonical lifts does not apply for supersingular surfaces, since the Picard group cannot be lifted to characteristic zero. Instead, the approach of [11, 31] is to exploit another instance of strange supersingular behavior, first observed by Artin [2]: in a connected family of supersingular K3 surfaces, although the Picard group can jump under specialization, the rank of the Picard group must remain constant. This observation is significant because, over $\mathbb{C}$, such phenomena never occur. Indeed, by a result of Green ([47, 4.2]) and Oguiso [37], given a non-isotrivial family of K3 surfaces over a complex base B, the set of points on B where the Picard group jumps in rank is *dense* in the complex-analytic topology. The approach to showing the Tate conjecture in the supersingular case is to exploit the tension between these two behaviors to force the existence of extra cycles in supersingular families.

2.3. Automorphic forms. In order to to do this, we apply work of Borcherds [9, 10] on automorphic forms to find an arithmetic approach to the Green-Oguiso result. Let $\mathsf{M}_{2d,k}$ denote the moduli space of primitively polarized K3 surfaces over k of degree $2d$. Given a rank 2 lattice Λ with a distinguished basis vector v of self-intersection $2d$, the Noether-Lefschetz divisor D_Λ is the locus of polarized surfaces (X, L) with an inclusion of pairs $(\Lambda, v) \subset (\mathrm{Pic}(X), L)$ (this locus is non-empty for suitable lattices). Given such a lattice Λ, we can associate discrete invariants $(\mathrm{disc}(\Lambda), \delta)$ as follows. If we extend v to a basis $\langle v, w\rangle$, the congruence class $\delta \in H_d = \mathbb{Z}/2d$ of the pairing $\langle v, w\rangle$ is well-defined up to a sign.

Using an explicit construction of automorphic forms, Borcherds shows that the generating function (with coefficients in $\mathrm{Pic}(\mathsf{M}_{2d}) \otimes \mathbb{Z}[H_d]$)

$$\phi(q) = -[\lambda]q^0 \otimes e_0 + \sum_{\Lambda} [D_\Lambda] q^{-\mathrm{disc}(\Lambda)/2d} \otimes e_\delta \in \mathrm{Pic}(\mathsf{M}_{2d})[[q]] \otimes \mathbb{Z}[H_d]$$

is a holomorphic modular form of weight $21/2$. Here, the constant term is the Hodge bundle $\lambda = c_1(\pi_*(\Omega^2_\pi))$ associated to the universal family of K3 surfaces $\pi : \mathcal{X} \to \mathsf{M}_{2d,k}$. This result has already found many applications in the enumerative geometry of K3 surfaces over the complex numbers (see [19, 32]).

Although Borcherds' work is in characteristic zero, we can still apply it to characteristic p by spreading out and intersecting these divisors with a proper curve in the locus of supersingular surfaces. In order to find proper curves in the supersingular locus, we can either study degenerations of K3 surfaces (which leads to the constraint on the prime in [31]) or instead use degenerations of abelian varieties, as in [11]. The result is a modular form whose coefficients encode the Noether-Lefschetz degrees of our family. One also needs to ensure this modular form isn't zero; by looking at the constant term, this follows from positivity properties of the Hodge bundle (see Corollary 3.3 in the next section). Once the constant term is nonzero, an elementary estimate shows that the modular coefficients grow very fast and we can deduce that Noether-Lefschetz degrees for sufficiently large discriminant are non-vanishing. This shows the following:

Proposition 2.2. *Let B be a proper curve and let $\mathcal{X} \to B$ denote a non-isotrivial family of K3 surfaces of degree $2d$ not divisible by p. For any suitable lattice Λ with $-\mathrm{disc}(\Lambda) \gg 0$, there exists $b \in B$ with $\mathrm{Pic}(\mathcal{X}_b) \supseteq \Lambda$.*

With this proposition in hand (and assuming we have proper curves to work with), the Tate conjecture follows easily. Indeed, if we consider lattices $\Lambda_k = \begin{pmatrix} 2d & k \\ k & 0 \end{pmatrix}$, these have arbitrarily large discriminant and correspond to elliptic fibrations. Therefore, the Tate conjecture holds for at least one fiber on any proper family by the result of Artin-Swinnerton-Dyer. As the Picard rank is constant in families, it must hold for every fiber in the family by Artin's observation.

One can push these techniques further. Using a more intricate analysis, Charles is able to prove the Tate conjecture for codimension two cycles on cubic fourfolds as well as divisors on certain reductions of higher-dimensional holomorphic symplectic varieties.

2.4. Noether-Lefschetz behavior. Over the complex numbers, the theorem of Green and Oguiso gives us more detailed information about how Picard ranks vary in a family of K3 surfaces; it says that these ranks jump when compared to the generic geometric rank on a dense set of points in the Euclidean topology (although note that by [33] p-adic density statements no longer hold). If we have a family over a more general one-dimensional base, we can at least ask for a Zariski dense (i.e. infinite) set of points in the jumping locus.

In what follows, let B denote either a smooth irreducible curve over an algebraically closed field k or let $B = \mathrm{Spec}\, \mathcal{O}_K[1/N]$ for a number field K; in either case, let $\overline{\eta}$ denote a geometric point lying over the generic point of B.

Question 2.3. *Let $\pi : \mathcal{X} \to B$ be a smooth non-isotrivial family of polarized K3 surfaces over B. Are there infinitely many closed points $b \in B$ such that $\mathrm{rk}\,\mathrm{Pic}(\mathcal{X}_b) > \mathrm{rk}\,\mathrm{Pic}(\mathcal{X}_{\overline{\eta}})$?*

As already stated, the answer is yes in equicharacteristic zero. Furthermore, since we have families of supersingular K3 surfaces (which have constant Picard rank 22), we know the answer is no in equicharacteristic p. Supersingular families are not the only counterexample; another example is the family of Kummer surfaces associated to the product of a fixed supersingular elliptic curve and a varying family of elliptic curves.

However, it seems reasonable to expect that this question has a positive answer if we put some further constraints on the family, e.g. that the generic fiber is ordinary or that the family does not meet the supersingular locus at all (even after passing to a partial compactification of B). In the mixed characteristic setting, it is expected that the answer is always yes; at least for certain Kummer surfaces, there is unpublished work of Charles along these lines.

Motivated by the last section, a more careful analysis of the modularity of Noether-Lefschetz degrees should help here. As before, the rapid growth of the coefficients of modular forms will typically produce infinitely many intersection points (although not always). Indeed, in the equicharacteristic p setting, one can often prove the answer to this question is yes for certain families using precisely this argument. In the arithmetic setting, there are conjectures of Kudla and collaborators (see [20] for an overview) predicting modular behavior for arithmetic Noether-Lefschetz degrees, in the sense of Arakelov theory; if provable, it seems natural to apply these modularity conjectures for Question 2.3.

3. Moduli of K3 surfaces

In this section, we discuss moduli spaces of polarized K3 surfaces in mixed and finite characteristic, following the work of Madapusi Pera [27, 28]. We discuss extensions of the period map, defined using the classical Kuga-Satake construction, and its structure in odd characteristic. From this, we can deduce many useful geometric corollaries.

3.1. Constructions in characteristic zero. Given an even positive integer $2d$, let M_{2d} denote the moduli space of K3 surfaces over $\mathbb{Z}[1/2]$, equipped with a primitive polarization of degree $2d$. It is a Deligne-Mumford stack, smooth over $\mathbb{Z}[1/2d]$ [13] (and has mild singularities at other odd primes).

Over $\mathbb{C}$, many global properties of the moduli space can be deduced via Hodge theory and the global Torelli theorem [40]. We recall briefly how this goes. Let U be the two-dimensional hyperbolic lattice, and let

$$M = U^{\oplus 3} \oplus E_8(-1)^{\oplus 2}$$

denote the abstract K3 lattice. If we write (e, f) for the standard basis of the first copy of U, then we set

$$L_{2d} = \langle e + df \rangle^{\perp} \subset M,$$

i.e. we take the orthogonal complement of a fixed primitive vector of degree $2d$.

Let $G = \mathrm{SO}(L_{2d})$ and $G' = \mathrm{CSpin}(L_{2d})$ denote the orthogonal and spin groups associated to L_{2d} and consider the period domain

$$\Omega_{2d} = \{\omega \in \mathbb{P}(L_{2d,\mathbb{C}}) | \langle \omega, \omega \rangle = 0, \langle \omega, \overline{\omega} \rangle > 0\} = \mathrm{SO}(2, 19)/\mathrm{SO}(2) \times \mathrm{SO}(19).$$

Let $\Gamma_{2d} \subset G(\mathbb{Z})$ be the subgroup which acts trivially on the discriminant group $L_{2d}^{\vee}/L_{2d}$. The analytic quotient $[\Omega_{2d}/\Gamma_{2d}]$ is algebraic via [5] and can be identified with the Shimura variety of orthogonal type $\mathrm{Sh}_{2d,\mathbb{C}} := \mathrm{Sh}(G_{\mathbb{Q}}, \Omega)_{\mathbb{C}}$.

Given a polarized complex K3 surface (X, L) and an isomorphism of its primitive cohomology with L_{2d}, the line $H^{2,0}(X)$ defines a point in Ω_{2d}. This map behaves well in families; after forgetting the marking, we obtain the period map

$$\mathsf{P}_{\mathbb{C}} : \widetilde{\mathsf{M}}_{2d,\mathbb{C}} \to \mathrm{Sh}_{2d,\mathbb{C}},$$

where $\widetilde{\mathsf{M}}_{2d}$ is an étale double cover corresponding to a choice of spin structure. It follows from the Torelli theorem that the period map is an open immersion, although not surjective. However, we can enlarge our moduli space to include pairs (X, L) where L is a big and nef line bundle on X (so-called quasi-polarized K3 surfaces). The period map extends to this larger space $\mathsf{M}^{\mathrm{qp}}_{2d,\mathbb{C}}$, and becomes surjective and étale (although not quite an isomorphism since $\mathsf{M}^{\mathrm{qp}}_{2d}$ is non-separated).

It follows from [1, 42] that this morphism descends to $\mathbb{Q}$: $\mathsf{P}_{\mathbb{Q}} : \widetilde{\mathsf{M}}_{2d,\mathbb{Q}} \to \mathrm{Sh}_{2d,\mathbb{Q}}$. After passing to a further finite étale cover, the analogous story holds with the Shimura variety of spin type $\mathrm{Sh}'_{2d} := \mathrm{Sh}(G'_{\mathbb{Q}}, \Omega)$.

The second classical idea I want to mention is the Kuga-Satake construction [21] which associates an abelian variety to a polarized K3 surface. Its definition is purely transcendental: given any polarized integral weight two Hodge structure V of K3 type, the Kuga-Satake construction defines a polarized weight one Hodge structure on the Clifford algebra $W =$

$\mathrm{Cl}(V)$. When V is the primitive cohomology of a polarized K3 surface (X, L), the abelian variety corresponding to W is its Kuga-Satake variety $\mathsf{KS}(X, L)$. As a consequence of its definition, we have an embedding of weight 0 Hodge structures

$$P^2(X, \mathbb{Z})(1) \hookrightarrow \mathrm{End}(H^1(\mathsf{KS}(X, L), \mathbb{Z})) \tag{*}$$

where P^2 is the primitive cohomology, which can be used to recover information about (X, L) from $\mathsf{KS}(X, L)$.

This construction can be rephrased via the Shimura variety Sh'_{2d} as follows. The spin representation defines a homomorphism from G' to the symplectic group $\mathrm{CSp}(\mathrm{Cl}(L_{2d}))$ associated to the Kuga-Satake variety. This map induces a morphism of Shimura varieties, and the Kuga-Satake construction comes from the composition

$$\widetilde{\mathsf{M}}_{2d,\mathbb{C}} \to \mathrm{Sh}'_{2d,\mathbb{C}} \to \mathcal{A}_{g,\mathbb{C}},$$

where the last space is the moduli space of principally polarized abelian varieties. One remarkable feature of this Shimura-theoretic approach is that the transcendental Kuga-Satake construction descends to $\mathbb{Q}$.

3.2. Extension to mixed characteristic. As far as I know, it was Deligne [12] who first applied a pointwise extension of the Kuga-Satake construction to characteristic p in his proof of the Weil conjectures for K3 surfaces.

In [17] and [27], Kisin and Madapusi Pera prove the following

Theorem 3.1. *There exists a regular integral canonical model* Sh_{2d} *over* $\mathbb{Z}[1/2]$ *for the Shimura varieties* $\mathrm{Sh}_{2d,\mathbb{Q}}$. *A similar result holds for* $\mathrm{Sh}'_{2d,\mathbb{Q}}$.

We briefly explain this terminology, referring the reader to the references for a precise statement. These integral canonical models are defined by an extension property analogous to Néron models for abelian schemes. For each prime p (after passage to infinite level) we require them to satisfy an extension property from certain regular test schemes over $\mathbb{Z}_p$. When p doesn't divide $2d$, Kisin constructs the integral model for Sh'_{2d} via taking the normalization of the closure inside a moduli space of abelian varieties with level structure, defined over $\mathrm{Spec}\,\mathbb{Z}[1/2d]$. To weaken the constraint on the prime, Madapusi Pera embeds Sh_{2d} inside a larger orthogonal type Shimura variety where Kisin's approach applies. The difficulty in both cases is proving regularity; this requires delicate deformation theory arguments.

Using the extension property, one can show that the period map defined on $\widetilde{\mathsf{M}}_{2d,\mathbb{Q}}$ can be extended over all primes, so we have the mixed-characteristic period map:

$$\mathsf{P} : \widetilde{\mathsf{M}}_{2d} \to \mathsf{Sh}_{2d}.$$

Since our integral models are defined via $\mathcal{A}_g$, we can think of the period map as being defined via its compatibility with the Kuga-Satake construction (and in particular this gives a definition of Kuga-Satake varieties in mixed characteristic). This approach to the extension of the period map and Kuga-Satake constructions was first pursued by Vasiu [46] and Rizov [42] with stronger constraints on the prime.

The key result is the following ([28], Theorem 4.2). For primes not dividing $2d$, the result was shown in [31].

Theorem 3.2. *The period map* P *is étale.*

For primes not dividing $2d$, this result was shown in [31] and [42] as well. The key step in all approaches is to extend the inclusion (*) to de Rham cohomology in characteristic p. Indeed, since deformations of K3 surfaces are detected by the Kodaira-Spencer map, this controls the differential of the period map. Since the period map is defined by extending from characteristic zero, this requires an *integral* comparison theorem between cohomology in characteristic p and characteristic 0 (Fontaine-Messing [14] in [31], Bloch-Kato [6] in [28]).

3.3. Geometric consequences. Theorem 3.2 has immediate geometric consequences:

Corollary 3.3.

(i) *For every $p > 2$, the moduli space $\mathsf{M}_{2d,\mathbb{F}_p}$ is a quasi-projective Deligne-Mumford stack over $\mathbb{F}_p$.*

(ii) *The Hodge bundle $\pi_*(\Omega^2_\pi)$ is an ample line bundle.*

(iii) *If p^2 does not divide d then $\mathsf{M}_{2d,\mathbb{F}_p}$ is geometrically irreducible.*

The last statement uses the existence of toroidal compactifications [29] in mixed characteristic, mimicking the Deligne-Mumford argument for the moduli space of curves. Madapusi Pera has further recent work which should weaken the constraint on the prime. I find the ampleness of the Hodge bundle especially interesting, since I know of no geometric approach to this question. Recent work of Patakfalvi [38] has provided semi-positivity statements in greater generality and it would be interesting to see if they apply here.

Lastly, one can ask for surjectivity of the period morphism, as in characteristic zero. Here, the results are currently limited. If the degree of the polarization is sufficiently small, we have the following results, from [30, 31].

Assuming $p > 18d+4$, in [31], I show that any one-parameter degeneration of polarized K3 surfaces can be replaced with a K-trivial central fiber with at worst normal-crossings and rational singularities, following a combinatorial classification due to Kulikov, Pinkham, and Persson [22, 39] over $\mathbb{C}$. In [30], Matsumoto uses this classification to prove a criterion for good reduction, analogous to the Néron-Ogg-Shafarevich condition for abelian schemes. The combination yields the following:

Corollary 3.4. *Assume $p > 18d + 4$. After extending to quasipolarized K3 surfaces, the period map*

$$\mathsf{P} : \widetilde{\mathsf{M}}^{\mathrm{qp}}_{2d,\overline{\mathbb{F}}_p} \to \mathsf{Sh}_{2d,\overline{\mathbb{F}}_p}$$

is surjective.

The condition on the prime is used to find semistable models for degenerations, so that we can run the semistable minimal model program using work of Kawamata [16]. Better results on semistable reduction for surfaces in characteristic p would allow us to remove this bound.

3.4. Application to Tate. Finally, we sketch briefly how the mixed-characteristic period map helps approach the Tate conjecture, independent of the arguments of the last section. Given a polarized K3 surface (X, L), the extended period map and its composition to $\mathcal{A}_g$ again gives us an abelian variety $A = \mathsf{KS}(X, L)$. The arguments above (in particular the proof of the étale property) show that we have ℓ-adic and crystalline versions of (*). Let $L_\ell(A) \subset \mathrm{End}(H^1(A, \mathbb{Q}_\ell)$ denote the image of the ℓ-adic version of (*) and $L_{\mathrm{cris}}(A)$ the

analogous crystalline subspace. We can define a subspace $\mathrm{LEnd}(A)$ of special endomorphisms of A to be those whose cycle classes (for all realizations) lie in $L_\ell(A)$ and $L_{\mathrm{cris}}(A)$. The Tate conjecture can then be rephrased as an isomorphism

$$\mathrm{LEnd}(A) \otimes \mathbb{Q}_\ell \to L_\ell(A)^{\mathrm{Gal}(\overline{k}/k)}.$$

After passing to a larger spin-type Shimura variety (leaving the world of K3 surfaces), one can assume the left-hand side is nonzero. Let $I \subset \mathrm{Aut}(A)_{\mathbb{Q}}$ be the largest algebraic subgroup whose realizations are contained in the image of CSpin. For a particular choice of ℓ, a result of Kisin [18] shows that this is a nonzero map between irreducible representations of I. Since the Tate conjecture is independent of ℓ, the result follows.

4. Applications

We close by explaining a pair of geometric applications.

4.1. Finiteness of K3 surfaces. We first explain a consequence of the Tate conjecture, shown in joint work with Lieblich and Snowden [25]:

Theorem 4.1. *Fix a finite field $k = \mathbb{F}_q$ of characteristic $p \geq 5$.*

(i) *The Tate conjecture holds for K3 surfaces over k implies that there are finitely many isomorphism classes of K3 surfaces defined over k that satisfy the Tate conjecture.*

(ii) *Conversely, the Tate conjecture for K3 surfaces over finite extensions of k is implied the finiteness of the set of isomorphism classes of K3 surfaces over extensions of k.*

As a consequence of the first implication and the Tate conjecture, we see that there are finitely many K3 surfaces over any finite field of characteristic ≥ 5. Even though the second direction is redundant now, I have stated it here, since the technique is quite different from the approach mentioned eariler.

For the first implication, notice that finiteness of the set of isomorphism classes does not follow from the existence of the moduli spaces of finite type, since we need to fix an auxiliary polarization and consider $\mathsf{M}_{2d,k}$ for all $d \geq 1$. A similar result for abelian varieties was shown by Zarhin (see [48]). The argument proceeds as follows. Using the Tate conjecture (in both the ℓ-adic and crystalline incarnations), for any K3 surface X defined over k, we can control $\mathrm{Pic}(X_{\overline{k}}) \otimes \mathbb{Z}_\ell$ and $\mathrm{Pic}(X_{\overline{k}}) \otimes \mathbb{Z}_p$ in terms of the action of geometric Frobenius on étale/crystalline cohomology. Since k is fixed, there are only finitely many possibilities for the eigenvalues of this action and thus we can control the discriminant of the Picard group. This allows us to find a polarization on X of degree bounded by a constant $C(k)$. One also needs to bound the number of non-isomorphic twists, but this follows from structure theory of $\mathrm{Aut}(X_{\overline{k}})$ due to Totaro [45].

In the other direction, the idea is to mimic the argument of Artin and Swinnerton-Dyer discussed earlier. Following [4], instead of considering twists of the elliptic structure, one can take Brauer classes and consider moduli spaces of twisted sheaves associated to these classes - for the right choice of discrete invariants, these will be twisted K3 surfaces. If the Brauer group were infinite, using period-index results of Lieblich [24] and the finiteness assumption, one can create an infinite sequence of (non-isomorphic) derived equivalent twisted

K3 surfaces over k. One can further lift this sequence (and the derived equivalences) to $\mathbb{C}$ which violates work of Huybrechts and Stellari [15].

While no longer strictly necessary, I think it is interesting to ask the following:

Question 4.2. *Is there an argument for finiteness without assuming the Tate conjecture?*

A positive answer to this question would be quite interesting; all the approaches to the Tate conjecture discussed above (including the work of Nygaard and Ogus) require crystalline techniques and global properties of the moduli space, while this approach is ℓ-adic in nature (although still technical in a different way).

4.2. Construction of rational curves. In this section, we mention briefly very interesting results of Bogomolov-Hassett-Tschinkel [7] and Li-Liedtke [23] on using the Tate conjecture to construct rational curves on K3 surfaces in characteristic zero.

The starting point is the following conjecture:

Conjecture 4.3. *Any K3 surface X over an algebraically closed field contains infinitely many integral rational curves.*

Over characteristic zero (and for non-supersingular surfaces in characteristic p), rational curves are necessarily isolated. The naive reason to expect this conjecture is a rough dimension count for any sufficiently ample linear system. One can make a stronger conjecture, asking for infinitely many rational curves contained in the linear systems $|nL|$ where $n \geq 1$ and L is a fixed polarization.

The first result along these lines is an old theorem of Bogomolov and Mumford [34] who prove the conjecture holds for a very general K3 surface, i.e. on the complement of a countable union of subvarieties of the moduli space. Their proof proceeds by writing reducible genus 0 curves C on Kummer surfaces X of increasing degree and proving the pair (X, C) can be deformed so that C is irreducible. By work of Bogomolov-Tschinkel [8], the conjecture also holds for elliptic K3 surfaces or surfaces with infinite automorphism group. Most recently, we have the following theorem [7, 23]:

Theorem 4.4. *In characteristic $\neq 2$, there are infinitely many integral rational curves on a K3 surface X with odd geometric Picard rank.*

We restrict to the characteristic zero case; the general case is similar. The strategy of the argument, first pursued in [7], is to execute a mixed-characteristic version of the Bogomolov-Mumford argument. That is, after reducing to the case of $\overline{\mathbb{Q}}$, one can spread X over the ring of integers R of a number field. Using the Tate conjecture (in fact, via a density argument, the finite height case suffices), it follows that the geometric Picard rank of specializations $X_{\overline{s}}$ is even and in particular larger than the generic Picard rank. This fact allows us to produce rational curves of increasing degree, contained in different specializations. In [23], the authors show that, after adding additional rational curve components, these curves can be lifted to characteristic zero. By taking appropriate irreducible components, they produce integral rational curves on X of ever-increasing degree.

The only situation not covered by this theorem and the Bogomolov-Tschinkel result are K3 surfaces with Picard lattices of rank 2 and 4. For example, the Picard lattice $\begin{pmatrix} -2 & 3 \\ 3 & -2 \end{pmatrix}$ is a simple example where we do not know how to produce enough rational curves in general. In all such cases, if we had a positive answer to Question 2.3, then the specialization strategy would apply here as well.

Acknowledgements. The author is partially supported by NSF Grant DMS-1159416. Most of the material in this note has been learned through conversation with François Charles, Brendan Hassett, Daniel Huybrechts, Max Lieblich, Christian Liedtke, Keerthi Madapusi Pera, and Andrew Snowden. I am grateful to all of them for these extremely useful discussions.

References

[1] Y. André, *On the Shafarevich and Tate conjectures for hyper-Kähler varieties*, Math. Ann., **305**, (1996), no. 2, 205–248.

[2] M. Artin, *Supersingular K3 surfaces*, Ann. Sci. ENS (4) **7** (1974), 543–567 (1975).

[3] M. Artin and B. Mazur, *Formal groups arising from algebraic varieties*, Ann. Sci. École Norm. Sup. (4) **10** (1977), 87–131.

[4] M. Artin and H.P.F. Swinnerton-Dyer, *The Shafarevich-Tate conjecture for pencils of elliptic curves on K3 surfaces*, Invent. Math. **20** (1973), 249–266.

[5] W. Baily and A. Borel, *Compactification of arithmetic quotients of bounded symmetric domains*, Ann. of Math., (2), **84**, (1966), 442–528.

[6] S. Bloch and K. Kato, *p-adic étale cohomology,* Inst. Hautes Études Sci. Publ. Math. **63** (1986), 107–152.

[7] F. Bogomolov, B. Hassett, and Y. Tschinkel, *Constructing rational curves on K3 surfaces,* Duke Math. J. **157** (2011), 535–550.

[8] F. Bogomolov and Y. Tschinkel, *Rational curves and points on K3 surfaces,* Amer. J. Math. **127** (2005), 825–835.

[9] R. Borcherds, *Automorphic forms with singularities on Grassmannians*, Invent. Math. **132** (1998), 491–562.

[10] ———, *The Gross-Kohnen-Zagier theorem in higher dimensions*, Duke J. Math. **97** (1999), 219–233.

[11] F. Charles, *The Tate conjecture for K3 surfaces over finite fields*, Invent. Math., **194** (2013), 119–145.

[12] P. Deligne, *La conjecture de Weil pour les surfaces K3*, Invent. Math. **15** (1972), 206–226.

[13] ———, *Relèvement des surfaces $K3$ en caractéristique nulle.*, Lecture Notes in Mathematics, 868: 58–79, 1981.

[14] J.-M. Fontaine and W. Messing, *p-adic periods and p-adic etale cohomology*, Current trends in arithmetical algebraic geometry (Arcata, Calif., 1985), 179–207, Contemp. Math. **67**, Amer. Math. Soc., Providence, RI, 1987.

[15] D. Huybrechts and P. Stellari, *Equivalences of twisted K3 surfaces*, Math. Ann. **332** (2005), no. 4, 901-936.

[16] Y. Kawamata, *Semistable minimal models of threefolds in positive or mixed characteristic*, J. algebraic Geom. **3**, (1994), no. 3, 463–491.

[17] M. Kisin, *Integral canonical models of Shimura varieties*, *Integral models for Shimura varieties of abelian type*, J. Amer. Math. Soc. **23** (2010) 967-1012.

[18] _____, *Mod p points on Shimura varieties of abelian type*, http://www.math.harvard.edu/?kisin/dvifiles/lr.pdf.

[19] A. Klemm, D. Maulik, R. Pandharipande, and E. Scheidegger, *Noether-Lefschetz theory and the Yau-Zaslow conjecture*, Journal of the AMS,

[20] S. Kudla, *Modular forms and arithmetic geometry,* Current Developments in Mathematics, 2002, 135–180, International Press, Boston, 2003.

[21] M. Kuga and I. Satake, *Abelian varieties attached to polarized K3 surfaces,* Math. Ann. **169** (1967) 239–242.

[22] V. Kulikov, *Degenerations of K3 surfaces and Enriques surfaces*, Izv. Akad. Nauk. SSSR Ser. Mat., **41**, (1977), no. 5, 1008–1042.

[23] J. Li and C. Liedtke, *Rational curves on K3 surfaces*, Invent. Math. **188** (2012), 713–727.

[24] M. Lieblich, *Twisted sheaves and the period-index problem,* Compos. Math. **144** (2008), no. 1, 1–31.

[25] M. Lieblich, D. Maulik, and A. Snowden, *Finiteness of K3 surfaces and the Tate conjecture*, Ann. Sci. ENS, to appear.

[26] C. Liedtke, *Supersingular K3 Surfaces are Unirational*, arXiv:1304.5623.

[27] K. Madapusi Pera, *Integral canonical models for Spin Shimura varieties*, arXiv:1212.1243.

[28] _____, *The Tate conjecture for K3 surfaces in odd characteristic*, arXiv:1301.6326.

[29] _____, *Toroidal compactifications of integral models of Shimura varieties of Hodge type*, arXiv:1211.1731.

[30] Y. Matsumoto, *Good reduction criterion for K3 surfaces*, arXiv:1401.1261.

[31] D. Maulik, *Supersingular K3 surfaces for large primes*, with an appendix by A. Snowden, Duke Math J., to appear.

[32] D. Maulik and R. Pandharipande, *Gromov-Witten theory and Noether-Lefschetz theory*.

[33] D. Maulik and B. Poonen, *Néron-Severi groups under specialization*, Duke Math. J. **161** (2012), no. 11, 2167-2206.

[34] S. Mori and S. Mukai, *The uniruledness of the moduli space of curves of genus 11*, Algebraic geometry (Tokyo/Kyoto, 1982), 334–353, Lecture Notes in Math. 1016, Springer, Berlin, 1983.

[35] N. Nygaard, *The Tate conjecture for ordinary K3 surfaces over finite fields*, Invent. Math. **74** (1983), 213–237.

[36] N. Nygaard and A. Ogus, *Tate's conjecture for K3 surfaces of finite height*, Ann. of Math. (2) **122** (1985), no. 3, 461–507

[37] K. Oguiso, *Local families of K3 surfaces and applications*, J. Algebraic Geom., **12** (2003), no. 3, 405-433.

[38] Z. Patakfalvi, *Semi-positivity in positive characteristics,* Annales de ENS, to appear.

[39] U. Persson and H. Pinkham, *Degeneration of surfaces with trivial canonical bundle*, Ann. of Math.,(2), **113**, (1981), no. 1, 45–66.

[40] I. Piatetski-Shapiro and I. Shafarevich, *Torelli's theorem for algebraic surfaces of type K3*, Izv. Akad. Nauk SSSR Ser. Mat., **35** (1971), 530–572.

[41] J. Rizov, *Moduli stacks of polarized K3 surfaces in mixed characteristic*, Serdica Math. J., **32**, (2006), no. 2-3, 131-178.

[42] _____, *Kuga–Satake abelian varieties of K3 surfaces in mixed characteristic*, J. Reine Agnew. Math., **648**, (2010), 13–67.

[43] A. Rudakov, T. Zink, and I. Shafarevich, *The influence of height on degenerations of algebraic surfaces of type K3*, Math. USSR Izv. **20** (1983), 119–135.

[44] J. Tate, *Algebraic cycles and poles of zeta functions*, Arithmetic Algebraic Geometry, p. 93–110, Harper and Row, New York, 1965.

[45] B. Totaro, *Algebraic surfaces and hyperbolic geometry*, Current Developments in Algebraic Geometry, ed. L. Caporaso, J. McKernan, M. Mustata, and M. Popa, MSRI Publications **59**, Cambridge (2012), 405-426.

[46] A. Vasiu, *Moduli schemes and the Shafarevich conjecture (the arithmetic case) for pseudo-polarized K3 surfaces*, draft available at http://www.math.binghamton.edu/adrian.

[47] C. Voisin, *Hodge theory and complex algebraic geometry II*, Translated from the French by Leila Schneps. Reprint of the 2003 English edition. Cambridge Studies in Advanced Mathematics, 77. Cambridge University Press, Cambridge, 2007.

[48] Ju.G. Zarhin, *Endomorphisms of abelian varieties and points of finite order in characteristic p* (Russian), Mat. Zametki **21** (1977), no. 6, 737–744.

Department of Mathematics, Columbia University, 2990 Broadway, New York, NY 10025
E-mail: dmaulik@math.columbia.edu

The dimension of jet schemes of singular varieties

Mircea Mustaţă

Abstract. Given a scheme X over k, a generalized jet scheme parametrizes maps $\operatorname{Spec} A \to X$, where A is a finite-dimensional, local algebra over k. We give an overview of known results concerning the dimensions of these schemes for $A = k[t]/(t^m)$, when they are related to invariants of singularities in birational geometry. We end with a discussion of more general jet schemes.

Mathematics Subject Classification (2010). Primary 14E18; Secondary 14B05.

Keywords. Jet scheme, log canonical threshold, minimal log discrepancy.

1. Introduction

Given a scheme X (say, of finite type over an algebraically closed field k), a tangent vector to X can be identified with a morphism $\operatorname{Spec} k[t]/(t^2) \to X$. The tangent vectors to X are the closed points of the first jet scheme $J_1(X)$ of X. More generally, for every $m \geq 1$ one can define a scheme of finite type $J_m(X)$ whose closed points parametrize all morphisms $\operatorname{Spec} k[t]/(t^{m+1}) \to X$. Explicitly, if X is defined in some affine space $\mathbf{A}^N$ by polynomials $f_1, \ldots, f_r$, then as a set $J_m(X)$ is identified to the set of solutions of $f_1, \ldots, f_r$ in $k[t]/(t^{m+1})$. Truncating induces natural maps $J_p(X) \to J_q(X)$ for every $p > q$. When X is a smooth, n-dimensional variety, then every such projection is locally trivial in the Zariski topology, with fiber $\mathbf{A}^{(p-q)n}$. In particular, $J_m(X)$ is a smooth variety of dimension $(m+1)n$. However, when X is singular, the jet schemes $J_m(X)$ have a more complicated behavior that reflects in a subtle way the singularities of X.

In fact, instead of the algebras $k[t]/(t^{m+1})$ we can consider an arbitrary local, finite k-algebra A. The morphisms $\operatorname{Spec}(A) \to X$ are parametrized by a generalized jet scheme that we denote $J_A(X)$. For example, if

$$A = k[t_1, \ldots, t_r]/(t_1^{m_1+1}, \ldots, t_r^{m_r+1}),$$

then $J_A(X)$ is isomorphic to the r-iterated jet scheme $J_{m_1}(J_{m_2}(\ldots J_{m_r}(X)))$. The construction and the formal properties of these more general jet schemes are very similar to those of the usual ones. We give an overview of the construction and of the basic properties of the generalized jet schemes in §2. By taking suitable projective limits, one can then define $J_A(X)$ when A is a local, complete k-algebra, with residue field k. We describe this construction in §3. The much-studied case is that of $A = k[\![t]\!]$, when $J_A(X)$ is known as the scheme of arcs of X.

Information on the schemes $J_m(X)$ is provided by the change of variable formula in motivic integration, see [8]. More precisely, if X is embedded in a smooth variety Y, then

Proceedings of the International Congress of Mathematicians, Seoul, 2014

one can use a log resolution of the pair (Y, X) to compute, for example, the dimensions of the schemes $J_m(X)$ in terms of the data of the resolution. In this way one can relate the behavior of these dimensions to some of the invariants of the pair (Y, X) that appear in birational geometry. We describe this story in §4.

One can ask questions with a similar flavor about the dimensions of the generalized jet schemes $J_A(X)$, but very little is known in this direction. We propose in §5 some invariants defined in terms of the asymptotic behavior of the dimensions of $J_A(X)$, when A varies over certain sequences of algebras of embedding dimension 2. We also discuss the irreducibility of iterated jet schemes for locally complete intersection varieties.

2. Generalized jet schemes

Let k be a field of arbitrary characteristic. All schemes we consider are schemes over k. Let LFA/k be the category whose objects are local finite k-algebras, with residue field k, with the maps being local homomorphisms of k-algebras. Given a scheme X and $A \in \mathrm{LFA}/k$, the scheme of A-jets of X is a scheme $J_A(X)$ that satisfies the following universal property: for every scheme Y, there is a functorial bijection

$$\mathrm{Hom}_{\mathrm{Sch}/k}(Y, J_A(X)) \simeq \mathrm{Hom}_{\mathrm{Sch}/k}(Y \times \mathrm{Spec}\, A, X). \tag{2.1}$$

Standard arguments imply that it is enough to have a functorial bijection as in (2.1) when Y is an affine scheme. In particular, by taking $Y = \mathrm{Spec}\, k$, we see that the set of k-valued points of $J_A(X)$ is in bijection with the set of A-*jets* of X, that is, A-valued points of X. Note that if $A = k$, then $J_A(X) = X$.

Before discussing existence, we make some general remarks. Suppose that $\phi\colon A \to B$ is a homomorphism in LFA/k and that $J_A(X)$ and $J_B(X)$ exist. We have a functorial transformation

$$\mathrm{Hom}_{\mathrm{Sch}/k}(Y \times \mathrm{Spec}\, A, X) \to \mathrm{Hom}_{\mathrm{Sch}/k}(Y \times \mathrm{Spec}\, B, X)$$

given by composition with $Y \times \mathrm{Spec}\, B \to Y \times \mathrm{Spec}\, A$. This is induced via the isomorphism (2.1) by a unique scheme morphism $\pi^X_{B/A}\colon J_A(X) \to J_B(X)$. In particular, if $A \in \mathrm{LFA}/k$, then the morphism to the residue field $A \to k$ induces $\pi^X_A = \pi^X_{A/k}\colon J_A(X) \to X$. Similarly, the structure morphism $k \to A$ induces a section $s^X_A\colon X \to J_A(X)$ of π^X_A. Given two morphisms $A \to B$ and $B \to C$ in LFA/k, if $J_A(X)$, $J_B(X)$, and $J_C(X)$ exist, then it is easy to see that $\pi^X_{C/B} \circ \pi^X_{B/A} = \pi^X_{C/A}$. When the scheme X is clear from the context, we simply write $\pi_{B/A}$, π_A, and s_A instead of $\pi^X_{B/A}$, π^X_A, and s^X_A.

When $A = k[t]/(t^{m+1})$, the A-jets are called m-jets and the corresponding scheme is denoted $J_m(X)$, the m^{th} *jet scheme* of X. It is not hard to deduce from the universal property that $J_1(X)$ is isomorphic as a scheme over X with the total tangent space $\mathcal{S}pec(\mathrm{Sym}(\Omega_{X/k}))$. We now sketch the proof of the existence of $J_A(X)$. The argument is the same as in the case of the usual jet schemes, hence we refer the reader to [12, Section 2] for details.

Consider first the case when $X = \mathrm{Spec}\, S$ is affine and consider a presentation

$$S \simeq k[x_1, \ldots, x_N]/(f_1, \ldots, f_r).$$

Let us fix a basis $(e_i)_{1\leq i\leq m}$ for A. We can thus write

$$e_i \cdot e_j = \sum_{\ell=1}^{m} c_{i,j,\ell} e_\ell. \tag{2.2}$$

We consider an affine scheme $Y = \operatorname{Spec} R$. Giving a morphism $Y \times \operatorname{Spec} A \to X$ is equivalent to giving a ring homomorphism

$$\phi\colon k[x_1, \ldots, x_N]/(f_1, \ldots, f_r) \to R \otimes_k A. \tag{2.3}$$

This is uniquely determined by $\phi(x_j)$, for $1 \leq j \leq N$, which can be written as

$$\phi(x_j) = \sum_{i=1}^{m} a_{i,j} \otimes e_i, \text{ with } a_{i,j} \in R \text{ for } 1 \leq i \leq m,\ 1 \leq j \leq N.$$

Furthermore, for any such choice of $\phi(x_j)$, we get a k-algebra homomorphism

$$\widetilde{\phi}\colon k[x_1, \ldots, x_N] \to R \otimes_k A$$

and this induces a k-algebra homomorphism (2.3) if and only if $\widetilde{\phi}(f_\alpha) = 0$ for $1 \leq \alpha \leq r$. On the other hand, the relations (2.2) imply that for every α, we can find polynomials $P_\alpha^{(i)}$ in $a_{j,\ell}$, with coefficients in k (these coefficients in turn are polynomials in the structure constants $c_{i,j,\ell}$) such that for $\widetilde{\phi}$ as above, we have

$$\widetilde{\phi}(f_\alpha) = \sum_{i=1}^{m} P_\alpha^{(i)}(a_{1,1}, \ldots, a_{m,N}) \otimes e_i.$$

This shows that $J_A(X)$ is cut out in $\mathbf{A}^{mN}$ by the equations $P_\alpha^{(i)}$ for $1 \leq \alpha \leq r$ and $1 \leq i \leq m$.

The above argument shows that $J_A(X)$ exists whenever X is affine. It is then easy to check that if X is any scheme such that $J_A(X)$ exists, then for every open subset U of X, the scheme $J_A(U)$ exists and it is isomorphic to $(\pi_A^X)^{-1}(U)$. Given now an arbitrary scheme X, consider an affine open cover $X = \cup_i U_i$. Note that $J_A(U_i)$ exists for every i. Moreover, for every i and j, the schemes $(\pi_A^{U_i})^{-1}(U_i \cap U_j)$ and $(\pi_A^{U_j})^{-1}(U_i \cap U_j)$ are canonically isomorphic, being isomorphic to $J_A(U_i \cap U_j)$. We can thus glue the schemes $J_A(U_i)$ along these open subsets and it is then straightforward to check that the resulting scheme satisfies the universal property of $J_A(X)$. We collect in the next proposition the conclusion of the above discussion.

Proposition 2.1. *For every $A \in \mathrm{LFA}/k$ and every scheme X, the scheme $J_A(X)$ of A-jets of X exists. Moreover, the following hold:*

i) *If X is of finite type over k, then $J_A(X)$ is of finite type over k.*

ii) *The canonical projection $\pi_A\colon J_A(X) \to X$ is affine.*

iii) *If X is an affine subscheme of $\mathbf{A}^N$ defined by r equations and $\dim_k(A) = m$, then $J_A(X)$ is defined in $J_A(\mathbf{A}^N) \simeq \mathbf{A}^{Nm}$ by rm equations. More generally, if $A \to A'$ is a surjective homomorphism in LFA/k, then*

$$J_A(X) \hookrightarrow (\pi_{A'/A}^{\mathbf{A}^N})^{-1}(J_{A'}(X))$$

is cut out by $r \cdot (\dim_k(A) - \dim_k(A'))$ equations.

In what follows we discuss some basic properties of generalized jet schemes. It is clear that if $f\colon X \to Z$ is a morphism and $A \in \mathrm{LFA}/k$, we have a unique morphism $J_A(f)\colon J_A(X) \to J_A(Z)$ such that the bijection (2.1) is functorial also in X. Therefore taking X to $J_A(X)$ gives a functor, such that if $A \to B$ is a morphism of finite k-algebras, then $\pi_{B/A}$ is a natural transformation.

Example 2.2. Iterated jet schemes can be described as schemes of A-jets. Indeed, given $m_1, \ldots, m_r \in \mathbf{Z}_{\geq 0}$, if we take $A = k[t_1, \ldots, t_r]/(t_1^{m_1+1}, \ldots, t_r^{m_r+1})$, then it follows from the universal property that for every scheme X, we have a canonical isomorphism of schemes over X

$$J_A(X) \simeq J_{m_1}(J_{m_2}(\ldots J_{m_r}(X))).$$

Remark 2.3. It follows from the explicit description of the scheme of A-jets of an affine scheme X that if $\iota\colon Z \hookrightarrow X$ is a closed immersion, then for every $A \in \mathrm{LFA}/k$, the induced morphism $J_A(\iota)$ is a closed immersion.

Remark 2.4. If X is any scheme over k and K/k is a field extension, for every $A \in \mathrm{LFA}/k$ we have $A \otimes_k K \in \mathrm{LFA}/K$ and there is a canonical isomorphism

$$J_{A\otimes_k K}(X \times_{\operatorname{Spec} k} \operatorname{Spec} K) \simeq J_A(X) \times_{\operatorname{Spec} k} \operatorname{Spec} K.$$

The assertion follows easily from the isomorphism (2.1).

The following proposition describing the behavior of generalized jet schemes with respect to étale morphisms is an immediate consequence of (2.1) and of the fact that étale morphisms are formally étale.

Proposition 2.5. *If $f\colon X \to Y$ is an étale morphism of schemes of finite type over k, then for every $A \in \mathrm{LFA}/k$ we have a Cartezian diagram*

$$\begin{array}{ccc} J_A(X) & \xrightarrow{J_A(f)} & J_A(Y) \\ {\scriptstyle \pi_A^X}\downarrow & & \downarrow{\scriptstyle \pi_A^Y} \\ X & \xrightarrow{f} & Y. \end{array}$$

Using Proposition 2.5 and the description of the scheme of A-jets for an affine space, we obtain the following

Corollary 2.6. *If X is a smooth n-dimensional variety*[1] *over k, then for every surjective morphism $A \to B$ in LFA/k, the induced morphism $J_A(X) \to J_B(X)$ is locally trivial in the Zariski topology, with fiber $\mathbf{A}^{dn}$, where $d = \dim_k(A) - \dim_k(B)$. In particular, $J_A(X)$ is a smooth variety of dimension $n \cdot \dim_k(A)$.*

Corollary 2.7. *If X is a scheme of finite type over k, then for every $A \in \mathrm{LFA}/k$, we have*

$$\dim(X) \leq \frac{\dim(J_A(X))}{\dim_k(A)} \leq \max_{x\in X} \dim(T_x X).$$

[1] A *variety* is assumed to be reduced, irreducible, and of finite type over k.

Proof. The first inequality follows from Corollary 2.6 and the fact that there is a locally closed immersion $Y \hookrightarrow X$, with Y smooth and $\dim(Y) = \dim(X)$. The second inequality also follows from Corollary 2.6 since for every $x \in X$, if $n = \dim(T_x X)$, then there is an open neighborhood U of x in X and a closed immersion $U \hookrightarrow Z$, where Z is a smooth variety of dimension n. □

Example 2.8. If we consider arbitrary $A \in \mathrm{LFA}/k$, then we can get arbitrarily close to the upper-bound in Corollary 2.7. Indeed, if $x \in X$ is such that $n = \dim(T_x X)$ and $(A, \mathfrak{m}) \in \mathrm{LFA}/k$ is such that $\mathfrak{m}^2 = 0$ and $\dim_k(A) = r$, then

$$(\pi_A)^{-1}(x) \simeq \mathbf{A}^{(r-1)n} \text{ hence } \frac{\dim(J_A(X))}{\dim_k(A)} \geq n - \frac{n}{r}.$$

Remark 2.9. If G is a group scheme over k, then it is easy to see that $J_A(G)$ is a group scheme over k for every $A \in \mathrm{LFA}/k$. Moreover, if G acts on a scheme X, then $J_A(G)$ acts on $J_A(X)$.

Remark 2.10. Suppose that $A \in \mathrm{LFA}/k$ is a graded[2] k-algebra. The grading of A induces a morphism $\phi\colon \mathbf{A}^1 \times \mathrm{Spec}(A) \to \mathrm{Spec}(A)$ that corresponds to the k-algebra homomorphism $A \to A[t]$ that takes a homogeneous element $a \in A$ of degree m to at^m. If X is any scheme, then we obtain an induced morphism $\Phi_A^X\colon \mathbf{A}^1 \times J_A(X) \to J_A(X)$ that takes any morphism $(u, v)\colon Y \to \mathbf{A}^1 \times J_A(X)$, with

$$v \in \mathrm{Hom}(Y, J_A(X)) \simeq \mathrm{Hom}(Y \times \mathrm{Spec}(A), X)$$

to the composition

$$Y \times \mathrm{Spec}(A) \xrightarrow{(\mathrm{id}, u \circ \mathrm{pr}_1)} Y \times \mathrm{Spec}(A) \times \mathbf{A}^1 \xrightarrow{(\mathrm{id}_Y, \phi)} Y \times \mathrm{Spec}(A) \xrightarrow{\quad v \quad} X.$$

It is easy to see that the restriction of Φ_A^X to $J_A(X) \simeq \{0\} \times J_A(X)$ is equal to $s_A^X \circ \pi_A^X$. In particular, for such A we see that if Z is an irreducible component of $J_A(X)$, then Φ_A^X induces a morphism $\mathbf{A}^1 \times Z \to Z$, hence $s_A^X \circ \pi_A^X(Z) \subseteq Z$.

3. Generalized arc schemes

We now generalize the construction in the previous section to complete local algebras. More precisely, let LCA/k be the category of complete local Noetherian k-algebras, with residue field k (the maps being local morphisms of k-algebras). Note that LFA/k is a full subcategory of LCA/k.

Given $(A, \mathfrak{m}) \in \mathrm{LCA}/k$ and a scheme X over k, we define a scheme $J_A(X)$

$$J_A(X) := \varprojlim_{A \to B} J_B(X), \tag{3.1}$$

where the projective limit is over all surjective maps $A \to B$ in LCA/k, with B lying in LFA/k. Note that a map from $\phi\colon A \to B$ to $\psi\colon A \to C$ is a map $f\colon B \to C$ such that $f \circ \phi = \psi$ and such a map induces a morphism $\pi_{C/B}^X\colon J_B(X) \to J_C(X)$. Since all $J_B(X)$

[2] All graded algebras we consider are graded by $\mathbf{Z}_{\geq 0}$.

are affine schemes over X, it follows that the projective limit (3.1) exists. In fact, if $U \subseteq X$ is an affine open subset, then

$$\Gamma(J_A(U), \mathcal{O}_{J_A(U)}) \simeq \varinjlim_{A \to B} \Gamma(J_B(U), \mathcal{O}_{J_B(U)}).$$

Note that if $A \in \mathrm{LFA}/k$, then we recover the previous definition. It is clear that if $h\colon X \to Y$ is a morphism of schemes, we obtain an induced morphism $J_A(h)\colon J_A(X) \to J_A(Y)$ and in this way we get a functor from the category of schemes over k to itself. If $g\colon A_1 \to A_2$ is a morphism in LCA/k, we obtain a functorial transformation

$$\pi^X_{A_2/A_1}\colon J_{A_1}(X) \to J_{A_2}(X).$$

Indeed, if $\phi\colon A_2 \to B_2$ is a surjective map in LCA/k, with B_2 finite over k, then $\phi \circ g$ factors through a quotient B_1 of A_1 by a power of the maximal ideal, hence we have a map $J_{A_1}(X) \to J_{B_1}(X) \to J_{B_2}(X)$, where the first map is given by the definition of projective limit and the second map is $\pi^X_{B_2/B_1}$. Note that this definition has the following two properties:

i) If $\phi\colon A \to B$ is a surjective map in LCA/k, with B finite over k, then the map $J_A(X) \to J_B(X)$ given by the projective limit definition is the same as $\pi^X_{B/A}$.

ii) If $A \to B \to C$ are maps in LCA/k, then $\pi^X_{C/B} \circ \pi^X_{B/A} = \pi^X_{C/A}$.

Remark 3.1. Suppose that X is a scheme and $A \in \mathrm{LCA}/k$. For every k-algebra R, consider the completion $R\widehat{\otimes}_k A$ of $R \otimes_k A$ with respect to the topology induced by A (more precisely, if $\mathfrak{m}_A$ is the maximal ideal in A, then the topology on $R \otimes_k A$ is the $\mathfrak{m}_A \cdot (R \otimes_k A)$-adic topology). In this case we have a canonical functorial map

$$\mathrm{Hom}(\mathrm{Spec}\, R\widehat{\otimes}_k A, X) \to \mathrm{Hom}(\mathrm{Spec}\, R, J_A(X)). \tag{3.2}$$

It is easy to see that this is a bijection if $R = k$ or if X is affine.

As we see in the next example, even when X is a finite type, the scheme $J_A(X)$ is not, in general, of finite type.

Example 3.2. If $A = k[\![t]\!]$, the scheme $J_A(X) = \varprojlim_m J_m(X)$ is the *scheme of arcs* of X, denoted by $J_\infty(X)$. For example, if $X = \mathbf{A}^n$, with $n \geq 1$, then $J_\infty(X)$ is an infinite-dimensional affine space, that is, $J_\infty(X) \simeq \mathrm{Spec}\, k[x_1, x_2, \ldots]$.

If $A = k[\![t_1, t_2]\!]$, then $J_A(X)$ is known as the *space of wedges* of X. It is easy to deduce from Remark 3.1 that this is canonically isomorphic to $J_\infty(J_\infty(X))$. More generally, if we put $A_n = k[\![x_1, \ldots, x_n]\!]$, then we have a canonical isomorphism $J_\infty(J_{A_n}(X)) \simeq J_{A_{n+1}}(X)$ for every $n \geq 1$.

Remark 3.3. It would be interesting to have explicit examples of schemes $J_A(X)$, when X is singular. Very few such examples are known and all of these only deal with $J_m(X)$ or $J_\infty(X)$. Moreover, in almost all cases one can only describe the reduced structure on these spaces. An easy example is that of schemes defined by monomial ideals in a polynomial ring (see [23, Proposition 4.10]). A more interesting example is that of $J_\infty(X)$, when X is a toric variety. In this case, if T is the torus acting on X, one can completely describe the orbits of the $J_\infty(T)$-action on $J_\infty(X)$, see [15]). It is much trickier to describe $J_m(X)$ for

a toric variety X; this is only understood in the 2-dimensional case, see [21]. One example in which both $J_m(X)$ and $J_\infty(X)$ are well understood is that of a generic determinantal variety X. In this case, X is a closed subscheme of a space of matrices $M = M_{m,n}(k)$ and the group $G = GL_m(k) \times GL_n(k)$ acts on M inducing an action on X. For a description of the orbits of $J_m(G)$ on $J_m(X)$ and of the orbits of $J_\infty(G)$ on $J_\infty(X)$, see [9].

4. Dimensions of jet schemes and invariants of singularities

Beyond the formal properties that we discussed, little is known about jet schemes in the generality that we considered in the previous two sections. We now restrict to the case of the "usual" jet schemes $J_m(X)$, for which a lot is known due to the connection to birational geometry that comes out of the theory of motivic integration.

In order to describe this connection, we first recall how one measures singularities in birational geometry. The idea is to use *all* divisorial valuations, suitably normalized by the order of vanishing along the relative Jacobian ideal. From now on we assume that the ground field k is algebraically closed, of characteristic 0, and we only consider the k-valued points of the schemes involved.

For simplicity, we assume that we work on an ambient smooth variety X. Let $\mathfrak{a}$ be a nonzero ideal on X (all ideals are assumed to be coherent). A *divisor over* X is a prime divisor on a normal variety Y that has a birational morphism to X (Y is a *birational model over* X). Each such divisor induces a valuation ord_E of the function field $K(Y) = K(X)$. We identify two such divisors if they give the same valuation. In particular, if $Z \to Y$ is a birational morphism, with Z normal, then we identify E with its strict transform on Z. Therefore we may always assume that Y is smooth (using a resolution of singularities of Y) and that Y is proper over X (using Nagata's compactification theorem). Given a divisor E over X, its center on X denoted $c_X(E)$ is the closure of the image of E in X (it is easy to see that this is independent of the chosen model).

Let E be a divisor over X. To a nonzero ideal $\mathfrak{a}$ on X, we can attach a nonnegative integer $\mathrm{ord}_E(\mathfrak{a})$, defined as follows. We may assume that E is a divisor on a smooth variety Y over X, such that the structural morphism $f\colon Y \to X$ factors through the blow-up along $\mathfrak{a}$. Therefore we may write $\mathfrak{a}\cdot\mathcal{O}_Y = \mathcal{O}_Y(-D)$ for an effective divisor D on Y and $\mathrm{ord}_E(\mathfrak{a}) = \mathrm{ord}_E(D)$ is the coefficient of E in D. For example, if $x \in X$ and E is the exceptional divisor on the blow-up of X at x, then $\mathrm{ord}_E(\mathfrak{a})$ is the *order of* $\mathfrak{a}$ at x, denoted by $\mathrm{ord}_x(\mathfrak{a})$. This is characterized by

$$\mathrm{ord}_x(\mathfrak{a}) := \max\{r \mid \mathfrak{a} \subseteq \mathfrak{m}_x^r\},$$

where $\mathfrak{m}_x$ is the ideal defining x.

The idea is to measure the singularities of $\mathfrak{a}$ using all invariants $\mathrm{ord}_E(\mathfrak{a})$, where E varies over the divisors over X. Very roughly, one thinks of the singularities of $\mathfrak{a}$ being "worse" if $\mathrm{ord}_E(\mathfrak{a})$ is larger. On the other hand, when we vary E, the numbers $\mathrm{ord}_E(\mathfrak{a})$ are unbounded, hence we need a normalizing factor. It turns out that the right factor to use is provided by the *log discrepancy*, which is defined as follows. If $f\colon Y \to X$ is a proper, birational morphism, with Y a smooth variety and E is a prime divisor on Y, then the *relative canonical class* $K_{Y/X}$ is the degeneracy locus of the morphism of vector bundles of the same rank $f^*(\Omega_X) \to \Omega_Y$. In other words, $K_{Y/X}$ is locally defined by the determinant of the Jacobian matrix of f. The log discrepancy of ord_E is $A(\mathrm{ord}_E) := \mathrm{ord}_E(K_{Y/X}) + 1$. It is easy to check that the definition does not depend on the model Y we have chosen.

There are various related invariants of singularities that one considers in birational geometry. In what follows, we focus on two such invariants, the *log canonical threshold* and the *minimal log discrepancy*. We begin by introducing the former invariant. If X is smooth and $\mathfrak{a}$ is a nonzero ideal on X, then the log canonical threshold of $\mathfrak{a}$ is

$$\mathrm{lct}(\mathfrak{a}) := \inf_E \frac{A(\mathrm{ord}_E)}{\mathrm{ord}_E(\mathfrak{a})},$$

where the infimum is over all divisors E over X. Note that this is finite whenever $\mathfrak{a} \neq \mathcal{O}_X$ and by convention we put $\mathrm{lct}(\mathcal{O}_X) = \infty$ and $\mathrm{lct}(0) = 0$. If W is the closed subscheme defined by $\mathfrak{a}$, we sometimes write $\mathrm{lct}(X, W)$ for $\mathrm{lct}(\mathfrak{a})$. Note that in the definition of $\mathrm{lct}(\mathfrak{a})$ we consider the reciprocals of the invariants $\mathrm{ord}_E(\mathfrak{a})$, hence "worse" singularities correspond to smaller log canonical thresholds.

While defined in terms of all divisors over X, it is a consequence of resolution of singularities that the log canonical threshold can be computed on suitable models. Recall that a log resolution of the pair $(X, \mathfrak{a})$ is a proper birational morphism $f\colon Y \to X$, with Y smooth, such that $\mathfrak{a} \cdot \mathcal{O}_Y = \mathcal{O}_Y(-D)$ for an effective divisor D on Y such that $D + K_{Y/X}$ has simple normal crossings[3]. The existence of such resolutions follows from Hironaka's fundamental results. A basic result about log canonical thresholds says that if $f\colon Y \to X$ is a log resolution of $(X, \mathfrak{a})$ as above, then $\mathrm{lct}(\mathfrak{a})$ is computed by the divisors on Y: if we write $D = \sum_{i=1}^r a_i E_i$ and $K_{Y/X} = \sum_{i=1}^r k_i E_i$, then

$$\mathrm{lct}(\mathfrak{a}) = \min_{i=1}^r \frac{k_i + 1}{a_i}.$$

A consequence of this formula is the fact, not apparent from the definition, that $\mathrm{lct}(\mathfrak{a})$ is a rational number.

The log canonical threshold is a fundamental invariant of singularities. It appeared implicitly already in Atiyah's paper [2] in connection with the meromorphic continuation of complex powers. The first properties of the log canonical threshold have been proved by Varchenko in connection with his work on asymptotic expansions of integrals and mixed Hodge structures on the vanishing cohomology, see [27], [28], and [29]. It was Shokurov who introduced the log canonical threshold in the context of birational geometry in [25]. From this point of view, $\mathrm{lct}(\mathfrak{a})$ is the largest rational number q such that the pair $(X, \mathfrak{a}^q)$ is *log canonical*. We mention that the notion of log canonical pairs is of central importance in the Minimal Model Program, since it gives the largest class of varieties for which one can hope to apply the program. In fact, in the context of birational geometry it is useful to not require that the ambient variety is smooth, but only that it has mild singularities, and it is in this more general setting that one can define the log canonical threshold. A remarkable feature of this invariant is that it is related to many points of view on singularities. We refer to [19] and [24] for an overview of some of these connections and for the basic properties of the log canonical threshold.

Example 4.1. In order to illustrate the behavior of the log canonical threshold, we list a few examples. When $\mathfrak{a}$ is generated by $f \in \mathcal{O}(X)$, we simply write $\mathrm{lct}(f)$ for the corresponding log canonical threshold.

[3] A divisor on a smooth variety has simple normal crossings if around every point we can find local algebraic coordinates $x_1, \ldots, x_n$ such that the divisor is defined by $x_1^{a_1} \cdots x_n^{a_n}$ for some nonnegative integers $a_1, \ldots, a_n$.

i) If the subscheme $V(\mathfrak{a})$ defined by $\mathfrak{a}$ has codimension r, then $\mathrm{lct}(\mathfrak{a}) \leq r$. This is an equality if $V(\mathfrak{a})$ is smooth.

ii) If $f = x_1^{a_1} \cdots x_n^{a_n} \in k[x_1, \ldots, x_n]$, then $\mathrm{lct}(f) = \min_i \frac{1}{a_i}$.

iii) If $f = x_1^{a_1} + \ldots + x_n^{a_n} \in k[x_1, \ldots, x_n]$, then $\mathrm{lct}(f) = \min\left\{1, \sum_i \frac{1}{a_i}\right\}$.

Remark 4.2. In terms of size, the log canonical threshold in a neighborhood of a point $x \in X$ is comparable to $\mathrm{ord}_x(\mathfrak{a})$. More precisely, given $x \in X$, there is an open neighborhood U of x such that the following inequalities hold:

$$\frac{1}{\mathrm{ord}_x(\mathfrak{a})} \leq \mathrm{lct}(\mathfrak{a}|_U) \leq \frac{\dim(X)}{\mathrm{ord}_x(\mathfrak{a})}.$$

It turns out that the log canonical threshold governs the growth of the dimensions of the jet schemes $J_m(W)$ of a scheme W. Note that by taking a finite affine open cover of W, we can reduce to the case when W can be embedded in a smooth variety (for example, in an affine space).

Theorem 4.3. *If X is a smooth n-dimensional variety and W is a proper, nonempty, closed subscheme of X, then*

$$\lim_{m\to\infty} \frac{\dim(J_m(W))}{m+1} = \max_m \frac{\dim(J_m(W))}{m+1} = n - \mathrm{lct}(X, W).$$

Moreover, the maximum above is achieved for all m such that $(m+1)$ is divisible enough.

This result was proved in [23] by making use of the change of variable formula in motivic integration (in fact, the formula had appeared implicitly earlier in [7]). We explain below, following [10], how this theorem follows from a more general result relating the approach to singularities via divisors over X and that using certain subsets in the space of arcs $J_\infty(X)$. Before doing this, we discuss another invariant of singularities, whose definition is similar to that of the log canonical threshold, but whose behavior turns out to be more difficult to study.

Suppose, as above, that X is a smooth variety, $\mathfrak{a}$ is a nonzero ideal on X, and q is a positive rational number. In order to avoid some pathologies of a trivial nature, we assume $\dim(X) \geq 2$. If Z is a proper, irreducible closed subset of X, then the *minimal log discrepancy* $\mathrm{mld}_Z(X, \mathfrak{a}^q)$ is defined as

$$\mathrm{mld}_Z(X, \mathfrak{a}^q) := \inf\{A(\mathrm{ord}_E) - q \cdot \mathrm{ord}_E(\mathfrak{a}) \mid c_X(E) \subseteq Z\}. \tag{4.1}$$

Note that "better" singularities correspond to larger minimal log discrepancies. It is a basic fact that the minimal log discrepancy is either $-\infty$ or it is nonnegative. We have $\mathrm{mld}_Z(X, \mathfrak{a}^q) \geq 0$ if and only if the pair $(X, \mathfrak{a}^q)$ is log canonical around Z, that is, there is an open neighborhood U of Z such that $\mathrm{lct}(\mathfrak{a}|_U) \geq q$. Like the log canonical threshold, the minimal log discrepancy can be computed on suitable models. More precisely, if I_Z is the ideal defining Z and $f\colon Y \to X$ is a log resolution of $(X, I_Z \cdot \mathfrak{a})$, then there is a prime divisor E on Y that achieves the infimum in (4.1), assuming that this infimum is not $-\infty$ (moreover, finitenes also can be tested just on the divisors on Y). For an introduction to minimal log discrepancies, we refer to [1]. The following result gives an interpretation of minimal log discrepancies in terms of jet schemes.

Theorem 4.4. *Let X be a smooth variety of dimension ≥ 2, $\mathfrak{a}$ a nonzero ideal on X defining the subscheme W, and q a positive rational number. For every proper, irreducible closed subset Z of X, we have*

$$\mathrm{mld}_Z(X, \mathfrak{a}^q) = \inf\{(m+1)(\dim(X) - q) - \dim(J_m(W) \cap (\pi_m^W)^{-1}(Z))\}.$$

Moreover, if the infimum is not $-\infty$, then it is a minimum.

The theorem was proved in [13] using the change of variable formula in motivic integration. In this special form in which X is assumed to be smooth, it can also be deduced from the main result in [10] that we discuss below.

We note that part of the interest in invariants of singularities like the log canonical threshold and the minimal log discrepancy comes from the fact that their behavior is related to one of the outstanding open problems in birational geometry, namely Termination of Flips (see [3] for the connection of this conjecture to log canonical thresholds and [26] for the connection to minimal log discrepancies). In this respect, there are two points to keep in mind:

- For all applications to birational geometry, it is important to work with ambient varieties that have mild (log canonical) singularities, and not just smooth, as above. In this context, one can still describe minimal log discrepancies in terms of properties of (certain subsets in) jet schemes, see [13]. However, this description is less effective in general. It was used to prove some open questions about minimal log discrepancies, such as Inversion of Adjunction and Semicontinuity in the case when the ambient variety is locally complete intersection, see [11] and [13]. However, this method has not been successful so far in dealing with more general ambient varieties.
- While properties of log canonical thresholds are better understood (see below for the ACC property), it is in fact the (conjectural) properties of minimal log discrepancies that would give a positive answer to the Termination of Flips Conjecture. More precisely, Shokurov has shown that two conjectures on minimal log discrepancies, the Semicontinuity Conjecture and the ACC Conjecture imply Termination of Flips; see [26] for the precise statements.

In order to give the flavor of ACC statements regarding invariants of singularities, we state the following result concerning log canonical thresholds.

Theorem 4.5. *For every $n \geq 1$, the set of rational numbers*

$$\{\mathrm{lct}(\mathfrak{a}) \mid \mathfrak{a} \subsetneq \mathcal{O}_X,\ X \text{ smooth},\ \dim(X) \leq n\}$$

satisfies ACC, that is, it contains no infinite strictly increasing sequences.

This result was conjectured by Shokurov and proved in [6]. A general version, in which the ambient variety is not assumed to be smooth (which, as we have already mentioned, is much more useful for birational geometry) has been recently proved in [14]. It is worth mentioning that a corresponding conjecture for minimal log discrepancies is widely open even in the case of smooth ambient varieties. For recent progress motivated by this question, see [17] and [18].

Besides the ACC Conjecture for minimal log discrepancies, the other important open problem about these invariants concerns their semicontinuity (this was conjectured by Ambro [1]). As we have already mentioned, when the ambient variety is smooth, this can be deduced from Theorem 4.4 using general properties of the dimension of algebraic varieties in families.

Corollary 4.6. *If X is a smooth variety, $\mathfrak{a}$ is a nonzero ideal on X, and q is a positive rational number, then the map*

$$X \ni x \to \mathrm{mld}_x(X, \mathfrak{a}^q)$$

is lower semicontinuous.

Our next goal is to describe more generally, following [10], a dictionary between the approach to singularities using divisorial valuations and that using certain subsets in the space of arcs. We fix a smooth variety X of dimension n. We will be concerned with certain subsets in the space of arcs $J_\infty(X)$. In what follows we restrict to the k-valued points of $J_\infty(X)$, considered as a topological space with the Zariski topology. Recall that we have canonical projections $\pi_m \colon J_\infty(X) \to J_m(X)$. A *cylinder* in $J_\infty(X)$ is a subset of the form $C = \pi_m^{-1}(S)$, where S is a constructible subset in $J_m(X)$. We say that C is closed, locally closed, or irreducible, if S has this property. Moreover, we put

$$\mathrm{codim}(C) := \mathrm{codim}(S, J_m(X)) = (m+1)n - \dim(S).$$

It is easy to see that all these notions do not depend on m, since the natural projections $J_{m+1}(X) \to J_m(X)$ are locally trivial with fiber $\mathbf{A}^n$.

The main examples arise as follows. Suppose that $\mathfrak{a}$ is a nonzero ideal on X, defining the subscheme W. Associated to W we have a function $\mathrm{ord}_W \colon J_\infty(X) \to \mathbf{Z}_{\geq 0} \cup \{\infty\}$ such that for $\gamma \colon \mathrm{Spec}\, k[\![t]\!] \to X$, we have

$$\mathrm{ord}_W(\gamma) = \mathrm{ord}_t(\gamma^{-1}(\mathfrak{a}))$$

(with the convention that this is ∞ if the ideal $\gamma^{-1}(\mathfrak{a})$ is 0). With this notation, the *contact locus*

$$\mathrm{Cont}^{\geq m}(W) := \mathrm{ord}_W^{-1}(\geq m)$$

is a closed cylinder, hence

$$\mathrm{Cont}^{m}(W) := \mathrm{ord}_W^{-1}(m) = \mathrm{Cont}^{\geq m}(W) \smallsetminus \mathrm{Cont}^{\geq (m+1)}(W)$$

is a locally closed cylinder. In fact, we have

$$\mathrm{Cont}^{\geq (m+1)}(W) = \pi_m^{-1}(J_m(W)).$$

In particular, we have

$$\mathrm{codim}(\mathrm{Cont}^{\geq (m+1)}(W)) = (m+1)n - \dim(J_m(W)).$$

The main point of the correspondence we are going to describe is that divisorial valuations correspond to cylinders in such a way that the log discrepancy function translates to the codimension of the cylinder. In order to simplify the exposition, let us assume that $X = \mathrm{Spec}(R)$ is affine. Note first that if C is an irreducible closed cylinder in $J_\infty(X)$, then we may define a map

$$\mathrm{ord}_C \colon R \to \mathbf{Z}_{\geq 0} \cup \{\infty\},\ \mathrm{ord}_C(f) := \min\{\mathrm{ord}_{V(f)}(\gamma) \mid \gamma \in C\}.$$

This satisfies

$$\mathrm{ord}_C(f+g) \geq \min\{\mathrm{ord}_C(f), \mathrm{ord}_C(g)\} \text{ and } \mathrm{ord}_C(fg) = \mathrm{ord}_C(f) + \mathrm{ord}_C(g).$$

Moreover, if C does not dominate X, then $\mathrm{ord}_C(f) < \infty$ for every nonzero f, hence ord_C extends to a valuation of the function field of X. The following is the main result from [10] concerning the description of divisorial valuations in terms of cylinders in the space or arcs. For an extension to singular varieties, see [4].

Theorem 4.7. *Let X be a smooth variety.*

i) *If C is an irreducible closed cylinder in $J_\infty(X)$ that does not dominate X, then there is a divisor E over X and a positive integer q such that* $\mathrm{ord}_C = q \cdot \mathrm{ord}_E$.

ii) *For every divisor E over X and every positive integer q, there is a closed irreducible cylinder C in $J_\infty(X)$ such that* $\mathrm{ord}_C = q \cdot \mathrm{ord}_E$. *Moreover, there is a unique maximal such cylinder $C = C_q(E)$, with respect to inclusion, and*

$$\mathrm{codim}(C_q(E)) = q \cdot A(\mathrm{ord}_E).$$

It is easy to see that this result implies both Theorems 4.3 and 4.4. The key step in the proof of Theorem 4.7 consists in analyzing the valuations corresponding to the irreducible components of $\mathrm{Cont}^{\geq m}(W)$, for a closed subscheme W. This is done by considering a log resolution $f\colon Y \to X$ of $(X, \mathfrak{a})$, where $\mathfrak{a}$ is the ideal defining W. In this case, we have an induced map $g = J_\infty(f)\colon J_\infty(Y) \to J_\infty(X)$. If f is an isomorphism over $X \smallsetminus A$, where A is a proper closed subset of X, then the valuative criterion for properness implies that g is a bijection over $J_\infty(Y) \smallsetminus J_\infty(A)$. While $J_\infty(A)$ is "small" in $J_\infty(X)$ (for example, if the ground field is uncountable, then no cylinder is contained in $J_\infty(A)$), the map g is far from being a homeomorphism over $J_\infty(Y) \smallsetminus J_\infty(A)$. In fact, it is a fundamental result that if C is a cylinder contained in $\mathrm{Cont}^e(K_{Y/X})$, then $g(C)$ is again a cylinder and

$$\mathrm{codim}(g(C)) = \mathrm{codim}(C) + e.$$

This is a consequence of the geometric statement behind the change of variable formula in motivic integration, due to Kontsevich [20]. For the relevant statement and its proof, as well as for an important generalization to the case when X is not smooth, we refer to [8]. Since $\mathrm{Cont}^{\geq m}(f^{-1}(W)) = g^{-1}(\mathrm{Cont}^{\geq m}(W))$, one can deduce the following formula for the codimension of $\mathrm{Cont}^{\geq m}(W)$, see [10] for details.

Theorem 4.8. *Let X be a smooth variety and W a proper closed subscheme of X, defined by the ideal $\mathfrak{a}$. If $f\colon Y \to X$ is a log resolution of the pair $(X, \mathfrak{a})$ and we write*

$$f^{-1}(W) = \sum_{i=1}^r a_i E_i \text{ and } K_{Y/X} = \sum_{i=1}^r k_i E_i,$$

then $\mathrm{codim}(\mathrm{Cont}^{\geq m}(W))$ *is equal to*

$$\min\left\{\sum_{i=1}^r (k_i+1)\nu_i \mid \nu = (\nu_1, \ldots, \nu_r) \in \mathbf{Z}^r_{\geq 0}, \bigcap_{\nu_i > 0} E_i \neq \emptyset, \sum_{i=1}^r a_i\nu_i \geq m\right\}.$$

Moreover, it is easy to see that in the setting of this theorem, every irreducible component C of $\mathrm{Cont}^{\geq m}(W)$ is the closure of the image of a multi-contact locus of the form $\cap_{i=1}^r \mathrm{Cont}^{\geq \nu_i}(E_i)$. Using this, one deduces that ord_C is equal to $q \cdot \mathrm{ord}_E$, where E is the

exceptional divisor on a suitable weighted blow-up of Y with respect to the simple normal crossing divisor $\sum_i E_i$. We note that since both log canonical thresholds and minimal log discrepancies can be computed using log resolutions, the statement of Theorem 4.8 is enough to imply Theorems 4.3 and 4.4. On the other hand, one can give a proof for Theorem 4.7 without using log resolutions, see [30]. An advantage of that approach is that one obtains the same result in positive characteristic.

Remark 4.9. It is an immediate consequence of Theorem 4.8 that if W is any scheme, then

$$\frac{\dim(J_{m-1}(W))}{m} \leq \frac{\dim(J_{mp-1}(W))}{mp}$$

for every positive integers m and p. It would be interesting to give a direct proof of this inequality without relying on log resolutions. Such an argument could then hopefully be extended to cover other jet schemes $J_A(X)$, for suitable A.

Remark 4.10. It was shown in [23] that one can use the description of the log canonical threshold in Theorem 4.3 to reprove some of the basic properties of this invariant. For example, one can use this approach to prove the following special case of Inversion of Adjunction: if X is a smooth variety and H is a smooth hypersurface in X, then for every ideal $\mathfrak{a}$ on X such that $\mathfrak{a} \cdot \mathcal{O}_H \neq 0$, there is an open neighborhood U of H such that

$$\mathrm{lct}(\mathfrak{a}|_U) \geq \mathrm{lct}(\mathfrak{a} \cdot \mathcal{O}_H). \tag{4.2}$$

The usual proof of this inequality makes use of vanishing theorems (see for example [19]). Since Theorem 4.8 also holds in positive characteristic, one deduces that the inequality (4.2) holds in this setting as well, see [30], in spite of the fact that vanishing theorems can fail.

Remark 4.11. It would be interesting to find a jet-theoretic proof of the following result of Varchenko. Suppose that T is a connected scheme and $\mathcal{W} \hookrightarrow \mathbf{A}^n \times T$ is an effective Cartier divisor, flat over T, such that for every (closed) point $t \in T$, the induced divisor $\mathcal{W}_t \hookrightarrow \mathbf{A}^n$ has an isolated singularity at 0. By using the connection between the log canonical threshold and Steenbrink's spectrum of a hypersurface singularity, Varchenko showed in [27] that if the Milnor number at 0 for each $\mathcal{W}_t$ is constant for $t \in T$, then $\mathrm{lct}_0(\mathbf{A}^n, \mathcal{W}_t)$ is independent of t (here $\mathrm{lct}_0(\mathbf{A}^n, \mathcal{W}_t) = \mathrm{lct}(U, \mathcal{W}_t \cap U)$, where U is a small neighborhood of 0). It would be interesting to deduce this fact from the behavior of jet schemes. It is easy to see that the jet schemes $J_m(\mathcal{W}_t)$ are the fibers of a relative jet scheme $J_m(\mathcal{W}/T)$ over T and a natural question is whether the constancy of Milnor numbers implies that this family is flat over T (in a neighborhood of the fiber over 0 via the natural map $J_m(\mathcal{W}/T) \to \mathbf{A}^n$).

Remark 4.12. Suppose, for simplicity, that $X = \mathrm{Spec}(R)$ is a smooth affine variety. Recall that if $\mathfrak{a}$ is an ideal in R, then its integral closure $\overline{\mathfrak{a}}$ consists of all $\phi \in R$ such that $\mathrm{ord}_E(\phi) \geq \mathrm{ord}_E(\mathfrak{a})$ for all divisors E over X. It follows from Theorem 4.7 that if $\mathfrak{b}$ is another ideal in R, then $\mathfrak{b} \subseteq \overline{\mathfrak{a}}$ if and only if $\mathrm{Cont}^{\geq m}(\mathfrak{a}) \subseteq \mathrm{Cont}^{\geq m}(\mathfrak{b})$ for all m. In particular, the integral closure $\overline{\mathfrak{a}}$ is determined by the contact loci $(\mathrm{Cont}^{\geq m}(\mathfrak{a}))_{m \geq 1}$. Given other invariants of an ideal that only depend on the integral closure, it would be interesting to find a direct description of these invariants in terms of the contact loci of that ideal. For example, if $\mathfrak{a}$ is supported at a point $x \in X$, it would be interesting to find a description of the Samuel multiplicity $e(\mathfrak{a} \cdot \mathcal{O}_{X,x}, \mathcal{O}_{X,x})$ in terms of the contact loci of $\mathfrak{a}$.

We now turn to a related connection between singularities and jet schemes. It turns out that in the case of locally complete intersection varieties, one can characterize various classes of singularities in terms of the behavior of the jet schemes. The following is the main result in this direction.

Theorem 4.13. *Let W be a locally complete intersection variety.*

i) *If W is normal, then W has log canonical singularities if and only if $J_m(W)$ has pure dimension for every $m \geq 0$. Moreover, in this case we have $\dim(J_m(W)) = (m+1) \cdot \dim(X)$ and $J_m(W)$ is a locally complete intersection.*

ii) *The variety W has rational (equivalently, canonical) singularities if and only if $J_m(W)$ is irreducible for every $m \geq 0$.*

iii) *The variety W has terminal singularities if and only if $J_m(W)$ is normal for every $m \geq 0$.*

The assertion in ii) was proved in [22] using motivic integration. It was then noticed in [10] that the main ingredient in the proof can also be deduced from Theorem 4.7. The result had been conjectured by David Eisenbud and Edward Frenkel. They use it in the appendix to [22] to give an analogue of a result of Kostant to the setting of loop Lie algebras. The assertions in i) and iii) follow using similar ideas once some basic Inversion of Adjunction statements are proved, see [11] and [13].

We end with an interpretation for the condition of having pure-dimensional or irreducible jet schemes under the assumptions of Theorem 4.13. It turns out that these considerations can be made in the context of generalized jet schemes and we will make use of this in the next section.

Proposition 4.14. *Let W be an n-dimensional locally complete intersection variety and let A be a local, finite k-algebra, with $\dim_k(A) = \ell$.*

i) *$J_A(W)$ is pure-dimensional if and only if all irreducible components of $J_A(W)$ have dimension ℓn. If this is the case, then $J_A(W)$ is locally a complete intersection.*

ii) *$J_A(W)$ is irreducible if and only if the inverse image in $J_A(W)$ of the singular locus of W has dimension $< \ell n$. If this is the case, then $J_A(W)$ is also reduced.*

iii) *Suppose that $A \to A'$ is a surjective k-algebra homomorphism, where A and A' are finite, local k-algebras, with A' being positively graded. If $J_A(W)$ is pure-dimensional or irreducible, then $J_{A'}(W)$ has the same property.*

Proof. If we write $W = U_1 \cup \ldots \cup U_m$, with each U_i open in X, then $J_A(W) = J_A(U_1) \cup \ldots \cup J_A(U_m)$. Using this, it is easy to see that if the assertions in the proposition hold for each U_i, then they hold for X. Therefore we may assume W is a closed subvariety of $X = \mathbf{A}^d$, whose ideal is generated by $r = d - n$ elements.

We have seen in Proposition 2.1 that $J_A(W)$ is cut out in $J_A(X) \simeq \mathbf{A}^{\ell d}$ by ℓr equations. Therefore each irreducible component of $J_A(X)$ has dimension $\geq \ell n$. On the other hand, it follows from Corollary 2.6 that $J_A(X_{\mathrm{sm}})$ is an open subset of $J_A(X)$ of dimension ℓn. We thus conclude that $J_A(X)$ is pure-dimensional if and only if all irreducible components of $J_A(X)$ have dimension ℓn. We also see that if this is the case, then $J_A(X)$ is itself a locally complete intersection. This proves i).

Note also that $J_A(W)$ is irreducible if and only if the inverse image of the singular locus W_{sing} in $J_A(W)$ has dimension $< \ell n$. If this is the case, then $J_A(W)$ is generically reduced

and being Cohen-Macaulay (recall that $J_A(W)$ has to be a locally complete intersection), it is reduced. This gives ii).

In order to prove iii), let $\ell' = \dim_k(A')$. It is enough to show that if Z is any closed subset of W, then

$$\dim\,(\pi^W_{A'})^{-1}(Z) \leq \dim\,(\pi^W_A)^{-1}(Z) - (\ell - \ell')n. \tag{4.3}$$

It follows from Proposition 2.1 that

$$J_A(W) \hookrightarrow (\pi^{\mathbf{A}^d}_{A'/A})^{-1}(J_{A'}(W)) \tag{4.4}$$

is cut out by $r(\ell - \ell')$ equations and the same holds if we restrict to the inverse images of Z. By putting these together, we obtain the inequality (4.3) if we can show that for every irreducible component R of $(\pi^W_{A'})^{-1}(Z)$, its inverse image $(\pi^{\mathbf{A}^d}_{A'/A})^{-1}(R)$ intersects $(\pi^W_A)^{-1}(Z)$. This is a consequence of the fact that if $x \in Z$, then $s^X_A(x)$ lies in this intersection by Remark 2.10. □

5. Some questions on generalized jet schemes

In this section we collect some questions and remarks concerning the behavior of the schemes $J_A(X)$, when $\mathrm{embdim}(A) \geq 2$. Very little is known in this context, partly due to a lack of examples. In what follows we work over an algebraically closed field k of characteristic zero.

Motivated by Theorem 4.3, we begin by proposing several invariants that measure the rate of growth of the dimensions of certain schemes $J_A(X)$. In order to simplify the notation, we restrict to the first unknown case, that when $\mathrm{embdim}(A) = 2$. We introduce three invariants, depending on the choice of algebras A.

We first consider the algebras $A_{p,q} = k[s,t]/(s^p, t^q)$, with $p, q \geq 1$. Note that $\dim_k(A) = pq$. Given a scheme X, let

$$\alpha_{p,q}(X) := \dim J_{A_{p,q}}(X)$$

and

$$\alpha(X) := \sup_{p,q\geq 1} \frac{\alpha_{p,q}(X)}{pq}.$$

Note that $\alpha(X) \leq \max_{x\in X} \dim(T_x X)$ by Corollary 2.7. Since $J_{A_{p,q}}(X) \simeq J_{p-1}(J_{q-1}(X))$, it follows from Remark 4.9 that for every positive integer m, we have

$$\frac{\alpha_{p,q}(X)}{pq} \leq \frac{\alpha_{mp,q}(X)}{mpq} \text{ and } \frac{\alpha_{p,q}(X)}{pq} \leq \frac{\alpha_{p,mq}(X)}{mpq}. \tag{5.1}$$

This clearly implies

$$\alpha(X) = \sup_{p\geq 1} \frac{\alpha_{p,p}(X)}{p^2} = \limsup_{p\to\infty} \frac{\alpha_{p,p}(X)}{p^2}. \tag{5.2}$$

On the other hand, it follows from Theorem 4.3 that if X is a closed subscheme of the smooth variety Y, then for every $q \geq 1$, we have

$$\lim_{p\to\infty} \frac{\alpha_{p,q}(X)}{pq} = \sup_{p\geq 1} \frac{\alpha_{p,q}(X)}{pq} = \dim(Y) - \frac{\mathrm{lct}(J_{q-1}(Y), J_{q-1}(X))}{q}. \tag{5.3}$$

It is easy to deduce from (5.1) and (5.3) the following proposition.

Proposition 5.1. *If X is a closed subscheme of the smooth variety Y, then*

$$\alpha(X) = \dim(Y) - \inf_{q \geq 1} \frac{\mathrm{lct}(J_{q-1}(Y), J_{q-1}(X))}{q}$$
$$= \dim(Y) - \liminf_{q \to \infty} \frac{\mathrm{lct}(J_{q-1}(Y), J_{q-1}(X))}{q}.$$

We now turn to another invariant, corresponding to a different sequence of algebras. For every $m \geq 1$, let $A_m = k[s,t]/(s,t)^m$. Note that $\dim_k(A_m) = \frac{m(m+1)}{2}$. For a scheme X over k, we put $\beta_m := \dim(J_{A_m}(X))$ and

$$\beta(X) := \sup_{m \geq 1} \frac{\beta_m(X)}{m(m+1)/2}. \tag{5.4}$$

It follows from Corollary 2.7 that $\beta(X) \leq \max_{x \in X} \dim(T_x X)$.

Question 5.2. Is the supremum in (5.4) also the limsup of the corresponding sequence? Is there any relation between $\alpha(X)$ and $\beta(X)$?

Example 5.3. The invariant $\beta(X)$ is slightly easier to compute than $\alpha(X)$. For example, suppose that X is defined in $\mathbf{A}^n = \operatorname{Spec} k[x_1, \ldots, x_n]$ by $x_1^{a_1} \ldots x_n^{a_n}$, for nonnegative integers $a_1, \ldots, a_n$, not all of them equal to 0. An element of $J_{A_m}(X)$ corresponds to a k-algebra homomorphism $\phi\colon k[x_1, \ldots, x_n] \to k[s,t]/(s,t)^m$, which is determined by $\phi(x_1), \ldots, \phi(x_n)$, such that $\prod_i \phi(x_i)^{a_i} = 0$. If we denote by ν_i the smallest power of (s,t) that contains $\phi(x_i)$, we see that we get a disjoint decomposition of $J_{A_m}(X)$ into locally closed subsets $J_{A_m}(X)_\nu$, parametrized by $\nu = (\nu_1, \ldots, \nu_n) \in \{0, 1, \ldots, m\}^n$ such that $\sum_{i=1}^n a_i \nu_i \geq m$. It is straightforward to see that

$$\dim(J_{A_m}(X)_\nu) = \frac{nm(m+1)}{2} - \sum_{i=1}^n \frac{\nu_i(\nu_i+1)}{2}$$

and an easy computation shows that since $\sum_{i=1}^n a_i \nu_i \geq m$, we have

$$\sum_{i=1}^n \nu_i(\nu_i+1) - \frac{m(m+1)}{\sum_{i=1}^n a_i^2} \geq 0. \tag{5.5}$$

Moreover, if we take $\nu_i = a_i \ell$ for some $\ell \geq 1$ and $m = \ell \cdot \sum_i a_i^2$, then the left-hand side of (5.5) is equal to $\ell\,(\sum_{i=1}^n a_i - 1)$. This implies that

$$\beta(X) = \limsup_{m \to \infty} \frac{\beta_m(X)}{m(m+1)/2} = n - \frac{1}{\sum_{i=1}^n a_i^2}.$$

Yet another invariant of a similar flavor is the following. If X is a scheme over k, then let

$$\gamma(X) := \sup_A \frac{\dim(J_A(X))}{\dim_k(A)},$$

where the supremum is over all algebras $A \in \mathrm{LFA}/k$ which are graded[4] and such that $\mathrm{embdim}(A) \leq 2$. It is clear from definition that $\alpha(X), \beta(X) \leq \gamma(X)$, while Corollary 2.7 implies that $\gamma(X) \leq \max_{x \in X} \dim(T_x X)$.

We now turn to a different question, motivated by Theorem 4.13 and inspired by [16]. Suppose that X is a locally complete intersection variety and r is a positive integer. We say that X has irreducible (resp. pure-dimensional) r-iterated jet schemes if the jet scheme $J_{m_1}(J_{m_2}(\ldots J_{m_r}(X)))$ is irreducible (resp. pure-dimensional) for all nonnegative integers $m_1, \ldots, m_r$.

Proposition 5.4. *For a locally complete intersection variety X, the following are equivalent:*

i) *X has irreducible (resp. pure-dimensional) r-iterated jet schemes.*

ii) *$J_A(X)$ is irreducible (resp. pure-dimensional) for every $A \in \mathrm{LFA}/k$ which is graded and such that $\mathrm{embdim}(A) \leq r$.*

iii) *$J_{A_m}(X)$ is irreducible (resp. pure-dimensional) for every $m \geq 1$, where $A_m = k[x_1, \ldots, x_r]/(x_1, \ldots, x_r)^m$.*

Furthermore, X has irreducible r-iterated jet schemes if and only if all $(r-1)$-iterated jet schemes of X have rational singularities. If this is the case, then all r-iterated jet schemes of X are reduced locally complete intersections.

The equivalence of i)-iii) is an easy consequence of Proposition 4.14, which together with Theorem 4.13 also gives the last assertions in the proposition. When $r \geq 2$, it is not easy to give examples of singular varieties which have all r-iterated jet schemes pure-dimensional. We discuss below the case of cones over smooth projective hypersurfaces, which provides the only nontrivial class of examples for $r = 2$.

Remark 5.5. Note that if X is a locally complete intersection variety, then it follows from Proposition 5.4 that X has pure-dimensional 2-iterated jet schemes if and only if $\alpha(X) = \dim(X)$, which is also equivalent to saying that either $\beta(X) = \dim(X)$ or that $\gamma(X) = \dim(X)$.

Remark 5.6. We note that Example 2.8 implies that if $x \in X$ is a singular point such that

$$r > \frac{\dim(X)}{\dim(T_x X) - \dim(X)},$$

then X does not have pure-dimensional r-iterated jet schemes.

Suppose now that $X \subseteq \mathbf{A}^n$, with $n \geq 3$, is defined by a homogeneous polynomial f of degree $d > 0$. We assume that $X \smallsetminus \{0\}$ is smooth, which is the case for general f. It is well-known that X has rational (log canonical) singularities if and only if $d < n$ (resp., $d \leq n$). Using Theorem 4.13, we conclude that X has irreducible (pure-dimensional) 1-iterated jet schemes if and only if $d < n$ (resp., $d \leq n$). We give a direct argument for this, in the spirit of [5, §3].

Example 5.7. We show that if $X = V(f) \subset \mathbf{A}^n$, with f homogeneous of degree $d > 0$, such that X has an isolated singularity at 0, then $J_m(X)$ is irreducible (pure-dimensional)

[4]Of course, it might make sense to remove the condition that A is graded. We do not know whether this would give a different invariant.

for all $m \geq 1$ if and only if $d < n$ (resp., $d \leq n$). Let us denote by $\pi_m \colon J_m(X) \to X$ and $\pi'_m \colon J_m(\mathbf{A}^n) \to \mathbf{A}^n$ the canonical projections. It follows from Proposition 4.14 that we need to show the following: we have $\dim(\pi_m^{-1}(0)) < (m+1)(n-1)$ (resp., $\dim(\pi_m^{-1}(0)) \leq (m+1)(n-1)$) for all $m \geq 1$ if and only if $d < n$ (resp., $d \leq n$). We use the following two facts:

i) If $m \leq (d-1)$, then the closed embedding $J_m(X) \hookrightarrow J_m(\mathbf{A}^n)$ induces an isomorphism $\pi_m^{-1}(0) \simeq {\pi'_m}^{-1}(0) \simeq \mathbf{A}^{mn}$.

ii) If $m \geq d$, then we have an isomorphism $\pi_m^{-1}(0) \simeq J_{m-d}(X) \times \mathbf{A}^{n(d-1)}$.

Both assertions follow from the universal property defining $J_m(X)$, together with the following observations: if R is a k-algebra and $u_1, \ldots, u_n \in tR[t]/(t^{m+1})$, then $f(u_1, \ldots, u_n) = 0$ whenever $m \leq d-1$; for $m \geq d$, we have $f(u_1, \ldots, u_n) = 0$ if and only if when we write $u_i = tv_i$, we have $f(\overline{v_1}, \ldots, \overline{v_n}) = 0$ in $R[t]/(t^{m+1-d})$, where $\overline{v_i}$ is the class of v_i in $R[t]/(t^{m+1-d})$.

In particular, it follows from i) that $\dim(\pi_{d-1}^{-1}(0)) = (d-1)n$. If $\dim(\pi_{d-1}^{-1}(0)) < d(n-1)$, we deduce that $d < n$. Conversely, suppose that $d < n$. We prove by induction on m that $\dim(\pi_m^{-1}(0)) < (m+1)(n-1)$. If $0 \leq m \leq d-1$, this follows easily from i). On the other hand, if $m \geq d$, then the assertion follows from ii) and the inductive hypothesis. One similarly shows that $d \leq n$ if and only if $\dim(\pi_m^{-1}(0)) \leq (m+1)(n-1)$ for all m.

With the above notation, we are interested in when the r-iterated jet schemes of X are irreducible or pure-dimensional for $r \geq 2$. It is easy to give a necessary condition, arguing as in Example 5.7.

Proposition 5.8. *Let $X \subset \mathbf{A}^n$ be defined by a homogeneous polynomial f of degree $d > 0$. If $r \geq 2$ and the r-iterated jet schemes of X are pure-dimensional, then $d^r \leq n$.*

Proof. For every positive integer j, we consider $A_j = k[t_1, \ldots, t_r]/(t_1, \ldots, t_r)^{jd}$. Note that for every k-algebra R and every $u_1, \ldots, u_n \in (t_1, \ldots, t_r)^j/(t_1, \ldots, t_r)^{jd} \subset R \otimes_k A_j$, we have $f(u_1, \ldots, u_n) = 0$. This shows that $J_{A_j}(X)$ contains a closed subscheme Z_j, with

$$\dim(Z_j) = n \cdot \dim_k (t_1, \ldots, t_r)^j A_j = n \left(\binom{jd+r-1}{r} - \binom{j+r-1}{r} \right).$$

Using Proposition 4.14 and the fact that $\dim_k(A_j) = \binom{jd+r-1}{r}$, we deduce from our assumption that

$$n \left(\binom{jd+r-1}{r} - \binom{j+r-1}{r} \right) \leq (n-1) \cdot \binom{jd+r-1}{r}.$$

This gives

$$\binom{jd+r-1}{r} \leq n \cdot \binom{j+r-1}{r}, \tag{5.6}$$

which we can rewrite as

$$n \geq d \cdot \prod_{i=1}^{r-1} \frac{jd+i}{j+i}. \tag{5.7}$$

Since each function $\phi_i(x) = \frac{dx+i}{x+i}$, with $1 \leq i \leq r-1$, is increasing, with $\lim_{x \to \infty} \phi_i(x) = d$, we conclude from (5.7) by letting j go to infinity that $n \geq d^r$. □

Remark 5.9. If in the above proposition we assume instead that the r-iterated jet schemes of X are irreducible, then in (5.7) we get strict inequality. However, this does not imply that $n > d^r$.

Question 5.10. Does the converse to the assertion in Proposition 5.8 hold when X has an isolated singularity? More precisely, suppose that X is defined in $\mathbf{A}^n$ by a homogeneous polynomial f of degree d, such that $X \smallsetminus \{0\}$ is smooth. If $r \geq 2$ is such that $n \geq d^r$, are all r-iterated jet schemes of X pure-dimensional? Are they irreducible? Do these assertions hold if f is general?

Remark 5.11. The only evidence for a positive answer to Question 5.10 is provided by the result from [16], saying that if $f = \sum_u a_u x^u \in k[x_1, \ldots, x_n]$, where the sum is over all $u = (u_1, \ldots, u_n) \in \mathbf{Z}^n_{\geq 0}$ with $\sum_i u_i = d$, then if $d^2 \leq n$ and the coefficients $(a_u)_u$ are algebraically independent over $\mathbf{Q}$, then the 2-iterated jet schemes of $X = V(f)$ are irreducible. The proof shows that all $J_m(X)$ have rational singularities by reducing to positive characteristic and using the theory of F-singularities. In particular, this shows that if the ground field k is uncountable, then for a very general polynomial, the 2-iterated jet schemes of $V(f)$ are irreducible.

Remark 5.12. It is interesting that the bound in Question 5.10 is (almost) the same as the bound that shows up in Lang's C_r condition on fields. Recall that if $r \geq 0$, then a field K satisfies the condition C_r if every homogeneous polynomial $f \in K[x_1, \ldots, x_n]$ of degree d, with $d^r < n$, has a nontrivial zero in K^n. For example, if k is algebraically closed, then it is known that the field $K = k(x_1, \ldots, x_r)$ satisfies condition C_r. It would be interesting if there was a connection between C_r fields and Question 5.10.

Acknowledgments. The author was partially supported by NSF grants DMS-1068190 and DMS-1265256. I am indebted to David Eisenbud and Edward Frenkel who introduced me to jet schemes and shared with me their conjecture on the irreducibility of jet schemes of locally complete intersection varieties. I am also grateful to Lawrence Ein, from whom I learned a great deal over the years, during our collaboration. In particular, several of the results discussed in this article have been obtained in joint work with him.

References

[1] Florin Ambro, *On minimal log discrepancies*, Math. Res. Lett. **6** (1999), 573–580.

[2] Michael F. Atiyah, *Resolution of singularities and division of distributions*, Comm. Pure Appl. Math. **23** (1970), 145–150.

[3] Caucher Birkar, *Ascending chain condition for log canonical thresholds and termination of log flips*, Duke Math. J. **136** (2007), 173–180.

[4] Tommaso de Fernex, *Lawrence Ein, and Shihoko Ishii, Divisorial valuations via arcs*, Publ. Res. Inst. Math. Sci. **44** (2008), 425–448.

[5] T. de Fernex, L. Ein, and M. Mustaţă, *Bounds for log canonical thresholds with applications to birational rigidity*, Math. Res. Lett. **10** (2003), 219–236.

[6] Tommaso de Fernex, Lawrence Ein, and Mircea Mustaţă, *Shokurov's ACC conjecture for log canonical thresholds on smooth varieties*, Duke Math. J. **152** (2010), 93–114.

[7] Jan Denef and François Loeser, *Motivic Igusa zeta functions*, J. Algebraic Geom. **7** (1998), 505–537.

[8] ———, *Germs of arcs on singular algebraic varieties and motivic integration*, Invent. Math. **135** (1999), 201–232.

[9] Roi Docampo, *Arcs on determinantal varieties*, Trans. Amer. Math. Soc. **365** (2013), 2241–2269.

[10] Lawrence Ein, Robert Lazarsfeld, and Mircea Mustaţă, *Contact loci in arc spaces*, Compos. Math. **140** (2004), 1229–1244.

[11] Lawrence Ein and Mircea Mustaţă, *Inversion of adjunction for local complete intersection varieties*, Amer. J. Math. **126** (2004), 1355–1365.

[12] ———, *Jet schemes and singularities*, in Algebraic geometry–Seattle 2005, Part 2, 505–546, Proc. Sympos. Pure Math., 80, Part 2, Amer. Math. Soc., Providence, RI, 2009.

[13] Lawrence Ein, Mircea Mustaţă, and Takehiko Yasuda, *Jet schemes, log discrepancies and inversion of adjunction*, Invent. Math. **153** (2003), 519–535.

[14] Christopher Hacon, James M$^{\text{c}}$Kernan, and Chenyang Xu, *ACC for log canonical thresholds*, Ann of Math., to appear.

[15] Shihoko Ishii, *The arc space of a toric variety*, J. Algebra **278** (2004), 666–683.

[16] Shihoko Ishii, Akiyoshi Sannai, and Kei-ichi Watanabe, *Jet schemes of homogeneous hypersurfaces*, in Singularities in geometry and topology, 39–49, IRMA Lect. Math. Theor. Phys., 20, Eur. Math. Soc., Zürich, 2012.

[17] Masayuki Kawakita, *Ideal-adic semi-continuity of minimal log discrepancies on surfaces*, Michigan Math. J. **62** (2013), 443–447.

[18] ———, *Ideal-adic semi-continuity problem for minimal log discrepancies*, Math. Ann. **356** (2013), 1359–1377.

[19] János Kollár, *Singularities of pairs*, Algebraic geometry Santa Cruz 1995, 221–287, Proc. Sympos. Pure Math., 62, Part 1, Amer. Math. Soc., Providence, RI, 1997.

[20] Maxim Kontsevich, *Lecture at Orsay* (December 7, 1995).

[21] Hussein Mourtada, *Jet schemes of toric surfaces*, C. R. Math. Acad. Sci. Paris **349** (2011), 563–566.

[22] Mircea Mustaţă, *Jet schemes of locally complete intersection canonical singularities*, With an appendix by David Eisenbud and Edward Frenkel, Invent. Math. **145** (2001), 397–424.

[23] ———, *Singularities of pairs via jet schemes*, J. Amer. Math. Soc. **15** (2002), 599–615.

[24] _____, *IMPANGA lecture notes on log canonical thresholds, Notes by Tomasz Szemberg*, in Contributions to algebraic geometry, 407–442, EMS Ser. Congr. Rep., Eur. Math. Soc., Zürich, 2012.

[25] Vyacheslav V. Shokurov, *Three-dimensional log perestroikas. With an appendix*, in English by Yujiro Kawamata, Izv. Ross. Akad. Nauk Ser. Mat. **56** (1992), 105–203, translation in Russian Acad. Sci. Izv. Math. **40** (1993), 95–202.

[26] _____, *Letters of a bi-rationalist. VII. Ordered termination*, Tr. Mat. Inst. Steklova **264** (2009), Mnogomernaya Algebraicheskaya Geometriya, 184–208; translation in Proc. Steklov Inst. Math. **264** (2009), 178–200.

[27] Alexander N. Varchenko, *The complex singularity index does not change along the stratum μ=const* (Russian), Funktsional. Anal. i Prilozhen. **16** (1982), 1–12.

[28] _____, *Asymptotic Hodge structures in the vanishing cohomology*, Math. USSR Izv. **18** (1982), 469–512.

[29] _____, *Semicontinuity of the complex singularity exponent* (Russian), Funksional. Anal. i Prilozhen. **17** (1983), 77–78.

[30] Zhixian Zhu, *Log canonical thresholds in positive characteristic*, preprint, arXiv:1308.5445.

Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA
E-mail: mmustata@umich.edu

Some aspects of explicit birational geometry inspired by complex dynamics

Keiji Oguiso

Abstract. Our aim is to illustrate how one can effectively apply the basic ideas and notions of topological entropy and dynamical degrees, together with recent progress of minimal model theory in higher dimension, for an explicit study of birational or biregular selfmaps of projective or compact Kähler manifolds, through concrete examples.

Mathematics Subject Classification (2010). 14E05, 14E09, 37F99.

Keywords. Entropy, dynamical degree, primitive automorphism, rational manifolds, Calabi-Yau manifolds.

1. Introduction

This is a survey of some aspects of recent progress on birational and biregular complex algebraic geometry inspired by complex dynamics in several variables. Our aim is to illustrate how one can apply the basic ideas and notions of topological entropy and dynamical degrees, together with recent progress of minimal model theory in higher dimension, for an explicit study of birational or biregular selfmaps of projective or compact Kähler manifolds. Especially, we focus on the following one of the most basic, natural problems:

Problem 1.1. Find many examples of projective or Kähler manifolds M admitting *interesting* birational automorphisms of *infinite order*, or more preferably, *primitive biregular* automorphisms of *positive entropy*.

There are so many interesting works in this area since the breakthrough results due to Serge Cantat [19] and Curtis T. McMullen [58], and this note is definitely far from being a complete panorama of this area. Also, needless to say, there is no universally acceptable mathematical definition of the term *interesting* and the choice of topics and materials owes much to my own flavour and ability, and probably not the one that everyone agrees with. For instance, the terms *of infinite order* ignore very beautiful aspects of finite group actions on manifolds.

Throughout this note, we work over the complex number field $\mathbf{C}$. We assume some familiarity with basics on complex geometry and algebraic geometry. Unless stated otherwise, the topology we use is the Euclidean topology (not Zariski topology), a point means a closed point, and manifolds and varieties are connected. By abuse of language, we call a (bi)meromorphic map also a (bi)rational map even under non-algebraic settings. We denote

Proceedings of the International Congress of Mathematicians, Seoul, 2014

by $I(f)$ the indeterminacy locus of a rational map $f : M \cdots \to N$, i.e., the complement of the maximal, necessarily Zariski open dense, subset $U \subset M$ such that $f|U$ is holomorphic. $I(f)$ is a Zariski closed subset of M of codimension ≥ 2 if M is normal. The set of birational selfmaps (resp. biregular selfmaps, resp. birational selfmaps being isomorphic in codimension one) of M form a group under the natural composition. We denote these groups by $\mathrm{Bir}\,(M)$ (resp. $\mathrm{Aut}\,(M)$, resp. $\mathrm{PsAut}\,(M)$). Here we call f isomorphic in codimension one if f neither contracts nor extracts any divisors, in other words, if there are Zariski open dense subsets U, U' of M such that $\mathrm{codim}\, M \setminus U \geq 2$, $\mathrm{codim}\, M \setminus U' \geq 2$ for which the restriction map $f|U : U \to U'$ is an isomorphism. We call an element of $\mathrm{Bir}\,(M)$ (resp. $\mathrm{PsAut}\,(M)$, resp. $\mathrm{Aut}\,(M)$) a birational automorphism (resp. a pseudo-automorphism, resp. an automorphism or a biregular automorphism) of M. We call a compact complex variety of class $\mathcal{C}$ if it is birational to a compact Kähler manifold. Unless stated otherwise, we denote the golden number $(\sqrt{5}+1)/2$ by η, the cyclic group of order n by $\mathbf{Z}_n$, the free product of groups G_1 and G_2 by $G_1 * G_2$.

2. What kind of manifolds we are interested in?

Let M be a compact Kähler manifold of $\dim M = l > 0$. We are interested in a *birational automorphism* f of M, in particular, *of infinite order*.

2.1. Primitive birational automorphisms after De-Qi Zhang. If $f_i \in \mathrm{Bir}\,(M_i)$ $(i = 1, 2)$, then $f_1 \times f_2 \in \mathrm{Bir}\,(M_1 \times M_2)$ and it is of infinite order if so is one of f_i. We are more interested in birational automorphisms *not coming from lower dimensional pieces*, more precisely, *primitive* ones in the sense of De-Qi Zhang [90]:

Definition 2.1. A birational automorphism f of M is *imprimitive* if there are a dominant rational map $\varphi : M \cdots \to B$ to a compact complex variety B with $0 < \dim B < \dim M = l$ and a rational map $g : B \cdots \to B$, necessarily a birational automorphism of B, such that $\varphi \circ f = g \circ \varphi$. A birational automorphism that is not imprimitive is *primitive*.

Here we may assume that B is smooth and $\varphi : M \to B$ is holomorphic whenever it is more convenient. Indeed, we may resolve B first and then resolve the indeternimacy of φ, by the fundamental result of Hironaka.

> *What kind of manifolds can have primitive birational automorphisms of infinite order?*

We can not answer this question completely, but if we *assume* that minimal model program [43, 49, 52] in higher dimensional projective manifolds work, then one has the following rough but quite nice picture at least in the projective case. This beautiful observation is due to De-Qi Zhang (see [90] and [65] for more precise results):

Theorem 2.2. *Let M be a projective manifold of dimension l. Assume that the minimal model conjecture (MMP) and the weak abundance (WA) in dimension l hold, in the sense that any l-dimensional projective manifold V is birational to either a minimal model with weak abundance or a Mori fiber space, i.e., there is a projective variety $V_{\min}$, birational to V, with only normal, $\mathbf{Q}$-factorial terminal singularities such that either one of the following two holds:*

(MMP-1) *in addition, $K_{V_{\min}}$ is nef (minimal model); or*

(MMP-2) *in addition, there exists an extremal contraction $\varphi : V_{\min} \to B$, with respect to $K_{V_{\min}}$ onto a normal projective variety such that $0 \leq \dim B < \dim V$ (Mori fiber space),*

and additionally,

(WA) *In the case* (MMP-1)*, $|mK_{V_{\min}}|$ is non-empty for some $m > 0$.*

Then any l-dimensional projective manifold M with a primitive birational automorphism f of infinite order is birational to either:

(RC) *a rationally connected manifold, in the sense that any two points can be connected by a finite chain of rational curves;*

(WCY) *a minimal Calabi-Yau variety, in the sense that it is a minimal variety with numerically trivial canonical divisor and of irregularity 0; or*

(T) *an abelian variety, i.e., a projective complex torus.*

Example 2.3.

(1) When $l = 1$, (RC) is $\mathbf{P}^1$, (T) is an elliptic curve and no (WCY).

(2) When $l = 2$, (RC) is a rational surface, (T) is an abelian surface, and (WCY) is a K3 surfaces or an Enriques surface (see eg. [2]).

Since the proof of Theorem 2.2 provides a good introduction on objects we are interested in, we give here a fairly complete proof, following [86, 90].

Proof. The most essential point is that any *canonically defined maps* are preserved by $\mathrm{Bir}(M)$.

Let $\kappa(M) \in \{-\infty, 0, 1, \cdots, l\}$ be the Kodaira dimension. $\kappa(M)$ is the maximal dimension of the images W_m under the pluri-canonical maps associated to the complete linear system $|mK_M|$ $(m = 1, 2, 3, \cdots)$:

$$\Phi_m := \Phi_{|mK_M|} : M \cdots \to W_m := \mathrm{Im}\, \Phi_{|mK_M|} \subset |mK_M|^* = \mathbf{P}^{\dim\, |mK_M|}\ ,$$

if $|mK_M| \neq \emptyset$ for some $m > 0$ and $\kappa(M) = -\infty$ otherwise. It is a birational invariant.

Consider first the case where $\kappa(M) \geq 0$. We may assume that Φ_m is regular. We can and do choose m so that $\dim W_m = \kappa(M)$ and Φ_m is of connected fibers. We write $W = W_m$. Note that any birational automorphism f preserves the set of global holomorphic pluri-canonical forms, as $\mathrm{codim}\, I(f) \geq 2$. Then the induced projective linear map $f_* \in \mathrm{PGL}(|mK_M|^*) = \mathrm{Aut}(|mK_M|^*)$ preserves W and is equivariant to f with respect to Φ_m. Hence f is imprimitive if $1 \leq \kappa(M) \leq l - 1$.

If $\kappa(M) = l$, then the same is true but M and W are birational. So

$$\mathrm{Bir}(M) = \mathrm{Aut}(W) \subset \mathrm{PGL}(|mK_M|^*).$$

It is Zariski closed in the affine noetherian group $\mathrm{PGL}(|mK_M|^*)$, as it is the stabilizer of the point $[W]$ of the action of $\mathrm{PGL}(|mK_M|^*)$ on $\mathrm{Hilb}(|mK_M|^*)$. Hence it is finite. Indeed, if otherwise, $\dim \mathrm{Aut}(W) \geq 1$ and we can choose a one dimensional algebraic subgroup, which is necessarily isomorphic to $\mathbf{C}$ or $\mathbf{C}^\times$. The Zariski closures of the orbits of general points of W under this 1-dimensional algebraic subgroup are necessarily rational, and cover W and M. Then again by the Hilbert scheme, we have a dominant holomorphic

maps $\pi : Y \to M$ from the fiber space $Y \to X$ whose general fibers Y_x are isomorphic to $\mathbf{P}^1$. Since we work over the field of characteristic 0, the map π is separable so that $|mK_Y| \neq \emptyset$ as well. Hence for general Y_x, we have $\deg mK_{Y_x} \geq 0$ by the adjunction formula, a contradiction to the fact that $Y_x \simeq \mathbf{P}^1$. Hence $\mathrm{Bir}\,(M)$ is finite and in particular, f is of finite order when $\kappa(M) = l$.

Hence $\kappa(M) = 0$ or $-\infty$. So far we did not use the assumption (MMP), (WA). Also, our assumption that f is of infinite order is used only to conclude $\kappa(M) \neq l$.

Assume that $\kappa(M) = 0$. Consider next the irregularity $q(M) := h^0(M, \Omega^1_M)$. If $q(M) > 0$, then we have the albanese morphism

$$\mathrm{alb}_M : M \to \mathrm{Alb}\,(M) = H^0(M, \Omega^1_M)^*/H_1(M, \mathbf{Z})\ .$$

It is classical that $\mathrm{Alb}\,(M)$ is an abelian variety. Since $\kappa(M) = 0$, a fundamental theorem due to Kawamata [46] (again free from (MMP), (WA)) says that alb_M is surjective with connected fibers, in particular $q(M) \leq l$. For the same reason as before, the action $\mathrm{Bir}\,(M)$ descends to the biregular action of $\mathrm{Alb}\,(M)$ equivariantly with respect to the albanese map. Hence either $q(M) = 0$ or $q(M) = l$. In the second case, M is birational to $\mathrm{Alb}\,(M)$. If $q(M) = 0$, then by our assumption (MMP), M is birational to a minimal Calabi-Yau variety. (Here, if one prefers, one also stops at the stage $\kappa(M) = 0$ and $q(M) = 0$. Then the conjectural (MMP) is not required here.)

It remains to treat the case $\kappa(M) = -\infty$. This is the most subtle case where we really use our assumptions (MMP) and (WA) (but we do not use our assumption that f is of infinite order anymore). (WA) is one way to conclude (MMP-1) and (MMP-2) are exclusive. Let us consider $W := M_{\min}$ in Theorem 2.2, whose existence needs (MMP). If K_W would be nef, then by (WA), $|mK_M| = |mK_W| \neq \emptyset$ for some $m > 0$, a contradiction to $\kappa(M) = -\infty$. Hence, the case (MMP-1) does not happen and therefore (MMP-2) happens by our assumptions (MMP). In (MMP-2), by the property of an extremal contraction, the fibers of the Mori fiber space are covered by rational curves. Now we consider the maximal rationally connected fibration, MRC fibration, for short [51]. The MRC fibration $r : M \cdots \to R$ is an almost holomorphic, rational dominant map such that for general $x \in M$, the fiber $M_p \ni x$ is the maximal rationally connected submanifold of M containing x, and birationally preserved by $\mathrm{Bir}\,(M)$, hence by f. Since $0 \leq \dim\, R \leq l - 1$ for our M, it follows that R is a point, i.e., M is rationally connected. □

(MMP) and (WA) hold in dimension ≤ 3, finally due to Mori [62] (MMP) and Kawamata [47] (WA) in the strongest form that nef $K_{V_{\min}}$ is actually semi-ample. So, Theorem 2.2 is unconditional in dimension ≤ 3. In higher dimension, both conjectures are expected to be true (cf. [13, 43, 64]). In dimension 2, Theorem 2.2 is essentially the same as the breakthough observation due to Cantat [19] (see also [22]), in terms of *topological entropy*.

2.2. Three classes of manifolds in Theorem 2.2. In this subsection, we discuss basic manifolds belonging to the three classes in Theorem 2.2, which are indeed the main objects in this note.

Rational manifolds. An excellent reference of rationally connected manifolds (RC manifolds) is [51]. Most basic examples of RC manifolds are rational manifolds, i.e., manifolds which are birational to $\mathbf{P}^l$. Note that $\mathrm{Aut}\,(\mathbf{P}^l) = \mathrm{PGL}(l+1, \mathbf{C})$. It is obvious that generic $g \in \mathrm{Aut}\,(\mathbf{P}^l)$ is of infinite order. On the other hand, any $g \in \mathrm{Aut}\,(\mathbf{P}^l)$ is imprimitive

if $l \geq 2$. Indeed, g has a fixed point $P \in \mathbf{P}^l$, corresponding to the eigenvector of a lift $\tilde{g} \in \mathrm{GL}(l+1, \mathbf{C})$. The family of lines through P is then stable under g. Let $\mathrm{Bl}_P \mathbf{P}^l$ be the blow up of $\mathbf{P}^l$ at P. The lines through P are the fibers of the natural morphism $\mathrm{Bl}_P \mathbf{P}^l \to \mathbf{P}^{l-1} = \mathbf{P}(T_{P,\mathbf{P}^l})$, and this fibration is stable under the natural (biregular) action of g.

So, *in our view*, $\mathrm{Aut}\,(\mathbf{P}^l) = \mathrm{PGL}\,(l+1, \mathbf{C})$ is not so interesting. However, the group $\mathrm{PGL}\,(l+1, \mathbf{C})$ has a very deep aspect in birational geometry, for instance, the following striking result due to Cantat [21]:

Theorem 2.4. *Let M be an l-dimensional projective manifold. If there is an injective group homomorphism $\mathrm{PGL}\,(n+1, \mathbf{C}) \to \mathrm{Bir}\,(M)$, as abstract groups, then $n \leq l$ and the equality $n = l$ holds if and only if M is rational. In particular, $\mathrm{Bir}\,(\mathbf{P}^l) \simeq \mathrm{Bir}\,(\mathbf{P}^{l'})$ as abstract groups if and only if $l = l'$.*

The *standard Cremona transformation*

$$\mathrm{cr}_l : \mathbf{P}^l \cdots \to \mathbf{P}^l \ , \ [x_0 : x_1 : \cdots : x_l] \mapsto \left[\frac{1}{x_0} : \frac{1}{x_1} : \cdots : \frac{1}{x_l}\right]$$

is the most basic birational non-biregular automorphism of $\mathbf{P}^l$ ($l \geq 2$). The indeterminacy locus of cr_l are $\cup_{i \neq j} L_{ij}$, where $L_{ij} := (x_i = x_j = 0)$. Let $\mathrm{SCr}_l := \langle \mathrm{PGL}(l+1, \mathbf{C}), \mathrm{cr}_l \rangle < \mathrm{Cr}_l := \mathrm{Bir}\,(\mathbf{P}^l)$. We have $\mathrm{SCr}_2 = \mathrm{Cr}_2$ (Noether's theorem, [35]). If $l \geq 3$, then SCr_l is much smaller than Cr_l but SCr_l is rich enough. One of unexpected applications of SCr_l is the following result due to Lesieutre [55]. In [55], the group SCr_3 and its complex dynamical aspect are effectively applied to prove the following derived categorical result:

Theorem 2.5. *There is a smooth rational threefold with infinitely many birational non-isomorphic Fourier-Mukai partners.*

Note that cr_l maps the coordinate hyperplane $H_i = (x_i = 0)$ to the standard coordinate point $e_i = [0 : \cdots : 0 : 1 : 0 : \cdots 0]$, where 1 is at the ith coordinate. So, $\mathrm{cr}_l \in \mathrm{Bir}\,(\mathbf{P}^l) \backslash \mathrm{PsAut}\,(\mathbf{P}^l)$. Actually $\mathrm{PsAut}\,(\mathbf{P}^l) = \mathrm{Aut}\,(\mathbf{P}^l)$ by $\mathrm{Pic}\,\mathbf{P}^l = \mathbf{Z}H$. However, if we blow-up $\mathbf{P}^l$ at the $(l+1)$ standard coordinate points e_i, then cr_l gives rise to a pseudo-automorphism $\tilde{\mathrm{cr}}_i$ of $\mathrm{Bl}_{\{e_i\}}\mathbf{P}^l$, and performing further blowing-ups, we can make it a biregular automorphism, of order 2. Let $S, T \in \mathrm{PGL}(l+1, \mathbf{C})$. Then $f = S \circ \mathrm{cr}_l \circ T^{-1} \in \mathrm{SCr}_l$ is of infinite order for almost all choices of S and T, and f lifts to a *pseudo-automorphism* of some blowing-ups of $\mathbf{P}^l$, under some periodicity condition for the indeterminacy loci $I(f^{\pm n})$ [6, 7, 9]. In this way, Bedford and Kim construct many interesting rational surface automorphisms as well as pseudo-automorphisms of rational threefolds very explicitly. However, when $l = 3$, none of them seems to be realized as a *biregular* automorphism (cf. Question 5.6).

A few properties of pseudo-automorphisms. Before entering two other classes, we recall a few basic properties of $\mathrm{PsAut}\,(M)$. The group $\mathrm{PsAut}\,(M)$ naturally acts on the Néron-Severi group $\mathrm{NS}\,(M) := \mathrm{Im}(c_1 : \mathrm{Pic}\,(M) \to H^2(M, \mathbf{Z}))$ as well as on $H^2(M, \mathbf{Z})$. This action is functorial, in the sense that $(f \circ g)^* = g^* \circ f^*$ on $H^2(M, \mathbf{Z})$ and preserves the Hodge decomposition of $H^2(M, \mathbf{Z})$ (but not the intersection (x^l) in general). For a minimal model in the sense (MMP-1), we have the following factorization. This fundamental result is due to Kawamata [48]:

Theorem 2.6. *Let M be a minimal model. Then* $\mathrm{Bir}\,(M) = \mathrm{PsAut}\,(M)$*, and any* $f \in \mathrm{Bir}\,(M)$ *is decomposed as* $f = \varphi \circ \iota_{m-1} \circ \cdots \circ \iota_0$*, where* $M_0 = M = M_m$ *and* $\iota_i : M_i \cdots \to M_{i+1}$ *(*$0 \leq i \leq m-1$*) are flops betweeen minimal models* M_i *and* M_{i+1} *and* $\varphi \in \mathrm{Aut}\,(M)$.

Complex tori, CY manifolds and HK manifolds. The case of complex tori is very much known (See [37, 78] for some dynamically interesting features in tori). Most basic examples of minimal Calabi-Yau varieties are CY manifolds and (projective) HK manifolds as defined below. CY manifolds, HK manifolds together with rational manifolds are the main objects in this note.

Definition 2.7. Let M be an l-dimensional simply-connected compact Kähler manifold.

(1) M is a Calabi-Yau manifold in the strict sense (*CY manifold*) if $l \geq 3$ and $H^0(M, \Omega^j_M) = 0$ for $0 < j \leq l-1$ and $H^0(M, \Omega^l_M) = \mathbf{C}\omega_M$, where ω_M is a nowhere vanishing holomorphic l-form.

(2) M is a compact hyperkähler manifold (*HK manifold*) if $H^0(M, \Omega^2_M) = \mathbf{C}\sigma_M$, where σ_M is an everywhere nondegenerate holomorphic 2-form.

Good references of CY manifolds and HK manifolds are [41, 56]. By definition, HK manifolds are of even dimension and K3 surfaces are nothing but HK manifolds of dimension 2. CY manifolds M are always projective by $h^0(\Omega^2_M) = 0$ (as $l \geq 3$). On the other hands, both projective HK manifolds and non-projective HK manifolds are dense both in the Kuranishi space and in the global moduli space of marked HK manifolds [38, 44]. Examples with interesting (birational) automorphisms in our view will be given in Sections 4, 5, 6.

The importance of CY manifolds and HK manifolds in complex algebraic geometry lies in the fact, called the Bogomolov decomposition theorem [4], that these two classes of manifolds together with complex tori form the building blocks of compact Kähler manifolds with trivial first Chern class.

We close this section by the following:

Remark 2.8. Let M be a CY manifold or a projective HK manifold and $G < \mathrm{Bir}\,(M) = \mathrm{PsAut}\,(M)$. Assume that there is an ample divisor H such that $f^*H = H$ in $\mathrm{Pic}\,(M) \simeq \mathrm{NS}\,(M)$ for all $f \in G$. Then $G < \mathrm{Aut}\,(M)$ and G is a finite group. In particular, if $\rho(M) := \mathrm{rank}\,\mathrm{NS}\,(M) = 1$, then $\mathrm{Bir}\,(M) = \mathrm{Aut}\,(M)$ and it is a finite group. *So, in our view, interesting cases are* $\rho(M) \geq 2$.

Indeed, the same argument as in Theorem 2.2, applied for the G-equivariant embedding $\Phi_{|mH|} : M \to |mH|^*$ for large $m > 0$, shows that $G < \mathrm{Aut}\,(M)$ and at the same time G is a Zariski closed algebraic subgroup of $\mathrm{PGL}\,(|mH|^*)$. Since $\dim G = 0$ by $H^0(M, TM) = 0$, the result follows.

3. Topological entropy and Dynamical degrees

3.1. Topological entropy. References of this subsection are [40, 45].
Let $X = (X, d)$ be a compact metric space and $f : X \to X$ be a continuous surjective selfmap of X. We denote by f^n the n-th iterate of f. The *topological entropy* of f is the fundamental invariant that measures *how fast two general points spread out under the*

action of the semi-group $\{f^n | n \in \mathbf{Z}_{\geq 0}\}$, hence, presents a kind of complexity of f. For the definition, we define the new distance $d_{f,n}$ on X by

$$d_{f,n}(x,y) = \max_{0 \leq j \leq n-1} d(f^j(x), f^j(y)) \text{ for } x, y \in X.$$

Under the identification $x \leftrightarrow \mathbf{x}^{(n)} := (x, f(x), \cdots, f^{n-1}(x))$, the new distance $d_{f,n}(x,y)$ is the distance of the graph

$$\Gamma_{f,n} := \{\mathbf{x}^{(n)} = (x, f(x), \cdots, f^{n-1}(x)) \,|\, x \in X\} \subset X^n$$

induced by the product distance on X^d. The first projection $\mathrm{pr}_1 : (\Gamma_{f,n}, d_{f,n}) \to (X, d_{f,n})$ is an isometry and $\mathrm{pr}_1 : (\Gamma_{f,n}, d_{f,n}) \to (X, d)$ is a homeomorphism.

Let $\epsilon > 0$ be a positive real number. We call two points $x, y \in X$ (n, ϵ)-*separated* if $d_{f,n}(y,x) \geq \epsilon$, and a subset $F \subset X$ (n, ϵ)-*separated* if any two distinct points of F are (n, ϵ)-separated. Let

$$N_d(f, n, \epsilon) := \mathrm{Max}\,\{|F| \mid F \subset X \text{ is } (n,\epsilon) - \text{separated } \}\ .$$

Note that $N_d(f, n, \epsilon)$ is a well-defined positive integer, because X is compact.

Remark 3.1. Please imagine that $\epsilon > 0$ is "very very small", so that we can *not* distinguish two points $x, y \in X$ with $d(x,y) < \epsilon$ by "our eyes" but can do if $d(x,y) \geq \epsilon$. Then, we can not distinguish x, y if they are not $(1, \epsilon)$-separated but we can distinguish them by performing f if they are $(2, \epsilon)$-separated. Similarly, we can distinguish x, y at some stage, say $f^j(x), f^j(y)$ $(0 \leq j \leq n-1)$, if they are (n, ϵ)-separated. In this sense, $N_d(f, 1, \epsilon)$ is the maximal number of points of X distinguished by eyes and $N_d(f, n, \epsilon)$ is the maximal number of points of X distinguished by eyes after performing f^j $(0 \leq j \leq n-1)$. So, roughly, the growth of the sequence $\{N_d(f, n, \epsilon)\}_{n \geq 1}$ measures how fast general points spread out under the iterations of f to be distinguishable by our eyes (if it will be).

Definition 3.2. The *topological entropy*, or *entropy* for short, of f is:

$$h_{\mathrm{top}}(f) := h_d(f) := \lim_{\epsilon \to +0} h_d(f, \epsilon)\ ,$$

where

$$h_d(f, \epsilon) := \limsup_{n \to \infty} \frac{\log N_d(f, n, \epsilon)}{n}\ .$$

Since $\log N_d(f, n, \epsilon) \geq 0$ is an increasing function of $\epsilon > 0$, the limit exists in $[0, \infty]$ (possibly ∞). More or less from the definition, we obtain:

Corollary 3.3.

(1) $h_{\mathrm{top}}(f)$ *is a topological invariant, in the sense that* $h_{d'}(f) = h_d(f)$ *for any distance* d' *of* X *such that* (X, d') *and* (X, d) *are homeomorphic.*

(2) *If* $h_{\mathrm{top}}(f) > 0$*, then* $f^m \neq id_X$ *for all* $m \geq 1$*, i.e.,* $\mathrm{ord}\,(f) = \infty$.

(3) *If* f *is an isometry, for instance a translation of a torus, then* $h_{\mathrm{top}}(f) = 0$*. In particular, the converse of (2) is not necessarily true.*

(4) $h_{\mathrm{top}}(f \times f') = h_{\mathrm{top}}(f) + h_{\mathrm{top}}(f')$ *where* f' *is a surjective selfmap of a compact metric space* $X' = (X', d')$.

Example 3.4. Let E be a 1-dimensional complex torus. Consider the abelian surface $A = E \times E$ and its surjective endomorphism $f_M(x) = Mx$ given by $M \in M(2, \mathbf{Z})$ with $\det M \neq 0$. Note that $f_M \in \mathrm{Aut}\,(A)$ if $\det M = \pm 1$. Let α, β be the eigenvalues of M such that $|\alpha| \leq |\beta|$. Then, according to the three cases (i) $|\alpha| \geq |\beta| \geq 1$, (ii) $|\alpha| \geq 1 \geq |\beta|$, (iii) $1 \geq |\alpha| \geq |\beta|$, the entropy $h_{\mathrm{top}}(f_M)$ is (i) $\log |\alpha\beta|^2$, (ii) $\log |\alpha|^2$, (iii) $\log 1 = 0$. In particular, $h_{\mathrm{top}}(f_M) = \log \eta^2 > 0$, the natural logarithm of (the square of) the golden number, for the Lie automorphism $f_M \in \mathrm{Aut}\,(A)$ given by

$$M = \begin{pmatrix} 2 & 1 \\ -1 & -1 \end{pmatrix} .$$

Very rough idea is as follows. For simplicity, we further assume that M is diagonalizable in $M(2, \mathbf{C})$. We fix the flat distance d on A from the universal cover $\mathbf{C}^2$. Let $\epsilon > 0$ be a very small number. Let us cover A by N mutually disjoint complex 2-dimensional ϵ-"parallelograms" (actual real dimension is 4) that are parallel to the complex eigenvectors of M. Then $N(f_M, 1, \epsilon)$ is about N. Next divide each of N ϵ-parallelograms into mutually disjoint ϵ-parallelograms with respect to the distance $d_{f_M,2}$. In case (i), each original parallelogram is devided into about $|\alpha\beta|^2$ new parallelograms, because $|\alpha| \geq 1$ and $|\beta| \geq 1$ (and real dimension is $2 + 2$). Therefore, $N(f_M, 2, \epsilon)$ is about $|\alpha\beta|^2 N$. In case (ii), each parallelogram is devided into $|\alpha|^2$ new parallelogram, because $|\alpha| \geq 1$ but $|\beta| \leq 1$. Therefore, $N(f_M, 2, \epsilon)$ is about $|\alpha|^2 N$. Similarly, in case (iii), $N(f_M, 2, \epsilon)$ remains N. Repeating this, we see that $N(f_M, n, \epsilon)$ is about $|\alpha\beta|^{2(n-1)} N$, $|\alpha|^{2(n-1)} N$, N according to the three cases (i), (ii), (iii). This implies the result.

Note that in each case, the entropy is the natural logarithm of the spectral radius of $f_M^* | H^*(A, \mathbf{Z})$. Actually, this is *not accidental* as we will explain in the next subsection.

3.2. Fundamental theorem of Gromov-Yomdin. References of this subsection are [33, 34, 39, 40, 89] (see also [22]).

Let M be a compact Kähler manifold of dimension l and η be any Kähler form on M. Then M is a compact metric space by the distance defined by η. Let $f : M \to M$ be a surjective holomorphic map. Then f^* naturally acts on the k-th cohomology group $H^k(M, \mathbf{Z})$ as well as each Hodge component $H^{p,q}(M)$. We define $r_p(f)$ to be the *spectral radius* of $f^* | H^{p,p}(M)$, that is, the maximum absolute value of eigenvalues of $f^* | H^{p,p}(M)$. Similarly, we denote by $r(f)$ (resp. $r^{\mathrm{even}}(f)$) the spectral radius of f^* on $\oplus_{k=0}^{2l} H^k(M, \mathbf{Z})$ (resp. $\oplus_{p=0}^{l} H^{2p}(M, \mathbf{Z})$).

We define the *p-th dynamical degree* $d_p(f)$ by

$$d_p(f) := \lim_{n \to \infty} (\delta_p(f^n))^{\frac{1}{n}} ,$$

where

$$\delta_p(f^n) := (\int_M (f^n)^*(\eta^p) \wedge \eta^{l-p}) = ([(f^n)^*(\eta^p)].[\eta^{l-p}])_M .$$

Here $(*, **)_M$ is the intersection number. The limit does not depend on the choice of η once the existence is guaranteed. Indeed, for two Kähler forms η and η', there are positive real number C and a Kähler form η" such that $C[\eta] = [\eta'] + [\eta"]$ in $H^{1,1}(M, \mathbf{R})$. The fact that the limit exists is non-trivial. There are many ways to see it. For instance, it is an immediate consequence of the following crucial observation by Dinh-Sibony, which holds also for *rational* dominant selfmaps [33, 34]:

Theorem 3.5. *There is a constant $C = C_{M,\eta}$ depending only on M and η (but not on f and g) such that*

$$\delta_p(f \circ g) \leq C\delta_p(f)\delta_p(g) \ ,$$

for any two dominant holomorphic selfmaps $f : M \to M$, $g : M \to M$.

The logarithmic volume $\mathrm{lov}\,(f)$, introduced by Gromov, is

$$\mathrm{lov}\,(f) := \mathrm{limsup}_{n\to\infty} \frac{\log \mathrm{Volume}(\Gamma_{f,n})}{n} \ ,$$

where

$$\mathrm{Volume}(\Gamma_{f,n}) := \frac{1}{l!}\int_{\Gamma_{f,n}} (\sum_{i=1}^{n} \mathrm{pr}_i^* \eta_M)^l \ .$$

The following fundamental theorem is due to Gromov-Yomdin:

Theorem 3.6. *Let M be a compact Kähler manifold of dimension l and $f : M \to M$ be a surjective holomorphic map. Then, $d_p(f) = r_p(f)$ and*

$$h_{\mathrm{top}}(f) = \mathrm{lov}\,(f) = \log \mathrm{max}_{0\leq p\leq l}\, d_p(f) = \log \mathrm{max}_{0\leq p\leq l}\, r_p(f) = \log r^{\mathrm{even}}(f) = \log r(f) \ .$$

Moreover, if M is projective, then $h_{\mathrm{top}}(f)$ is also equal to the natural logarithm of the spectral radius of $f^*|\oplus_p H^{p,p}(M, \mathbf{Z})$.

This theorem opens the door to compute the entropy of a biregular automorphism by algebro-geometric methods. For instance, Example 3.4 is immediate from this theorem; one may just compute $r_p(f)$ for $p = 0, 1, 2$. Moreover, $d_p(f)$ and $r_p(f)$ can be regarded as *finer* invariants of f than $h_{\mathrm{top}}(f)$, while geometric meanings become less apparent.

Corollary 3.7.

(1) $h_{\mathrm{top}}(f) = 0$ *if* $\dim\, M = 1$, *and also* $h_{\mathrm{top}}(f) = 0$ *for* $f \in \mathrm{Aut}^0(M)$ *(the identity component of* $\mathrm{Aut}\,(M)$*). For instance,* $h_{\mathrm{top}}(f) = 0$ *if* $f \in \mathrm{Aut}\,(\mathbf{P}^d)$ *or again if f is a translation automorphism of a complex torus.*

(2) *The topological entropy is the natural logarithm of an algebraic integer.*

Indeed, (1) is clear by Theorem 3.6. Since the eigenvalues of $f^*|H^*(M, \mathbf{Z})$ are algebraic integers, (2) follows from Theorem 3.6.

Corollary 3.8.

(1) $d_0(f) = 1$ *and* $d_l(f) = \deg f$, *the topological degree of f.*

(2) *The sequence* $\{d_p(f)\}_{0\leq p\leq l}$ *is log-concave, i.e.,* $d_{p-1}(f)d_{p+1}(f) \leq d_p(f)^2$.

(3) $d_p(f) \geq 1$ *for all p and,*

(4) $h_{\mathrm{top}}(f) > 0$ *if and only if* $d_1(f) > 1$.

(1) is clear by definition. We have $\delta_{p-1}(f)\delta_{p+1}(f) \leq \delta_p(f)^2$ by the Hodge index theorem (when M is projective and η is chosen to be a Hodge metric) and by the Tessier-Khovanski inequality in general case. Then (2) follows from (1), and (2) implies (3), (4).

Very brief outline of the proof of Theorem 3.6 is as follows [33, 40]. Note the *obvious* relation $(f^n)^* = (f^*)^n$ for $f \in \mathrm{Aut}\,(M)$. Then $d_p(f) = r_p(f)$ follows from linear algebra

plus the Perron-Frobenious theorem on the linear maps preserving a strict convex cone. So, $d_p(f) = r_p(f) \leq r^{\rm even}(f) \leq r(f)$. The deepest part is $h_{\rm top}(f) \geq \log r(f)$ for any compact oriented Riemannian C^∞-manifolds and surjective oriented C^∞-map $f : M \to M$. This is due to Yomdin [89] (see also [40]). Gromov [39] proved the reverse inequality $h_{\rm top}(f) \leq {\rm lov}\,(f) = \log \max_{0\leq p\leq k} d_p(f)$. The essential part of this inequality is that if $F \subset M$ is (n, ϵ)-separated, then the corresponding subset $\mathbf{F}^{(n)}$ in the graph $\Gamma_{f,n}$ is $(1, \epsilon)$-separated, and therefore the balls $\Gamma_{f,n} \cap B(\mathbf{x}^{(n)}, \epsilon/2)$ ($\mathbf{x}^{(n)} \in \mathbf{F}^{(n)}$) are mutually disjoint. This gives an obvious estimate of ${\rm Volume}\,(\Gamma_{f,n})$ from the below and leads the first inequality, via Lelong's theorem. ${\rm lov}\,(f) = \max_{0\leq p\leq d} d_p(f)$ is non-trivial but doable by fairly straightforward computations of the differential forms, *at least when f is holomorphic*.

See also [28] for derived categorical approach for the topological entropy.

3.3. Generalization for rational mappings after Dinh-Sibony. References of this subsection are [33, 34, 42] (see also [22]).

Let M be a compact Kähler manifold of dimension l, and $f : M \cdots \to M$ be a dominant rational selfmap. The topological entropy $h_{\rm top}(f)$ is defined in the same way as in the holomorphic case, just by considering the well-defined orbits, i.e., $\{f^k(x)\}_{k=0}^{n-1}$ with $f^k(x) \notin I(f)$ at each n-th step. The logarithmic volume ${\rm lov}\,(f)$ is defined again in the same way as above, just by taking the graph $\Gamma^0_{f,n}$ over $M \setminus \cup_{k=0}^{n-1} I(f^k)$ at each n-th step [33, 42].

The pullback operation $f^* : H^{p,p}(M) \to H^{p,p}(M)$ is well-defined if one uses *currents*. Let $\tilde{M}$ be a resolution of the indeterminacy of f and $p_i : \tilde{M} \to M$ $(i = 1, 2)$ be the natural projections. Then for any closed (p, p)-form α, we define $f^*(\alpha) = (p_1)_* p_2^*(\alpha)$, where p_2^* is the natural pullback as forms and $(p_1)_*$ is the natural pushforward as currents, i.e., $\langle (p_1)_*(p_2^*(\alpha)), \beta\rangle_M := \langle p_2^*(\alpha), p_1^*(\beta)\rangle_{\tilde{M}}$. The action f^*naturally descends to the linear action on $H^{p,p}(M)$. So, the definitions of $\delta_p(f)$ and the p-th dynamical degree $d_p(f)$ make sense without any change. $d_p(f)$ does not depend on the choice of η for the same reason as before, again once the existence of the limit is guaranteed. However, there is one crucial difference from the holomorphic case; $(f\circ g)^* \neq g^*\circ f^*$ *and* $(f^n)^* \neq (f^*)^n$ *in general.* This already happens for the standard Cremona transformation ${\rm cr}_l$ of $\mathbf{P}^l$ and makes outlined proof in the holomorphic case delicate at all the places where we freely use them. For instance, in general, there is no way to compare $d_p(f)$ and $r_p(f)$. Dinh and Sibony [33, 34] proved:

Theorem 3.9. *Let X be a compact Kähler manifold of dimension l and $f : M \cdots \to M$ be a dominant, rational map (= meromorphic map, by our convention). Then, Theorem 3.5 holds for rational dominat maps and*

$$h_{\rm top}(f) \leq {\rm lov}\,(f) = \log \max_{0\leq p\leq d} d_p(f)\ .$$

Moreover $d_p(f)$ are birational invariants in the sense that $d_p(f) = d_p(\varphi \circ f \circ \varphi^{-1})$ for any birational map $\varphi : M \cdots \to M'$ between compact Kähler manifolds.

On the other hand, $h_{\rm top}(f)$ is *not* a birational invariant by Guedj [42]:

Example 3.10. Let $f : \mathbf{C}^2 \to \mathbf{C}^2$ be a morphism defined by $(x, y) \to (x^2, y + 1)$. Then f naturally extends to a rational selfmap f_1 of $\mathbf{P}^2$ and a holomorphic selfmap f_2 of $\mathbf{P}^1 \times \mathbf{P}^1$. Then, $h_{\rm top}(f_1) = 0$ but $h_{\rm top}(f_2) = \log 2 > 0$.

Because of the birational invariance of dynamical degrees, one can define dynamical degrees for a dominant selfmap of a singular compact complex space of class $\mathcal{C}$ in an obvious

manner, but not the topological entropy in this way. For this reason, dynamical degrees fit well more with birational geometry than the entropy. They are also useful when we study *biregular* automorphisms in Problem 1.1, as we shall see in concrete cases in Section 5.

The essential part of the proof of Theorems 3.5, 3.9 and their variants later is a deep theory of semi-regularization of currents, very roughly, a method for approximating currents well by sequences of smooth forms. Once such semi-regularization results are well established, then the proof goes along the same line as in the holomorphic case *if one carefully replaces all necessary estimates for currents by those of semi-regularizing smooth forms*.

For a rational dominant selfmap f, Corollary 3.7 (2) *is expected to be true but unknown*. For instance, (2) is true for $d_1(f)$ if $f \in \mathrm{PsAut}\,(M)$. This is because then $(f^n)^* = (f^*)^n$ on $H^2(M, \mathbf{R})$, hence $d_1(f) = r_1(f)$ for the same reason as in the holomorphic case. Corollary 3.8 (1), (2), (3), being free from $h_{\mathrm{top}}(f)$, is clearly true, but (4) is *not true* as Example 3.10 shows.

3.4. Relative dynamical degrees after Dinh-Nguyên-Truong. References of this subsection are [30] and [31]. Corollary 3.3 (3) or dynamical degrees of the product map $f \times f' : X \times X' \to X \times X'$ suggests a good notion of relative dynamical degrees with nice properties. If exists, then it will provide a useful numerical criterion for the primitivity of a selfmap, as we shall test in some concrete cases in Section 5.

Setting I. *Let $f : M \cdots \to M$, $g : B \cdots \to B$ be dominant rational maps such that $\pi \circ f = g \circ \pi$. Here $\pi : M \to B$ is a surjective holomorphic map between compact Kähler manifolds M and B of dimensions l and b (necessarily $l \geq b$) with Kähler forms η_M and η_B.*

In Setting I, we define *the relative dynamical degrees* $d_p(f|\pi)$ by

$$d_p(f|\pi) := \lim_{n\to\infty} \left(\int_M (f^n)^*(\eta_M^p) \wedge \pi^*(\eta_B^b) \wedge \eta_M^{l-p-b} \right)^{1/n} , \ 0 \leq p \leq l-b \ .$$

This definition is due to Dinh and Nguyên [30]. If we take η_B so that η_B^b is the Poincaré dual of a point and *virtually identify* all the fibers M_b and regard then $f|M_b : M_b \cdots \to M_{g(b)}$ as the *virtual selfmap* of M_b, then $d_p(f|\pi)$ is the same form as the p-th dynamical degree of the *virtual* $f|M_b$. Note also that $\pi^*(\eta_B^b) \wedge \eta_M^{l-p-b}$ is a form as π is *holomorphic* and $(f^n)^*(\eta_M^p)$ is a current of proper bidegree. So the integration in the right hand side makes sense. This is the reason why we assume that π is holomorphic. The existence of the limit is again non-trivial, but settled by [30] and [31]. The following fundamental result is due to Dinh-Nguyên-Truong [30, 31]:

Theorem 3.11. *In Setting I, for all $0 \leq p \leq l$,*

$$d_p(f) = \max_j d_j(g) d_{p-j}(f|\pi) \ .$$

Here j runs through all the integers for which the integrations defining $d_j(g)$ and $d_{p-j}(f|\pi)$ are meaningful, i.e., j runs through $\max\{0, p-l+b\} \leq j \leq \min\{p, b\}$. Moreover, $\{d_p(f|\pi)\}_p$ satisfy $d_{p-1}(f|\pi) d_{p+1}(f|\pi) \leq d_p(f|\pi)^2$ (the log-concavity) and they are birational invariants in an obvious sense, within Setting I.

That π is holomorphic in Theorem 3.11 appears slightly restrictive, compared with usual situations, and the most natural setting is probably the following:

Setting II. *$\pi' : M' \cdots \to B'$ is a dominant rational map from an l-dimensional compact complex variety M' of class $\mathcal{C}$ to a compact complex variety B' of dimension b, equivariant with rational dominant selfmaps f' and g' of M' and B'.*

In Setting II, B' is of class $\mathcal{C}$ by the original definition of class $\mathcal{C}$, and therefore, there is a birational morphism $\varphi : B \to B'$ from a compact Kähler manifold B as well. Then resolving the indeterminacy of the rational map $\varphi^{-1} \circ \pi'$ from M' to B, we obtain a *holomorphic* surjective morphism between compact Kähler *manifolds* $\pi : M \to B$. Moreover, π is equivariant to the rational dominant selfmaps f and g of M and B, naturally induced from f' and g'. This is exactly Setting I in Theorem 3.11. By the birational invariance of dynamical degrees, we have $d_p(f) = d_p(f')$, $d_p(g) = d_p(g')$. Moreover, by the birational invariance of the relative dynamical degrees in Setting I in Theorem 3.11, we *can define* $d(f'|\pi') := d(f|\pi)$ which is independent of the choice of models $\pi : M \to B$. Then, the equation in Theorem 3.11 is nothing but the equation $d_p(f') = \max_j d_j(g')d_{p-j}(f'|\pi')$ in Setting II. Therefore:

Corollary 3.12. *Theorem 3.11 is true also in the* Setting II.

Note that $d_0(f'|\pi') = 1$ and $d_{l-b}(f'|\pi')$ is the topological degree of $f'|M_t : M_t \cdots \to M_{g'(t)}$ for a generic fiber M'_t ($t \in B$). The log-concavity then implies that $d_p(f'|\pi') \geq 1$ for any meaningful p as before.

Since only $d_0(f|\pi) = 1$ is the meaningful relative dynamical degree for a generically finite map, we obtain the following [30, 31] from Theorem 3.12:

Corollary 3.13.

(1) *The dynamical degrees are invariant under any equivariant generically finite dominant maps, i.e., if $\pi : M \cdots \to B$ is a generically finite dominant rational map equivariant to the rational dominant selfmaps f, g of M, B, then $d_p(f) = d_p(g)$ for every p.*

(2) *The topological entropy of dominant holomorphic selfmaps of compact Kähler manifolds are invariant under equivariant generically finite dominant rational maps. More precisely, in (1), if both M and B are compact Kähler manifolds and f and g are holomorphic, then $h_{\mathrm{top}}(f) = h_{\mathrm{top}}(g)$.*

(2) follows from (1) and Theorem 3.6.

Our primary interest in Theorem 3.11 is its applicability for primitivity of rational selfmaps. When $l = \dim M \leq 3$, we can deduce the following fairly useful *numerical* criterion for the primitivity of $f \in \mathrm{Bir}\, M$ from Theorem 3.11. (1) is known before Theorem 3.11 and (2) is due to Truong and myself [79]:

Corollary 3.14. *Let M be a compact Kähler manifold and $f \in \mathrm{Bir}\,(M)$.*

(1) *Assume that $\dim M = 2$. Then f is primitive if $d_1(f) > 1$. In particular, $f \in \mathrm{Aut}\,(M)$ is primitive if $h_{\mathrm{top}}(f) > 0$.*

(2) *Assume that $\dim M = 3$. Then f is primitive if $d_1(f) \neq d_2(f)$.*

Outline of (2) is as follows. Assume that f is imprimitive. Then there are a compact Kähler manifold B, dominant rational maps $\pi : M \cdots \to B$, $g : B \cdots \to B$ such that $\pi \circ f = g \circ \pi$. Here $0 < \dim B < 3 = \dim M$. We consider the case $\dim B = 2$ (the case $\dim B = 1$ is similar). Then by Corollary 3.12, we have

$$d_1(f) = \max\{d_1(g), d_1(f|\pi)\}\ ,\ d_2(f) = \max\{d_1(f|\pi)d_1(g), d_2(g)\}\ .$$

Since f and g are birational, $d_3(f) = d_2(g) = 1$. Thus, by Corollary 3.12, $1 = d_3(f) = d_2(g)d_1(f|\pi)$, hence, $d_1(f|\pi) = 1$. So, $d_1(f) = \max\{d_1(g), 1\} = d_2(f)$.

4. Surface automorphisms in the view of entropy

In this section, we take a closer look at surface automorphisms of positive entropy. They are primitive by Corollary 3.14(1). We assume some familiarity with classification of surfaces. A good reference is [2] with [35] for rational surfaces. Throughout this section, S is a smooth compact Kähler surface.

4.1. Surface automorphisms of positive entropy. We note that a birational automorphism $f \in \mathrm{Bir}\,(S)$ naturally induces a *biregular* automorphism of the minimal model $S_{\min}$ of the same dynamical degrees (Theorem 3.9) if $\kappa\,(S) \geq 0$. In this way, one can almost recover from Theorem 2.2 the following breakthrough observation due to Cantat [19]:

Theorem 4.1. *Assume that S admits an automorphism $f \in \mathrm{Aut}\,(S)$ of positive entropy, i.e., $d_1(f) > 1$. Then S is birational to either (i) $\mathbf{P}^2$, (ii) a K3 surface, (iii) a 2-dimensional complex torus, or (iv) an Enriques surface. In the case (i), S is a blow up of $\mathbf{P}^2$ at 10 or more points, possibly infinitely near [63].*

Recall that $d_1(f) = r_1(f) > 1$ is an algebraic integer (Corollary 3.7 (2)). It turns out to be a special algebraic integer of even degree, called a *Salem number*:

Definition 4.2. An irreducible monic polynomial $S(x) \in \mathbf{Z}[x]$ is called a *Salem polynomial* if the complex roots are of the following form (*possibly* $d = 1$):

$$a \in (1, \infty)\ ,\ 1/a \in (0,1)\ ,\ \alpha_i, \overline{\alpha_i} \in S^1 := \{z \in \mathbf{C}\,|\,|z| = 1\} \setminus \{\pm 1\} (1 \leq i \leq d-1)\ .$$

The unique root $a > 1$ is called a *Salem number* of degree $2d$ $(= \deg S(x))$.

The smallest *known* Salem number is the *Lehmer number* which is the unique root > 1 of the following Salem polynomial of degree 10:

$$x^{10} + x^9 - x^7 - x^6 - x^5 - x^4 - x^3 + x + 1\ .$$

It is approximately 1.17628 and conjectured to be the minimum among all Salem numbers. So far, this conjecture is neither proved nor disproved.

The following theorem is due to McMullen [57–59]:

Theorem 4.3. *Let $f \in \mathrm{Aut}\,(S)$ and assume that $d_1(f) > 1$. Then, $d_1(f)$ is a Salem number, and $d_1(f)$ is always greater than or equal to the Lehmer number.*

See also [60, 61, 73, 81, 85] for relevant results.

4.2. Examples of surface automorphisms of positive entropy. There is a huge number of works concerning automorphisms of surfaces. Here among many examples, I present four examples which are smoothly connected to the topics in the next sections.

Example 1 - Rational surface automorphisms. Recall that any birational automorphism of $\mathbf{P}^2$ is expressed by two rational functions of the affine coordinates $(x = x_2/x_1, y = x_3/x_1)$ of $\mathbf{P}^2_{[x_1:x_2:x_3]}$. Consider the birational automorphism of the following special form ([59], also compare with an earlier form in [6]):

$$f^*(x,y) := f^*_{(a,b)}(x,y) := (a+y, b+\frac{y}{x}) \ , \ a,b \in \mathbf{C} \ .$$

$I(f)$ is the set of coordinate points $\{e_1, e_2, e_3\}$ and $I(f^{-1})$ is $\{e_2, e_3, e_4\}$, where $e_4 := (a,b)$. Set $e_{k+4} := f^k(e_4)$. Choose (a,b) so that $e_k \notin e_1e_2 \cup e_2e_3 \cup e_3e_1$ for all $4 \leq k \leq n$ with $n \geq 10$ and $e_{n+1} = e_1$ (periodicity condition of indeterminacies). Then one can realize f as an automorphism of $S = S_n$, the blowing-ups of $\mathbf{P}^2$ at the n points e_k. This f also realizes the Coxeter element c_n of the Weyl group $W(E_n)$ in the sense that $f^* = c_n$ on $H^2(S, \mathbf{Z})$ under the natural identification $E_n = K_S^{\perp}$ and $W(E_n) < \mathrm{O}\,(H^2(S, \mathbf{Z}))$, hence of positive entropy. For instance, f with $n = 10$ and $(a,b) = (0.4995..., -0.0837...)$ realizes the Lehmer first dynamical degree and f with $n = 11$ and (a,b) approximately $(0.0444 - 0.4422i, 0.0444 + 0.4422i)$ has a Siegel domain (cf. Example 4). The actual construction in [59] is not merely the numerical one but is based on an explicit marked Torelli type result for log K3 surfaces (S, C) with C being a unique cuspidal rational curve in $|-K_S|$.

Example 2 - Birational automorphisms after Diller-Favre. Reference here is [27]. Let $c \in \mathbf{C}$ and consider the birational automorphism of $\mathbf{P}^1 \times \mathbf{P}^1$ defined by the following affine form:

$$f_c(x,y) := (y+1-c, x\frac{y-c}{y+1}) \ .$$

[27] computes the first dynamical degree of f_c and observes many interesting features, depending on $c \in \mathbf{C}$. For instance, if c is irrational, then $d_1(f_c) = \eta$, the golden number. Note that the golden number is *not* a Salem number, so that f_c with irrational c can *never* be realized as a biregular automorphism of any smooth birational models.

Example 3 - Cayley's K3 surface after Festi, Garbagnati, van Geemen and van Luijk. In the long history of automorphisms of K3 surfaces or more specifically those of smooth quartic surfaces, Cayley seems the first who suggested the existence of automorphisms of infinite order. Here we explain his beautiful, very explicit construction, following a modern elegant account [36]. This example will be also used to construct higher dimensional HK example in Section 6.

Let a_{ijk} $(1 \leq i,j,k \leq 4)$ be 4^3 generic complex numbers. Let us consider the following determinantal quartic surface in $\mathbf{P}^3$ with homogeneous coordinates $\mathbf{x} = [x_1 : x_2 : x_3 : x_4]$:

$$S_0 := (\det M_0(\mathbf{x}) = 0) \subset \mathbf{P}^3_{\mathbf{x}} \ ,$$

where $M_0 = M_0(\mathbf{x}) := (\sum_i a_{ijk}x_i)_{k,j}$ is the 4×4 matrix whose (k,j) entry is $\sum_i a_{ijk}x_i$. By our genericity assumption, $\mathrm{rank}\, M_0(\mathbf{x}) = 3$ for all $\mathbf{x} \in S_0$ and S_0 is a smooth quartic K3 surface. One can also construct two more smooth determinantal quartic K3 surfaces from a_{ijk}:

$$S_1 := (\det M_1(\mathbf{y}) = 0) \subset \mathbf{P}^3_{\mathbf{y}} \ , \ S_2 := (\det M_2(\mathbf{z}) = 0) \subset \mathbf{P}^3_{\mathbf{z}} \ .$$

Here $M_1 = M_1(\mathbf{y}) := (\sum_i a_{ijk}y_j)_{i,k}$, $M_2 = M_2(\mathbf{z}) := (\sum_i a_{ijk}z_k)_{j,i}$.

Let P_i be the cofactor matrix of M_i. Then,

$$P_iM_i = M_iP_i = \det(M_i)\cdot I_4 \text{ and } \mathbf{x}M_1(\mathbf{y}) = (\sum_{i,j} a_{ijk}x_iy_j)_k = \mathbf{y}M_0(\mathbf{x})^t \ .$$

Recall that $\mathrm{rank}\,(M_0(\mathbf{x})) \geq 3$ for each $\mathbf{x} \in \mathbf{P}^3$ and the same for $M_1(\mathbf{y})$, $M_2(\mathbf{z})$. Thus, the j-th column $(p_{ij}(\mathbf{x}))_i$ of $P_0 = P_0(\mathbf{x})$ gives a *Cremona transformation* $\varphi_0 : \mathbf{P}^3_{\mathbf{x}} \cdots \to \mathbf{P}^3_{\mathbf{y}}$ that maps S_0 to S_1, hence, an isomorphism $\varphi_0|S_0 : S_0 \to S_1$. In the same way, we have two more Cremona transformations $\varphi_1 : \mathbf{P}^3_{\mathbf{y}} \cdots \to \mathbf{P}^3_{\mathbf{z}}$, $\varphi_2 : \mathbf{P}^3_{\mathbf{z}} \cdots \to \mathbf{P}^3_{\mathbf{x}}$ and isomorphisms $\varphi_1|S_1 : S_1 \to S_2$, $\varphi_2|S_2 : S_2 \to S_0$. In this way, we obtain an *explicit* automorphism of S_0: $g := \varphi_2 \circ \varphi_1 \circ \varphi_0$. This is the automorphism that *Cayley first found around 1870 and said that "The process may be indefinitely repeated"* [25, 36].

As observed by [36], in modern terminologies, a characterization of linear determinantal varieties [5] with our genericity assumption says that S_0 is nothing but a K3 surface with

$$\mathrm{NS}\,(S_0) = (\mathbf{Z}h_1 \oplus \mathbf{Z}h_2, \begin{pmatrix} 4 & 6 \\ 6 & 4 \end{pmatrix}) = (\mathbf{Z}[\eta], 4\mathrm{Nm}(*)) \ , \ \mathrm{Nm}(a+b\eta) = a^2 + ab - b^2 \ .$$

Here η is the golden number and the lattice identification is given by $h_1 \leftrightarrow 1$ and $h_2 \leftrightarrow \eta^2$. Under this identification, the action of g on $\mathrm{NS}\,(S)$ is the multiplication by η^6 on $\mathbf{Z}[\eta]$. So, as predicted by Cayley, g is actually of infinite order and $h_{\mathrm{top}}(g) = \log \eta^6 > 0$. [36] further shows that $\mathrm{Aut}\,(S_0) = \langle g \rangle$. They also give the explicit *integers* $a_{ijk} \in \mathbf{Z}$ with desired properties. I re-discovered Cayley's automorphism in answering a question of Kawaguchi ([74], see also [12]):

Theorem 4.4. *Let W be a smooth compact Kähler surface with automorphism f such that f is of positive entropy and has no fixed point. Then W is birational to a projective K3 surface, and the pair of Cayley's K3 surface S_0 and its automorphism g is one of such examples.*

Example 4 - Non-projective K3 surface automorphism with Siegel domain after McMullen. Let f be an automorphism of a smooth surface S. We call a domain $U \subset S$ a *Siegel domain* of f if $f(U) = U$ and U is biholomorphic to the 2-dimensional unit disk Δ^2 with coordinates (z_1, z_2) such that the induced action of f on Δ^2 is of the form $f^*(z_1, z_2) = (\alpha_1 z_1, \alpha_2 z_2)$ for some multiplicatively independent complex numbers α_1 and α_2 on the unit circle S^1, i.e., $\alpha_1^{m_1}\alpha_2^{m_2} \neq 1$ for any integers $(m_1, m_2) \neq (0, 0)$ and $|\alpha_1| = |\alpha_2| = 1$.

If S is a K3 surface and f is an automorphism with Siegel domain as above, then $f^*\sigma_S = \alpha_1\alpha_2\sigma_S$. Here $\sigma_S \neq 0$ is a global holomorphic 2-form on S. Note that $\alpha_1\alpha_2$ is not root of unity. Thus S is necessarily non-projective, as the pluri-canonical representation of $\mathrm{Bir}\,(M)^*|H^0(M, \mathcal{O}_M(mK_M))$ is always finite if M is projective [86]. The next very surprising result due to McMullen [58] gave me a strong motivation to study birational automorphisms from the view of this note:

Theorem 4.5. *There is a K3 surface S of Picard number 0 with $\mathrm{Aut}\,(S) = \langle f \rangle$ such that f is of positive entropy and has a Siegel domain. In particular, the canonical representation $f^*|H^0(\mathcal{O}(K_S)) = f^*|H^0(\Omega^2_S)$ is of infinite order, and there is no point $Q \in S$ such that the orbit $\mathrm{Aut}\,(S)\cdot Q$ is topological dense (even though f is of positive entropy). Slightly more explicitly, one of such (S, f) is realizable so that the characteristic polynomial of $f^*|H^2(S, \mathbf{Z})$ is the following Salem polynomial of degree 22;*

$$S_{22}(X) := x^{22} + x^{21} - x^{19} - 2x^{18} - 3x^{17} - 3x^{16} - 2x^{15} + 2x^{13} + 4x^{12}$$

$$+5x^{11}+4x^{10}+2x^9-2x^7-3x^6-3x^5-2x^4-x^3+x+1\ .$$

In this case, $h_{\mathrm{top}}(f) = \log a$*, where* a *is the Salem number of* $S_{22}(x)$*, approximately,* 1.37289.

Unlike the examples above, construction is highly implicit, based on the surjectivity of the period map and golobal Torelli theorem for K3 surfaces, and the existence of Siegel domain is based on a deep transcendental number theoretical result [58], from which one can deduce the transcendency of π and e in one line. See also [73] for a slightly different example.

We close this section with a few remarks relevant to Theorem 4.5:

Remark 4.6. As mentioned, there are smooth rational surfaces with an automorphism with Siegel domain [7, 8, 59]. Rational surfaces are always projective, but this does not contradict the finiteness of pluri-canonical representation, because

$$\kappa(S) = -\infty,\ \text{i.e.}, H^0(S, \mathcal{O}_S(mK_S)) = 0 \quad \text{for all} \quad m > 0.$$

Remark 4.7. Let S, f be as in Theorem 4.5 and P be the center of the Siegel domain. Let M be the blowing-ups of $N := S \times S$, first at the intersection point (P, P) of $S \times \{P\}$, $\{P\} \times S$, the diagonal Δ, and the graph Γ_f, and next along the proper transforms of $S \times \{P\}$, $\{P\} \times S$, Δ, Γ_f. Then M is a simply-connected compact Kähler fourfold which can not be deformed into projective manifolds under any small proper deformation ([72], also compare with [88]).

Remark 4.8. Let S be a projective K3 surface. Then $1 \leq \rho(S) \leq 20$, and projective K3 surfaces with $\rho(S) \geq \rho$ form countable union of $(20-\rho)$-dimensional families. The automorphism group of S *tends* to be larger if $\rho(S)$ becomes larger (see [67] for the precise statement in terms of deformation). If $\rho(S) = 20$, for instance if S is the Fermat quartic surface, then $\mathrm{Aut}\,(S)$ contains the free subgroup $\mathbf{Z} * \mathbf{Z}$ with many elements of positive entropy, and the orbit $\mathrm{Aut}\,(S) \cdot P$ is topologically dense in S for generic $P \in S$ ([20, 69] see also [82]).

5. Rational and CY threefolds with primitive automorphisms of positive entropy

5.1. Biregular automorphisms vs. birational automorphisms. Some experiences show that in dimension ≥ 3, the biregular automorphisms tend to be drastically fewer than birational automorphisms. So, finding manifolds with "interesting" biregular automorphisms is more challenging in some sense. Here I present a few examples of this tendency.

Example 1 - CY manifolds in Fano manifolds. Cayley's K3 surfaces are smooth anti-canonical members of the Fano threefold $\mathbf{P}^3$. Smooth anti-canonical members of higher dimensional Fano manifolds are CY manifolds. However,

Theorem 5.1. *Let* $l \geq 3$ *and* M *be a smooth member of* $|-K_V|$ *of a smooth Fano manifold* V *of dimension* $l+1$*. Then* M *is a CY manifold of dimension* $l \geq 3$*, but* $|\mathrm{Aut}\,(X)|$ *is finite.*

Lefschetz hyperplane theorem shows that $\iota^* : H^2(V, \mathbf{Z}) \simeq H^2(M, \mathbf{Z})$ under the inclusion $\iota : M \to V$ for $l \geq 3$. Kollár [16] shows that $\overline{\mathrm{Amp}\,(M)} \simeq \overline{\mathrm{Amp}\,(V)}$ under

ι^*. Hence $\overline{\mathrm{Amp}\,(M)}$ is a finite rational polyhedral cone as so is $\overline{\mathrm{Amp}\,(V)}$. This implies $|\mathrm{Aut}\,(M)| < \infty$ by Remark 2.8.

Example 2 - CY manifolds of smaller Picard numbers. Recall that Cayley's K3 surfaces are of Picard number 2 and have automorphism of positive entropy. On the other hand, we have ([75]; see also [53, 54]):

Theorem 5.2. $|\mathrm{Aut}\,(M)| < \infty$ *for an odd dimensional CY manifold of* $\rho(M) = 2$.

Example 3. Let M be a general complete intersection in $\mathbf{P}^3 \times \mathbf{P}^3$ of 2 hypersurfaces of bidegree $(1,1)$ and a hypersurface of bidegree $(2,2)$. Then M is a CY threefold of Picard number 2 (hence an example of both Theorems 5.1, 5.2). Let ι_k be the covering involution of the k-th projection $\mathrm{pr}_k : M \cdots \to \mathbf{P}^3$ of degree 2. Set $f := \iota_2 \circ \iota_1 \in \mathrm{Bir}\,(M)$. Then, $d_1(f) = 17 + 12\sqrt{2} > 1$, $\langle f \rangle \simeq \mathbf{Z}$ and $[\mathrm{Bir}\,(M) : \langle f \rangle] < \infty$ even though $\mathrm{Aut}\,(M)$ is finite.

Recall that $\mathrm{Bir}\,(M) = \mathrm{PsAut}\,(M)$ for CY manifolds. Example 3 is an application of Theorem 2.6 with an explicit analysis of the movable cone [75].

The following theorem also shows a sharp contrast in dimension 2 and ≥ 3:

Theorem 5.3. *Let* $l \geq 2$ *and* $M = (2, \cdots, 2) \subset (\mathbf{P}^1)^{l+1} = \mathbf{P}^1_1 \times \mathbf{P}^1_2 \times \cdots \times \mathbf{P}^1_{l+1}$ *be a smooth generic hypersurface of multi-degree* $(2, \cdots, 2)$. *Then* M *is a K3 surface if* $l = 2$ *and a CY manifold of dimension* l *if* $l \geq 3$, *and:*

(1) *If* $l = 2$, *then* $\mathrm{Bir}\,(M) = \mathrm{Aut}\,(M) = \langle \iota_1, \iota_2, \iota_3 \rangle \simeq \mathbf{Z}_2 * \mathbf{Z}_2 * \mathbf{Z}_2$, *while*

(2) *If* $l \geq 3$, *then* $\mathrm{Aut}\,(M) = \{id_M\}$ *and* $\mathrm{Bir}\,(M) = \langle \iota_1, \cdots, \iota_{l+1} \rangle \simeq \mathbf{Z}_2 * \cdots * \mathbf{Z}_2$ *($(l+1)$-times free product).*

Here ι_k *is the covering involutions of the natural projection to the product* $(\mathbf{P}^1)^l$ *in which the* k*-th factor* $\mathbf{P}^1_k$ *of* $(\mathbf{P}^1)^{l+1}$ *removed. Moreover, there are (many) elements* f *with* $d_1(f) > 1$.

So, $\mathrm{Bir}\,(M)$ becomes larger and larger according to the dimension, but $\mathrm{Aut}\,(M)$ suddenly disappears in dimension ≥ 3. This is proved by Cantat and myself [23]. The essential algebro-geometric part of (2) is that in $l \geq 3$, the covering involutions ι_k are (no longer automorphism but) a birational involutions and at the same time all possible flops of M. We then apply Theorem 2.6.

Example 4 - CY manifold automorphisms of positive entropy. Only one "series" of examples with automorphisms of positive entropy that I know is:

Theorem 5.4. *Let* M *be the universal cover of the punctual Hilbert scheme* $\mathrm{Hilb}^l\,(S)$ *of length* $l \geq 2$ *of an Enriques surface* S. *Then* M *is a CY manifold of dimension* $2l$, *and* M *admits (many) biregular automorphisms of positive entropy if* S *is generic.*

See [77] and [23]. It is interesting to ask:

Question 5.5. *Does a CY manifold* M *in Theorem 5.4 admit a primitive automorphism of positive entropy?*

Higher dimensional rational manifold automorphisms of positive entropy. Finding "interesting" *biregular* automorphisms of rational manifolds seems much more difficult. Sur-

prisingly, the following most basic question, posed by Bedford, is still unsolved (See also [84] for many negative evidences):

Question 5.6. *Is there a biregular automorphism of positive entropy on a smooth rational threefold obtained by blowing-ups of* $\mathbf{P}^3$ *along smooth centers?*

On the other hand, in [9–11, 80], there are constructed many examples of rational manifolds with interesting pseudo-automorphisms, by generalizing constructions of rational surface automorphisms. Especially, the following result due to Bedford-Cantat-Kim [10] is quite remarkable and also strongly supports a negative answer for Question 5.6:

Theorem 5.7. *There is a smooth rational threefold* M*, obtained by blowing-ups of* $\mathbf{P}^3$ *at smooth centers, with* $f \in \mathrm{PsAut}\,(M)$ *such that* f *is primitive, of* $d_1(f) > 1$*, and* (M, f) *has no equivariant smooth birational model* (M', f') *with* $f' \in \mathrm{Aut}\,(M')$*.*

Their f does not preserve even any foliation. In their construction, $\pi : M \to \mathbf{P}^3$ is a partial resolution of indeterminacy of $\tau \circ \mathrm{cr}_3 \in \mathrm{Bir}\,(\mathbf{P}^3)$ and their iterates, for suitably chosen $\tau \in \mathrm{PGL}\,(4, \mathbf{C})$ with periodicity conditions of indeterminacy. Then the birational automorphism $f = \pi \circ (\tau \circ \mathrm{cr}_3) \circ \pi^{-1}$ becomes a pseudo-automorphism of M. In their construction, there is a rational surface $S \subset M$ preserved by f^4 in the sense that $f^4|S \in \mathrm{Bir}\,(S)$. Their crucial observation for the non-existence of (M', f') is that $d_1(f^4|S)$ is *neither* a Salem number *nor* 1. On the other hand, if f could be regularized, then so is $f^4|S$, possibly on other regularized models on which S survives. But, then $d_1(f^4|S)$ must be a Salem number or 1 by the birational invariance of the dynamical degrees (Theorem 3.9) and by Theorem 4.3.

5.2. First examples of rational and CY threefolds with primitive automorphisms of positive entropy. Main reference is [79]. The essential idea is the quotient construction from a manifold M with rich automorphisms: If $G < \mathrm{Aut}\,(M)$ is a "small" finite subgroup with "big" normalizer $N < \mathrm{Aut}\,(M)$, then the "big" group N/G acts biregularly on the quotient variety M/G and on its equivariant resolution as well.

Our actual construction is as follows. Let $E_\tau = \mathbf{C}/(\mathbf{Z} + \mathbf{Z}\tau)$ be the elliptic curve of period τ. There are exactly two elliptic curves with a Lie automorphism other than ± 1. They are $E_{\sqrt{-1}}$ and E_ω, where $\omega := (-1 + \sqrt{-3})/2$. Let X_4 (resp. X_6, resp. X_3) be the canonical resolutions of the quotient threefolds

$$E_{\sqrt{-1}} \times E_{\sqrt{-1}} \times E_{\sqrt{-1}}/\langle \sqrt{-1} I_3 \rangle \ , \ E_\omega \times E_\omega \times E_\omega/\langle -\omega I_3 \rangle \ , E_\omega \times E_\omega \times E_\omega/\langle \omega I_3 \rangle$$

i.e., the blow up at the maximal ideals of singular points. As is well known, X_3 is a CY threefold [3]. It is analytically rigid, but plays an important role in the classification of CY threefolds in the view of the second Chern class [66, 76]. Our X_3, X_6, X_4 provide the *first* examples of a Calabi-Yau threefold and smooth rational threefolds with *primitive biregular* automorphisms of *positive entropy*:

Theorem 5.8.

(1) *Both* X_6 *and* X_4 *are rational.*

(2) *Moreover,* X_3*,* X_6*,* X_4 *admit primitive biregular automorphisms of positive entropy.*

(1) is proved by Truong and myself for X_6 [79], and by Colliot-Thélène [26] for X_4 via [24] both answering a question of Ueno and Campana [17, 86]). (2) is proved by Truong and myself [79].

The most crucial parts are the rationality of X_6 and X_4 and finding primitive automorphisms. One of the key steps for the rationality is the following result ([79], [24]) shown via determination of the rational function fields:

Theorem 5.9. *Let (s,t,z,w) be the standard affine coordinates of $\mathbf{C}^4$. Then:*

(1) *X_4 is birational to the hypersurface H_4 in $\mathbf{C}^4$ defined by*

$$(t^2-z)(s^2-w^3)=(s^2-w)(t^2-z^3)\ .$$

(2) *X_6 is birational to the hypersurface H_6 in $\mathbf{C}^4$ defined by*

$$(w^3-1)(t^2-1)=(z^3-1)(s^2-1)\ .$$

The projection $p_{34}:(t,s,z,w)\mapsto(z,w)$ gives the conic bundle structures on H_4 and H_6; $p_{34}: H_4\to\mathbf{C}^2$, $p_{34}: H_6\to\mathbf{C}^2$. It is clear that $(t,s,z,w)=(1,1,z,w)$ is a section of $p_{34}: H_6\to\mathbf{C}^2$, and therefore H_6 is rational. $p_{34}: H_4\to\mathbf{C}^2$ does not admit a rational section. However, Colliot-Thélène [26] shows that the conic bundle $p_{34}: H_4\to\mathbf{C}^2$ is birational to the conic bundle $p_{34}: (H_4)'\to\mathbf{C}^2$ over the same base. Here $(H_4)'$ is the affine hypersurface defined by $t^2-zs^2-w=0$. This process is not explicit, but a consequence of the fact that these two conic bundles define the same element of the Brauer group $\mathrm{Br}\,(\mathbf{C}(z,w))$ of the base space $\mathbf{C}^2$ [26]. $(H_4)'$ is rational as $w=t^2-zs^2$, whence so is H_4.

Let us give an example of primitive biregular automorphisms of positive entropy of X_3, X_6, X_4. Let us consider the matrix

$$P=P_a=\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -1 & 3a^2 & 0\end{pmatrix}\ ,$$

where a is an integer such that $a\geq 2$. Since $\det P=1$, P naturally defines Lie automorphisms $g_3=g_6$, g_4 of $(E_\omega)^3$ and $(E_{\sqrt{-1}})^3$. By the construction and by the universality of blowing-up, the automorphisms g_k descends to the *biregular* automorphisms, say f_k, of X_k. The eigenvalues of α, β, γ of P are real numbers with $|\alpha|>|\gamma|>1>|\beta|$. By using α, β, γ, one can compute that $d_1(g_k)=\alpha^2$ and $d_2(g_k)=\alpha^2\gamma^2$. Thus $d_2(g_k)>d_1(g_k)>1$. Then $d_2(f_k)>d_1(f_k)>1$ by Corollary 3.13. Hence f_k provide desired automorphisms by Corollary 3.14 (2).

Question 5.10. *It is interesting to connect X_6 and X_4 to $\mathbf{P}^3$ by explicit blowing-ups and blowing-downs along smooth centers. This is in principle possible by [1]. In the view of Question 5.6, it is quite interesting to see if one can obtain X_6 and/or X_4 only by blowing-ups of $\mathbf{P}^3$ along smooth centers or not.*

6. Birational automorphisms of HK manifolds

6.1. Some generalities. We assume some familiarity with basics on HK manifolds. Excellent references are [41], Part III before Verbitsky's Torelli theorem [87] and [56] after that. We only recall that any HK manifold admits a non-degenerate integral symmetric bilinear

form, called Beauville-Bogomolov form (BB-form, for short), $b_M(*,**) : H^2(M,\mathbf{Z}) \times H^2(M,\mathbf{Z}) \to \mathbf{Z}$ with signature $(3, b_2(M)-3)$, being compatible with Hodge decomposition and invariant under deformation and $\mathrm{Bir}\,(M)$. We denote by $S^{[n]} = \mathrm{Hilb}^n(S)$ the Hilbert scheme of the 0-dimensional closed subschemes of lengths $n \geq 2$ on a K3 surface S. $S^{[n]}$ is a HK manifold of dimension $2n$ and of $\rho(S^{[n]}) = \rho(S)+1$. Through the Hilbert-Chow morphism $S^{[n]} \to \mathrm{Sym}^n S$, we have a natural identification as $\mathbf{Z}$-modules, $H^2(S^{[n]},\mathbf{Z}) = H^2(S,\mathbf{Z}) \oplus \mathbf{Z}e$, where $e = [E]/2$, the half of the exceptional divisor E of the Hilbert-Chow morphism. Under the BB-form $b(*,**)$, the above isomorphism is also an isometry with $(e^2) = -2(n-1)$.

6.2. Rough structure theorem on birational automorphisms. Let M be a HK manifold of $\dim M = 2n$. We have the following Tits' alternatives:

Theorem 6.1. *Let $G < \mathrm{Bir}\,(M)$. Then:*

(1) *For each (M,G), either one of the following two holds:*

  (i) *G is an almost abelian group of rank r, i.e., G is isomorphic to $\mathbf{Z}^r$ ($r \geq 0$) up to finite kernel and cokernel, or*

  (ii) *G is essentially non-commutative, i.e., G contains a free subgroup $\mathbf{Z} * \mathbf{Z}$.*

(2) *(ii) happens only if M is projective and $\rho(M) \geq 3$. Moreover, in case (ii), there are (many) $f \in G$ such that $d_1(f) > 1$. In particular, if in addtion $G \subset \mathrm{Aut}\,(M)$, then there are (many) $f \in G$ with $h_{\mathrm{top}}(f) > 0$ in the case (ii).*

This is proved by [68–70]. The essential part is as follows. The natural representation $r : G \to \mathrm{GL}\,(H^2(M,\mathbf{Z}))$ has a finite kernel (Remark 2.8) for projective case and [44] for general case). Thus G is well approximated by the image $G^* := r(G)$. Then, the fundamental result of Tits (Theorem 6.2) reduces the problem to showing that if G^* is virtually solvable, then it is an almost abelian group of finite rank. This can be done by using the additional strong condition that G^* is a subgroup of $\mathrm{O}_{\mathrm{Hodge}}(H^2(M,\mathbf{Z}))$.

Theorem 6.2. *Any group $H < \mathrm{GL}\,(n,k)$ (k is a field of characteristic 0) satisfies either one of the following two:*

(1) *H has a solvable subgroup of finite index (virtually solvable), or*

(2) *H is essentially non-commutative, i.e., H contains $\mathbf{Z} * \mathbf{Z}$.*

Remark 6.3. Tits' alternative type result with the same form as in Theorem 6.1 does not hold in general. However, some meaningful different formulation is proposed and proved for the biregular automorphism group of any compact Kähler manifold ([50, 91], see also [29, 32]).

One can also compute the dynamical degrees and entropy [71]:

Theorem 6.4. *For any $f \in \mathrm{Aut}\,(M)$ of any HK manifold M, the dynamical degrees $d_k(f)$ are all Salem numbers or 1. More precisely, $d_{2n-k}(f) = d_k(f) = d_1(f)^k$ for all $0 \leq k \leq n = \dim M/2$. In particular, if $d_1(f) > 1$, then*

$$1 = d_0(f) < d_1(f) < \cdots < d_{n-1}(f) < d_n(f) > d_{n+1}(f) > \cdots > d_{2n}(f) = 1\ ,$$

and $h_{\mathrm{top}}(f) = n \log\, d_1(f) > 0$ (resp. 0) if $d_1(f) > 1$ (resp. $d_1(f) = 1$).

6.3. A few examples. By the definition of $S^{[n]}$, we have a natural inclusion $\mathrm{Aut}\,(S) \subset \mathrm{Aut}\,(S^{[n]})$. This shows that if $\mathrm{Aut}\,(S)$ is infinite, then so is $\mathrm{Aut}\,(S^{[n]})$. So, contrary to the case of CY manifolds and rational manifolds, there are many examples of HK manifold with many biregular automorphisms.

Example 1 - Non-projective primitive automorphism of positive entropy. Let (S, f) be as in Theorem 4.5. Set $M = S^{[n]}$ and denote by $f_M \in \mathrm{Aut}\,(M)$ the automorphism naturally induced by f. We have $d_1(f_M) = d_1(f) = a$, the Salem number, and therefore $h_{\mathrm{top}}(f_M) = n \log\, a > 0$. Since S has no non-constant global meromorphic function (as $\rho(S) = 0$), the same is true for M. Then M has no rational fibration [18]. Hence f_M is primitive as well. We also see that $\mathrm{Bir}\,(M) = \mathrm{Aut}\,(M) = \langle f_M \rangle \simeq \mathrm{Aut}\,(S) \simeq \mathbf{Z}$.

Example 2 - The case where Picard number 2. We have the following:

Theorem 6.5.

(1) *Let S be a projective K3 surface with $\rho(S) = 1$. Then, $\rho(S^{[n]}) = 2$ but $\mathrm{Bir}\,(S^{[n]})$ is a finite group.*

(2) *There is a projective HK fourfold M deformation equivalent to $S^{[2]}$ such that $\rho(M) = 2$, $\mathrm{Aut}\,(M) = \mathrm{Bir}\,(M)$ is almost abelian group of rank 1 with element of positive entropy. More specifically, M with $\mathrm{NS}\,(M) \simeq (\mathbf{Z}[\eta], 4\mathrm{Nm}(*))$, "the same Néron-Severi lattice as Cayley's K3 surface", gives such an example. In particular, 2 is the minimal Picard number of projective HK manfolds of dimension ≥ 4 with automorphism of positive entoropy (cf. Remark 2.8).*

(1) is observed by [75]. Unlike Cayley's K3 surfaces, our M in (2) is highly non-constructible. *However, it is likely true that M in (2) has a primitive automorphism of positive entropy (not yet settled).*

Example 3 - Projective HK manifold of Picard number 3. Let $S \subset \mathbf{P}^3$ be a smooth quartic surface. Then for two general points P, Q in S, the line PQ in $\mathbf{P}^3$ meets S in four points, say, P, Q, P', Q'. The correspondence $\{P, Q\} \mapsto \{P', Q'\}$ defines a birational automorphism ι_S of $S^{[2]}$ of order 2, called the *Beauville involution* [3]. If S has no line, then ι_S is biregular. Note that $\iota_S \in \mathrm{Bir}\,(S^{[2]}) \setminus \mathrm{Aut}\,(S)$ under $\mathrm{Aut}\,(S) \subset \mathrm{Aut}\,(S^{[2]})$.

Let S be a Cayley's K3 surface. Identifying $S = S_0 \subset \mathbf{P}^3$, our S has three different embeddings $\Phi_k : S \to S_k \subset \mathbf{P}^3$ ($k = 0, 1, 2$) under the notation in Example 3 in Subsection (4.2). Let ι_k be the Beauville involution with respect to the embedding Φ_k. We have the following theorem similar to Theorem 5.3:

Theorem 6.6. *Let S be a Cayley's K3 surface. Then,*

$$\mathrm{Bir}\,(S^{[2]}) = \mathrm{Aut}\,(S^{[2]}) = \langle \iota_0, \iota_1, \iota_2 \rangle \text{ and } g = \iota_0 \circ \iota_1 \circ \iota_2\ ,$$

under the natural inclusion $\langle g \rangle = \mathrm{Aut}\,(S) \subset \mathrm{Aut}\,(S^{[2]})$. Moreover, $\mathrm{Aut}\,(S^{[2]})$ has a subgroup isomorphic to the free product $\mathbf{Z} * \mathbf{Z}$, hence admits an automorphism of positive entoropy (Theorem 6.1). In particular, $[\mathrm{Aut}\,(S^{[2]}) : \mathrm{Aut}\,(S)] = \infty$ and 3 is the minimal Picard number of projective HK manfolds of dimension ≥ 4 with essentially non-commutative automorphism group.

One of interesting fact is that we have the second factorization $g = \iota_0 \circ \iota_1 \circ \iota_2$ in $S^{[2]}$, which looks similar to , but completely different from, the factorization that Cayley found

in $\mathbf{P}^3$ (Example 3 in Subsection (4.2)). Another interesting fact is that $\mathrm{Aut}\,(S^{[2]})$ becomes much bigger than $\mathrm{Aut}\,(S)$ in this example, which makes a sharp contrast to the following open question, called the naturality question, posed by Boissière and Sarti [14, 15]):

Question 6.7. *Is* $\mathrm{Aut}\,(S) = \mathrm{Aut}\,(S^{[m]})$ *under the natural inclusion for* $m \geq 3$*?*

Acknowledgements. First of all, I would like to express my sincere thanks to Professor Yujiro Kawamata for his continuous, warm encouragement, support and proper advices, both in mathematics and in life, since I was his graduate student on 1987. I would like to express my thanks to all my collaborators and teachers, especially Professors Tuyen Truong, De-Qi Zhang, Jun-Muk Hwang and Professors Fabrizio Catanese, Akira Fujiki, Heisuke Hironaka, JongHae Keum, Nessim Sibony, Tetsuji Shioda, Shing-Tung Yau, and Late Professors Eiji Horikawa, Eckart Viehweg, and to Professors Serge Cantat, Tien-Cuong Dinh for several valuable comments. *It is my great honor to dedicate this note to Professor Doctor Thomas Peternell on the occasion of his sixtieth birthday. He continues to inspire me through many interesting problems with his brilliant ideas since 1993*.

References

[1] Abramovich, D., Karu, K., Matsuki, K., and Wlodarczyk, J., *Torification and factorization of birational maps*, J. Amer. Math. Soc. **15** (2002), 531–572.

[2] Barth, W., Hulek, K., Peters, C., and Van de Ven, A., *Compact complex surfaces*, Springer-Verlag, Berlin, 2004.

[3] Beauville, A., *Some remarks on Kähler manifolds with* $c_1 = 0$*,* 1–26, Progr. Math., **39** Birkhäuser 1983.

[4] ———, *Variétés Kähleriennes dont la première classe de Chern est nulle,* J. Differential Geom. **18** (1983), 755–782 (1984).

[5] ———, *Determinantal hypersurfaces*, Michigan Math. J. **48** (2000), 39–64.

[6] Bedford, E. and Kim, K.-H., *Periodicities in linear fractional recurrences: Degree growth of birational surface maps*, Michigan Math. J. **54** (2006), 647–670.

[7] ———, *Dynamics of rational surface automorphisms: linear fractional recurrences,* J. Geom. Anal. **19**. (2009), 553–583.

[8] ———, *Dynamics of rational surface automorphisms: rotation domains,* Amer. J. Math. **134** (2012), 379–405.

[9] ———, *Dynamics of (pseudo) automorphisms of 3-space: periodicity versus positive entropy*, Publ. Mat. **58** (2014), 65–119.

[10] Bedford, E., Cantat, S., and Kim, K.-H., *Pseudo-automorphisms with no invariant foliation*, arXiv:1309.3695.

[11] Bedford, E., Diller,, J., and Kim, K.-H., *Pseudoautomorphisms with invariant elliptic curves*, arXiv:1401.2386.

[12] Bhargava, M., Ho, W., and Kumar, A., *Orbit Parametrizations for K3 Surfaces*, arXiv:1312.0898.

[13] Birkar, C., Cascini, P., Hacon, C. D., and McKernan, J., *Existence of minimal models for varieties of log general type*, J. Amer. Math. Soc. **23** (2010), 405–468.

[14] Boissière, S., *Automorphismes naturels de l'espace de Douady de points sur une surface*, Canad. J. Math. **64** (2012), 3–23.

[15] Boissière, S., Nieper-Wisskirchen, M., and Sarti, A., *Higher dimensional Enriques varieties and automorphisms of generalized Kummer varieties*, J. Math. Pures Appl. (9) **95** (2011), 553–563.

[16] Borcea, C., *Homogeneous vector bundles and families of Calabi-Yau threefolds. II.*, Proc. Sympos. Pure Math. **52** Part 2, Amer. Math. Soc. (1991), 83–91.

[17] Campana, F., *Remarks on an example of K. Ueno*, Series of congress reports, Classification of algebraic varieties, European Math. Soc., (2011), 115–121.

[18] Campana, F., Oguiso, K., and Peternell, Th., *Non-algebraic hyperkähler manifolds*, J. Differential Geom. **85** (2010), 397–424.

[19] Cantat, S., *Dynamique des automorphismes des surfaces projectives complexes,* C. R. Acad. Sci. Paris Sér. I Math.**328** (1999), 901–906.

[20] ———, *Dynamique du groupe d'automorphismes des surfaces K3*, Transform. Groups **6** (2001), 201–214.

[21] ———, *Morphisms between Cremona groups and a characterization of rational varieties*, to appear in Compositio Math., avaiable at his HP.

[22] Cantat, S., Chambert-Loir, A., and Guedj, V., *Quelques aspects des systèmes dynamiques polynomiaux*, Panoramas et Synthèses, 30, S.F.M. (2010).

[23] Cantat, S. and Oguiso, K., *Birational automorphism groups and the movable cone theorem for Calabi-Yau manifolds of Wehler type via universal Coxeter groups*, arXiv:1107.5862.

[24] Catanese, F., Oguiso, K., and Truong, T.-T., *Unirationality of Ueno-Campana's threefold*, arXiv:1310.3569.

[25] Cayley, A., *A memoir on quartic surfaces,* Proc. London. Math. Soc. **3**, (1869–71), 19–69.

[26] Colliot-Thélène, J.-L. *Rationalité d'un fibré en coniques*, arXiv:1310.5402.

[27] Diller, J. and Favre, C., *Dynamics of bimeromorphic maps of surfaces,* Amer. J. Math. **123** (2001), 1135–1169.

[28] Dimitrov, G., Haiden, F., Katzarkov, L., and Kontsevich, M., *Dynamical systems and categories*, arXiv:1307.8418.

[29] Dinh, T.-C., *Tits alternative for automorphism groups of compact Kähler manifolds*, Acta Math. Vietnam. **37** (2012), 513–529.

[30] Dinh, T.-C. and Nguyên V.-A., *Comparison of dynamical degrees for semi-conjugate meromorphic maps,* Comment. Math. Helv. **86** (2011), 817–840.

[31] Dinh, T.-C., Nguyên V.-A., and Truong, T.-T., *On the dynamical degrees of meromorphic maps preserving a fibration,* Commun. Contemp. Math. **14** (2012), 18 pp, arXiv:1108.4792.

[32] Dinh, T.-C. and Sibony, N., *Groupes commutatifs d'automorphismes d'une variété kählérienne compacte*, Duke Math. J. **123** (2004), 311–328.

[33] _____, *Une borne supérieure de l'entropie topologique d'une application rationnelle,* Ann. of Math., **161** (2005), 1637–1644.

[34] _____, *Green currents for holomorphic automorphisms of compact Kähler manifolds,* J. Amer. Math. Soc. **18** (2005), 291–312.

[35] Dolgachev, I. V., *Lectures on Cremona transformations*, avaiable at his HP.

[36] Festi, D., Garbagnati, A., van Geemen, B., and van Luijk, R., *The Cayley-Oguiso automorphism of positive entropy on a K3 surface*, J. Mod. Dyn. **7**, (2013) 75–97, arXiv:1208.1016.

[37] Fu, B. and Zhang, D.-Q., *A characterization of compact complex tori via automorphism groups*, Math. Ann. **357** (2013), 961–968.

[38] Fujiki, A., *On primitively symplectic compact Kähler V -manifolds of dimension four*, 71–250, Progr. Math., **39**, Birkhäuser 1983.

[39] Gromov, M., *On the entropy of holomorphic maps,* Enseign. Math., **49** (2003), 217–235.

[40] _____, *Entropy, homology and semialgebraic geometry*, Astérisque **145**–**146**, (1987) 225–240.

[41] Gross M., Huybrechts D., and Joyce D., *Calabi-Yau manifolds and related geometries*, Universitext. Springer-Verlag, 2003.

[42] Guedj, V. *Entropie topologique des applications méromorphes*, Ergodic Theory Dynam. Systems **25**, (2005) 1847–1855.

[43] Hacon, C. D. and McKernan, J., *Flips and flops*, Proceedings of the International Congress of Mathematicians. Volume II, pp. 513–539, 2010.

[44] Huybrechts, D., *Compact hyper-Kähler manifolds: basic results*, Invent. Math. **135** (1999) 63–113, and Erratum **152** (2003), 209–212.

[45] Katok, A. and Hasselblatt, B., *Introduction to the modern theory of dynamical systems*, Cambridge University Press, 1995.

[46] Kawamata, Y., *Characterization of abelian varieties*, Compositio Math. **43** (1981), 253–276.

[47] _____, *Abundance theorem for minimal threefolds*, Invent. Math. **108** (1992), 229–246.

[48] ———, *Flops connect minimal models,* Publ. Res. Inst. Math. Sci. **44** (2008) 419–423.

[49] Kawamata, Y., Matsuda, K., and Matsuki, K., *Introduction to the minimal model problem,* 283–360, Adv. Stud. Pure Math., **10**, 1987.

[50] Keum, J.-H., Oguiso, K., and Zhang, D.-Q., *Conjecture of Tits type for complex varieties and theorem of Lie-Kolchin type for a cone*, Math. Res. Lett. **16** (2009), 133–148.

[51] Kollár, J., *Rational curves on algebraic varieties*, Springer-Verlag, 1996.

[52] Kollár J. and Mori, S., *Birational geometry of algebraic varieties*, Cambridge University Press, 1998.

[53] Lazić, V., Oguiso, K., and Peternell, Th., *Automorphisms of Calabi-Yau threefolds with Picard number three*, to appear in Adv. Stud. Pure Math., arXiv:1310.8151.

[54] Lazić, V. and Peternell, Th., *On the Cone conjecture for Calabi-Yau manifolds with Picard number two*, to appear in Math. Res. Lett., arXiv:1207.3653.

[55] Lesieutre, J., *Derived-equivalent rational threefolds*, arXiv:1311.0056.

[56] Markman, E., *A survey of Torelli and monodromy results for holomorphic-symplectic varieties*, pp. 257–322, Springer Proc. Math., **8**, Springer, 2011, arXiv:1101.4606.

[57] McMullen, C. T., *Coxeter groups, Salem numbers and the Hilbert metric*, Publ. Math. Inst. Hautes Études Sci. **95** (2002), 151–183.

[58] ———, *Dynamics on K3 surfaces: Salem numbers and Siegel disks,* J. Reine Angew. Math. **545** (2002), 201–233.

[59] ———, *Dynamics on blowups of the projective plane,* Publ. Math. Inst. Hautes Études Sci. **105** (2007), 49–89.

[60] ———, *K3 surfaces, entropy and glue,* J. Reine Angew. Math. **658** (2011), 1–25.

[61] ———, *Automorphisms of projective K3 surfaces with minimum entropy,* preprint, avaiable at his HP.

[62] Mori, S., *Flip theorem and the existence of minimal models for 3 -folds*, J. Amer. Math. Soc. **1** (1988), 117–253.

[63] Nagata, M., *On rational surfaces, I*, Mem. Coll. Sei. Kyoto (A) **33** (1960), 352–370.

[64] Nakayama, N., *Zariski-Decomposition and Abundance*, MSJ Memoirs, **14**, Math. Soc. Japan, 2004.

[65] Nakayama, N. and Zhang, D.-Q., *Building blocks of étale endomorphisms of complex projective manifolds,* Proc. Lond. Math. Soc. **99** (2009), 725–756.

[66] Oguiso, K., *On algebraic fiber space structures on a Calabi-Yau 3-fold*, Internat. J. Math. **4** (1993), 439–465.

[67] ———, *Local families of K3 surfaces and applications*, J. Algebraic Geom. **12** (2003), 405–433.

[68] ———, *Tits alternative in hypekähler manifolds*, Math. Res. Lett. **13** (2006), 307–316.

[69] ———, *Automorphisms of hyperkähler manifolds in the view of topological entropy*, Contemp. Math., **422** (2007), 173–185.

[70] ———, *Bimeromorphic automorphism groups of non-projective hyperkähler manifolds – a note inspired by C. T. McMullen*, J. Differential Geom. **78** (2008) 163–191.

[71] ———, *A remark on dynamical degrees of automorphisms of hyperkähler manifolds,* Manuscripta Math. 130 (2009), 101–111.

[72] ———, *Salem polynomials and the bimeromorphic automorphism group of a hyper-Kähler manifold,* Amer. Math. Soc. Transl. Ser. 2, **230**, Amer. Math. Soc. (2010), 201–227.

[73] ———, *The third smallest Salem number in automorphisms of K3 surfaces,* 331–360, Adv. Stud. Pure Math., **60**, 2010.

[74] ———, *Free automorphisms of positive entropy on smooth Kähler surfaces,* to appear in Adv. Stud. Pure Math., arXiv:1202.2637.

[75] ———, *Automorphism groups of Calabi-Yau manifolds of Picard number two,* to appear in J. Algebraic Geom., arXiv:1206.1649.

[76] Oguiso, K. and Sakurai, J., *Calabi-Yau threefolds of quotient type,* Asian J. Math. **5** (2001), 43–77.

[77] Oguiso, K. and Schröer, S., *Enriques manifolds*, J. Reine Angew. Math. **661** (2011), 215–235.

[78] Oguiso, K. and Truong, T.-T., *Salem numbers in dynamics of Kähler threefolds and complex tori*, to appear in Math. Zeit., arXiv:1309.4851.

[79] ———, *Explicit Examples of rational and Calabi-Yau threefolds with primitive automorphisms of positive entropy*, arXiv:1306.1590.

[80] Perroni, F. and Zhang, D.-Q., *Pseudo-automorphisms of positive entropy on the blowups of products of projective spaces,* to appear in Math. Ann., arXiv:1111.3546.

[81] Reschke, P., *Salem numbers and automorphisms of complex surfaces,* Math. Res. Lett., **19** (2012), 475–482.

[82] Shioda, T. and Inose, H., *On singular K3 surfaces*, Complex analysis and algebraic geometry, pp. 119–136, Iwanami Shoten, Tokyo, 1977.

[83] Tits, J., *Free subgroups in linear groups*, J. Algebra **20** (1972), 250–270.

[84] Truong, T.-T., *On automorphisms of blowups of* $\mathbb{P}^3$*,* arXiv:1202.4224.

[85] Uehara, T., *Rational surface automorphisms with positive entropy,* arXiv:1009.2143.

[86] Ueno, K., *Classification theory of algebraic varieties and compact complex spaces,* Lecture Notes in Mathematics, **439**, Springer-Verlag, 1975.

[87] Verbitsky, M., *A global Torelli theorem for hyperkähler manifolds*, Duke Math. J. **162** (2013), 2929–2986. arXiv:0908.4121.

[88] Voisin, C., *On the homotopy types of compact Kähler and complex projective manifolds,* Invent. Math., **157** (2004) 329–343.

[89] Yomdin, Y., *Volume growth and entropy,* Israel J. Math. **57** (1987), 285–300.

[90] Zhang, D.-Q., *Dynamics of automorphisms on projective complex manifolds*, J. Differential Geom. **82** (2009), 691–722.

[91] _____, *A theorem of Tits type for compact Kähler manifolds*, Invent. Math. **176** (2009), 449–459.

Department of Mathematics, Osaka University, Toyonaka 560-0043, Osaka, Japan and Korea Institute for Advanced Study, Hoegiro 87, Seoul, 133-722, Korea

E-mail: oguiso@math.sci.osaka-u.ac.jp

Local mirror symmetry in the tropics

Mark Gross and Bernd Siebert

Abstract. We discuss how the reconstruction theorem of [20] applies to local mirror symmetry [11]. This theorem associates to certain combinatorial data a degeneration of (log) Calabi-Yau varieties. While in this case most of the subtleties of the construction are absent, an important normalization condition already introduces rich geometry. This condition guarantees the parameters of the construction are canonical coordinates in the sense of mirror symmetry. The normalization condition is also related to a count of holomorphic disks and cylinders, as conjectured in [20] and partially proved in [7–9]. We sketch a possible alternative proof of these counts via logarithmic Gromov-Witten theory.

There is also a surprisingly simple interpretation via rooted trees marked by monomials, which points to an underlying rich algebraic structure both in the relevant period integrals and the counting of holomorphic disks.

Mathematics Subject Classification (2010). Primary 14J33; Secondary 14J32, 14T05.

Keywords. local mirror symmetry, tropical curves, Gross-Siebert program.

1. Introduction

In [18, 20], we proposed a mirror construction as follows. We begin with a polarized degenerating flat family $\mathcal{X} \to T = \operatorname{Spec} R$ of n-dimensional Calabi-Yau varieties where R is a complete local ring. We consider only degenerations of a special sort which we term *toric degenerations*, see [18], Def. 4.1. Roughly, these are degenerations for which the central fibre is a union of toric varieties glued along toric strata, and such that the map $\mathcal{X} \to T$ is locally given by a monomial near the zero-dimensional strata of the central fibre X_0. Associated to this degeneration we construct the *dual intersection complex* $(B, \mathscr{P}, \varphi)$, where

(a) B is an n-dimensional integral affine manifold with singularities (possibly with boundary). In other words, B is a topological manifold with an open subset B_0 with $\Delta := B \setminus B_0$ of codimension ≥ 2, such that B_0 has an atlas of coordinate charts whose transition maps lie in $\operatorname{Aff}(\mathbb{Z}^n)$, the group of integral affine transformations.

(b) $\mathscr{P}$ is a decomposition of B into convex lattice polyhedra (possibly unbounded). The singular locus Δ is typically the union of codimension two cells of the first barycentric subdivision of $\mathscr{P}$ not intersecting the interior of a maximal cell of $\mathscr{P}$ nor containing a vertex of $\mathscr{P}$. There is a one-to-one inclusion reversing correspondence between elements of $\mathscr{P}$ and toric strata of X_0. The local structure of $\mathscr{P}$ near a vertex is determined by the fan defining the corresponding irreducible component. The maximal cells of $\mathscr{P}$ are determined by the toric structure of the map $\mathcal{X} \to T$ near the corresponding zero-dimensional strata of X_0.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

(c) φ is a *multi-valued piecewise affine function*. This is a collection $\{(U_i, \varphi_i)\}$ of $\mathbb{R}$-valued functions φ_i on an open cover $\{U_i\}$ of B, with each φ_i piecewise affine linear with respect to the polyhedral decomposition $\mathscr{P}$, and $\varphi_i - \varphi_j$ being affine linear on $U_i \cap U_j$. We assume the slopes of the φ_i on cells of $\mathscr{P}$ to be integral. In this case, φ is determined by the polarization on $\mathcal{X}$, with local representatives near vertices given by a piecewise linear function defined by restricting the polarization to the corresponding irreducible component.

Given this data, we obtain the mirror to the degeneration $\mathcal{X} \to T$ by reinterpreting $(B, \mathscr{P}, \varphi)$ as the *intersection complex* of another polarized toric degeneration $\mathcal{Y} \to \operatorname{Spec} \Bbbk[\![t]\!]$ (in the projective case). This time, there is a one-to-one inclusion preserving correspondence between cells of $\mathscr{P}$ and toric strata of X_0, the central fibre of this new degeneration. The cells of $\mathscr{P}$ are the Newton polytopes for the polarization restricted to the various strata of X_0, and φ is determined by the local toric structure of the map near zero-dimensional strata.

The prime difficulty in the program lies in reconstructing $\mathcal{Y} \to \operatorname{Spec} \Bbbk[\![t]\!]$ from the data $(B, \mathscr{P}, \varphi)$. The main result of [20] gives an algorithm for constructing a *structure* $\mathscr{S}$ of walls which tell us how to construct the degeneration.

More recently [15] has considered families constructed using the technology of [20] over higher dimensional base schemes. This represents a modification of the above procedure. In the typical example, instead of choosing a fixed polarization on $\mathcal{Y}$, one chooses a monoid P of polarizations. Let $Q = \operatorname{Hom}(P, \mathbb{N})$ be the dual monoid. Then this data determines a multi-valued piecewise linear function φ taking values in $Q^{\mathrm{gp}}_{\mathbb{R}} := Q^{\mathrm{gp}} \otimes_{\mathbb{Z}} \mathbb{R}$. If $\mathfrak{m}$ is the maximal monomial ideal of $\Bbbk[Q]$, and $\widehat{\Bbbk[Q]}$ denotes the completion of $\Bbbk[Q]$ with respect to this ideal, then the construction gives a family $\mathcal{Y} \to \operatorname{Spec} \widehat{\Bbbk[Q]}$.

The history of the problem of associating a geometric object (complex manifold, non-Archimedean space, toric degeneration...) to an integral affine manifold with singularities began with work of Fukaya [12]. Fukaya gave a heuristic suggesting that one should be able to construct the mirror to a K3 surface using objects that look like structures in two dimensions (in two dimensions, we can think of a structure as just consisting of a possibly infinite number of unbounded rays). Fukaya observed that holomorphic disks with boundary on fibres of an SYZ fibration ([27]) gave similar pictures of structures on the mirror side. In 2004, Kontsevich and Soibelman in [26] gave the first construction of a structure, showing how given a two-dimensional affine sphere with singularities one could construct a consistent structure and from this structure a non-Archimedean K3 surface. We combined the picture of toric degenerations we had been developing independently of the above-mentioned authors with some ideas from [26], allowing us to construct degenerations from structures in all dimensions in [20].

In the first two sections of this paper, we shall illustrate the program by carrying it out completely for toric Calabi-Yau manifolds, a case usually referred to as local mirror symmetry [11]. This particular case can be viewed as being complementary to the case that the ideas of [26] were able to handle. In the remaining sections, we shall analyze enumerative meaning and a tropical interpretation of this construction.

2. Degenerations of toric Calabi-Yau varieties

Our running example is the construction of the mirror of what is called "local $\mathbb{P}^2$", the total space X of the canonical bundle $K_{\mathbb{P}^2}$ over $\mathbb{P}^2$. Since X itself is a toric variety, its anti-

canonical divisor $-K_X$ is linearly equivalent to the sum of toric divisors. There are four toric divisors, the zero section $S \subset X$, which is the maximal compact subvariety of X, and the preimages F_0, F_1, F_2 of the three coordinate lines in $\mathbb{P}^2$ under the bundle projection $X \to \mathbb{P}^2$. Toric methods show that $S + F_0 + F_1 + F_2 \sim 0$ and hence X is a non-compact Calabi-Yau threefold. The normal bundle $N_{S|X} = \mathcal{O}_{\mathbb{P}^2}(-3)$ is determined by the adjunction formula from the Calabi-Yau condition and it is the dual of an ample line bundle. Hence, by a result of Grauert [13], any embedded $\mathbb{P}^2$ in a Calabi-Yau threefold has an analytic neighbourhood biholomorphic to an analytic neighbourhood of S in X.

For the general description, fix throughout $M = \mathbb{Z}^n$, $M_{\mathbb{R}} = M \otimes_{\mathbb{Z}} \mathbb{R}$, $N = \mathrm{Hom}_{\mathbb{Z}}(M, \mathbb{Z})$.[1] Let $\sigma \subseteq M_{\mathbb{R}}$ be a compact lattice polytope, and assume $0 \in \sigma$. Define

$$C(\sigma) = \{(rm, r) \,|\, m \in \sigma, r \in \mathbb{R}_{\geq 0}\} \subseteq M_{\mathbb{R}} \oplus \mathbb{R}.$$

The cone $C(\sigma)$ viewed as a fan defines an affine toric variety X_σ. A polyhedral decomposition $\overline{\mathscr{P}}$ of σ into standard simplices leads to a fan $\Sigma = \{C(\tau) \,|\, \tau \in \overline{\mathscr{P}}\}$ which is a refinement of $C(\sigma)$. This yields a toric resolution of singularities $X_\Sigma \to X_\sigma$. Assume also that the fan Σ supports at least one strictly convex piecewise linear function.

For the case of local $\mathbb{P}^2$ take $n = 2$ and $\sigma = \mathrm{Conv}\{(1,0),(0,1),(-1,-1)\}$, where $\mathrm{Conv}(S)$ denotes the convex hull of the set S. Then the dual cone $C(\sigma)^\vee$ is $C(\sigma^*)$, the cone over the polar polytope σ^* with vertices $(-1,-1)$, $(2,-1)$, $(-1,2)$. It turns out that $X_\sigma = \mathrm{Spec}(\mathbb{C}[C(\sigma)^\vee \cap \mathbb{Z}^3])$ is the cyclic quotient $\mathbb{A}^2/\mathbb{Z}_3$ with $\mathbb{Z}_3$ acting diagonally on the coordinates by multiplication with third roots of unity. Taking the polyhedral decomposition as shown in Figure 2.1 yields for X_Σ the blowing up of the origin of X_σ. One can show that X_Σ is the total space of $K_{\mathbb{P}^2}$ and the map to X_σ is the contraction of the zero section. Note also that the projection $C(\sigma) \to M_{\mathbb{R}}$ defines a map from Σ to the fan of $\mathbb{P}^2$, which indeed corresponds to the bundle projection $X_\Sigma \to \mathbb{P}^2$.

In general, the map $X_\Sigma \to X_\sigma$ has a reducible exceptional locus, with one component for each vertex of $\overline{\mathscr{P}}$ that is not a vertex of σ, and the explicit description of the geometry is more complicated.

It turns out that constructing a mirror to X_Σ does not fit well with our program. The reason is that X_Σ does not seem to possess a fibration by Lagrangian tori of the kind expected by mirror symmetry [16]. Rather, such a fibration will exist only after removal of a hypersurface in X_Σ that is disjoint from the exceptional fibre of $X_\Sigma \to X_\sigma$. To run our program we could give an ad hoc construction of an affine manifold with singularities derived from the fan Σ or write down a toric degeneration of X_Σ. The local $\mathbb{P}^2$ case has been discussed from the former point of view in [21], Examples 5.1 and 5.2. Since it can be done easily in the present case we follow the latter method here. This method is motivated by the construction of toric degenerations of hypersurfaces in toric varieties in [17].

To exhibit X_Σ as an anticanonical hypersurface in a toric variety we embed the fan Σ in $M_{\mathbb{R}} \oplus \mathbb{R}$ as a subfan of a fan $\tilde{\Sigma}$ in $M_{\mathbb{R}} \oplus \mathbb{R}^2$. For each maximal cone $C \in \Sigma$ the fan $\tilde{\Sigma}$ has two maximal cones

$$C_1 = C \times 0 + \mathbb{R}_{\geq 0} \cdot (0,1,-1), \quad C_2 = C \times 0 + \mathbb{R}_{\geq 0} \cdot (0,0,1).$$

Then Σ is the subfan of $\tilde{\Sigma}$ consisting of cones lying in the hyperplane $M_{\mathbb{R}} \oplus \mathbb{R} \oplus 0 \subset M_{\mathbb{R}} \oplus \mathbb{R}^2$. The fan $\tilde{\Sigma}$ only has two rays not contained in Σ, with generators $(0,0,1)$ and $(0,1,-1)$. The inclusion $M_{\mathbb{R}} \oplus \mathbb{R} \oplus 0 \subset M_{\mathbb{R}} \oplus \mathbb{R}^2$ induces a map of fans from Σ to $\tilde{\Sigma}$, hence an

[1] Since M, N will eventually be treated as data for the mirror side our conventions in this section are opposite to the usual ones in toric geometry.

embedding $j : X_\Sigma \hookrightarrow X_{\tilde\Sigma}$ identifying X_Σ with the closure of the orbit of the subtorus defined by this inclusion through the distinguished point (the unit of the toric variety). Note that the projection to $\mathbb{R}^2$ maps $\tilde\Sigma$ to the fan $\Sigma_{\widehat{\mathbb{A}}^2}$ of the toric blowing up $\widehat{\mathbb{A}}^2$ of $\mathbb{A}^2$, with rays generated by $(0,1),(1,0),(1,-1)$. Under this map the subfan $\Sigma \subset \tilde\Sigma$ maps to the interior ray $\mathbb{R}_{\geq 0}\cdot(1,0)$. Viewing this interior ray as giving a map of fans, from the one-dimensional fan defining $\mathbb{A}^1$ to the two-dimensional fan defining $\widehat{\mathbb{A}}^2$, we obtain an embedding $i : \mathbb{A}^1 \hookrightarrow \widehat{\mathbb{A}}^2$.

We thus obtain a cartesian diagram of toric morphisms

$$\begin{array}{ccc} X_\Sigma & \xrightarrow{\ j\ } & X_{\tilde\Sigma} \\ {\scriptstyle p}\downarrow & & \downarrow{\scriptstyle q} \\ \mathbb{A}^1 & \xrightarrow{\ i\ } & \widehat{\mathbb{A}}^2 \end{array}$$

The left vertical arrow is induced by the projection $M_\mathbb{R} \oplus \mathbb{R} \to \mathbb{R}$, hence is given by the pull-back to X_Σ of the distinguished monomial x on X_σ defining the toric boundary as a reduced subscheme.

Explicitly, write x, y for the toric coordinates on $\mathbb{A}^2$ and $\widehat{\mathbb{A}}^2 = (xu - yv = 0) \subset \mathbb{A}^2 \times \mathbb{P}^1$ for the blowing up. Then $\mathrm{im}(i)$ is the strict transform of the diagonal $x = y$. Dehomogenizing $u = 1$ or $v = 1$ we obtain the usual two coordinate patches with coordinates y, v and x, u respectively with the transitions $v = u^{-1}$ and $x = yv$ or $y = xu$. We use the same notation for the pull-back of x, y, u, v to the corresponding two types of affine patches with $u \neq 0$ or $v \neq 0$ of $X_{\tilde\Sigma}$.

To describe $X_{\tilde\Sigma}$ let $C \in \Sigma$ be a maximal cone. Then if $(m, a) \in N \oplus \mathbb{Z}$ defines a facet $C' \subset C$, that is, (m, a) generates an extremal ray of $C^\vee$, the element $(m, a, a) \in N \oplus \mathbb{Z}^2$ defines the facet $C' + \mathbb{R}_{\geq 0}(0, 1, -1)$ of C_1. There is only one more facet of C_1, namely C itself, defined by $(0, 0, -1)$, and hence

$$C_1^\vee = \{(m, a, a) \,|\, (m, a) \in C^\vee\} + \mathbb{R}_{\geq 0} \cdot (0, 0, -1).$$

The rays of $C_2^\vee$ are generated by $(m, a, 0)$ for (m, a) an extremal ray of $C^\vee$, and by $(0, 0, 1)$, so $C_2^\vee = C^\vee \times 0 + \mathbb{R}_{\geq 0}(0, 0, 1)$. In either case, we have an identification

$$\mathrm{Spec}\, \Bbbk[C_i^\vee \cap (N \oplus \mathbb{Z}^2)] = \mathrm{Spec}\, \Bbbk[C^\vee \cap N] \times \mathbb{A}^1 \subset X_\Sigma \times \mathbb{A}^1.$$

The toric coordinate for $\mathbb{A}^1$ is $v = z^{(0,0,-1)}$ for C_1 and $u = z^{(0,0,1)}$ for C_2. From this description it is clear that the embedding of X_Σ in $X_{\tilde\Sigma}$ is given by $u = 1$ in affine patches with $v \neq 0$ and by $v = 1$ in the affine patches with $u \neq 0$.

To write down a degeneration of X_Σ to the toric boundary $\partial X_{\tilde\Sigma} \subset X_{\tilde\Sigma}$ view u, v as sections of the line bundle $q^*\mathcal{O}(-E)$ where $E \subset \widehat{\mathbb{A}}^2$ is the exceptional curve. Then X_Σ is the zero locus of $s := u - v$. On the other hand, $xu = yv$ defines a section s_0 of $q^*\mathcal{O}(-E)$ with zero locus $\partial X_{\tilde\Sigma}$. Thus the hypersurface $\mathcal{X} \subset X_{\tilde\Sigma} \times \mathbb{A}^1$ with equation

$$ts + s_0 = 0$$

defines a pencil in $X_{\tilde\Sigma}$ with members X_Σ at $t = \infty$ and with the toric boundary $\partial X_{\tilde\Sigma}$ at $t = 0$. Note this pencil is the preimage of the pencil on $\widehat{\mathbb{A}}^2$ defined by the same equations. In particular, by direct computation $\mathcal{X}_t$ is completely contained in either type of coordinate patch for $t \neq 0$. Working in a patch with $v \neq 0$ we have $s = u - 1$, $s_0 = xu$ and the equation

$$0 = ts + s_0 = t(u - 1) + xu = u(t + x) - t$$

shows $u(t+x) = t \neq 0$. Thus $t + x \neq 0$ and u can be eliminated. In other words, $\mathcal{X}_t \simeq X_\Sigma \setminus Z_t$ with $Z_t \subset X_\Sigma$ the hypersurface $x = -t$. Note also that our notation is consistent in that x indeed descends to the defining equation of the toric boundary of X_σ.

It is not difficult to show that $\mathcal{X} \to \mathbb{A}^1$ is a toric degeneration. Indeed, we have already checked that $\mathcal{X}_0$ is the toric boundary of $X_{\tilde{\Sigma}}$. Some harder work shows that locally near the zero-dimensional strata of X_0, the projection $\mathcal{X} \to \mathbb{A}^1$ is toric. We omit the details, but this can be done similarly to arguments given in [17].

The dual intersection complex is then easily described along the lines given in [17], where, for a Calabi-Yau hypersurface in a toric variety, B was described as the boundary of a reflexive polytope, with the cones over the faces of the polytope yielding the fan defining the ambient toric variety. Topologically, we can write $B \subseteq M_\mathbb{R} \oplus \mathbb{R}^2$ as

$$B = \tilde{\sigma}_1 \cup \tilde{\sigma}_2$$

where

$$\begin{aligned} \tilde{\sigma}_1 &= \operatorname{Conv}\big((0,1,-1) \cup (\sigma \times \{(1,0)\})\big), \\ \tilde{\sigma}_2 &= \operatorname{Conv}\big((0,0,1) \cup (\sigma \times \{(1,0)\})\big). \end{aligned}$$

Note that the support of the fan $\tilde{\Sigma}$ above is the cone over $\tilde{\sigma}_1 \cup \tilde{\sigma}_2$. We then take $\mathscr{P} = \{C \cap B \,|\, C \in \tilde{\Sigma}\}$.

Finally, the affine structure on B is defined as follows. Identify σ with $\sigma \times \{(1,0)\} \subseteq B$, and take the discriminant locus Δ to be the union of cells of the first barycentric subdivision of $\overline{\mathscr{P}}$ not containing vertices of $\overline{\mathscr{P}}$, see Figure 2.1. We then define affine charts as follows. First, we define affine charts $\iota_i : \tilde{\sigma}_i \setminus \sigma \hookrightarrow \mathbb{A}_i$ as the inclusions, where $\mathbb{A}_i$ denotes the affine hyperplane in $M_\mathbb{R} \oplus \mathbb{R}^2$ spanned by $\tilde{\sigma}_i$. Second, for each vertex $v \in \overline{\mathscr{P}}$, choose a neighbourhood U_v of $(v,1,0) \in B$. These neighbourhoods can be chosen so that $U_v \cap U_{v'} = \emptyset$ if $v \neq v'$ and the two sets $\tilde{\sigma}_i \setminus \sigma$ along with the open sets U_v cover $B \setminus \Delta$. Define a chart $\iota_v : U_v \to (M_\mathbb{R} \oplus \mathbb{R}^2)/\mathbb{R}(v,1,0)$ via the inclusion followed by the projection. It is easy to check that these charts give an integral affine structure. This again precisely follows the procedure for Calabi-Yau hypersurfaces in toric varieties considered in [17]. This gives rise to the pair $(B, \mathscr{P})$.

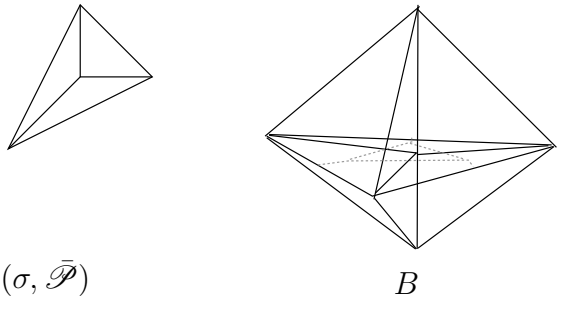


Figure 2.1. On the left is the initial polytope with its decomposition. We take the vertices of σ to be $(1,0)$, $(0,1)$ and $(-1,-1)$. On the right is the resulting B with its discriminant locus Δ, indicated by the dotted line.

In general, a pair $(B, \mathscr{P})$ can be described by specifying the lattice polytopes in $\mathscr{P}$ and specifying a *fan structure* at each vertex v, that is, the identification of a neighbourhood of

each vertex with the neighbourhood of 0 in a fan Σ_v. This identification gives a one-to-one inclusion preserving correspondence between cells of $\mathscr{P}$ containing v and cones of Σ_v, along with integral affine identifications of the tangent wedges of each cell $\tau \in \mathscr{P}$ containing v with the corresponding cone of Σ_v. These identifications patch together to give an affine chart in a neighbourhood of the vertex v.

In our example, it is worth describing the fan structure at a vertex $v \in \sigma$. Since the fan structure at a vertex must be the fan yielding the corresponding irreducible component of $\mathcal{X}_0$, toric geometry tells us this fan structure must be given as the quotient fan obtained from $\tilde{\Sigma}$ by dividing out by the ray generated by this vertex. Explicitly, we use the chart

$$\iota_v : U_v \to (M_{\mathbb{R}} \oplus \mathbb{R}^2)/\mathbb{R} \cdot (v, 1, 0) \cong M_{\mathbb{R}} \oplus \mathbb{R}, \tag{2.1}$$

the latter isomorphism given by $(m, r_1, r_2) \mapsto (m - r_1 v, r_2)$. The fan Σ_v can then be described as the fan of tangent wedges to images of cells $\tau \in \mathscr{P}$ containing v. The set of maximal cones of this fan, described as subsets of $M_{\mathbb{R}} \oplus \mathbb{R}$, is

$$\{T_v\tau + \mathbb{R}_{\geq 0}(-v, -1) \,|\, v \in \tau \in \overline{\mathscr{P}}_{\max}\} \cup \{T_v\tau + \mathbb{R}_{\geq 0}(0, 1) \,|\, v \in \tau \in \overline{\mathscr{P}}_{\max}\}, \tag{2.2}$$

where $T_v\tau$ denotes the tangent wedge to $v \in \tau$ in $M_{\mathbb{R}} \oplus 0$. Figure 2.2 shows some of the fan structures when σ is an interval $[-1, 1]$ of length two.

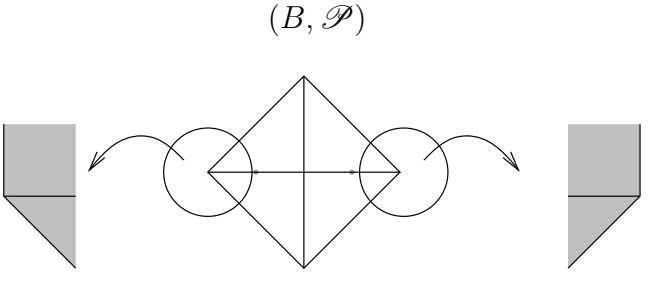


Figure 2.2.

One can understand the nature of the singularities of B by studying the local system Λ of integral vector fields on B_0. Given integral affine coordinates $y_1, \ldots, y_n$, Λ is locally the family of lattices in the tangent bundle of B_0 generated by $\partial/\partial y_1, \ldots, \partial/\partial y_n$. If $v, v' \in \overline{\mathscr{P}}$ are adjacent vertices, consider a path γ passing from v through $\check{\sigma}_1$ to v' and then through $\check{\sigma}_2$ back to v. To identify Λ_v, we can use the chart (2.1), which gives an identification of Λ_v with $M \oplus \mathbb{Z}$. It is an easy exercise to check that parallel transport in Λ around γ yields a *monodromy transformation*

$$\begin{aligned} T_{vv'} : \Lambda_v &\to \Lambda_v \\ (m, r) &\mapsto (m + r(v - v'), r) \end{aligned} \tag{2.3}$$

The final piece of data for the dual intersection complex $(B, \mathscr{P}, \varphi)$ of $\mathcal{X} \to \mathbb{A}^1$ is a multi-valued piecewise linear function φ describing aspects of the Kähler geometry of the situation.

In the next section, we will build the mirror family over some base scheme. The natural choice for this base is related to the Kähler cone of X_Σ, or the Picard group. By toric

geometry, $\mathrm{Pic}(X_\Sigma)$ equals piecewise linear functions on Σ modulo linear functions. It will thus be convenient to normalize the piecewise linear functions as follows. Choose a maximal cell $\tau \in \overline{\mathscr{P}}_{\max}$ which has 0 as a vertex. Let P be the monoid of integral convex piecewise linear functions on the fan Σ which take the value 0 on the cone $C(\tau)$. Note that $P^{\mathrm{gp}} \cong \mathrm{Pic}\, X_\Sigma$. Setting $Q := \mathrm{Hom}(P, \mathbb{N})$, there is a universal piecewise linear function $\psi : |\Sigma| \to Q^{\mathrm{gp}}_{\mathbb{R}}$, with

$$\psi(x) = (P \ni \varphi \mapsto \varphi(x)).$$

This function is strictly convex in the sense of [14], Definition 1.12. In the local $\mathbb{P}^2$ case normalized piecewise linear functions are determined by the value at the one remaining vertex of σ not contained in τ and hence $Q = \mathbb{N}$.

The multi-valued piecewise linear function φ comes from the universal piecewise linear function ψ on $|\Sigma|$ by descent to a quotient fan, or rather from a choice of extension of this function to $\widetilde{\Sigma}$. This choice can be made by choosing an element $q \in Q \setminus \{0\}$. While the choice of q affects the family of polarizations on $X_{\tilde{\Sigma}}$, it does not affect the family after restriction to $\mathcal{X}_t$ for $t \neq 0$. However, it does affect the polarization on $\mathcal{X}$, and hence will play some role in the mirror, seen explicitly in (3.7). We take $\widetilde{\psi}$ to be the $Q^{\mathrm{gp}}_{\mathbb{R}}$-valued piecewise linear extension of ψ which takes the value 0 at $(0, 1, -1)$ and the value q at $(0, 0, 1)$. One can check that this function is strictly convex in the sense of [14], Definition 1.12.

We can then construct φ from $\widetilde{\psi}$ as follows. For each $C \in \tilde{\Sigma}$, let $\tau = C \cap B$ be the corresponding cell of $\mathscr{P}$. The function $\widetilde{\psi}$ induces a function on the quotient fan of $\widetilde{\Sigma}$ along C (this quotient fan determining the fan structure of B along τ) as follows. Let $\widetilde{\psi}_\tau \in \mathrm{Hom}(M \oplus \mathbb{Z}^2, Q)$ be a linear extension of $\widetilde{\psi}|_C$. Then $\widetilde{\psi} - \widetilde{\psi}_\tau$ is a piecewise affine function on $\tilde{\Sigma}$ vanishing on C, hence descending to the quotient fan of $\tilde{\Sigma}$ along C. We take $(\widetilde{\psi} - \widetilde{\psi}_\tau)|_B$ as a representative of φ on a small open neighbourhood of $\mathrm{Int}(\tau)$ in B; this is clearly the pull-back of the corresponding function on the quotient fan of $\tilde{\Sigma}$ along C under the projection to $(M_{\mathbb{R}} \oplus \mathbb{R}^2)/\mathbb{R}C$.

3. The mirror degeneration and slab functions

Having described $(B, \mathscr{P}, \varphi)$ in our example as the dual intersection complex of a degeneration of the local Calabi-Yau X_Σ, we turn to the construction of the mirror, which shall be a family $\mathcal{Y} \to \mathrm{Spec}\, \widehat{\Bbbk[Q]}$ over a generally higher-dimensional base.

This family is constructed by constructing families $\mathcal{Y}_k \to \mathrm{Spec}\, \Bbbk[Q]/\mathfrak{m}^{k+1}$ to each order k, giving rise to a formal scheme $\widehat{\mathcal{Y}} \to \mathrm{Spf}\, \widehat{\Bbbk[Q]}$. As the case at hand will be projective, the Grothendieck existence theorem gives rise to the desired family. Alternatively, $\mathcal{Y}$ can be constructed using a graded ring of theta functions, following [15].

Here is a brief summary of the construction. The central fibre Y_0 can be described as

$$Y_0 = \bigcup_{\sigma \in \mathscr{P}_{\max}} \mathbb{P}_\sigma$$

where $\mathscr{P}_{\max}$ denotes the maximal cells of $\mathscr{P}$ and $\mathbb{P}_\sigma$ is the toric variety (projective if σ is compact) determined by the polyhedron σ. These toric varieties are glued together in a manner reflecting the combinatorics of $\mathscr{P}$: if $\sigma_1 \cap \sigma_2 = \tau$, then the strata $\mathbb{P}_\tau \subseteq \mathbb{P}_{\sigma_1}$, $\mathbb{P}_\tau \subseteq \mathbb{P}_{\sigma_2}$ are identified.

Local models for the k^{th} order deformation of Y_0 are determined by the function φ. A key point of the construction involves an invariant description for the local models, which we explain here. The function φ, defined on an open cover $\{U_i\}$ by single-valued functions $\varphi_i : U_i \to Q^{\mathrm{gp}}_{\mathbb{R}}$, determines an extension of Λ by $\underline{Q}^{\mathrm{gp}}$, the constant sheaf with coefficients in Q^{gp}. Indeed, on $U_i \cap B_0$, this extension will split as $\Lambda|_{U_i} \oplus \underline{Q}^{\mathrm{gp}}$, and on the overlap, (m, r) as a section of $\Lambda|_{U_i} \oplus \underline{Q}^{\mathrm{gp}}$ is identified on $U_i \cap U_j$ with $(m, r + d(\varphi_j - \varphi_i)(m))$ as a section of $\Lambda|_{U_j} \oplus \underline{Q}^{\mathrm{gp}}$, interpreting $d(\varphi_j - \varphi_i) \in \mathrm{Hom}(\Lambda|_{U_i}, Q^{\mathrm{gp}})$. We then have an exact sequence

$$0 \to \underline{Q}^{\mathrm{gp}} \to \mathcal{P} \to \Lambda \to 0 \tag{3.1}$$

on B_0. We write the map $\mathcal{P} \to \Lambda$ as $m \mapsto \bar{m}$. After choosing a representative φ_i of φ in a neighbourhood of a point $x \in B_0$, the stalk $\mathcal{P}_x$ is identified with $\Lambda_x \oplus Q^{\mathrm{gp}}$. There is a fan $\Sigma_x = \{T_x\sigma \,|\, x \in \sigma \in \mathscr{P}\}$ (of not-necessarily strictly convex cones), where $T_x\sigma$ denotes the tangent wedge to σ at x. This allows us to define a convex PL function $\varphi_x : |\Sigma_x| \to Q^{\mathrm{gp}}_{\mathbb{R}}$ whose slope on $T_x\sigma$ coincides with the slope of $\varphi_i|_\sigma$. We then set

$$P_x := \{(m, q) \,|\, m \in \Lambda_x \cap |\Sigma_x|, q \in Q^{\mathrm{gp}}, q - \varphi_x(m) \in Q\} \subseteq \mathcal{P}_x \tag{3.2}$$

While this definition as described inside of $\Lambda_x \oplus Q$ depends on the choice of representative, in fact it is independent of this choice when viewed as a submonoid of $\mathcal{P}_x$.

Note that Q acts naturally on P_x, giving $\Bbbk[P_x]$ a $\Bbbk[Q]$-algebra structure. For a vertex v, we can now view $\mathrm{Spec}\,\Bbbk[P_v]/\mathfrak{m}^{k+1}$ as a local model for the k^{th} order deformation of Y_0 in a neighbourhood of the stratum of Y_0 corresponding to v. In addition, the local system $\mathcal{P}$ gives a method of defining parallel transport of monomials.

Let us describe certain aspects of this construction for our local mirror symmetry example. Using the fan structure given by (2.2), we can describe the monoid $P_v \subseteq \mathcal{P}_v$ as $\{(m, r, s) \,|\, s - \varphi_v(m, r) \in Q\} \subseteq M \oplus \mathbb{Z} \oplus Q^{\mathrm{gp}}$ using the identifications

$$\mathcal{P}_v \cong \Lambda_v \oplus Q^{\mathrm{gp}} \cong M \oplus \mathbb{Z} \oplus Q^{\mathrm{gp}} \tag{3.3}$$

induced first by the representative φ_v at v and second by the affine coordinate chart on U_v. In particular, for the purposes of the discussion below, we can describe the most relevant part of P_v as follows. First, we choose the representative φ_v by choosing the linear function $\widetilde{\psi}_v$ to be $(0, \bar{\psi}(v), 0) \in (N \oplus \mathbb{Z}^2) \otimes_{\mathbb{Z}} Q^{\mathrm{gp}}$, with $\bar{\psi} = \psi|_{\sigma \times \{1\}}$. Let $\bar{P}_v \subseteq P_v$ be the submonoid consisting of $m \in P_v$ with $\bar{m}$ tangent to σ. Then $\bar{P}_v$ is naturally described in terms of $\bar{\psi}$. Indeed, consider the convex hull of the graph of $\bar{\psi}$,

$$\Xi_{\bar{\psi}} := \{(m, 0, s) \,|\, m \in \sigma, s - \bar{\psi}(m) \in Q\} \subseteq M_{\mathbb{R}} \oplus \mathbb{R} \oplus Q^{\mathrm{gp}}_{\mathbb{R}},$$

an unbounded polyhedron with vertices mapping to vertices of $\overline{\mathscr{P}}$ under the projection $M_{\mathbb{R}} \oplus \mathbb{R} \oplus Q^{\mathrm{gp}}_{\mathbb{R}} \to M_{\mathbb{R}}$. Then we can identify $\bar{P}_v$ with the integral points in the tangent wedge of $\Xi_{\bar{\psi}}$ at $(v, 0, \bar{\psi}(v))$.

We also note that under the identification (3.3) of $\mathcal{P}_v$, the monodromy of Λ described in (2.3) lifts to a monodromy transformation of $\mathcal{P}_v$ given by

$$\begin{aligned} &T_{vv'} : \mathcal{P}_v \to \mathcal{P}_v \\ &(m, r, q) \to (m + r(v - v'), r, q + r(\bar{\psi}(v) - \bar{\psi}(v'))) \end{aligned} \tag{3.4}$$

The key additional (and usually most complex) ingredient for constructing $\mathcal{Y}_k$ is a *structure* $\mathscr{S}$. A structure encodes data about how certain forms of these local models are glued

together. We will explain this structure in our example, but not go into too much detail. A more detailed explanation for how this works is given in the expository paper [21].

The structure takes a particularly simple form here. In general, a structure is a collection of walls, polyhedral cells in B of codimension one each contained in a cell of $\mathscr{P}$ carrying the additional data of certain formal power series. In [20] we distinguish a special sort of wall, namely those contained in codimension one cells, and call them *slabs*. They tend to have a different behaviour. In the case at hand, only slabs appear, and these cover σ. The functions attached to the slabs are determined from the monodromy around the discriminant locus Δ.

In this example, the slabs are the sets $\tau \times \{(1,0)\}$ for $\tau \in \overline{\mathscr{P}}_{\max}$. For a slab $\mathfrak{b}$, associated to any point $x \in \mathfrak{b} \setminus \Delta$ is a formal power series $f_{\mathfrak{b},x} = \sum_{m \in P_x} c_m z^m$. This should only depend on the connected component of $\mathfrak{b} \setminus \Delta$ containing x, so there is in fact one such expression for each vertex v of τ, and we can write $f_{\mathfrak{b},v} = \sum_{m \in P_v} c_m z^m$. Furthermore, $c_m \neq 0$ implies $\bar{m}$ is tangent to $\mathfrak{b}$, so in fact the sum is over $m \in \bar{P}_v$.

The series $f_{\mathfrak{b},v}$ are completely determined by a number of simple properties. This follows in the case under consideration from having chosen $\overline{\mathscr{P}}$ to consist of *standard* simplices. In what follows we will want to compare $f_{\mathfrak{b},v}$ with $f_{\mathfrak{b},v'}$ for different vertices v, v' of $\mathfrak{b}$. To do so, we use parallel transport in $\mathcal{P}$ from v to v'. Given the identification $\mathcal{P}_v$ with $M \oplus \mathbb{Z} \oplus Q^{\mathrm{gp}}$ used to give the formula (3.4) and noting that only monomials of the form $z^{(m,0,p)}$ can appear in $f_{\mathfrak{b},v}$, we see that the particular path chosen between v and v' is irrelevant.

We can now state the conditions determining the $f_{\mathfrak{b},v}$:

1. The constant term of each $f_{\mathfrak{b},v}$ is 1.
2. If v and v' are adjacent vertices of $\mathfrak{b}$, then the corresponding slab functions are related by
$$f_{\mathfrak{b},v'} = z^{(v-v',0,\bar{\psi}(v)-\bar{\psi}(v'))} f_{\mathfrak{b},v}. \tag{3.5}$$
Here the equality makes sense after parallel transport of the exponents from v to v' in the local system $\mathcal{P}$, and $(v - v', 0, \bar{\psi}(v) - \bar{\psi}(v')) \in P_{v'}$ using the identification of $\mathcal{P}_{v'}$ given by (3.3).
3. $\log f_v$ contains no terms of the form z^q for $q \in Q \setminus \{0\}$. Here we view $Q \subseteq P_v$ via the natural inclusion $Q^{\mathrm{gp}} \subseteq \mathcal{P}_v$.
4. If v lies in slabs $\mathfrak{b}, \mathfrak{b}'$, then $f_{\mathfrak{b},v} = f_{\mathfrak{b}',v}$.

Item 1 is a normalization which originated in [18], Def. 4.23. However, we shall see its enumerative importance in §4. Item 2 is the crucial point of slabs: they allow us to define parallel transport of monomials through slabs in a way which cancels the effects of monodromy. We shall say more about this shortly. The condition 3 is interpreted by writing $f_v = 1 + \cdots$ and using the Taylor expansion for $\log(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} x^i / i$. This can be interpreted inside some suitably completed ring. After expanding out each expression $(\cdots)^i$, one demands that no monomials of the form z^q appear for any $q \in Q \setminus \{0\}$. Finally, 4 tells us how expressions propagate across $\sigma \times \{1\}$.

To see the significance of the second condition, let $w \in \mathrm{Int}(\tilde{\sigma}_1)$, $w' \in \mathrm{Int}(\tilde{\sigma}_2)$. Suppose we want to compare monomials defined at w (that is, monomials with exponent in P_w) with monomials defined at w' (that is, monomials with exponent in $P_{w'}$). If we parallel transport from $\mathcal{P}_w$ to $\mathcal{P}_{w'}$, the result depends on the path. For example, let v, v' be adjacent vertices of $\tau \in \overline{\mathscr{P}}_{\max}$. Let $T_v, T_{v'}$ denote parallel transport in Λ from w to w' via the vertices v and

v' respectively. Then from (3.4), it follows that for $(m,r,q) \in \mathcal{P}_w = M \oplus \mathbb{Z} \oplus Q$,

$$T_{v'}(m,r,q) - T_v(m,r,q) = \big(r(v-v'), 0, r(\bar\psi(v) - \bar\psi(v'))\big).$$

For convenience, we can identify $\mathcal{P}_w$ and $\mathcal{P}_{w'}$ with $\mathcal{P}_v$ so that T_v is the identity. This difference between T_v and $T_{v'}$ creates problems for comparing the rings $\Bbbk[P_w]$ and $\Bbbk[P_{w'}]$. However, we can follow the rule that if we wish to transport a monomial $z^{(m,r,q)}$ along a path between w and w' which crosses a slab $\mathfrak{b}$ in a connected component of $\mathfrak{b} \setminus \Delta$ containing a vertex v, we apply an automorphism

$$z^{(m,r,q)} \mapsto z^{(m,r,q)} f_{\mathfrak{b},v}^{-r}. \tag{3.6}$$

Here r represents the result of projecting $\overline{(m,r,q)} = (m,r)$ via the projection $\pi : \Lambda_v \to \mathbb{Z}$ obtained by dividing out by the tangent space to the slab. If instead we pass through the slab $\mathfrak{b}$ via the connected component of $\mathfrak{b} \setminus \Delta$ containing v', we get

$$z^{(m,r,q)} \longmapsto z^{(m+r(v-v'),r,q+r(\bar\psi(v)-\bar\psi(v')))} f_{v'}^{-r} = z^{(m,r,q)} f_v^{-r},$$

coinciding with (3.6). Here we use the above expression for $T_v - T_{v'}$ and (3.5). Hence we see that the ambiguity produced by monodromy is resolved by the slab functions.

Examples 3.1. In the following examples, we express the various functions $f_{\mathfrak{b},v}$ as formal power series with exponents appearing in $\bar P_v$, using the representation of $\bar P_v$ as the integral points of the tangent wedge of $\Xi_{\bar\psi}$ at $(v,0,\bar\psi(v))$.

(1) Take σ to be the interval $[-1,1]$ as in Figure 2.2, with $\mathscr{P}$ as given there. The monoid of convex piecewise linear functions on Σ is generated by the function which takes the values 0, 0 and 1 respectively at $(-1,1)$, $(0,1)$ and $(1,1)$. Thus we have $Q = \mathbb{N}$, and the universal piecewise linear function ψ coincides with the above generator. For a vertex v, with $\bar P_v \subseteq M \oplus 0 \oplus Q^{\mathrm{gp}}$, write $x = z^{(1,0,0)}$, $t = z^{(0,0,1)}$, t being the generator of $\Bbbk[Q]$. Then we have

$$\begin{aligned} f_{[-1,0],-1} &= 1 + x + x^2t + xt, \\ f_{[-1,0],0} = f_{[0,1],0} &= 1 + x^{-1} + xt + t, \\ f_{[0,1],1} &= 1 + x^{-1}t^{-1} + x^{-2}t^{-1} + x^{-1}. \end{aligned}$$

Note that $\log f_{[-1,0],-1}$, $\log f_{[0,1],1}$ are clearly devoid of pure powers of t as any power, say, of $x+x^2t+xt$ clearly produces only terms with positive powers of x. On the other hand, $f_{[-1,0],0} = (1+x^{-1})(1+xt)$, and taking logs we get $\log(1+x^{-1})+\log(1+xt)$ which will again involve no pure t power. The t term in $f_{[0,1],0}$ was necessary to achieve this.

(2) Take σ to be as in Figure 2.1. Again, the monoid of convex piecewise linear functions on the fan Σ is generated by, say, the function taking the values 0 at $(0,0,1)$, $(1,0,1)$ and $(0,1,1)$ and the value 1 at $(-1,-1,1)$. So again $Q = \mathbb{N}$, with the universal function ψ agreeing with this generator. Writing $x = z^{(1,0,0,0)}$, $y = z^{(0,1,0,0)}$, $t = z^{(0,0,0,1)}$, it is easy to see that the terms of the slab function $f_{\mathfrak{b},(0,0)}$ (independent of $\mathfrak{b}$ by the fourth condition) required by conditions 1 and 2 are $1 + x + y + tx^{-1}y^{-1}$. The normalization condition forces us to add some additional terms:

$$f_{\mathfrak{b},(0,0)} = 1 + x + y + tx^{-1}y^{-1} + \sum_{k\geq 1} a_k t^k,$$

where the a_k are uniquely determined by the requirement that

$$\sum_{i=1}^{\infty}(-1)^{i+1}\frac{(x+y+tx^{-1}y^{-1}+\sum_{k\geq 1}a_kt^k)^i}{i}$$

contains no pure powers of t. This series in t begins as

$$-2t+5t^2-32t^3+286t^4-3038t^5+\cdots.$$

(3) Let σ be the convex hull of the points $(\pm 1,0)$, $(0,\pm 1)$ and take $\mathscr{P}$ to be the star subdivision at the origin. Now the monoid P of convex piecewise linear functions which are 0 on $(0,0,1)$, $(1,0,1)$ and $(0,1,1)$ is isomorphic to $\mathbb{N}^2$, determined by the values α_1,α_2 of the function at generators of the other two rays. Thus we can write $Q=\mathbb{N}^2$, $t_1=z^{(1,0)}\in\Bbbk[Q]$, $t_2=z^{(0,1)}\in\Bbbk[Q]$. Using x,y as defined in the previous example, one can check that for any slab $\mathfrak{b}$,

$$f_{\mathfrak{b},0}:=1+x+y+t_1x^{-1}+t_2y^{-1}+t_1+t_2+3t_1t_2+5t_1^2t_2+5t_1t_2^2+\cdots.$$

The additional terms represented by $\cdots$ give a power series in t_1,t_2.

We now describe the degeneration $\mathcal{Y}\to\operatorname{Spec}\widehat{\Bbbk[Q]}$ produced by the above data. In fact, it is not difficult to do this in terms of equations, as follows. First, define

$$C(\Xi_{\bar\psi}):=\overline{\{((um,0,uq,u)\,|\,(m,0,q)\in\Xi_{\bar\psi},\ u\in\mathbb{R}_{\geq 0}\}}\subseteq M_{\mathbb{R}}\oplus\mathbb{R}\oplus Q^{\mathrm{gp}}_{\mathbb{R}}\oplus\mathbb{R}.$$

Here the closure is necessary because $\Xi_{\bar\psi}$ is unbounded. We then obtain a graded ring

$$S_{\bar\psi}:=\Bbbk[C(\Xi_{\bar\psi})\cap(M\oplus\mathbb{Z}\oplus Q^{\mathrm{gp}}\oplus\mathbb{Z})]$$

where the grading is given by the projection from $M\oplus\mathbb{Z}\oplus Q^{\mathrm{gp}}\oplus\mathbb{Z}$ onto the last copy of $\mathbb{Z}$. Note the closure in the definition of cone adds the cone $\{0\}\times\{0\}\times\mathbb{R}_{\geq 0}Q\times\{0\}$ to the set, so we see the degree 0 part of $S_{\bar\psi}$ is $\Bbbk[Q]$. We can then complete, with

$$\widehat{S}_{\bar\psi}:=S_{\bar\psi}\otimes_{\Bbbk[Q]}\widehat{\Bbbk[Q]}.$$

It is then natural to think of the slab functions as being given by a single degree 1 element of $\widehat{S}_{\bar\psi}$. Indeed, given a vertex $v\in\overline{\mathscr{P}}$, we obtain from $f_{\mathfrak{b},v}$ an element of degree 1 by multiplying all monomials of $f_{\mathfrak{b},v}$ by $z^{(v,0,\bar\psi(v),1)}$. It follows from (3.5) that this is independent of the choice of v and gives an element $F\in\widehat{S}_{\bar\psi}$ of degree 1. One can then show that

$$\mathcal{Y}=\operatorname{Proj}\widehat{S}_{\bar\psi}[U,W]/(UW-z^qV_0F). \tag{3.7}$$

Here U,W are of degree 1, $V_0\in\widehat{S}_{\bar\psi}$ is the element corresponding to $(0,0,0,1)$ (which lies in $\Xi_{\bar\psi}$ by the assumption that $0\in\sigma$ and ψ has been chosen so that $\bar\psi(0)=0$). The element $q\in Q$ is the element chosen in the definition of $\widetilde{\psi}$ at the end of §2. This can be shown in much the way the special case discussed in [21], Example 5.2, being the case of Examples 3.1, (2). Note that after localizing at z^q, this family does not depend on the choice of q up to isomorphism, just as the choice of q did not affect the polarization on the general fibres of $\mathcal{X}\to\mathbb{A}^1$.

The homogeneous coordinate ring of $\mathcal{Y}$ is generated in degree 1 by *theta functions*, as explored in [15]. Each point of $B(\mathbb{Z})$ (the set of points of B with integral coordinates) corresponds to a generator of this ring as a $\widehat{\Bbbk[Q]}$-algebra. Explicitly, the integral points in this example are the integral points of σ and the apexes of the pyramids $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$. If v is an integral point of σ, then $z^{(v,0,\bar{\psi}(v),1)} \in \widehat{S}_{\bar{\psi}}$ is the corresponding theta function. On the other hand, the monomials U and W correspond to the two apexes.

This description of $\mathcal{Y}$ can be related to the more traditional mirror to X_Σ as described in [11]. Here $\mathcal{Y}$ can be decompactified by setting $V_0 = 1$, obtaining an open subset $\mathcal{Y}^o$. Passing to the generic fibre $\mathcal{Y}^o_\eta$ of $\mathcal{Y}^o \to \operatorname{Spec} \widehat{\Bbbk[Q]}$, we obtain a variety defined over the field of fractions K of $\widehat{\Bbbk[Q]}$. We can describe $\mathcal{Y}^o$ as a subvariety of $\mathbb{A}^2 \times (N \otimes_{\mathbb{Z}} \mathbb{G}_m)$ over the field K given by the equation

$$uw = z^q f_{\mathfrak{b},0}, \tag{3.8}$$

where $\mathfrak{b}$ is any slab containing $0 \in \sigma$. Here u, w are coordinates on $\mathbb{A}^2$. Without the normalization condition, we could take $f_{\mathfrak{b},0} = \sum_{m \in \sigma \cap M} z^{(m,\bar{\psi}(m))}$, which would lead to the mirror of X_Σ being precisely that given in [11].

Remark 3.2. The crucial feature of the mirror family we have just described, as opposed to the one given in [11], is that the monomial coordinates on the base $\operatorname{Spec} \widehat{\Bbbk[Q]}$ are canonical in the sense of mirror symmetry. To describe this briefly, we work over the field $\Bbbk = \mathbb{C}$, and assume that the power series $f := f_{\mathfrak{b},0}$ is convergent in some analytic neighbourhood U of the zero-dimensional stratum in $\operatorname{Spec} \mathbb{C}[Q]$. Let $U^* = U \setminus \partial \operatorname{Spec} \mathbb{C}[Q]$, the complement of the union of toric divisors. Thus we can view $\mathcal{Y}^o$ as giving an analytic family $\mathcal{Y}^o \to U^*$. We write $\mathcal{Y}^o_t$ for the fibre over $t \in U^*$. On such a fibre, one has the holomorphic volume form on the fibres of $\mathcal{Y}^o \to \operatorname{Spec} \widehat{\Bbbk[Q]}$ given by

$$\Omega = (2\pi i)^{-n-1} d\log u \wedge d\log x_1 \wedge \cdots \wedge d\log x_n.$$

One then finds that there is a monodromy invariant cycle $\alpha_0 \in H_{n+1}(\mathcal{Y}^o_t, \mathbb{Z})$ such that $\int_{\alpha_0} \Omega = 1$, so that Ω is a normalized holomorphic form in the sense of mirror symmetry. Further, if $q_1, \ldots, q_r \in Q^{\mathrm{gp}}$ are a basis for Q^{gp}, one can find (multi-valued) flat families of $(n+1)$-cycles $\alpha_1, \ldots, \alpha_r$ with $\int_{\alpha_i} \Omega = \log z^{q_i}$. The key point of this calculation is to take the logarithmic derivative of these period integrals and reduce the resulting integral to an integral on the hypersurface $f_{\mathfrak{b},v} = 0$ in $N \otimes_{\mathbb{Z}} \mathbb{G}_m$. Via residues, this is translated into an integral of the derivative of $\log f_{\mathfrak{b},0}$ over various tori in $N \otimes_{\mathbb{Z}} \mathbb{G}_m$. The fact that these integrals are then constant follows precisely from the normalization condition on $f_{\mathfrak{b},0}$.

4. Enumerative predictions

So far we have seen two interpretations of the slab functions and the normalization condition. The first came from the desire to write down a correction to the patching of the naive toric models for the mirror degeneration $\mathcal{Y} \to \operatorname{Spec} \widehat{\Bbbk[Q]}$ in a way consistent with local monodromy of the affine structure on B. We discussed in §2 how this condition along with the normalization condition determines the slab functions uniquely. Then in Remark 3.2 we saw that the normalization condition is responsible for making our families canonically parametrized in the sense of mirror symmetry. Both of these arguments concern the *complex geometry* of the mirror degeneration $\mathcal{Y} \to \operatorname{Spec} \widehat{\Bbbk[Q]}$.

In the following two sections we will give two related interpretations of normalized slab functions related to the *symplectic geometry* of the degeneration $\mathcal{X} \to \mathbb{A}^1$ of the local Calabi-Yau variety X_Σ we started with. The interpretation supports the view that the degenerations constructed by structures are indeed the ones expected from homological mirror symmetry and open-closed string theory.

Since the completion of [20], a clearer idea emerged as to the precise meaning of structures. This picture has arisen from several converging points of view: (1) The heuristic correspondence between tropical Morse trees and Floer homology emerging in discussions between us and Mohammed Abouzaid. Some of these ideas were discussed in [3] and [22]. (2) Auroux's work [4] on T-duality on complements of anti-canonical divisors, describing the complex structure on the SYZ dual of a Lagrangian fibration using counts of Maslov index two disks. This has inspired quite a bit of work, which is realising Fukaya's original dream of correcting the complex structure of the mirror via counts of holomorphic disks. (3) [24] made explicit the enumerative content of the key part of the algorithm of [26] (or the two-dimensional version of [20]). In particular, this established an enumerative meaning for functions attached to walls of a structure.

Heuristically, one expects the following interpretation in the SYZ picture of mirror symmetry. Suppose given a (special) Lagrangian fibration $f : X \to B$ from a Calabi-Yau X, with the general fibre being a torus. Consider Maslov index zero holomorphic disks with boundary a fibre of f. For dimensional reasons the expectation is that the set of points $x \in B$ such that $f^{-1}(x)$ bounds a Maslov index zero holomorphic disk is real codimension one in B, forming a collection of walls. These walls should determine the structure necessary to build the mirror to X, but one needs to attach functions to the walls. Again, heuristically, these functions are expected to take the shape, at a point $x \in B$ with $L = f^{-1}(x)$,

$$\exp\left(\sum_{\substack{\beta \in \pi_2(X,L) \\ \partial\beta \neq 0}} k_\beta n_\beta z^\beta\right). \tag{4.1}$$

Here the sum is over all relative homotopy classes β such that $\partial\beta \in \pi_1(L)$ is non-zero, k_β is the index of $\partial\beta \in \pi_1(L)$ and n_β is some count of Maslov index zero disks with boundary on L. This series should be defined as a formal power series in some suitable ring. One can note that as $x \in B$ varies, the groups $\pi_2(X, L)$ vary forming a local system on B_0 (where $B_0 = \{x \in B \mid f^{-1}(x) \text{ is non-singular}\}$). This local system is analogous to the sheaf $\mathcal{P}$ of §2, with the exact sequence of homotopy groups

$$\pi_2(L) = 0 \to \pi_2(X) \to \pi_2(X, L) \to \pi_1(L) \to \pi_1(X)$$

being analogous to the exact sequence (3.1).

It is difficult to give exact definitions for the numbers n_β. There have been several approaches to dealing with this. For example, Auroux [4] pioneered, in the case of an effective anti-canonical divisor, the use of counts of Maslov index two disks to define holomorphic coordinates which are then transformed by wall-crossing automorphisms as we cross walls in B over which Maslov index zero disks live.

A different approach, using log geometry, originates in [24]. There, working with Pandharipande, we used relative Gromov-Witten invariants to make sense of the formula (4.1). The situation there was effectively that of a rational surface with an anti-canonical divisor D, and the n_β of (4.1) are replaced with counts of curves meeting the divisor D in one point. This was used for a general mirror symmetry construction for such surfaces in [14].

It is interesting to note how these two points of view apply to the case of local mirror symmetry considered in this paper. Auroux's point of view was used effectively in a sequence of papers [7–9] to study the same local mirror symmetry situation as discussed in this paper. Wall-crossing formulas for counts of Maslov index two disks are used to obtain what should be the same slab functions as discussed in this paper. The count of Maslov index two disks is reduced to a closed Gromov-Witten invariant on a toric variety, which can then be calculated via known mirror symmetry results. This allows one to show show that the slab functions defined using their counts give rise to canonical coordinates just as our slab functions do.

On the other hand, generalising the idea of replacing holomorphic disks with relative curves, one should be able to work with a certain kind of logarithmic curve called a *punctured curve*, the theory of which is currently being developed in a joint project with Abramovich and Chen [2]. These curves will live in the central fibre of the toric degeneration $\mathcal{Y} \to \mathbb{A}^1$ constructed in §2, and can be viewed as a substitute for holomorphic curves with boundary in an algebro-geometric context. Then (4.1) can be used to define slab functions, where now n_β is a count of genus 0 logarithmic curves with one puncture.

We do not propose calculating the slab functions in this way. Rather, we should be able to show that the slab functions defined in this way satisfy the same determining properties that the slab functions of §2 did. This is done by probing slabs by broken lines (see [6, 14, 15]) and interpreting these enumeratively using a different type of punctured curve, roughly corresponding to cylinders. These punctured curves play the same role that Maslov index two disks play in the analysis of slab functions of [7–9]. Crucially, we need to use the gluing formula of [2] to relate broken lines and punctured curves.

While the details of this approach will be given elsewhere, let us demonstrate this using the simple example from Examples 3.1, (1). We depict in Figure 4.1 the central fibre of the degeneration $\mathcal{X} \to \mathbb{A}^1$ constructed in §2 in this case. The total space $\mathcal{X}$ has two ordinary double points, situated on the singular locus of $\mathcal{X}_0$, where the map $\mathcal{X} \to \mathbb{A}^1$ is not normal crossings. The inclusion $\mathcal{X}_0 \subseteq \mathcal{X}$ induces a log structure on $\mathcal{X}_0$, but the log structure is not well-behaved at the two points (not *fine* in the sense of log geometry). In particular, the theory of log Gromov-Witten invariants as developed in [1, 10, 23] cannot be used directly. While a theory of invariants which can deal directly with this poorly behaved log structure is under development, for the moment we will deal with it via a small resolution of the ordinary double points. There are four choices of such resolutions, one of which is shown on the right in Figure 4.1. These choices can be thought of in terms of the affine geometry of the dual intersection complex B, with the resolutions corresponding to sliding the two singularities of the affine structure along σ to various choices of vertices.

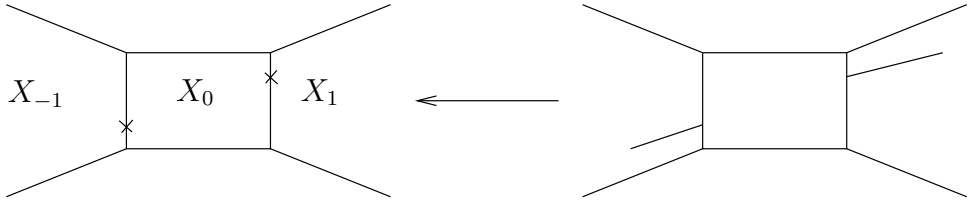


Figure 4.1. The left-hand figure shows the five irreducible components of $\mathcal{X}_0$, with the three labelled components indexed by the vertices of σ. Here $X_0 \cong \mathbb{P}^1 \times \mathbb{P}^1$ and X_{-1}, X_1 are isomorphic to the blow-up of $\mathbb{A}^2$ at a point.

We have different slab functions $f_{\mathfrak{b},v}$ for the vertices $v = -1, 0, 1$. To identify the slab function at a vertex v as a generating function, we choose a small resolution $\widetilde{\mathcal{X}} \to \mathcal{X}$ so that the irreducible component indexed by v remains toric. This effectively slides the singularities away from the vertex. The resolution in Figure 4.1 is used for the vertex $v = 0$. The slab function is given by (4.1) where n_β is a count of log curves of genus 0 with one puncture mapping to the boundary of the component X_v indexed by v. In Figure 4.2 we show the two obvious such curves for $v = 0$. However, multiple covers of these curves totally ramified at the puncture points are also possible, and a d-fold cover will contribute with multiplicity $(-1)^{d+1}/d^2$. The slab function is then, following (4.1),

$$(1 + x^{-1})(1 + xt) = \exp\left(\sum_{d=1}^{\infty} d \cdot \frac{(-1)^{d+1}}{d^2} x^{-d} + \sum_{d=1}^{\infty} d \cdot \frac{(-1)^{d+1}}{d^2} (tx)^d\right),$$

with the monomials x^{-1} and tx and their powers playing the role of z^β.

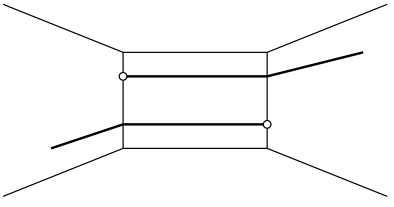


Figure 4.2. The two punctured curves corresponding to holomorphic disks. The curves include the exceptional divisors of the small resolution, and the punctures are represented by the white circles.

To prove this formula without a direct calculation, we show the slab functions defined by these counts satisfy conditions 1-4 of §3. Conditions 1 and 3 are obvious from (4.1), as the statement that only monomials z^β with $\partial\beta \neq 0$ appear is analogous to the statement that no terms of the form z^q for $q \in Q \setminus \{0\}$ appear inside the exponential. Condition 4 is automatic because in this situation the slab function only depends on the vertex. It remains to show condition 2, and we use broken lines for this, which can be reviewed in [22]. A broken line is a piecewise linear path with monomials $c_L z^{m_L}$ attached to each domain of linearity, and the derivative of the line in the domain L is $-\bar{m}_L$. When the broken line crosses a wall, we may change the monomial by applying the wall-crossing automorphism (3.6) to the monomial and choosing a new monomial being one of the terms in the expression obtained after applying this automorphism. In Figure 4.3, we consider germs of broken lines which come vertically from below with initial monomial z^m with $\bar{m} = (0, -1) \in M_{\mathbb{R}} \oplus \mathbb{R}$. We define $\mathrm{Lift}_Q(m)$ to be the sum over all broken lines ending at a basepoint Q of the final attached monomials. Note that if Q is near a vertex v of σ, then in fact $\mathrm{Lift}_Q(m) = z^m f_{\mathfrak{b},v}$. It is then not difficult to show that (3.5) holds for all pairs of adjacent vertices if and only if $\mathrm{Lift}_Q(m)$ is independent of Q chosen above the slabs as in Figure 4.3.

Broken lines can be viewed as a purely combinatorial (tropical) way to count holomorphic cylinders. But we can actually count logarithmic curves of genus 0 with two punctures to emulate cylinders, and there will be a correspondence between such twice-punctured logarithmic curves and broken lines. Varying the basepoint can be achieved by varying a point constraint for one of the punctures. The key point is that various ways of degenerating the

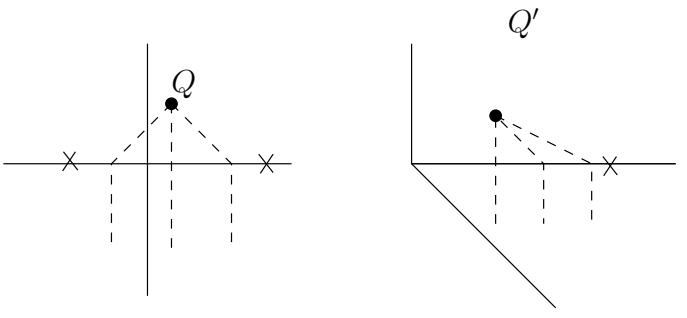


Figure 4.3. There are four broken lines with endpoints Q on the left, two of which don't bend. All have initial monomial z^m with $\bar{m} = (0, -1)$. Once the broken line crosses the slab, there are four possible attached monomials: z^m, tz^m, $x^{-1}z^m$, xtz^m. The right-hand picture shows a different choice of basepoint Q', and there are again four broken lines.

point constraint can lead to different broken lines with different endpoints. However, the count of these punctured curves will be independent of the constraint.

To see this explicitly, let's look at the example of the straight line in the left-hand diagram in Figure 4.3 with attached monomial z^m. To understand what happens when we move this broken line through the singularity, it is helpful to move the singularity to the vertex 0 by using the small resolution depicted in Figure 4.4. Consider the family of twice-punctured curves given by the vertical fibres of X_0, the blowup of $\mathbb{P}^1 \times \mathbb{P}^1$. Any curve in this family has a tropicalization (see [23], §3). The tropicalization of a general curve in this family is just the vertical line through the singularity on the right-hand side of Figure 4.4. Combinatorially, this just indicates that the curve intersects the upper and lower boundary divisors of X_0. However, the family has two special members which are degenerate with respect to the log structure on the central fibre. The tropicalization of the punctured curve when it falls into $X_0 \cap X_1$ is the straight broken line depicted in Figure 4.4 to the right of the vertex. If on the other hand we move the punctured curve to the left, it becomes reducible, the union of $X_{-1} \cap X_0$ and the exceptional curve of the small resolution. This curve tropicalizes to the tropical curve depicted on the left, now with two vertices corresponding to the two components. The bend is a consequence of gluing the once-punctured curve with support the exceptional curve to the thrice-punctured curve with support $X_0 \cap X_{-1}$. The broken line is then a subset of this tropical curve.

The point is that the two broken lines make the same contribution to $\mathrm{Lift}_Q(m)$ as Q varies because they can both be viewed as counting the number of curves in the one-parameter family described passing through some point in X_0. The point can degenerate into $X_{-1} \cap X_0$ or $X_0 \cap X_1$, giving the two types of broken line behaviour. Thus the invariance of $\mathrm{Lift}_Q(m)$ can be viewed as the fact that these lifts are generating functions for counts of certain types of punctured curves.

The key point for the argument is then to prove that broken lines really calculate Gromov-Witten invariants of punctured curves. This shall be shown using a general gluing formula [2].

Note so far we have not actually calculated the Gromov-Witten invariants of X_Σ. These should be extracted in the B-model from some additional period integrals past the ones discussed in Remark 3.2. A significant challenge remaining is to give a tropical description

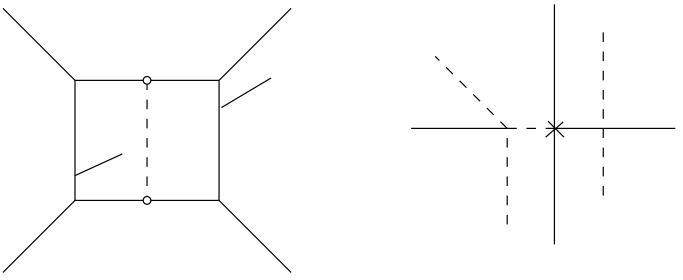


Figure 4.4.

for these period integrals and Gromov-Witten invariants.

Remark added in proof. After the initial release of this paper, Lau in [25] completed the proof that the analogue of the slab functions considered in [7–9] in fact satisfy our normalization condition, and hence agree with our slab functions.

5. Tropical disks and slab functions

The picture of counting holomorphic disks and cylinders from §4 suggests an interpretation of the slab functions in terms of tropical curves. In this section we give a surprisingly simple interpretation of this sort. The arguments are by algebraic manipulations of the slab functions. We are thus lead to the challenge of interpreting the tropical counts in terms of the counting of holomorphic disks on X_Σ.

We study the collection of slab functions at a vertex $v \in \overline{\mathscr{P}}$ with $v \in \operatorname{Int}\sigma$. By Condition (4) of slab functions all the $f_{\mathfrak{b},v}$ for slabs $\mathfrak{b}$ containing v agree. Dehomogenizing (3.7) at v we are thus left with the local model $uw - ft = 0$ for the mirror degeneration for some $f \in \widehat{\Bbbk[P]}$. Here $P = \bar{P}_v$ is a toric submonoid of $M \oplus Q^{\mathrm{gp}}$ with $P^\times = \{0\}$ and the completion is with respect to $P \setminus \{0\}$. Recall also the projection

$$M \oplus Q^{\mathrm{gp}} \longrightarrow M, \quad m \longmapsto \bar{m}.$$

For example, for the mirror of local $\mathbb{P}^2$ we had $Q = \mathbb{N}$, $P \subset \mathbb{N}^3$ generated by $(1,0,0)$, $(0,1,0)$, $(-1,-1,1)$, hence $\widehat{\Bbbk[P]} = \Bbbk[x,y,z][\![t]\!]/(xyz - t)$, and

$$f = 1 + x + y + z - 2t + 5t^2 - 32t^3 + 286t^4 - 3038t^5 + \cdots.$$

In general we assume $f = 1 + \sum_{i=1}^r z^{m_i} + g$ with $\bar{m}_i \neq 0$ for all i and $g = \sum_q b_q \cdot z^q \in \widehat{\Bbbk[Q]}$ taking care of the normalization condition.[2] Under this assumption we are going to give an infinite product expansion

$$f = \prod_{\{m \,|\, \bar{m} \neq 0\}} (1 + a_m z^m)$$

in $\widehat{\Bbbk[P]}$, with each a_m having an interpretation in terms of tropical disks in $M_{\mathbb{R}}$ with root

[2]Note that by the universal nature of Q the sum over z^{m_i} implicitly comprises a universal choice of coefficients.

weight m. Moreover, each coefficient b_q of g has an interpretation in terms of pointed tropical curves of genus zero.[3]

To this end consider the following definition of the *type of a tropical disk*. A *rooted tree* is a partially ordered finite set with a unique maximal element, called the *root vertex*, which is connected and without cycles when viewed as a graph. The *predecessors* of a vertex v are the adjacent vertices that are smaller than v. The minimal elements of a tree are called its *leaves*, so these are the elements without predecessors. We require that there are no elements with only one predecessor. In graph theory language this means that the interior vertices are at least trivalent and the leaves are the unique univalent vertices.

We now define types of tropical trees weighted by elements of P. Note that Q can be identified with the submonoid $\{m \in P \mid \bar{m} = 0\}$ of P.

Definition 5.1. The *type of a P-labelled tropical disk* is a rooted tree Γ with sets V_Γ of vertices and E_Γ of edges along with a *vertex-labeling* map

$$w : V_\Gamma \longrightarrow P \setminus Q, \quad v \longmapsto m_v$$

fulfilling the following conditions:

1. For any non-leaf vertex $v \in V_\Gamma$ with predecessors $v_1, \ldots, v_\ell$ the balancing condition

$$m_v = m_{v_1} + \cdots + m_{v_\ell}$$

holds.

2. For any vertex v the weights $m_1, \ldots, m_\ell$ of the adjacent predecessor vertices are pairwise distinct.

By abuse of notation we suppress the labelling function in the notation and write just Γ for the type of a tropical disk. The set of non-leaf vertices is denoted $\hat{V}_\Gamma$.

If we take the weight m_Γ of the root vertex in Q rather than in $P \setminus Q$ and otherwise leave the definition unchanged we arrive at the notion of *type of P-labelled pointed rational tropical curve*.

Each type of tropical disk or rational tropical curve determines an isotopy class of traditional tropical curves in $M_\mathbb{R}$ with edges labelled by lifts of the direction vector (an element of M) to P, the labelling of the predecessor vertex. In the disk case one may add another edge to force the balancing condition at the root vertex.

The balancing condition for a tropical disk implies that the labelling function is uniquely determined by its values on the leaf vertices $v_1, \ldots, v_\ell$. In particular, for the weight of the root vertex we have

$$m_\Gamma = m_{v_1} + \cdots + m_{v_\ell}.$$

Let now $S = \{m_1, \ldots, m_r\}$ be the set of exponents m occuring in f with $\bar{m} \neq 0$. For $m \in P$ with $\bar{m} \neq 0$ denote by $\mathscr{T}_m(S)$ the set of types of P-labelled tropical disks Γ with $m_\Gamma = m$ and with leaf labels in S. Similarly $\mathscr{R}_q(S)$ denotes the set of types of P-labelled pointed rational tropical curves with leaf labels in S and $m_\Gamma = q$.

[3] If σ has several interior integral points the change of vertex formula (3.4) provides a non-trivial identity between expressions labelled by different sets of tropical trees. It would be interesting to give an interpretation of this formula within the following discussion.

Proposition 5.2. *For $f = 1 + \sum_{m\in S} z^m + g$ with $g = \sum_{q\neq 0} b_q z^q$ as above it holds*

$$f = \prod_{\{m\,|\,\bar m\neq 0\}} (1 + a_m z^m) \tag{5.1}$$

with

$$a_m = \sum_{\Gamma\in\mathscr{T}_m(S)} (-1)^{|\hat V_\Gamma|} \quad \textit{and} \quad b_q = \sum_{\tilde\Gamma\in\mathscr{R}_q(S)} (-1)^{|V_{\tilde\Gamma}|-1}.$$

Proof. Expanding the infinite product in the statement and gathering according to monomials yields

$$\prod_{\{m\,|\,\bar m\neq 0\}} (1 + a_m z^m) = \sum_{m\in P} \left(\sum_{\ell=1}^{\infty} \sum_{\substack{m=m_1+\cdots+m_\ell \\ \Gamma_i\in\mathscr{T}_{m_i}(S)}} (-1)^{|\hat V_{\Gamma_1}|}\cdots(-1)^{|\hat V_{\Gamma_\ell}|} \right) z^m. \tag{5.2}$$

In this expansion ℓ is the number of a_m-terms in the infinite product to be multiplied. Thus the third sum on the right-hand side is over all decompositions $m = m_1 + \cdots + m_\ell$ of m into ℓ *pairwise distinct* summands in P. Recall that $\hat V_\Gamma$ is the set of non-leaf vertices. Fix m with $\bar m \neq 0$ now and consider the coefficient of z^m. Then for $\ell \geq 2$ any collection $\Gamma_1, \ldots, \Gamma_\ell$ of types of tropical disks with $m = m_{\Gamma_1} + \cdots + m_{\Gamma_\ell}$ can be merged into a new type of tropical disk $\Gamma \in \mathscr{T}_m(S)$ by connecting the root vertex of each Γ_i by one edge to the root vertex $v_0 \in V_\Gamma$. Thus the root vertex of Γ is ℓ-valent with adjacent predecessor trees $\Gamma_1, \ldots, \Gamma_\ell$. Now this merged tree Γ contributes to the coefficient of z^m as one term for $\ell = 1$. Since the vertices of Γ other than the root vertex are in bijection with the vertices of $\Gamma_1, \ldots, \Gamma_\ell$ it holds

$$(-1)^{|\hat V_\Gamma|} = -(-1)^{|\hat V_{\Gamma_1}|}\cdots(-1)^{|\hat V_{\Gamma_\ell}|}.$$

Thus each term with $\ell \geq 2$ in the sum of the right-hand side of (5.2) cancels with one term for $\ell = 1$. Conversely, if the root vertex of the type of a tropical disk Γ has valency $\ell \geq 2$ then Γ is obtained by this merging procedure. On the right-hand side of (5.2) we are thus left only with those m with $\bar m = 0$ and in addition with those trees with only one vertex. The latter condition means that the root vertex is also a leaf vertex. These terms yield the sum $\sum_{m\in S} z^m$. The terms with $\bar m = 0$ define a power series $1 + h \in \widehat{\Bbbk[Q]}$. We have thus shown

$$\prod_{\{m\,|\,\bar m\neq 0\}} (1 + a_m z^m) = 1 + \sum_{i=1}^{r} z^{m_i} + h,$$

with $h \in \widehat{\Bbbk[Q]}$. Since the left-hand side of this equation is clearly normalized we see that $h = g$. Tropically, the coefficient of z^q in g is the weighted sum of types of P-labelled pointed rational tropical curves, with the marked point (of valency $\ell \geq 2$) the merging point of ℓ tropical trees $\Gamma_1, \ldots, \Gamma_\ell$. The balancing condition of an underlying tropical curve at the marked point is the statement

$$\bar m_{\Gamma_1} + \cdots + \bar m_{\Gamma_\ell} = \bar m = 0. \qquad \square$$

For $f = 1 + x + y + z + g$ the expansion up to order 4 is

$$(1 + x)(1 + y)(1 + z)(1 - xy)(1 - yz)(1 - xz)(1 + x^2 y)$$

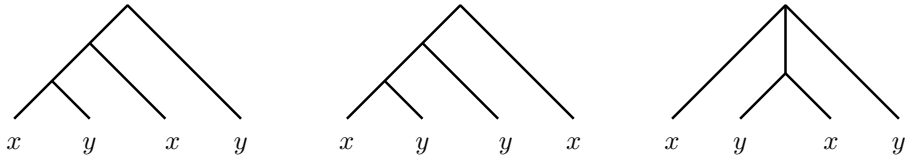


Figure 5.1.

$$\cdot(1+xy^2)\dots(1+yz^2)(1-x^2y^2)(1-y^2z^2)(1-x^2z^2)$$
$$\cdot(1-x^2yz)(1-xy^2z)(1-xyz^2)(1-xz^3)\dots(1-yz^3)$$

Figure 5.1 shows the tropical trees contributing to the coefficient $-1=(-1)^3+(-1)^3+(-1)^2$ of x^2y^2. Note that many labelled trees with four leaves are ruled out because of the third condition in Definition 5.1 that no two predecessor subtrees at some vertex be isomorphic.

We finish this section with two remarks on a possible enumerative interpretation of the expansion in terms of tropical disks and trees. First, according to (4.1) we should write the product expansion (5.1) in exponential form. Indeed, we can also write

$$f=\exp\left(\sum_{\{m\,|\,\bar{m}\neq 0\}}\sum_{\Gamma\in\tilde{\mathscr{T}}_m(S)}\frac{(-1)^{|\hat{V}_\Gamma|}}{|\operatorname{Aut}(\Gamma)|}z^m\right).$$

Here the sum is over the space $\tilde{\mathscr{T}}_m(S)$ of tropical disks with the stability condition Definition 5.1,2 dropped. Expanding $\exp$ in a Taylor series the proof is largely the same as the one given, with extra care taken concerning automorphisms.

Second, in log Gromov-Witten theory the log structure on the moduli space only depends on the type of tropical curve associated to a stable log map [23]. It is tempting to believe in a formulation of the counting problem by a symmetric obstruction theory [5] on a moduli space with a log structure stratified by types of tropical disks, with each stratum contributing $(-1)^{|\hat{V}_\Gamma|}/|\operatorname{Aut}(\Gamma)|$.

Acknowledgements. M.G. was partially supported by NSF grant 1105871 and 1262531. We would like to thank all people who influenced our way of thinking about various aspects of our program. Special thanks go to Mohammed Abouzaid, Paul Hacking, Sean Keel, Diego Matessi and Rahul Pandharipande.

References

[1] D. Abramovich and Q. Chen, *Stable logarithmic maps to Deligne–Faltings pairs II,* preprint, 2011.

[2] D. Abramovich, Q. Chen, M. Gross, and B. Siebert, *Gluing project*, in progress.

[3] P. Aspinwall, T. Bridgeland, A. Craw, and M. Douglas, M. Gross, A. Kapustin, G. Moore, G. Segal, B. Szendroi, and P. Wilson, *Dirichlet branes and mirror symme-*

try, Clay Mathematics Monographs, **4**. American Mathematical Society, Providence, RI; Clay Mathematics Institute, Cambridge, MA, 2009. x+681 pp.

[4] D. Auroux, *Mirror symmetry and T-duality in the complement of an anticanonical divisor*, J. Gökova Geom. Topol. GGT **1** (2007), 51–91.

[5] K. Behrend and B. Fantechi, *Symmetric obstruction theories and Hilbert schemes of points on threefolds*, Algebra Number Theory **2** (2008), 313–345.

[6] M. Carl, M. Pumperla, and B. Siebert, *A tropical view of Landau-Ginzburg models*, available at http://www.math.uni-hamburg.de/home/siebert/preprints/LGtrop.pdf

[7] K. Chan, C-H. Cho, S-C. Lau, and H-H. Tseng, *Gross fibrations, SYZ mirror symmetry, and open Gromov-Witten invariants for toric Calabi-Yau orbifolds*, preprint, 2013.

[8] K. Chan, S-C. Lau, and N.C. Leung, *SYZ mirror symmetry for toric Calabi-Yau manifolds*, J. Differential Geom. **90** (2012), 177–250.

[9] K. Chan, S-C. Lau, and H-H. Tseng, *Enumerative meaning of mirror maps for toric Calabi-Yau manifolds*, Adv. Math. **244** (2013), 605–625.

[10] Q. Chen, *Stable logarithmic maps to Deligne–Faltings pairs I,* preprint, 2010.

[11] T.-M. Chiang, A. Klemm, S.-T. Yau, and E. Zaslow, *Local mirror symmetry: calculations and interpretations*, Adv. Theor. Math. Phys. **3**, (1999), 495–565.

[12] K. Fukaya, *Multivalued Morse theory, asymptotic analysis and mirror symmetry,* in *Graphs and patterns in mathematics and theoretical physics*, 205–278, Proc. Sympos. Pure Math., **73**, Amer. Math. Soc., Providence, RI, 2005.

[13] H. Grauert, *Über Modifikationen und exzeptionelle Mengen*, Math. Ann. **146** 1962, 331–368.

[14] M. Gross, P. Hacking, and S. Keel, *Mirror symmetry for log Calabi-Yau surfaces I*, preprint, 2011.

[15] M. Gross, P. Hacking, S. Keel, and B. Siebert, *Theta functions on varieties with effective anticanonical class*, preprint, 2014.

[16] M. Gross, *Examples of special Lagrangian fibrations,* in *Symplectic geometry and mirror symmetry* (Seoul, 2000), 81–109, World Sci. Publishing, River Edge, NJ, 2001.

[17] ———, *Toric Degenerations and Batyrev-Borisov Duality*, Math. Ann. **333** (2005), 645-688.

[18] M. Gross and B. Siebert, *Mirror symmetry via logarithmic degeneration data I*, J. Differential Geom. **72** (2006), 169–338.

[19] ———, *Mirror symmetry via logarithmic degeneration data II*, J. Algebraic Geom. **19** (2010), 679–780.

[20] ———, *From real affine to complex geometry*, Ann. of Math. **174** (2011), 1301–1428.

[21] ———, *An invitation to toric degenerations*, Surv. Differ. Geom. **16**, 43–78, Int. Press 2011.

[22] ———, *Theta functions and mirror symmetry*, preprint arXiv:1204.1991 [math.AG], 43pp.

[23] ———, *Logarithmic Gromov-Witten invariants*, J. of the AMS, **26** (2013), 451–510.

[24] M. Gross, R. Pandharipande, and B. Siebert, *The tropical vertex*, Duke Math. J. **153** (2010), 297–362.

[25] S.-C. Lau, *Gross-Siebert's slab functions and open GW invariants for toric Calabi-Yau manifolds,* preprint, 2014, arXiv:1405:3863.

[26] M. Kontsevich and Y. Soibelman, *Affine structures and non-archimedean analytic spaces*, The unity of mathematics, 321–385, Progr. Math., 244, Birkhäuser Boston, Boston, MA, 2006.a

[27] A. Strominger, S.-T. Yau, and E. Zaslow, *Mirror Symmetry is T-duality,* Nucl. Phys. **B479**, (1996) 243–259.

DPMMS, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CV3 0WB, United Kingdom

E-mail: mgross@dpmms.cam.ac.uk

FB Mathematik, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany

E-mail: siebert@math.uni-hamburg.de

Derived category of coherent sheaves and counting invariants

Yukinobu Toda

Abstract. We survey recent developments on Donaldson-Thomas theory, Bridgeland stability conditions and wall-crossing formula. We emphasize the importance of the counting theory of Bridgeland semistable objects in the derived category of coherent sheaves to find a hidden property of the generating series of Donaldson-Thomas invariants.

Mathematics Subject Classification (2010). Primary 14N35; Secondary 18E30.

Keywords. Donaldson-Thomas invariants, Bridgeland stability conditions.

1. Introduction

1.1. Moduli spaces and invariants. The study of the *moduli spaces* is a traditional research subject in algebraic geometry. They are schemes or stacks whose points bijectively correspond to fixed kinds of algebro-geometric objects, say curves, sheaves on a fixed variety, etc. These moduli spaces are interesting not only in algebraic geometry, but also in connection with other research fields such as number theory, differential geometry and string theory. In general it is not easy to study the geometric properties of the moduli spaces. Instead one tries to construct and study the invariants of the moduli spaces, e.g. their (weighted) Euler characteristics, virtual Poincaré or Hodge polynomials, integration of the virtual cycles via deformation-obstruction theory. It has been observed that the best way to study such invariants is taking the generating series. Sometimes the generating series defined from the moduli spaces have beautiful forms and properties. Let us observe this phenomenon for some rather amenable examples. For a quasi-projective variety X (in this article, we always assume that the varieties are defined over $\mathbb{C}$), the *Hilbert scheme of n-points* denoted by $\mathrm{Hilb}_n(X)$ is the moduli space of zero dimensional subschemes $Z \subset X$ such that $\chi(\mathcal{O}_Z) = n$. It contains an open subset corresponding to n-distinct points in X, and the geometric structures of its complement is in general complicated. Nevertheless if X is nonsingular, the generating series of the Euler characteristics of $\mathrm{Hilb}_n(X)$ have the following beautiful forms [21, 27]

$$\sum_{n\geq 0}\chi(\mathrm{Hilb}_n(X))q^n=\begin{cases}(1-q)^{-\chi(X)}, & \dim X=1\\ \prod_{m\geq 1}(1-q^m)^{-\chi(X)}, & \dim X=2\\ \prod_{m\geq 1}(1-q^m)^{-m\chi(X)}, & \dim X=3.\end{cases} \tag{1.1}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

In the case $X = \mathbb{C}^d$, the torus localization shows that $\chi(\mathrm{Hilb}_n(\mathbb{C}^d))$ coincides with the number of d-dimensional partitions of n. The resulting product formulas are consequences of enumerative combinatorics, as in [64]. A general case is reduced to the case $X = \mathbb{C}^d$.

1.2. Curve counting invariants. The study of the invariants of the moduli spaces of curves inside a variety is more important and interesting, because of its connection with *world sheet counting* in string theory. A particularly important case is when X is a *Calabi-Yau 3-fold*, i.e. X has a trivial canonical line bundle with $H^1(X, \mathcal{O}_X) = 0$, as the string theory predicts our universe to be the product of the four dimensional space time with a Calabi-Yau 3-fold. Similarly to $\mathrm{Hilb}_n(X)$, we denote by $\mathrm{Hilb}_n(X, \beta)$ the *Hilbert scheme of curves* inside X, that is the moduli space of projective subschemes $C \subset X$ with $\dim C \leq 1$, $[C] = \beta$ and $\chi(\mathcal{O}_C) = n$. The following result was obtained by the author in 2008, and plays a key role in this article:

Theorem 1.1 ([77, 78]). *Let X be a smooth projective Calabi-Yau 3-fold. Then for fixed $\beta \in H_2(X, \mathbb{Z})$, the quotient series*

$$\frac{\sum_{n \in \mathbb{Z}} \chi(\mathrm{Hilb}_n(X, \beta)) q^n}{\sum_{n \geq 0} \chi(\mathrm{Hilb}_n(X)) q^n} \tag{1.2}$$

is the Laurent expansion of a rational function of q, invariant under $q \leftrightarrow 1/q$.

Note that the denominator of (1.2) is given by the formula (1.1). A typical example of a rational function of q invariant under $q \leftrightarrow 1/q$ is $q/(1+q)^2 = q - 2q^2 + 3q^3 - \cdots$. We remark that the invariance of $q \leftrightarrow 1/q$ does not say the invariance of the generating series after the formal substitution $q \mapsto 1/q$, but so after taking the analytic continuation of the function (1.2) from $|q| \ll 1$ to $|q| \gg 1$. The above result was conjectured in [45] as the *unweighted version* of the rationality conjecture of rank one Donaldson-Thomas (DT) invariants by Maulik-Nekrasov-Okounkov-Pandharipande (MNOP) [48]. The rationality conjecture was proposed in order to formulate the *Gromov-Witten/Donaldson-Thomas correspondence conjecture* comparing two kinds of curve counting invariants on Calabi-Yau 3-folds.

The DT invariant was introduced by Thomas [67], as a holomorphic analogue of Casson invariants of real 3-manifolds. It counts stable coherent sheaves on a Calabi-Yau 3-fold, and is a higher dimensional generalization of Donaldson invariants on algebraic surfaces. For a Calabi-Yau 3-fold X, an ample divisor H on X and a cohomology class $v \in H^*(X, \mathbb{Q})$, the DT invariant $\mathrm{DT}_H(v) \in \mathbb{Z}$ is defined to be the degree of the zero dimensional virtual fundamental cycle on the moduli space of H-stable coherent sheaves E on X with $\mathrm{ch}(E) = v$. It also coincides with the weighted Euler characteristic with respect to the Behrend's constructible function on that moduli space [6]. The DT invariants were later generalized by Joyce-Song [36] and Kontsevich-Soibelman [40] so that they also count strictly H-semistable sheaves. The generalized DT invariants involve the Behrend functions and the motivic Hall algebras in the definition, and they are $\mathbb{Q}$-valued.

The Hilbert scheme of points or curves on a Calabi-Yau 3-fold is also interpreted as a moduli space of stable sheaves, by assigning a subscheme $C \subset X$ with its ideal sheaf $I_C \subset \mathcal{O}_X$. The resulting DT invariant is the weighted Euler characteristic of the Hilbert scheme of points or curves, and in particular it is independent of H. In this sense, the invariant $\chi(\mathrm{Hilb}_n(X, \beta))$ is the unweighted version of the DT invariant, which coincides with the honest DT invariant up to sign if $\mathrm{Hilb}_n(X, \beta)$ is non-singular. The result of Theorem 1.1 for the weighted version was later proved by Bridgeland [19]. The rationality property and the

invariance of $q \leftrightarrow 1/q$ of the series (1.2) are not visible if we just look at the moduli spaces of curves or points. Such hidden properties of the series (1.2) are visible after we develop new moduli theory and invariants of objects in the derived category of coherent sheaves.

1.3. Derived category of coherent sheaves. Recall that for a variety X, the bounded derived category of coherent sheaves $D^b\mathrm{Coh}(X)$ is defined to be the localization by quasi-isomorphisms of the homotopy category of the bounded complexes of coherent sheaves on X. The derived category is no longer an abelian category, but has a structure of a triangulated category. It was originally introduced by Grothendieck in 1960's in order to formulate the relative version of Serre duality theorem, known as Grothendieck duality theorem. Later it was observed by Mukai [51] that an abelian variety and its dual abelian variety, which are not necessary isomorphic in general, have the equivalent derived categories of coherent sheaves. This phenomena suggests that the category $D^b\mathrm{Coh}(X)$ has more symmetries than the category of coherent sheaves, as the latter category is known to reconstruct the original variety. Such a phenomena has drawn much attention since Kontsevich proposed the *Homological mirror symmetry conjecture* in [39]. It predicts an equivalence between the derived category of coherent sheaves on a Calabi-Yau manifold and the derived Fukaya category of its mirror manifold, based on an insight that the derived category $D^b\mathrm{Coh}(X)$ is a mathematical framework of D-branes of type B in string theory. There have been several developments in constructing Mukai type derived equivalences between non-isomorphic varieties [5, 14, 38, 55], and non-trivial autequivalences [34, 62], based on the ideas from mirror symmetry. Furthermore such Mukai type equivalences have been discovered beyond algebraic geometry. For instance, derived McKay correspondence [10] gives an equivalence between the derived category of finite group representations and the derived category of coherent sheaves on the crepant resolution of the quotient singularity. This is now interpreted as a special case of equivalences between usual commutative varieties and non-commutative varieties in the context of Van den Bergh's non-commutative crepant resolutions [22]. There also exists an Orlov's equivalence [56] between the derived category of coherent sheaves on a Calabi-Yau hypersurface in the projective space and the category of graded matrix factorizations of the defining equation of it. This result, called Landau-Ginzbrug/Calabi-Yau correspondence, was also motivated by mirror symmetry. Now it is understood that the derived categories have more symmetries than the categories of coherent sheaves. Our point of view is to make the hidden properties of the generating series of DT type invariants visible via symmetries in the derived categories.

1.4. Bridgeland stability conditions. The idea of applying derived categories to the study of generating series of DT type invariants suggests an importance of constructing moduli spaces and invariants of objects in the derived categories. Note that in constructing the original DT invariants, we need to fix an ample divisor on a Calabi-Yau 3-fold X, and the associated stability condition on $\mathrm{Coh}(X)$ in order to construct a good moduli space of stable sheaves. The notion of stability conditions on triangulated categories, in particular on derived categories of coherent sheaves, was introduced by Bridgeland [16] as a mathematical framework of Douglas's Π-stability [24] in string theory. For a triangulated category $\mathcal{D}$, a Bridgeland stability condition on it roughly consists of data $\sigma = (Z, \{\mathcal{P}(\phi)\}_{\phi\in\mathbb{R}})$ for a group homomorphism $Z\colon K(\mathcal{D}) \to \mathbb{C}$ called the *central charge*, and the collection of subcategories $\mathcal{P}(\phi) \subset \mathcal{D}$ for $\phi \in \mathbb{R}$ whose objects are called *σ-semistable objects with phase ϕ*. The main result by Bridgeland [16] is to show that the set of 'good' stability conditions

on $\mathcal{D}$ forms a complex manifold. This complex manifold is in particular important when $\mathcal{D} = D^b\mathrm{Coh}(X)$ for a Calabi-Yau manifold X. In this case, the space of stability conditions $\mathrm{Stab}(X)$ is expected to contain the universal covering space of the moduli space of complex structures of a mirror manifold of X. So far the space $\mathrm{Stab}(X)$ has been studied in several situations, e.g. X is a curve [16, 47], X is a K3 surface [17], X is a some non-compact Calabi-Yau 3-fold [11, 15, 74, 76]. On the other hand, there has been a serious issue in studying Bridgeland stability conditions on projective Calabi-Yau 3-folds which are likely to be the most important case: we are not able to prove the existence of Bridgeland stability conditions on smooth projective Calabi-Yau 3-folds. In [12], the existence problem is reduced to showing a conjectural Bogomolov-Gieseker type inequality evaluating the third Chern character of certain two term complexes of coherent sheaves. However proving that inequality conjecture seems to require a new idea.

1.5. New invariants via derived categories. Let X be a smooth projective Calabi-Yau 3-fold. We expect that, for a given $\sigma \in \mathrm{Stab}(X)$ and $v \in H^*(X, \mathbb{Q})$, there exists the DT type invariant $\mathrm{DT}_\sigma(v) \in \mathbb{Q}$ which counts σ-semistable objects $E \in D^b\mathrm{Coh}(X)$ with $\mathrm{ch}(E) = v$. As we mentioned, there is a serious issue in constructing a Bridgeland stability condition on projective Calabi-Yau 3-folds, but let us ignore this for a while. For an ample divisor H on X, we expect that the classical H-stability appears as a certain special limiting point in $\mathrm{Stab}(X)$ called the *large volume limit*. If we take $\sigma \in \mathrm{Stab}(X)$ near the large volume limit point, then we expect the equality $\mathrm{DT}_\sigma(v) = \mathrm{DT}_H(v)$. On the other hand, suppose that there is an autequivalence Φ of $D^b\mathrm{Coh}(X)$ and $\tau \in \mathrm{Stab}(X)$ so that the equality $\mathrm{DT}_\tau(v) = \mathrm{DT}_\tau(\Phi_* v)$ holds for any v. Then the generating series of the invariants $\mathrm{DT}_\tau(v)$ is preserved by the variable change induced by $v \mapsto \Phi_* v$. If we are able to relate $\mathrm{DT}_\sigma(v)$ and $\mathrm{DT}_\tau(v)$, then it would imply the hidden symmetry of the generating series of classical DT invariants $\mathrm{DT}_H(v)$ with respect $v \mapsto \Phi_* v$. The relationship between $\mathrm{DT}_\sigma(v)$ and $\mathrm{DT}_\tau(v)$ is studied by the wall-crossing phenomena: there should be a wall and chamber structure on the space $\mathrm{Stab}(X)$ so that the invariants $\mathrm{DT}_*(v)$ are constant on a chamber but jumps if $*$ crosses a wall. The wall-crossing formula of the invariants $\mathrm{DT}_*(v)$ should be described by a general framework established by Joyce-Song [36], Kontsevich-Soibelman [40], using stack theoretic Hall algebras.

However as we mentioned, we are not able to prove $\mathrm{Stab}(X) \neq \emptyset$, so the above story is the next stage after proving the non-emptiness. The idea of proving Theorem 1.1 was to introduce 'weak' Bridgeland stability conditions on triangulated categories, and apply the above story for the space of weak stability conditions on the subcategory of $D^b\mathrm{Coh}(X)$ generated by $\mathcal{O}_X$ and one or zero dimensional sheaves. The latter subcategory is called the category of *D0-D2-D6 bound states*. The notion of weak stability conditions is a kind of limiting degenerations of Bridgeland stability conditions, and it is a coarse version of Bayer's polynomial stability conditions [2], the author's limit stability conditions [75]. It is easier to construct weak stability conditions and enough to prove Theorem 1.1 applying the above story. The derived dual $E \mapsto \mathbf{R}\mathcal{H}om(E, \mathcal{O}_X)$, an autequivalence of $D^b\mathrm{Coh}(X)$, turned out to be responsible for the hidden symmetric property of $q \leftrightarrow 1/q$ of the series (1.2) in the above story.

The idea of proving Theorem 1.1 has turned out to be useful in proving several other interesting properties of DT type invariants, say DT/PT correspondence [19, 77] conjectured by Pandharipande-Thomas [59]. We refer to [20, 52, 65, 72, 73, 79, 80, 82–84] for other works relating the above story.

1.6. Plan of this article. In Section 2, we review and survey recent developments of Donaldson-Thomas theory. In Section 3, we survey the developments on Bridgeland stability conditions. In Section 4, we discuss open problems on DT theory and Bridgeland stability conditions.

2. Donaldson-Thomas theory

2.1. Moduli spaces of semistable sheaves. Let X be a smooth projective variety and H an ample divisor on X. For an object $E \in \mathrm{Coh}(X)$, its Hilbert polynomial is given by

$$\chi(E \otimes \mathcal{O}_X(mH)) = a_d m^d + a_{d-1} m^{d-1} + \cdots$$

for $a_i \in \mathbb{Q}$ by the Riemann-Roch theorem. Here $a_d \neq 0$ and d is the dimension of the support of E. The reduced Hilbert polynomial $\overline{\chi}_H(E,m)$ is defined to be $\chi(E \otimes \mathcal{O}_X(mH))/a_d$.

Definition 2.1. An object $E \in \mathrm{Coh}(X)$ is *H-(semi)stable* if for any subsheaf $0 \neq F \subsetneq E$, we have $\dim \mathrm{Supp}(F) = \dim \mathrm{Supp}(E)$ and the inequality $\overline{\chi}_H(F,m) < (\leq) \overline{\chi}_H(E,m)$ holds for $m \gg 0$.

Remark 2.2. Note that if E is torsion free, then $\overline{\chi}_H(F,m) = m^d + c \cdot \mu_H(E) m^{d-1} + O(m^{d-2})$ where $d = \dim X$, $\mu_H(E) = c_1(E)H^{d-1}/\mathrm{rank}(E)$, and c is some constant. Hence the H-(semi)stability is the refinement of H-slope (semi)stability defined by the slope function $\mu_H(*)$.

Let $\mathcal{C}oh(X)$ be the 2-functor from the category of complex schemes to the groupoid, whose S-valued points form the groupoid of flat families of coherent sheaves on X over S. The 2-functor $\mathcal{C}oh(X)$ forms a stack, which is known to be an Artin stack locally of finite type, but neither finite type nor separated. The situation becomes better if we consider the substacks for $v \in H^*(X,\mathbb{Q})$

$$\mathcal{M}_H^{\mathrm{s}}(v) \subset \mathcal{M}_H^{\mathrm{ss}}(v) \subset \mathcal{C}oh(X).$$

Here $\mathcal{M}_H^{\mathrm{s(ss)}}(v)$ is the substack of H-(semi)stable $E \in \mathrm{Coh}(X)$ with $\mathrm{ch}(E) = v$, which is an open substack of $\mathcal{C}oh(X)$. The stack $\mathcal{M}_H^{\mathrm{ss}}(v)$ is of finite type but not separated in general. The stack $\mathcal{M}_H^{\mathrm{s}}(v)$ is of finite type, separated, and a $\mathbb{C}^*$-gerb over a quasi-projective scheme $M_H^{\mathrm{s}}(v)$. The scheme $M_H^{\mathrm{s}}(v)$ is projective if $\mathcal{M}_H^{\mathrm{s}}(v) = \mathcal{M}_H^{\mathrm{ss}}(v)$.

2.2. Donaldson-Thomas invariants. Let X be a smooth projective 3-fold. We say it is a *Calabi-Yau 3-fold* if $K_X = 0$ and $H^1(X,\mathcal{O}_X) = 0$. A typical example is a quintic hypersurface in $\mathbb{P}^4$. Let H be an ample divisor on X, v an element in $H^*(X,\mathbb{Q})$, and consider the moduli scheme $M_H^{\mathrm{s}}(v)$. A standard deformation theory of sheaves (cf. [32]) shows that the tangent space at $[E] \in M_H^{\mathrm{s}}(v)$ is given by $\mathrm{Ext}^1(E,E)$, and the obstruction space is given by $\mathrm{Ext}^2(E,E)$. The Calabi-Yau condition and the Serre duality implies that the latter space is dual to $\mathrm{Ext}^1(E,E)$. Hence the virtual dimension at $[E]$, defined to be the dimension of the tangent space minus the dimension of the obstruction space, is zero which is independent of E. Based on this observation, Thomas [67] constructed two term complex of vector bundles $\mathcal{E}^\bullet$ on $M_H^{\mathrm{s}}(v)$ and a morphism $\mathcal{E}^\bullet \to L_{M_H^{\mathrm{s}}(v)}$ in $D(M_H^{\mathrm{s}}(v))$, giving a symmetric perfect obstruction theory in the sense of Behrend-Fantechi [8, 9]. By

the construction in [8], there is the associated zero dimensional virtual cycle $[M_H^{\mathrm{s}}(v)]^{\mathrm{vir}}$ on $M_H^{\mathrm{s}}(v)$, and we are able to take its degree if $M_H^{\mathrm{s}}(v)$ is projective. The DT invariant is defined as follows:

Definition 2.3. If $\mathcal{M}_H^{\mathrm{s}}(v) = \mathcal{M}_H^{\mathrm{ss}}(v)$ holds, we define $\mathrm{DT}_H(v) \in \mathbb{Z}$ to be the degree of $[M_H^{\mathrm{s}}(v)]^{\mathrm{vir}}$.

The above construction via the virtual cycle easily shows that the DT invariant is invariant under deformations of complex structures of X. However in practice, it is more convenient to describe the DT invariant in terms of Behrend's constructible function [6]. The Behrend function is easily described if we use the following result by Joyce-Song [36]:

Theorem 2.4 ([36]). *for any $p \in M_H^{\mathrm{s}}(v)$, there is an analytic open subset $p \in U \subset M_H^{\mathrm{s}}(v)$, a complex manifold V and a holomorphic function $f \colon V \to \mathbb{C}$ such that U is isomorphic to $\{df = 0\}$.*

Using the above result, the Behrend function ν_B on $M_H^{\mathrm{s}}(v)$ is described as

$$\nu_B(p) = (-1)^{\dim V}(1 - \chi(M_p(f)))$$

where $M_p(f)$ is the Milnor fiber of f at p. The function ν_B is shown to a be well-defined constructible function on $M_H^{\mathrm{s}}(v)$.

Theorem 2.5 ([6]). *If $\mathcal{M}_H^{\mathrm{s}}(v) = \mathcal{M}_H^{\mathrm{ss}}(v)$ holds, we have the equality*

$$\mathrm{DT}_H(v) = \sum_{k \in \mathbb{Z}} k \cdot \chi(\nu_B^{-1}(k)). \tag{2.1}$$

In particular if $M_H^{\mathrm{s}}(v)$ is non-singular and connected, the invariant $\mathrm{DT}_H(v)$ coincides with $\chi(M_H^{\mathrm{s}}(v))$ up to sign.

Based on the above description of the DT invariant, Joyce-Song [36] and Kontsevich-Soibelman [40] constructed the generalized DT invariant $\mathrm{DT}_H(v) \in \mathbb{Q}$ without the condition $\mathcal{M}_H^{\mathrm{s}}(v) = \mathcal{M}_H^{\mathrm{ss}}(v)$. The construction uses the stack theoretic Hall algebra $H(\mathrm{Coh}(X))$ of $\mathrm{Coh}(X)$, and its well-definedness is highly non-trivial. A very rough description of it may be

$$\mathrm{DT}_H(v) = \int_{\log \mathcal{M}_H^{\mathrm{ss}}(v)} \nu_B \cdot d\chi.$$

The 'log' is taken in the algebra $H(\mathrm{Coh}(X))$. Some more explanation of a specific case is available in [81].

Remark 2.6. We can define another invariant $\mathrm{DT}_H^{\chi}(v) \in \mathbb{Q}$ by formally putting $\nu_B \equiv 1$ in the definition of $\mathrm{DT}_H(v)$. If $\mathcal{M}_H^{\mathrm{s}}(v) = \mathcal{M}_H^{\mathrm{ss}}(v)$, it coincides with the usual Euler characteristic $\chi(M_H^{\mathrm{s}}(v))$. When we say a result as a *weighted (resp. an unweighted) version*, it means the result for the invariants $\mathrm{DT}_H(v)$ (resp. $\mathrm{DT}_H^{\chi}(v)$).

2.3. Rank one DT invariants. In what follows, we identify $H^4(X, \mathbb{Q})$, $H^6(X, \mathbb{Q})$ with $H_2(X, \mathbb{Q})$, $\mathbb{Q}$ respectively by the Poincaré duality. Given $\beta \in H_2(X, \mathbb{Z})$ and $n \in \mathbb{Z}$, it

is easy to show that $\mathrm{Hilb}_n(X,\beta)$ is isomorphic to $M_H^{\mathrm{s}}(v)$ for $v=(1,0,-\beta,-n)$ by the assignment $C\mapsto I_C$. The resulting invariant

$$I_{n,\beta}=\mathrm{DT}_H(1,0,-\beta,-n)\in\mathbb{Z}$$

is independent of H, and it counts one or zero dimensional subschemes $C\subset X$ with $[C]=\beta$, $\chi(\mathcal{O}_C)=n$. For $\beta\in H_2(X,\mathbb{Z})$, the series $I_\beta(X)$ is defined to be

$$I_\beta(X)=\sum_{n\in\mathbb{Z}}I_{n,\beta}q^n.$$

Example 2.7. (i) If $\beta=0$, we have [9, 42, 44]

$$I_0(X)=\prod_{k\geq 1}(1-(-q)^k)^{-k\chi(X)}.$$

(ii) If $f\colon X\to Y$ is a birational contraction whose exceptional locus is $C\cong\mathbb{P}^1$ with normal bundle $\mathcal{O}_C(-1)^{\oplus 2}$, we have [3]

$$\sum_{m\geq 0}I_{m[C]}(X)t^m=\prod_{k\geq 1}(1-(-q)^k)^{-k\chi(X)}\prod_{k\geq 1}(1-(-q)^kt)^k.$$

The above example indicates that the quotient series $I_\beta(X)/I_0(X)$ is the honest curve counting series with homology class β. The following conjecture was proposed by MNOP [48]:

Conjecture 2.8 ([48]).

(i) *The quotient series $I_\beta(X)/I_0(X)$ is the Laurent expansion of a rational function of q, invariant under $q\leftrightarrow 1/q$.*

(ii) *After the variable change $q=-e^{i\lambda}$, we have the equality*

$$\sum_{\beta\geq 0}\frac{I_\beta(X)}{I_0(X)}t^\beta=\exp\left(\sum_{g\geq 0,\beta>0}\mathrm{GW}_{g,\beta}(X)\lambda^{2g-2}t^\beta\right).$$

Here $\mathrm{GW}_{g,\beta}(X)\in\mathbb{Q}$ is the Gromov-Witten invariant counting stable maps $f\colon C\to X$ from projective curves C with at worst nodal singularities with $g(C)=g$, $f_*[C]=\beta$. The variable change $q=-e^{i\lambda}$ makes sense by the rationality conjecture (i). The above conjecture was first proved for toric Calabi-Yau 3-folds in [48].

2.4. Developments on MNOP conjecture. As we mentioned in the introduction, the result of Theorem 1.1 is the unweighted version of Conjecture 2.8 (i). The weighted version was proved in [19]. We have the following result [77, 78] (unweighted version), [19] (weighted version):

Theorem 2.9. *There exist invariants $N_{n,\beta}\in\mathbb{Q}$, $L_{n,\beta}\in\mathbb{Q}$ satisfying*

- *$N_{n,\beta}=N_{-n,\beta}=N_{n+H\beta,\beta}$ for any ample divisor H on X,*
- *$L_{n,\beta}=L_{-n,\beta}$, and it is zero for $|n|\gg 0$,*

such that we have the following formula:

$$\sum_{\beta\geq 0} I_\beta(X)t^\beta = \prod_{n>0,\beta\geq 0} \exp((-1)^{n-1}nN_{n,\beta}q^n t^\beta) \left(\sum_{n,\beta} L_{n,\beta}q^n t^\beta\right).$$

Remark 2.10. The proofs for the unweighted version in the author's papers [77, 78] can be modified to show the weighted version, if once a similar result of Theorem 2.4 for the moduli spaces of complexes in [35, 43] is shown to be true (cf. [81]). This is also applied for the results below.

The rationality conjecture is an easy consequence of Theorem 2.9:

Corollary 2.11. *Conjecture 2.8 (i) is true.*

There exist geometric meanings of $N_{n,\beta}$ and $L_{n,\beta}$. The former invariant is nothing but the generalized DT invariant $\mathrm{DT}_H(0,0,\beta,n)$, which counts one or zero dimensional H-semistable sheaves F on X with $[F]=\beta$, $\chi(F)=n$. A priori, $N_{n,\beta}$ is defined using the ample divisor H, but the resulting invariant is shown to be independent of H. The latter invariant $L_{n,\beta}$ is more interesting. It counts certain two term complexes $E\in D^b\mathrm{Coh}(X)$ (indeed they are perverse coherent sheaves in the sense of [7, 37]) satisfying $\mathrm{ch}(E)=(1,0,-\beta,-n)$, which are semistable with respect to a derived self dual weak stability condition on it. The result of Theorem 2.9 is proved along with the idea stated in Subsection 1.5.

A similar idea also proves Pandharipande-Thomas conjecture [59] relating the quotient series of rank one DT invariants with the invariants counting stable pairs. The definition of stable pairs is given as follows:

Definition 2.12 ([59]). A *stable pair* is data (F,s) where F is a pure one dimensional sheaf on X, $s\colon \mathcal{O}_X\to F$ is a morphism which is surjective in dimension one.

A typical example of a stable pair is $(\mathcal{O}_C(D),s)$, where $C\subset X$ is a smooth curve, $D\subset C$ is an effective divisor and s is a natural composition $\mathcal{O}_X\twoheadrightarrow\mathcal{O}_C\subset\mathcal{O}_C(D)$. For given $\beta\in H_2(X,\mathbb{Z})$ and $n\in\mathbb{Z}$, the moduli space $P_n(X,\beta)$ of stable pairs (F,s) with $[F]=\beta$, $\chi(F)=n$ is a projective scheme with a symmetric perfect obstruction theory [59]. The PT invariant $P_{n,\beta}\in\mathbb{Z}$ is defined to be the degree of the zero dimensional virtual fundamental cycle $[P_n(X,\beta)]^{\mathrm{vir}}$ on $P_n(X,\beta)$. The invariant $P_{n,\beta}$ is deformation invariant, and coincides with the weighted Euler characteristic with respect to the Behrend function on $P_n(X,\beta)$. The following conjecture was proposed by [59], its unweighted version was proved in [63, 77], and the weighted version was proved in [19]:

Theorem 2.13. *For fixed $\beta\in H_2(X,\mathbb{Z})$, we have the equality of the generating series*

$$\frac{I_\beta(X)}{I_0(X)}=\sum_{n\in\mathbb{Z}}P_{n,\beta}q^n.$$

Finally in [58], Pandharipande-Pixton proved Conjecture 2.8 (ii) for large class of Calabi-Yau 3-folds including quintic hypersurfaces in $\mathbb{P}^4$:

Theorem 2.14 ([58]). *Conjecture 2.8 (ii) is true if X is a complete intersection Calabi-Yau 3-fold in the product of projective spaces.*

Indeed what they proved is the correspondence between Gromov-Witten invariants and stable pair invariants. Combined with Theorem 2.13, the result of Theorem 2.14 was proved. Their proof relies on the degeneration formula of GW and PT invariants, and the torus localization formula.

2.5. Non-commutative DT theory and flops. The DT theory can be also constructed for non-commutative varieties or algebras. Let Y be a quasi-projective 3-fold which admits two crepant small resolutions giving a *flop*:

$$\phi\colon X \xrightarrow{f} Y \xleftarrow{f^{\dagger}} X^{\dagger}. \tag{2.2}$$

In this situation, Van den Bergh [22] constructed sheaf of non-commutative algebras A_Y on Y and derived equivalences

$$D^b\mathrm{Coh}(X^{\dagger}) \xrightarrow{\Psi} D^b\mathrm{Coh}(A_Y) \xrightarrow{\Phi} D^b\mathrm{Coh}(X) \tag{2.3}$$

so that their composition gives Bridgeland's flop equivalence [16]. For $n \in \mathbb{Z}$ and $\beta \in H_2(X,\mathbb{Z})$, let $\mathrm{Hilb}_n(A_Y,\beta)$ be the moduli space of surjections $A_Y \twoheadrightarrow F$ in $\mathrm{Coh}(A_Y)$ such that F has at most one dimensional support and $[\Phi(F)] = \beta$, $\chi(\Phi(F)) = n$. If X is a smooth projective Calabi-Yau 3-fold, there is a symmetric perfect obstruction theory on $\mathrm{Hilb}_n(A_Y,\beta)$, and the degree of its zero dimensional virtual fundamental cycle defines the *non-commutative DT (ncDT) invariant* $A_{n,\beta} \in \mathbb{Z}$. Alternatively, $A_{n,\beta}$ is defined to be the weighted Euler characteristic of the Behrend function on $\mathrm{Hilb}_n(A_Y,\beta)$. We set $I_\beta(A_Y)$ to be

$$I_\beta(A_Y) = \sum_{n\in\mathbb{Z}} A_{n,\beta} q^n.$$

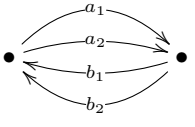
The following result was proved in [84] for the unweighted version, and [20] for the weighted version, basically along with the argument in Subsection 1.5:

Theorem 2.15. *We have the following identities:*

$$\sum_{f_*\beta=0} I_\beta(A_Y)t^\beta = \prod_{k\geq 1}(1-(-q)^k)^{k\chi(X)} \left(\sum_{f_*\beta=0} I_\beta(X)t^\beta\right)\left(\sum_{f_*\beta=0} I_{-\beta}(X)t^\beta\right)$$

$$\frac{\sum_\beta I_\beta(X)t^\beta}{\sum_{f_*\beta=0} I_\beta(X)t^\beta} = \frac{\sum_\beta I_\beta(A_Y)t^\beta}{\sum_{f_*\beta=0} I_\beta(A_Y)t^\beta} = \frac{\sum_\beta I_{\phi_*\beta}(X^{\dagger})t^\beta}{\sum_{f_*\beta=0} I_{\phi_*\beta}(X^{\dagger})t^\beta}.$$

Example 2.16. Let $Y = (xy+zw=0) \subset \mathbb{C}^4$ be the conifold singularity, and take two crepant small resolutions (2.2) by blowing up at the ideals (x,z) and (x,w). In this case, the algebra A_Y is the path algebra of the following quiver



with relation given by the derivations of the super potential $W = a_1b_1a_2b_2 - a_1b_2a_2b_1$. Although X is not projective in this case, the ncDT invariant $A_{n,m[C]} \in \mathbb{Z}$ makes sense, and coincides with the weighted Euler characteristic of the moduli space of framed A_Y-representations with dimension vector $(n, m+n)$. The proof of Theorem 2.15 also works in this situation. Using Example 2.7 (ii), the first identity of Theorem 2.15 becomes

$$\sum_{n,m} A_{n,m[C]} q^n t^m = \prod_{k\geq 1}(1-(-q)^k)^{-2k} \prod_{k\geq 1}(1-(-q)^k t)^k \prod_{k\geq 1}(1-(-q)^k t^{-1})^k.$$

The above formula was first conjectured by Szendrői [66], and later proved by Young [88], Nagao-Nakajima [54].

Remark 2.17. In general for a quiver Q with a super potential W, we are able to define the ncDT theory for (Q,W). A mutation of the pair (Q,W) defines another quiver with a super potential $(Q^{\dagger},W^{\dagger})$. The relationship between ncDT invariants on (Q,W) and $(Q^{\dagger},W^{\dagger})$ is described in terms of cluster transformations. We refer to [40, 53] for the detail.

3. Bridgeland stability conditions

3.1. Definitions. We recall the definition of Bridgeland stability conditions on a triangulated category $\mathcal{D}$. We fix a finitely generated free abelian group Γ with a norm $\| * \|$ on $\Gamma_{\mathbb{R}}$ together with a group homomorphism $\mathrm{cl}\colon K(\mathcal{D}) \to \Gamma$. A typical example is that $\mathcal{D}=D^b\mathrm{Coh}(X)$ for a smooth projective variety X, Γ is the image of the Chern character map $\mathrm{ch}\colon K(X) \to H^*(X,\mathbb{Q})$, and $\mathrm{cl}=\mathrm{ch}$. By taking the dual of cl, we always regard a group homomorhpism $\Gamma \to \mathbb{C}$ as a group homomorphism $K(\mathcal{D}) \to \mathbb{C}$.

Definition 3.1 ([16]). A *stability condition* on $\mathcal{D}$ is data $\sigma=(Z,\{\mathcal{P}(\phi)\}_{\phi\in\mathbb{R}})$, where $Z\colon \Gamma \to \mathbb{C}$ is a group homomorphism (called *central charge*), $\mathcal{P}(\phi)\subset \mathcal{D}$ is a full subcategory (called *σ-semistable objects with phase ϕ*) satisfying the following conditions:

- For $0\neq E\in \mathcal{P}(\phi)$, we have $Z(E)\in \mathbb{R}_{>0}\exp(\sqrt{-1}\pi\phi)$.
- For all $\phi\in\mathbb{R}$, we have $\mathcal{P}(\phi+1)=\mathcal{P}(\phi)[1]$.
- For $\phi_1>\phi_2$ and $E_i\in\mathcal{P}(\phi_i)$, we have $\mathrm{Hom}(E_1,E_2)=0$.
- (Harder-Narasimhan property): For each $0\neq E\in\mathcal{D}$, there is a collection of distinguished triangles $E_{i-1}\to E_i\to F_i\to E_{i-1}[1]$, $E_N=E, E_0=0$ with $F_i\in\mathcal{P}(\phi_i)$ and $\phi_1>\phi_2>\cdots>\phi_N$.

Another way defining a stability condition is to use a t-structure as follows:

Lemma 3.2 ([16]). *Giving a stability condition on $\mathcal{D}$ is equivalent to giving data $(Z,\mathcal{A})$, where $Z\colon \Gamma\to\mathbb{C}$ is a group homomorphism, $\mathcal{A}\subset\mathcal{D}$ is the heart of a bounded t-structure, satisfying*

$$Z(\mathcal{A}\setminus\{0\})\in\{r\exp(i\pi\phi): r>0, 0<\phi\leq 1\} \tag{3.1}$$

together with the Harder-Narasimhan property: for any $E\in\mathcal{A}$, there exists a filtration $0=E_0\subset E_1\subset\cdots\subset E_N=E$ such that $F_i=E_i/E_{i-1}$ is Z-semistable with $\arg Z(F_i)>\arg Z(F_{i+1})$ for all i. Here $E\in\mathcal{A}$ is Z-semistable if for any subobject $0\neq F\subsetneq E$, we have $\arg Z(F)<(\leq)\arg Z(E)$.

Proof. The correspondence is as follows: given $(Z,\{\mathcal{P}(\phi)\}_{\phi\in\mathbb{R}})$, the corresponding heart $\mathcal{A}$ is the extension closure of $\mathcal{P}(\phi)$ for $0<\phi\leq 1$. Conversely given $(Z,\mathcal{A})$, the category $\mathcal{P}(\phi)$ is defined to be the category of Z-semistable objects $E\in\mathcal{A}$ with $Z(E)\in\mathbb{R}_{>0}\exp(i\pi\phi)$. Other $\mathcal{P}(\phi)$ are defined by the rule $\mathcal{P}(\phi+1)=\mathcal{P}(\phi)[1]$. □

Example 3.3. Let C be a smooth projective curve, $\mathcal{D}=D^b\mathrm{Coh}(C)$, $\Gamma=\mathbb{Z}^{\oplus 2}$ and $\mathrm{cl}=(\mathrm{rank},\deg)$. We set $Z\colon\Gamma\to\mathbb{C}$ to be $(r,d)\mapsto -d+ir$. Then $(Z,\mathrm{Coh}(C))$ is a stability condition, whose Z-semistable objects coincide with classical semistable sheaves on C.

3.2. The space of stability conditions. The space of stability conditions is defined as follows:

Definition 3.4. We define $\mathrm{Stab}_\Gamma(\mathcal{D})$ to be the set of stability conditions on $\mathcal{D}$ satisfying the *support property*, i.e. there is a constant $C > 0$ such that $\|\mathrm{cl}(E)\|/|Z(E)| < C$ holds for any $0 \neq E \in \cup_{\phi\in\mathbb{R}}\mathcal{P}(\phi)$.

The main result of Bridgeland [16] shows that the set $\mathrm{Stab}_\Gamma(\mathcal{D})$ has a structure of a complex manifold. If $\mathcal{D} = D^b\mathrm{Coh}(X)$ for a smooth projective variety X, $\Gamma = \mathrm{Im}(\mathrm{ch})$ and $\mathrm{cl} = \mathrm{ch}$, we set $\mathrm{Stab}(X) = \mathrm{Stab}_\Gamma(\mathcal{D})$. Let $\mathrm{Auteq}(X)$ be the group of exact autequivalences of $D^b\mathrm{Coh}(X)$. The space $\mathrm{Stab}(X)$ admits a left action of $\mathrm{Auteq}(X)$ and a right action of $\mathbb{C}$. The latter action is given by $(Z, \{\mathcal{P}(\phi)\}_{\phi\in\mathbb{R}}) \cdot \lambda = (e^{-i\pi\lambda}Z, \{\mathcal{P}(\phi + \mathrm{Re}\lambda)\}_{\phi\in\mathbb{R}})$ for $\lambda \in \mathbb{C}$. We are interested in the double quotient stack

$$[\mathrm{Auteq}(X)\backslash\mathrm{Stab}(X)/\mathbb{C}]\,. \tag{3.2}$$

The conjecture by Bridgeland [18] is that if X is a Calabi-Yau manifold, the above double quotient stack contains the stringy Kähler moduli space of X, that is the moduli space of complex structures of a mirror manifold of X.

Example 3.5.

1. If C is an elliptic curve, then (3.2) is shown in [16] to be isomorphic to the modular curve $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. This is compatible with the fact that C is self mirror.
2. Let $\pi\colon X \to \mathbb{P}^2$ be the total space of $\omega_{\mathbb{P}^2}$, which is a non-compact Calabi-Yau 3-fold. In this case, $\mathrm{Stab}(X)$ is defined to be $\mathrm{Stab}_\Gamma(\mathcal{D})$ where $\mathcal{D}$ is the bounded derived category of compact supported coherent sheaves on X, Γ is the image of $\mathrm{ch}\circ\pi_*$ in $H^*(\mathbb{P}^2, \mathbb{Q})$, and $\mathrm{cl} = \mathrm{ch}\circ\pi_*$. Then the quotient stack (3.2) contains $[(\mathbb{C}\setminus\mu_3)/\mu_3]$ by [11]. The latter stack is the parameter space ψ^3 of the mirror family of X given by
$$\{y_0^3 + y_1^3 + y_2^3 - 3\psi y_1y_2y_3 = 0\} \subset \mathbb{P}^2.$$

The space (3.2) is most interesting for projective Calabi-Yau 3-folds, e.g. quintic hypersurfaces in $\mathbb{P}^4$. Even in the quintic 3-fold case, the space (3.2) is very difficult to study. In this case, Bridgeland's conjecture [18] is stated in the following way:

Conjecture 3.6. *Let $X \subset \mathbb{P}^4$ be a smooth quintic 3-fold, and set $\mathcal{M}_K = [(\mathbb{C}\setminus\mu_5)/\mu_5]$. Then there is an embedding*

$$\mathcal{M}_K \hookrightarrow [\mathrm{Auteq}(X)\backslash\mathrm{Stab}(X)/\mathbb{C}]\,.$$

The above embedding should be given by the solutions of the Picard-Fuchs equation which the period integrals of the mirror family of X satisfy. Its explicit description is available in [69]. However in the projective Calabi-Yau 3-fold case, it is even not known that whether $\mathrm{Stab}(X)$ is non-empty or not. The first issue in solving Conjecture 3.6 is to construct stability conditions, which we discuss in the next subsection.

3.3. Existence problem. It has been a serious issue to construct Bridgeland stability conditions on projective Calabi-Yau 3-folds. Contrary to the one dimensional case, it turns out that there is no stability condition $\sigma \in \mathrm{Stab}(X)$ of the form $\sigma = (Z, \mathrm{Coh}(X))$ if $\dim X \geq 2$. From the arguments in string theory, we expect the following conjecture:

Conjecture 3.7. *Let X be a smooth projective variety and take $B+i\omega \in H^2(X,\mathbb{C})$ with ω ample class. Then there exists the heart of a bounded t-structure $\mathcal{A}_{B,\omega}$ on $D^b\mathrm{Coh}(X)$ such that the pair $\sigma_{B,\omega}=(Z_{B,\omega},\mathcal{A}_{B,\omega})$ determines a point in $\mathrm{Stab}(X)$, where $Z_{B,\omega}$ is given by*

$$Z_{B,\omega}(E)=-\int_X e^{-i\omega}\mathrm{ch}^B(E).$$

Here $\mathrm{ch}^B(E)$ is defined to be $e^{-B}\mathrm{ch}(E)$.

The resulting stability conditions are expected to form a *neighborhood at the large volume limit* in terms of string theory. The above conjecture is known to hold if $\dim X\leq 2$. In the $\dim X=2$ case, the heart $\mathcal{A}_{B,\omega}$ is constructed to be a certain tilting of $\mathrm{Coh}(X)$, which we are going to review.

Let X be a d-dimensional smooth projective variety and take $B+i\omega\in H^2(X,\mathbb{C})$ with ω ample. The ω-slope function on $\mathrm{Coh}(X)$ is defined to be

$$\mu_\omega(E)=\frac{\mathrm{ch}_1(E)\cdot\omega^{d-1}}{\mathrm{rank}(E)}\in\mathbb{R}\cup\{\infty\}.$$

Here $\mu_\omega(E)=\infty$ if $\mathrm{rank}(E)=0$.

Definition 3.8. An object $E\in\mathrm{Coh}(X)$ is *μ_ω-(semi)stable* if for any non-zero subobject $0\neq F\subsetneq E$, we have $\mu_\omega(F)<(\leq)\mu_\omega(E/F)$.

We define the pair of subcategories $(\mathcal{T}_{B,\omega},\mathcal{F}_{B,\omega})$ of $\mathrm{Coh}(X)$ to be

$$\mathcal{T}_{B,\omega}=\langle E\in\mathrm{Coh}(X): E \text{ is } \mu_\omega\text{-semistable with } \mu_\omega(E)>B\omega^{d-1}\rangle$$
$$\mathcal{F}_{B,\omega}=\langle E\in\mathrm{Coh}(X): E \text{ is } \mu_\omega\text{-semistable with } \mu_\omega(E)\leq B\omega^{d-1}\rangle.$$

Here $\langle *\rangle$ means the extension closure. The existence of Harder-Narasimhan filtrations with respect to the μ_ω-stability implies that the pair $(\mathcal{T}_{B,\omega},\mathcal{F}_{B,\omega})$ is a torsion pair (cf. [33]) of $\mathrm{Coh}(X)$. Its tilting defines another heart

$$\mathcal{B}_{B,\omega}=\langle\mathcal{F}_{B,\omega}[1],\mathcal{T}_{B,\omega}\rangle\subset D^b\mathrm{Coh}(X).$$

The following result is due to [1, 17, 85].

Proposition 3.9. *If $\dim X=2$, then $(Z_{B,\omega},\mathcal{B}_{B,\omega})\in\mathrm{Stab}(X)$.*

Proof. Here is a rough sketch of the proof: if $\dim X=2$, then $Z_{B,\omega}(E)$ is written as

$$Z_{B,\omega}(E)=-\mathrm{ch}_2^B(E)+\mathrm{ch}_0^B(E)\omega^2/2+i\mathrm{ch}_1^B(E)\omega.$$

The construction of $\mathcal{B}_{B,\omega}$ immediately implies $\mathrm{Im}Z_{B,\omega}(E)\geq 0$ for any $0\neq E\in\mathcal{B}_{B,\omega}$. We need to check that $\mathrm{Im}Z_{B,\omega}(E)=0$ implies $\mathrm{Re}Z_{B,\omega}(E)<0$. This property can be easily deduced from the the classical Bogomolov-Gieseker (BG) inequality in Theorem 3.10 below. □

The following BG inequality played an important role:

Theorem 3.10 ([13, 29]). *Let X be a d-dimensional smooth projective variety, and E a torsion free μ_ω-semistable sheaf on X. Then we have the following inequality:*

$$\left(\mathrm{ch}_1^B(E)^2-2\mathrm{ch}_0^B(E)\mathrm{ch}_2^B(E)\right)\cdot\omega^{d-2}\geq 0.$$

3.4. Double tilting construction for 3-folds. Suppose that X is a smooth projective 3-fold, and B, ω are defined over $\mathbb{Q}$. In this case, the central charge $Z_{B,\omega}$ is written as

$$Z_{B,\omega}(E) = -\mathrm{ch}_3^B(E) + \mathrm{ch}_1^B(E)\omega^2/2 + i\left(\mathrm{ch}_2^B(E)\omega - \mathrm{ch}_0^B(E)\omega^3/6\right).$$

Contrary to the surface case, the heart $\mathcal{B}_{B,\omega}$ does not fit into a stability condition with central charge $Z_{B,\omega}$. In [12], Bayer, Macri and the author constructed a further tilting of $\mathcal{B}_{B,\omega}$ in order to give a candidate of $\mathcal{A}_{B,\omega}$ in Conjecture 3.7. The key observation is the following lemma, which also relies on Theorem 3.10:

Lemma 3.11 ([12]). *For any $0 \neq E \in \mathcal{B}_{B,\omega}$, one of the following conditions hold:*

(i) $\mathrm{ch}_1^B(E)\omega^2 > 0$.
(ii) $\mathrm{ch}_1^B(E)\omega^2 = 0$ *and* $\mathrm{Im}Z_{B,\omega}(E) > 0$.
(iii) $\mathrm{ch}_1^B(E)\omega^2 = \mathrm{Im}Z_{B,\omega}(E) = 0$ *and* $\mathrm{Re}Z_{B,\omega}(E) < 0$.

The above lemma indicates that the vector $(\mathrm{ch}_1^B(E)\omega^2, \mathrm{Im}Z_{B,\omega}(E), -\mathrm{Re}Z_{B,\omega}(E))$ behaves as if it were $(\mathrm{rank}, c_1, \mathrm{ch}_2)$ on coherent sheaves on surfaces. In [12], this observation led to the following slope function on $\mathcal{B}_{B,\omega}$:

$$\nu_{B,\omega}(E) = \frac{\mathrm{Im}Z_{B,\omega}(E)}{\mathrm{ch}_1^B(E)\omega^2} \in \mathbb{Q} \cup \{\infty\}.$$

Here $\nu_{B,\omega}(E) = \infty$ if $\mathrm{ch}_1^B(E)\omega^2 = 0$. The above lemma shows that $\nu_{B,\omega}$ satisfies the weak see-saw property, and it defines a slope stability condition on $\mathcal{B}_{B,\omega}$. In [12], it was called *tilt-stability*:

Definition 3.12. An object $E \in \mathcal{B}_{B,\omega}$ is tilt (semi)stable if for any subobject $0 \neq F \subsetneq E$, we have $\nu_{B,\omega}(F) < (\leq)\nu_{B,\omega}(E/F)$.

We can show the existence of Harder-Narasimhan filtrations with respect to the tilt stability. Similarly to the surface case, the pair of subcategories $(\mathcal{T}'_{B,\omega}, \mathcal{F}'_{B,\omega})$ of $\mathcal{B}_{B,\omega}$ defined to be

$$\mathcal{T}'_{B,\omega} = \langle E \in \mathcal{B}_{B,\omega} : E \text{ is tilt semistable with } \nu_{B,\omega}(E) > 0 \rangle$$
$$\mathcal{F}'_{B,\omega} = \langle E \in \mathcal{B}_{B,\omega} : E \text{ is tilt semistable with } \nu_{B,\omega}(E) \leq 0 \rangle$$

is a torsion pair. By tilting, we have another heart

$$\mathcal{A}_{B,\omega} = \langle \mathcal{F}'_{B,\omega}[1], \mathcal{T}'_{B,\omega} \rangle \subset D^b\mathrm{Coh}(X).$$

By the construction, we have $\mathrm{Im}Z_{B,\omega}(E) \geq 0$ for any $E \in \mathcal{A}_{B,\omega}$. In [12], we proposed the following conjecture:

Conjecture 3.13 ([12]). *If $\dim X = 3$, we have $(Z_{B,\omega}, \mathcal{A}_{B,\omega}) \in \mathrm{Stab}(X)$.*

3.5. Conjectural BG inequality for 3-folds. Our double tilting construction led to a BG type inequality conjecture evaluating the third Chern characters of tilt semistable objects.

Conjecture 3.14 ([12]). *Let X be a smooth projective 3-fold. Then for any tilt semistable object $E \in \mathcal{B}_{B,\omega}$ with $\nu_{B,\omega}(E) = 0$, i.e. $\mathrm{ch}_2^B(E)\omega = \mathrm{ch}_0^B(E)\omega^3/6$, we have the inequality*

$$\mathrm{ch}_3^B(E) \leq \frac{1}{18}\mathrm{ch}_1^B(E)\omega^2.$$

Remark 3.15. In order to show $(Z_{B,\omega}, \mathcal{A}_{B,\omega})$ satisfies the property (3.1), it is enough to show the weaker inequality $\mathrm{ch}_3^B(E) < \mathrm{ch}_1^B(E)\omega^2/2$. If this is true, the existence of HN filtrations is proved in [12], while the support property remains open. The stronger bound in Conjecture 3.14 was obtained by the requirement that the equality is achieved for tilt semistable objects with zero discriminant.

It seems to be a hard problem to show Conjecture 3.14 even in concrete examples. So far, it is proved when $X = \mathbb{P}^3$ by Macri [46], X is a quadric 3-fold by Schmidt [61], and X is a principally polarized abelian 3-fold with Picard rank one by Maciocia-Piyaratne [49, 50]. Another kind of evidence is that assuming Conjecture 3.14 implies some open problems in other research fields. In [4], it was proved that Conjecture 3.14 implies (almost) Fujita conjecture for 3-folds: for any polarized 3-fold (X, L), $K_X + 4L$ is free and $K_X + 6L$ is very ample. In [83], it was also proved that Conjecture 3.14 implies a conjectural relationship between two kinds of DT type invariants inspired by string theory. This result will be reviewed in Theorem 4.5. It may be worth pointing out that, in both of the above applications, assuming a weaker inequality, say $\mathrm{ch}_3^B(E) < \mathrm{ch}_1^B(E)\omega^2/2$, does not imply anything. The stronger evaluation in Conjecture 3.14 is crucial for the proofs of the applications.

3.6. The space of weak stability conditions. Although the existence of Bridgeland stability conditions on projective Calabi-Yau 3-folds remains open, we are able to modify the definition of Bridgeland stability conditions so that the story in Subsection 1.5 works. The notion of *weak stability conditions* in [77] is one of them. In the situation of Subsection 3.1, we further fix a filtration $0 \subsetneq \Gamma_0 \subsetneq \cdots \subsetneq \Gamma_N = \Gamma$ such that each subquotient Γ_j/Γ_{j-1} is a free abelian group. Instead of considering a group homomorphism $Z \colon \Gamma \to \mathbb{C}$, we consider an element

$$Z = \{Z_i\}_{j=0}^N \in \prod_{j=0}^N \mathrm{Hom}(\Gamma_j/\Gamma_{j-1}, \mathbb{C}). \tag{3.3}$$

Given an element (3.3), we set $Z(v) \in \mathbb{C}$ for $v \in \Gamma$ as follows: there is a unique $0 \leq m \leq N$ such that $v \in \Gamma_m \setminus \Gamma_{m-1}$, where $\Gamma_{-1} = \emptyset$. Then $Z(v)$ is defined to be $Z_m([v]) \in \mathbb{C}$ where $[v]$ is the class of v in Γ_m/Γ_{m-1}.

Definition 3.16. A weak stability condition on $\mathcal{D}$ with respect to the filtration $\Gamma_\bullet$ is data $(Z, \{\mathcal{P}(\phi)\}_{\phi \in \mathbb{R}})$, where Z is as in (3.3), $\mathcal{P}(\phi) \subset \mathcal{D}$ is a full subcategory, satisfying the same axiom in Definition 3.1.

Similarly to Lemma 3.2, giving a weak stability condition is equivalent to giving $(Z, \mathcal{A})$, where Z is as in (3.3), $\mathcal{A} \subset \mathcal{D}$ is the heart of a bounded t-structure, satisfying the same conditions in Lemma 3.2. We denote by $\mathrm{Stab}_{\Gamma_\bullet}(\mathcal{D})$ the set of weak stability conditions on $\mathcal{D}$ with respect to $\Gamma_\bullet$ with a support property. This set also has a structure of a complex

manifold, and coincides with $\mathrm{Stab}_\Gamma(\mathcal{D})$ if $N = 0$, i.e. the filtration $\Gamma_\bullet$ is trivial. In general, it is easier to show the non-emptiness for the space $\mathrm{Stab}_{\Gamma_\bullet}(\mathcal{D})$ with a non-trivial filtration $\Gamma_\bullet$. The result of Theorem 2.9 was obtained by the wall-crossing formula in the space of weak stability conditions on the following triangulated category

$$\mathcal{D}_X = \langle \mathcal{O}_X, \mathrm{Coh}_{\leq 1}(X)\rangle_{\mathrm{tr}} \subset D^b\mathrm{Coh}(X).$$

Here $\mathrm{Coh}_{\leq 1}(X)$ is the category of one or zero dimensional sheaves on X, and $\langle * \rangle_{\mathrm{tr}}$ is the triangulated closure. The relevant data is

$$\Gamma_0 = \mathbb{Z} \oplus H_2(X,\mathbb{Z}) \oplus \{0\} \subset \Gamma_1 = \Gamma = \mathbb{Z} \oplus H_2(X,\mathbb{Z}) \oplus \mathbb{Z}$$

with the map cl given by $\mathrm{cl}(E) = (\mathrm{ch}_3(E), \mathrm{ch}_2(E), \mathrm{ch}_0(E))$. Here $H_2(X,\mathbb{Q})$ is identified with $H^4(X,\mathbb{Q})$ via Poincaré duality. The result of Theorem 2.9 is proved along with the wall-crossing argument of Subsection 1.5 with respect to the one parameter family of weak stability conditions on $\mathcal{D}_X$

$$\sigma_\theta = (Z_{\omega,\theta}, \mathcal{A}_X) \in \mathrm{Stab}_{\Gamma_\bullet}(\mathcal{D}_X), \quad 1/2 \leq \theta < 1$$

from $\theta = 1/2$ to $\theta \to 0$. Here ω is an ample divisor on X, $Z_{\omega,\theta,j}$ are given by

$$Z_{\omega,\theta,0} \colon \Gamma_0 \ni (n,\beta) \mapsto n - (\omega\cdot\beta)i, \quad Z_{\omega,\theta,1} : \mathbb{Z} \ni r \mapsto r\exp(i\pi\theta).$$

The heart $\mathcal{A}_X \subset \mathcal{D}_X$ is obtained as the extension closure of objects $\mathcal{O}_X$ and $\mathrm{Coh}_{\leq 1}(X)[-1]$. We are able to construct DT type invariant

$$\mathrm{DT}_{\sigma_\theta}(1,0,-\beta,-n) \in \mathbb{Q}$$

which counts σ_θ-semistable objects $E \in \mathcal{A}_X$ with $\mathrm{ch}(E) = (1,0,-\beta,-n)$. It is shown that

$$\mathrm{DT}_{\theta\to 1}(1,0,-\beta,-n) = P_{n,\beta}, \quad \mathrm{DT}_{\theta=1/2}(1,0,-\beta,-n) = L_{n,\beta}$$

where $L_{n,\beta}$ is the invariant in Theorem 2.9. The wall-crossing formula describes the difference between $P_{n,\beta}$ and $L_{n,\beta}$. A similar wall-crossing phenomena also implies the relationship between $I_{n,\beta}$ and $P_{n,\beta}$ in Theorem 2.13. Combined them, we obtain the result of Theorem 2.9. Some more detail is also available in [81].

4. Further results and conjectures

4.1. Multiple cover formula conjecture. Although Conjecture 2.8 (i) is proved, a stronger version of the rationality conjecture remains open. It was proposed by Pandharipande-Thomas [59], and predicts the product expansion (called *Gopakumar-Vafa form*) of the generating series of PT invariants:

$$\begin{aligned} &1 + \sum_{n\in\mathbb{Z},\beta>0} P_{n,\beta} q^n t^\beta \\ &= \prod_{\beta>0}\left(\prod_{j=1}^{\infty}(1-(-q)^j t^\beta)^{jn_0^\beta} \prod_{g=1}^{\infty}\prod_{k=0}^{2g-2}(1-(-q)^{g-1}t^\beta)^{(-1)^{k+g}n_g^\beta\binom{2g-2}{k}}\right) \end{aligned}$$

for some $n_g^\beta \in \mathbb{Z}$. Using Theorem 2.9 and Theorem 2.13, the above strong rationality conjecture is proved in [81] to be equivalent to the following conjecture:

Conjecture 4.1 ([36, 81]). *We have the following identity:*

$$N_{n,\beta}=\sum_{k\in\mathbb{Z}_{\geq 1},k|(n,\beta)}\frac{1}{k^2}N_{1,\beta/k}.$$

The invariant $N_{1,\beta}$ is always integer, and the above conjecture is stronger then the integrality conjecture by Kontsevich-Soibelman [40].

4.2. Gepner type stability conditions. Let $W\in A=\mathbb{C}[x_1,\cdots,x_n]$ be a homogeneous polynomial of degree d. By definition, a graded matrix factorization consists of data

$$P^0\stackrel{p^0}{\to}P^1\stackrel{p^1}{\to}P^0(d)$$

where P^i are graded free A-modules of finite rank, p^i are homomorphisms of graded A-modules, $P^i\mapsto P^i(1)$ is the shift of the grading, satisfying $p^1\circ p^0=p^0\circ p^1=\cdot W$. The triangulated category $\mathrm{HMF}(W)$ is defined to be the homotopy category of graded matrix factorizations of W. It has a structure of a triangulated category, and related to $D^b\mathrm{Coh}(X)$ for the hypersurface $X=(W=0)\subset\mathbb{P}^{n-1}$. For instance if $d=n$, there is an equivalence [56]

$$\mathrm{HMF}(W)\stackrel{\sim}{\to}D^b\mathrm{Coh}(X). \tag{4.1}$$

As an analogy of Gieseker stability on $\mathrm{Coh}(X)$, we expect the existence of a natural stability condition on $\mathrm{HMF}(W)$. Based on the earlier works [41, 87], the following conjecture is proposed in [71]:

Conjecture 4.2. *There is a Bridgeland stability condition $\sigma_G=(Z_G,\{\mathcal{P}_G(\phi)\}_{\phi\in\mathbb{R}})$ on $\mathrm{HMF}(W)$ whose central charge Z_G is given by*

$$Z_G\left(\bigoplus_{i=1}^N A(m_i)\rightleftarrows\bigoplus_{i=1}^N A(n_i)\right)=\sum_{i=1}^N\left(e^{\frac{2\pi m_i\sqrt{-1}}{d}}-e^{\frac{2\pi n_i\sqrt{-1}}{d}}\right)$$

and the set of semistable objects satisfy $\tau\mathcal{P}_G(\phi)=\mathcal{P}_G(\phi+2/d)$, where τ is the graded shift functor $P^\bullet\mapsto P^\bullet(1)$.

If $n=d=5$, i.e. X is a quintic 3-fold, a stability condition above is expected to correspond to the orbifold point $0\in\mathcal{M}_K$ in Conjecture 3.6 called Gepner point. By this reason, a stability condition in Conjecture 4.2 is called *Gepner type*. Some evidence of Conjecture 4.2 is available in [41, 70, 71]. Suppose that Conjecture 4.2 is true for $n=d=5$. Then as an analogy of Fan-Jarvis-Ruan-Witten theory [26] in GW theory, we may define the DT type invariant

$$\mathrm{DT}_G(\gamma)\in\mathbb{Q},\ \gamma\in\mathrm{HH}_0(W) \tag{4.2}$$

which counts σ_G-semistable graded matrix factorizations $P^\bullet$ with $\mathrm{ch}(P^\bullet)=\gamma$. Here $\mathrm{HH}_0(W)$ is the zero-th Hochschild homology of $\mathrm{HMF}(W)$, and ch is the Chern character map on graded matrix factorizations (cf. [60]). Because of the property of σ_G, the invariant (4.2) should satisfy $\mathrm{DT}_G(\gamma)=\mathrm{DT}_G(\tau_*\gamma)$. Under the Orlov equivalence (4.1), the equivalence τ on the LHS corresponds to the equivalence $\mathrm{ST}_{\mathcal{O}_X}\circ\mathcal{O}_X(1)$ on the RHS, where $\mathrm{ST}_{\mathcal{O}_X}$ is the Seidel-Thomas twist [62] associated to $\mathcal{O}_X$. Along with the argument in Subsection 1.5, the existence of the invariant (4.2) should imply a hidden symmetry of the generating series of the original DT invariants on the quintic hypersurface $X=(W=0)$ with respect to the equivalence $\mathrm{ST}_{\mathcal{O}_X}\circ\mathcal{O}_X(1)$.

4.3. S-duality conjecture for DT invariants. Let us recall the original S-duality conjecture by Vafa-Witten [86]. It predicts the (at least almost) modularity of the generating series of Euler characteristics of moduli spaces of stable torsion free sheaves on algebraic surfaces with a fixed rank and a first Chern class. We refer to [28] for the developments on the S-duality conjecture so far. Instead of stable torsion free sheaves on algebraic surfaces, we consider semistable pure two dimensional torsion sheaves on Calabi-Yau 3-folds, and DT invariants counting them. Let X be a smooth projective Calabi-Yau 3-fold, H an ample divisor on X and fix a divisor class $P \in H^2(X, \mathbb{Z})$. We consider the following generating series

$$\mathrm{DT}_H(P) = \sum_{\beta \in H_2(X), n \in \mathbb{Q}} \mathrm{DT}_H(0, P, -\beta, -n - P \cdot c_2(X)/24) q^n t^\beta. \tag{4.3}$$

Here each coefficient counts H-semistable $E \in \mathrm{Coh}(X)$ whose Mukai vector (*not* Chern character) satisfies $\mathrm{ch}(E) \cdot \sqrt{\mathrm{td}_X} = (0, P, -\beta, -n)$. As a 3-fold version of the S-duality conjecture, we expect that the series (4.3) satisfies a modular transformation property of (almost) Jacobi forms. (We refer to [25] for a basic of Jacobi forms.) Some computations of the invariants $\mathrm{DT}_H(0, P, -\beta, -n)$ are available in [30, 31]. Also the transformation formula of the series (4.3) under a flop is obtained in [68]. Let us consider a flop diagram (2.2) with Y projective, and ω an ample divisor on Y. We assume that the exceptional locus C, $C^\dagger$ of f, $f^\dagger$ are isomorphic to $\mathbb{P}^1$ with $p = f(C) = f^\dagger(C^\dagger)$. Let l be the scheme theoretic length of $f^{-1}(p)$ at the generic point of C.

Theorem 4.3 ([68]). *There exist $n_j \in \mathbb{Z}_{\geq 1}$ for $1 \leq j \leq l$ such that we have the following formula:*

$$\begin{aligned}\mathrm{DT}_{f^{\dagger *}\omega}(\phi_* P) &= \phi_* \mathrm{DT}_{f^*\omega}(P) \\ &\quad \cdot \prod_{j=1}^{l} \left\{ i^{jP\cdot C - 1} \eta(q)^{-1} \vartheta_{1,1}(q, ((-1)^{\phi_* P} t)^{jC^\dagger}) \right\}^{j n_j P \cdot C}.\end{aligned}$$

Here ϕ_* is the variable change $(n, \beta) \mapsto (n, \phi_* \beta)$, $\eta(q)$ is the Dedekind eta function and $\vartheta_{1,1}(q, t)$ is the Jacobi theta function, given as follows:

$$\eta(q) = q^{\frac{1}{24}} \prod_{k \geq 1} (1 - q^k), \quad \vartheta_{1,1}(q, t) = \sum_{k \in \mathbb{Z}} q^{\frac{1}{2}\left(k + \frac{1}{2}\right)^2} (-t)^{k + \frac{1}{2}}. \tag{4.4}$$

Although $f^*\omega$ is not ample, it is shown that the invariants $\mathrm{DT}_{f^*\omega}(v)$ are well-defined. Recall that $\eta(q)$ is a modular form of weight $1/2$, $\vartheta_{1,1}(q, t)$ is a Jacobi form of weight $1/2$ and index $1/2$. The result of Theorem 4.3 shows that the series (4.3) transforms under a flop by a multiplication of a meromorphic Jacobi form, which gives evidence of the S-duality conjecture for DT invariants.

4.4. Mathematical approach toward OSV conjecture. In string theory, the OSV conjecture [57] predicts a certain approximation

$$\mathcal{Z}_{\mathrm{BH}} \sim |\mathcal{Z}_{\mathrm{top}}|^2 \tag{4.5}$$

where the LHS is the partition function of black hole entropy, and the RHS is the partition function of topological string. A version of the above conjecture is mathematically stated as

an approximation between the generating series of DT invariants counting torsion sheaves on Calabi-Yau 3-folds, and the generating series of GW invariants. In [23], Denef-Moore proposed a relationship among the series (4.3) and the generating series of $I_{n,\beta}$, $P_{n,\beta}$ in order to give a derivation of (4.5). A mathematical refinement of Denef-Moore conjecture is stated in [83]. For simplicity, suppose that $\mathrm{Pic}(X)$ is generated by an ample divisor H. For $m \in \mathbb{Z}_{>0}$, we define the following cut off series

$$I^m(q,t) = \sum_{(\beta,n)\in C(m)} I_{n,\beta} q^n t^\beta, \quad P^m(q,t) = \sum_{(\beta,n)\in C(m)} P_{n,\beta} q^n t^\beta.$$

Here $C(m) = \{(\beta,n) : \beta H < mH^3/2, |n| < m^2H^3/2\}$. Moreover, we define the cut off generating series of D6-anti-D6 brane counting

$$\mathcal{Z}_{\mathrm{D6}-\overline{\mathrm{D6}}}(q,t,w) = \sum_{m_2-m_1=m} q^{H^3(m_1^3-m_2^3)/6} t^{H^2(m_1^2-m_2^2)/2} w^{H^3m^3/6+Hc_2(X)m/12} I^m(qw^{-1}, q^{m_2H}tw^{-mH}) P^m(qw^{-1}, q^{-m_1H}t^{-1}w^{-mH}).$$

Conjecture 4.4 ([23, 83]). *For $m \gg 0$, we have the equality*

$$\mathrm{DT}_H(mH) = \frac{\partial}{\partial w} \mathcal{Z}_{\mathrm{D6}-\overline{\mathrm{D6}}}(q,t,w)|_{w=-1}$$

modulo terms of $q^n t^\beta$ with

$$-\frac{H^3}{24}m^3\left(1-\frac{1}{m}\right) \leq n + \frac{(\beta\cdot H)^2}{2mH^3}.$$

In [83], we proved the following:

Theorem 4.5 ([83]). *The unweighted version of Conjecture 4.4 is true if we assume Conjecture 3.14.*

Even if Conjecture 4.4 is proved, still the relationship (4.5) is not obvious. If we follow the arguments in [23], at least we need to prove S-duality conjecture for DT invariants in the previous subsection and MNOP conjecture. Moreover we need to make a mathematical understanding of the approximation $\sim$ in (4.5). Although the relationship (4.5) is motivated by string theory, it seems to involve deep and interesting mathematics.

Acknowldegements. The author is supported by World Premier International Research Center Initiative (WPI initiative), MEXT, Japan, and also by Grant-in Aid for Scientific Research grant (22684002) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] D. Arcara and A. Bertram, *Bridgeland-stable moduli spaces for K-trivial surfaces. With an appendix by Max Lieblich*, J. Eur. Math. Soc. **15** (2013), 1–38.

[2] A. Bayer, *Polynomial Bridgeland stability conditions and the large volume limit*, Geom. Topol. **13** (2009), 2389–2425.

[3] K. Behrend and J. Bryan, *Super-rigid Donaldson - Thomas invariants*, Math. Res. Lett. **14** (2007), 559–571.

[4] A. Bayer, A. Bertram, E. Macri, and Y. Toda, *Bridgeland stability conditions on 3-folds II: An application to Fujita's conjecture*, J. Algebraic Geom. (to appear), arXiv:1106.3430.

[5] L. Borisov and A. Caldararu, *The Pfaffian-Grassmannian derived equivalence*, J. Algebraic. Geom. **18** (2009), 201–222.

[6] K. Behrend, *Donaldson-Thomas invariants via microlocal geometry*, Ann. of Math **170** (2009), 1307–1338.

[7] R. Bezrukavnikov, *Perverse coherent sheaves (after Deligne)*, preprint, arXiv:0005152.

[8] K. Behrend and B. Fantechi, *The intrinsic normal cone*, Invent. Math. **128** (1997), 45–88.

[9] _____, *Symmetric obstruction theories and Hilbert schemes of points on threefolds*, Algebra Number Theory **2** (2008), 313–345.

[10] T. Bridgeland, A. King, and M. Reid, *The McKay correspondence as an equivalence of derived categories*, J. Amer. Math. Soc. **14** (2001), 535–554.

[11] A. Bayer and E. Macri, *The space of stability conditions on the local projective plane*, Duke Math. J. **160** (2011), 263–322.

[12] A. Bayer, E. Macri, and Y. Toda, *Bridgeland stability conditions on 3-folds I: Bogomolov-Gieseker type inequalities*, J. Algebraic Geom. **23** (2014), 117–163.

[13] F. A. Bogomolov, *Holomorphic tensors and vector bundles on projective manifolds*, Izv. Akad. Nauk SSSR Ser. Mat. **42** (1978), 1227–1287.

[14] T. Bridgeland, *Flops and derived categories*, Invent. Math **147** (2002), 613–632.

[15] _____, *Stability conditions on a non-compact Calabi-Yau threefold*, Comm. Math. Phys. **266** (2006), 715–733.

[16] _____, *Stability conditions on triangulated categories*, Ann. of Math **166** (2007), 317–345.

[17] _____, *Stability conditions on K3 surfaces*, Duke Math. J. **141** (2008), 241–291.

[18] _____, *Spaces of stability conditions*, Proc. Sympos. Pure Math. **80** (2009), 1–21, Algebraic Geometry-Seatle 2005.

[19] _____, *Hall algebras and curve-counting invariants*, J. Amer. Math. Soc. **24** (2011), 969–998.

[20] J. Calabrese, *Donaldson-Thomas invariants on Flops*, preprint, arXiv:1111.1670.

[21] J. Cheah, *On the cohomology of Hilbert schemes of points*, J. Algebraic Geom. **5** (1996), 479–511.

[22] M. Van den Bergh, *Three dimensional flops and noncommutative rings*, Duke Math. J. **122** (2004), 423–455.

[23] F. Denef and G. Moore, *Split states, Entropy Enigmas, Holes and Halos*, arXiv:hep-th/0702146.

[24] M. Douglas, *Dirichlet branes, homological mirror symmetry, and stability*, Proceedings of the 1998 ICM (2002), 395–408.

[25] M. Eichler and D. Zagier, *The theory of Jacobi forms*, Progress in Mathematics, vol. 55, Birkhäuser Boston, 1985.

[26] H. Fan, T. J. Jarvis, and Y. Ruan, *The Witten equation and its virtual fundamental cycle*, preprint, arXiv:0712.4025.

[27] L. Göttsche, *The Betti numbers of the Hilbert scheme of points on a smooth projective surface*, Math. Ann. **286** (1990), 193–207.

[28] ———, *Invariants of Moduli Spaces and Modular Forms*, Rend. Istit. Mat. Univ. Trieste **41** (2009), 55–76.

[29] D. Gieseker, *On a theorem of Bogomolov on Chern Classes of Stable Bundles*, Amer. J. Math. **101** (1979), 77–85.

[30] A. Gholampour and A. Sheshmani, *Donaldson-Thomas Invariants of 2-Dimensional sheaves inside threefolds and modular forms*, preprint, arXiv:1309.0050.

[31] ———, *Generalized Donaldson-Thomas Invariants of 2-Dimensional sheaves on local* $\mathbb{P}^2$, preprint, arXiv:1309.0056.

[32] D. Huybrechts and M. Lehn, *Geometry of moduli spaces of sheaves*, Aspects in Mathematics, vol. E31, Vieweg, 1997.

[33] D. Happel, I. Reiten, and S. O. Smal, *Tilting in abelian categories and quasitilted algebras*, Mem. Amer. Math. Soc, vol. 120, 1996.

[34] D. Huybrechts and R. P. Thomas, $\mathbb{P}$*-objects and autoequivalences of derived categories*, Math. Res. Lett. (2006), 87–98.

[35] M. Inaba, *Toward a definition of moduli of complexes of coherent sheaves on a projective scheme*, J. Math. Kyoto Univ. **42-2** (2002), 317–329.

[36] D. Joyce and Y. Song, *A theory of generalized Donaldson-Thomas invariants*, Mem. Amer. Math. Soc. **217** (2012).

[37] M. Kashiwara, t*-structures on the derived categories of holonomic* $\mathcal{D}$*-modules and coherent* $\mathcal{O}$*-modules*, Mosc. Math. J. **981** (2004), 847–868.

[38] Y. Kawamata, D*-equivalence and* K*-equivalence*, J. Differential Geom. **61** (2002), 147–171.

[39] M. Kontsevich, *Homological algebra of mirror symmetry*, Proceedings of ICM, vol. 1, Birkhäuser, Basel, 1995.

[40] M. Kontsevich and Y. Soibelman, *Stability structures, motivic Donaldson-Thomas invariants and cluster transformations*, preprint, arXiv:0811.2435.

[41] H. Kajiura, K. Saito, and A. Takahashi, *Matrix factorizations and representations of quivers II. Type ADE case*, Advances in Math **211** (2007), 327–362.

[42] J. Li, *Zero dimensional Donaldson-Thomas invariants of threefolds*, Geom. Topol. **10** (2006), 2117–2171.

[43] M. Lieblich, *Moduli of complexes on a proper morphism*, J. Algebraic Geom. **15** (2006), 175–206.

[44] M. Levine and R. Pandharipande, *Algebraic cobordism revisited*, Invent. Math. **176** (2009), 63–130.

[45] W. P. Li and Z. Qin, *On the euler numbers of certain moduli spaces of curves and points*, Comm. Anal. Geom. **14** (2006), 387–410.

[46] E. Macri, *A generalized Bogomolov-Gieseker inequality for the three-dimensional projective space*, preprint, arXiv:1207.4980.

[47] ———, *Stability conditions on curves*, Math. Res. Lett. (2007), 657–672.

[48] D. Maulik, N. Nekrasov, A. Okounkov, and R. Pandharipande, *Gromov-Witten theory and Donaldson-Thomas theory. I*, Compositio. Math **142** (2006), 1263–1285.

[49] A. Maciocia and D. Piyaratne, *Fourier-Mukai Transforms and Bridgeland Stability Conditions on Abelian Threefolds*, preprint, arXiv:1304.3887.

[50] ———, *Fourier-Mukai Transforms and Bridgeland Stability Conditions on Abelian Threefolds II*, preprint, arXiv:1310.0299.

[51] S. Mukai, *Duality between $D(X)$ and $D(\hat{X})$ with its application to picard sheaves*, Nagoya Math. J. **81** (1981), 101–116.

[52] K. Nagao, *On higher rank Donaldson-Thomas invariants*, preprint, arXiv:1002.3608.

[53] ———, *Donaldson-Thomas theory and cluster algebras*, Duke Math. J. **162** (2013), 1313–1367.

[54] K. Nagao and H. Nakajima, *Counting invariant of perverse coherent sheaves and its wall-crossing*, Int. Math. Res. Not. (2011), 3855–3938.

[55] D. Orlov, *On Equivalences of derived categories and $K3$ surfaces*, J. Math. Sci (New York) **84** (1997), 1361–1381.

[56] D. Orlov, *Derived categories of coherent sheaves and triangulated categories of singularities*, Algebra, arithmetic, and geometry: in honor of Yu. I. Manin, Progr. Math. **270** (2009), 503–531.

[57] H. Ooguri, A. Strominger, and C. Vafa, *Black hole attractors and the topological string*, Phys. Rev. D **70** (2004), arXiv:hep-th/0405146.

[58] R. Pandharipande and A. Pixton, *Gromov-Witten/Pairs correspondence for the quintic 3-fold*, preprint, arXiv:1206.5490.

[59] R. Pandharipande and R. P. Thomas, *Curve counting via stable pairs in the derived category*, Invent. Math. **178** (2009), 407–447.

[60] A. Polishchuk and A. Vaintrob, *Chern characters and Hirzebruch-Riemann-Roch formula for matrix factorizations*, Duke Math. J. **161** (2012), 1863–1926.

[61] B. Schmidt, *A generalized Bogomolov-Gieseker inequality for the smooth quadric threefold*, preprint, arXiv:1309.4265.

[62] P. Seidel and R. P. Thomas, *Braid group actions on derived categories of coherent sheaves*, Duke Math. J. **108** (2001), 37–107.

[63] J. Stoppa and R. P. Thomas, *Hilbert schemes and stable pairs: GIT and derived category wall crossings*, Bull. Soc. Math. France **139** (2011), 297–339.

[64] R. Stanley, *Enumerative combinatorics*, Cambridge University Press, 1999.

[65] J. Stoppa, *D0-D6 states counting and GW invariants*, Lett. Math. Phys. **102** (2012), 149–180.

[66] B. Szendrői, *Non-commutative Donaldson-Thomas theory and the conifold*, Geom. Topol. **12** (2008), 1171–1202.

[67] R. P. Thomas, *A holomorphic Casson invariant for Calabi-Yau 3-folds and bundles on $K3$-fibrations*, J. Differential. Geom **54** (2000), 367–438.

[68] Y. Toda, *Flops and S-duality conjecture*, preprint, arXiv:1311.7476.

[69] ———, *Gepner point and strong Bogomolov-Gieseker inequality for quintic 3-folds*, to appear in Professor Kawamata's 60th volume, arXiv:1305.0345.

[70] ———, *Gepner type stability condition via Orlov/Kuznetsov equivalence*, preprint, arXiv:1308.3791.

[71] ———, *Gepner type stability conditions on graded matrix factorizations*, preprint, arXiv:1302.6293.

[72] ———, *Birational Calabi-Yau 3-folds and BPS state counting*, Communications in Number Theory and Physics **2** (2008), 63–112.

[73] ———, *Moduli stacks and invariants of semistable objects on K3 surfaces*, Advances in Math **217** (2008), 2736–2781.

[74] ———, *Stability conditions and crepant small resolutions*, Trans. Amer. Math. Soc. **360** (2008), 6149–6178.

[75] ———, *Limit stable objects on Calabi-Yau 3-folds*, Duke Math. J. **149** (2009), 157–208.

[76] ———, *Stability conditions and Calabi-Yau fibrations*, J. Algebraic Geom. **18** (2009), 101–133.

[77] ———, *Curve counting theories via stable objects I: DT/PT correspondence*, J. Amer. Math. Soc. **23** (2010), 1119–1157.

[78] ———, *Generating functions of stable pair invariants via wall-crossings in derived categories*, Adv. Stud. Pure Math. **59** (2010), 389–434, New developments in algebraic geometry, integrable systems and mirror symmetry (RIMS, Kyoto, 2008).

[79] ———, *On a computation of rank two Donaldson-Thomas invariants*, Communications in Number Theory and Physics **4** (2010), 49–102.

[80] ———, *Curve counting invariants around the conifold point*, J. Differential Geom. **89** (2011), 133–184.

[81] ———, *Stability conditions and curve counting invariants on Calabi-Yau 3-folds*, Kyoto Journal of Mathematics **52** (2012), 1–50.

[82] ———, *Stable pairs on local K3 surfaces*, J. Differential. Geom. **92** (2012), 285–370.

[83] ———, *Bogomolov-Gieseker type inequality and counting invariants*, Journal of Topology **6** (2013), 217–250.

[84] ———, *Curve counting theories via stable objects II. DT/ncDT flop formula*, J. Reine Angew. Math. **675** (2013), 1–51.

[85] ———, *Stability conditions and extremal contractions*, Math. Ann. **357** (2013), 631–685.

[86] C. Vafa and E. Witten, *A Strong Coupling Test of S-Duality*, Nucl. Phys. B **431** (1994).

[87] J. Walcher, *Stability of Landau-Ginzburg branes*, Journal of Mathematical Physics **46** (2005), arXiv:hep-th/0412274.

[88] B. Young, *Computing a pyramid partition generating function with dimer shuffling*, J. Combin. Theory Ser. (2009), 334–350.

Kavli Institute for the Physics and Mathematics of the Universe, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, 277-8583, Japan.

E-mail: yukinobu.toda@ipmu.jp

Derived algebraic geometry and deformation quantization

Bertrand Toën

Abstract. This is a report on recent progress concerning the interactions between derived algebraic geometry and deformation quantization. We present the notion of derived algebraic stacks, of shifted symplectic and Poisson structures, as well as the construction of deformation quantization of shifted Poisson structures. As an application we propose a general construction of the quantization of the moduli space of G-bundles on an oriented space of arbitrary dimension.

Mathematics Subject Classification (2010). Primary 14A20; Secondary 16E35, 53D55.

Keywords. Derived algebraic geometry, deformation quantization, symplectic structures, Poisson structures, derived categories.

1. Introduction

Quantization is an extremely vast subject, particularly because it has a long-standing physical origin and history. Even from the more restrictive point of view of a pure mathematician, quantization possesses many facets and connects with a wide variety of modern mathematical domains. This variety of interactions explains the numerous mathematical incarnations that the expression *quantization* finds in the existing literature, though a common denominator seems to be a *perturbation of a commutative structure into a non-commutative structure*. For a commutative object X (typically a commutative algebra or a manifold), a quantization is most often realized as a family $X_\hbar$, of objects depending on a parameter $\hbar$, which recovers X when $\hbar = 0$ and which is non-commutative for general values of $\hbar$. The existence of the family $X_\hbar$ is in most cases related to the existence of certain additional geometric structures, such as symplectic or Poisson structures.

The purpose of this manuscript is to present a new approach to quantization, or more specifically to the construction and the study of interesting non-commutative deformations of commutative objects of geometrico-algebraic origins. This new approach is based on the *derived algebraic geometry*, a version of algebraic geometry that has emerged in the last decade (see [31, 32]), and which itself consists of a *homotopical perturbation of algebraic geometry*. Derived algebraic geometry not only leads to a unified geometric interpretation of most of the already existing quantum objects (e.g. it treats the quantum group and deformation quantization of a Poisson manifold on an equal footing), but also opens up a whole new world of quantum objects which, as far as we know, have not been identified in the past even though they seem to appear naturally in algebraic geometry, algebraic topology, or representation theory.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Convention. All varieties, algebras, schemes, stacks, algebraic groups etc ... will be over a ground field k of characteristic zero.

2. Quantization as deformed categories: three motivating examples

In this first section we briefly recall three well known important examples of "quantization in action" in different domains: quantum groups, skein algebras and Donaldson-Thomas invariants. We identify the natural moduli spaces behind each of these examples and explain how they all can be considered from the unified point of view of deformations of categories and monoidal categories of sheaves.

Quantum groups. Probably the most famous and most fundamental of quantum objects are quantum groups. For an algebraic group G, with lie algebra $\mathfrak{g}$, and a choice of a G-invariant element $p \in Sym^2(\mathfrak{g})$, Drinfeld constructs a quantum group (see [10]). Algebraically the quantum group is a deformation of the Hopf algebra $A = \mathcal{O}(G)$ of functions on G, into a non-commutative Hopf algebra $A_\hbar$.

Skein algebras. Skein algebras appear in low dimensional topology (see [38]). They are associated with a given Riemann surface Σ, and are explicitly defined in terms of generators and relations. The generators are given by simple curves traced on the surface Σ, and the relations are given by the so-called skein relations, which possess natural deformations by a parameter $q = e^{2i\pi\hbar}$. The skein algebra associated with Σ, $K_\hbar(\Sigma)$, is a non-commutative deformation of the ring of functions on the character variety of Σ for the group Sl_2 (i.e. the affine algebraic variety whose points describe Sl_2-representations of the fundamental group $\pi_1(\Sigma)$).

Donaldson-Thomas invariants. For X a Calabi-Yau algebraic variety of dimension 3, we denote by $\mathcal{M}_X$ the moduli space of stable vector bundles with fixed numerical invariants. It is a singular variety in general but with a very specific local structure. Indeed, it is known that locally around each point, X embeds into a smooth ambient variety Z as the critical points of a function $f : Z \to \mathbb{A}^1$. Each of these locally defined functions f define a (perverse) sheaf ν_f of vanishing cycles on X, which under an orientability assumption glue to a globally defined perverse sheaf $\mathcal{E}$ on X (see [2, 7, 8] for more on the subject). The sheaf $\mathcal{E}$ is a quantization of the space X, in the sense that it can be seen to be a deformation of the line bundle of *virtual half forms on* X (we refer here to the next section for more about virtual structures). This deformation is again a non-commutative deformation, but this time in a dramatic way as the multiplicative structure itself is lost and $\mathcal{E}$ only exists as a sheaf of (complexes of) vector spaces on X.

Despite their different origins and differences in appearance these three examples of quantization can be considered in a striking unified way: they all are deformations of categories of sheaves on natural moduli spaces, where the categories eventually come equipped with monoidal structures. More is true, these deformations are all induced by the same type of structures on the corresponding moduli spaces, at least when they are appropriately viewed as derived algebraic stacks as we will explain later. The moduli spaces related to these three examples are easy to guess: they are respectively the moduli space $Bun_G(*)$

of G-bundles on a point $*$, the moduli space $Bun_{Sl_2}(\Sigma)$ of Sl_2-bundles on the surface Σ, and the moduli space $\mathcal{M}_X$ of algebraic Gl_n-bundles on X, also denoted $Bun_{Gl_n}(X)$. In the case of the quantum group, there is no non-trivial G-bundles on a point, but the trivial G-bundle possesses many automorphisms. The moduli space $Bun_G(*)$ is thus trivial from the point of view of algebraic varieties but can be realized as a non-trivial algebraic stack BG. Quasi-coherent sheaves on BG are nothing else than linear representations of G, $QCoh(BG) = Rep(G)$. The quantum group $A_\hbar$ can then be realized as a deformation of $QCoh(BG)$ considered as a braided monoidal category. For the case of skein algebras, we already mentioned that $K_{\hbar=0}(\Sigma)$ is the ring of functions on the moduli space $\chi(\Sigma) = Bun_{Sl_2}(\Sigma)$. The moduli space $Bun_{Sl_2}(\Sigma)$ is an affine algebraic variety and thus its category of quasi-coherent sheaves is equivalent to modules over its ring of functions, $QCoh(Bun_{Sl_2}(\Sigma)) = K_{\hbar=0}(\Sigma) - Mod$. The deformation $K_\hbar(\Sigma)$ can thus be realized as a deformation of the category $QCoh(Bun_{Sl_2}(\Sigma))$, simply considered as a linear category. Finally, the perverse sheaf $\mathcal{E}$ on $\mathcal{M}_X$ can itself be considered as a deformation of a natural object $\omega_X^{1/2,virt}$, of *virtual half top forms on* X, which is almost a quasi-coherent sheaf on X (it is a complex of such). The quantized object $\mathcal{E}$ is thus not a deformation of $QCoh(X)$, but is rather a deformation of one of its objects.

To summarize, all of the three examples discussed above have an interpretation in terms of deformations of categories, possibly with monoidal structures, of quasi-coherent sheaves on certain moduli spaces. Monoidal categories can be organized in a hierarchy, corresponding to the degree of symmetry imposed on the monoidal structure. For instance, a monoidal category can come equipped with a braiding, or a symmetry constraint. Monoidal categories will be referred to as 1-fold monoidal categories, braided monoidal categories as 2-fold monoidal categories, and symmetric monoidal categories as ∞-fold monoidal categories. We will moreover see that when categories are replaced by ∞-categories there is a notion of n-fold monoidal ∞-categories for $2 < n < \infty$ (also called E_n-monoidal ∞-categories), interpolating between braided and symmetric monoidal categories. When $n = 0$, a 0-fold category can be defined to simply be a category, a (-1)-monoidal category can be declared to be an object in a category, and a (-2)-monoidal category can be defined as an endomorphism of an object in a category. This hierarchy is rather standard in the setting of higher category theory in which a monoidal category is often considered as a 2-category with a unique object, and a braided monoidal categories as a 3-category with unique object and unique 1-morphism (see for instance [26, §V-25]).

In our examples above, quantum groups are deformations of $QCoh(Bun_G(*))$ as a 2-fold monoidal category. Skein algebras are deformations of $QCoh(Bun_{Sl_2}(\Sigma))$ as 0-fold monoidal categories. Finally, the perverse sheaf $\mathcal{E}$ is a deformation of $QCoh(Bun_{Gl_n}(X))$ considered as (-1)-fold monoidal category. The purpose of the present paper is to explain that this is only a very small part of a bigger coherent picture, which we present here as a key principle.

Principle 2.1. *For any oriented manifold of dimension d (understood either in the topological or in the algebraic sense), and any reductive group G, the moduli space of G-bundles on X, $Bun_G(X)$, possesses a quantization which is a deformation of $QCoh(Bun_G(X))$ considered as an $(2-d)$-fold monoidal ∞-category.*

We will see how this principle can becomes a theorem, after a suitable interpretation of $Bun_G(X)$, $QCoh(Bun_G(X))$, and a suitable understanding of $(2-d)$-fold monoidal structures. We will also see how this principle follows from the general framework of symplectic

and poisson structures in derived algebraic geometry, and a general quantization procedure.

3. Moduli spaces as derived stacks

The concept of an algebraic variety is not quite enough to encompass all the aspects of the moduli problems appearing in algebraic geometry. Starting in the 50' and continuing until this day, several successive generalizations of algebraic varieties were introduced in order to understand more and more refined aspects of moduli spaces. As a first step nilpotent functions have been allowed as it is well known that many interesting moduli spaces are non-reduced and must be considered as schemes instead of algebraic varieties. Secondly algebraic stacks have been introduced in order to take into account the fact that in most examples, moduli spaces classify objects only up to isomorphisms, and in many situations non-trivial automorphisms prevent the existence of any reasonable moduli spaces. Unfortunately algebraic stacks are still not enough to capture all aspects of moduli problems, as even though they see non-trivial automorphisms the so-called *higher structures* remain invisible. We will explain in this section what the higher structures are and how the notion of *derived algebraic stacks* is needed in order to incorporate them as part of the refined moduli space.

3.1. Higher structures I: higher stacks. A first type of higher structure concerns *higher homotopies*, which appear naturally each time objects are classified not only up to isomorphism but up to a weaker notion of *equivalences*. A typical example is the extension of the moduli space of vector bundles on a given smooth and projective algebraic variety X, by also allowing complexes of vector bundles, now considered up to quasi-isomorphism[1]. The moduli space of vector bundles on X can be realized as an algebraic stack, but the moduli of complexes of vector bundles taken up to quasi-isomorphism can not be represented by an algebraic stack in the sense of [1]. The reason for this is the existence of higher homotopies between maps between complexes, which is reflected in the fact that a complex E on X can have non-trivial negative self extension groups $Ext^{-i}(E,E)$. The vector spaces $Ext^{-i}(E,E)$ for $i>0$ are higher analogues of automorphism groups of vector bundles and their existence prevent the representability by an algebraic stack for the exact same reason that the existence of non-trivial automorphisms of vector bundles prevent the representability of the moduli problem of vector bundles on X by a scheme. In his manuscript "Pursuing stacks", A. Grothendieck brought forward the idea of *higher stack*, which is an extension of the notion of stacks of groupoids usually considered in moduli theory to a higher categorical or higher homotopical setting. This idea has been made concrete in [27] by the introduction of a notion of *algebraic n-stacks* (see also [31]). These algebraic n-stacks behave in a very similar way to algebraic stacks, and most of the standard notions and techniques of algebraic geometry remain valid in this new setting (they have derived categories, cohomology, tangent spaces, dimensions ...). Fundamental examples of algebraic n-stacks include the Eilenberg-McLane stacks of the form $K(A,n)$, for A a commutative algebraic group, which are higher analogues of classifying stacks BG. Another important example for us are the so-called linear stacks: for a scheme X and a complex of vector bundles E_* on X concentrated in degrees $[-n,0]$, there is a linear stack $\mathbb{V}(E_*)\longrightarrow X$, which is a generalization of the total space of a vector bundle. Finally, for X a smooth and projective variety, there is an

[1] This appears typically in Donaldson-Thomas theory for which moduli spaces of objects in the bounded coherent derived category $D^b_{coh}(X)$ must be considered.

algebraic n-stack of complexes of vector bundles on X, which also possesses many possible non-commutative generalizations (see [31, 32] for more on the subject).

3.2. Higher structures II: derived algebraic stacks. A second type of higher structure attached to moduli problems is called the *derived structure*. These derived structures are somehow dual to the higher homotopies we have just mentioned and exist even in absence of any stacky phenomenon (i.e. even when there are no non-trivial automorphisms). They have been introduced through the eye of deformation theory and originally were only considered at the formal level of moduli spaces. The *derived deformation theory*, also referred to as DDT, is a collection of ideas going back to the 80's, stipulating that moduli spaces, formally around a given fixed point, can be described in terms of Mauer-Cartan elements in a suitable dg-Lie algebra associated to this point. The most famous example is the deformation theory of a given smooth projective variety X, for which the natural dg-Lie algebra is $C^*(X, \mathbb{T}_X)$, the cochain complex computing the cohomology of the tangent sheaf, endowed with its dg-Lie structure coming from the bracket of vector fields. This example is not special, and in fact all possible moduli problems come with natural dg-Lie algebras describing their formal completions.

A striking consequence of the DDT is the existence of *virtual sheaves* on moduli spaces. Indeed, according to the DDT, for a point $x \in \mathcal{M}$ in some moduli space $\mathcal{M}$, we can find a dg-Lie algebra $\mathfrak{g}_x$ controlling the formal local ring of $\mathcal{M}$ at x. There is moreover an explicit formula reconstructing formal functions at x:

$$\widehat{\mathcal{O}}_{\mathcal{M},x} \simeq H^0(C^*(\mathfrak{g}_x)) \simeq H^0(\mathfrak{g}_x, k),$$

where $C^*(\mathfrak{g}_x) = \widehat{Sym}_k(\mathfrak{g}_x^*[-1])$ is the (completed) Chevalley complex of $\mathfrak{g}_x$, which also computes the cohomology of k considered as a the trivial $\mathfrak{g}_x$-module. An important observation is that the Chevalley complex $C^*(\mathfrak{g}_x)$ is a commutative dg-algebra which can have non trivial cohomology in non-positive degrees. These cohomology groups, $H^i(C^*(\mathfrak{g}_x))$ for $i < 0$, provide non-trivial coherent sheaves over the formal neighborhood of x, which are by definition the *derived structures of $\mathcal{M}$ around x*. These local coherent sheaves are quite important, as they control for instance the smoothness defect of the moduli space $\mathcal{M}$, and lead to the so-called virtual fundamental class (see [16]). Incorporating these higher structures as an intrinsic part of the moduli space itself has lead to the theory of *derived algebraic geometry*, and to introduction of *derived schemes* and *derived algebraic $(n-)$stacks* as the correct geometrico-algebraic notion to fully represent moduli problems in algebraic geometry.

3.3. Derived schemes and derived algebraic stacks. The foundations of the theory of derived algebraic geometry can be found in [36, 37] and [18]. We will not give precise definitions here, as the details easily become technical, and will rather concentrate on some basic definitions and basic facts.

3.3.1. Derived schemes. As objects derived schemes are rather easy to define and understand. We display below one possible definition of derived schemes (specific to the characteristic zero case, recall that everything is over a base field k of zero characteristic).

Definition 3.1. *A* derived scheme *(over the field k) consists of a pair $(X, \mathcal{O}_X)$, where X is a topological space and $\mathcal{O}_X$ is a sheaf of commutative differential graded k-algebras on X, satisfying the following conditions.*

- *The sheaves $H^i(\mathcal{O}_X)$ vanish for $i > 0$.*
- *The ringed space $(X, H^0(\mathcal{O}_X))$ is a k-scheme.*
- *For all i, $H^i(\mathcal{O}_X)$, considered as a sheaf of $H^0(\mathcal{O}_X)$-modules, is a quasi-coherent sheaf.*

The above definition makes derived schemes look as rather simple objects, but things get more sophisticated when morphisms between derived schemes are introduced. The sheaf of dg-algebras $\mathcal{O}_X$ must only be considered up to quasi-isomorphisms, and quasi-isomorphic derived schemes have to be considered as equivalent. Therefore, there is a *derived category of derived schemes*, which is a non-linear analogue of the derived category of a ring, and for which quasi-isomorphic sheaves of dg-algebras define the same derived scheme. There are two possible constructions of the category of derived schemes, a first one relies on model category structures and includes the quasi-isomorphisms as the weak equivalences of a certain model category. A more modern approach, powerful in practice but more demanding in terms of foundations, is to use ∞-categories, and to define the category of derived schemes directly as an ∞-category (see for instance [26]). Concretely this means that morphisms between two given derived schemes do not form a set anymore but are topological spaces, or a simplicial sets. This allows to consider homotopies between morphisms of derived schemes, and thus to define equivalences between derived schemes as morphisms having inverses up to homotopy. The ∞-category $\mathbf{dSch}$ of derived schemes is then defined so that quasi-isomorphisms become homotopy equivalences in $\mathbf{dSch}$. We refer to [32, §2.1] for more details on these two approaches, and we will consider $\mathbf{dSch}$ as an ∞-category in what follows.

There have been a certain number of works on the notion of derived schemes, making many of the basic aspects of scheme theory available in the derived setting. Derived schemes behave in a very similar fashion to schemes, they have a notion of (quasi-coherent) sheaves, cohomology, smooth, flat and étale maps etc .... Special among the derived schemes are the affine derived schemes, which are completely characterized by their functions, which themselves form a non-positively graded cdga. There is a $\mathbf{Spec}$ construction, sending a cdga A to an affine derived scheme $\mathbf{Spec}\, A$, whose underlying space is $\mathbf{Spec}\, H^0(A)$ and whose structure sheaf is given by the various localizations $A[f^{-1}]$ in a very similar manner as for un-derived schemes. The $\mathbf{Spec}$ functor produces a full embedding of the (opposite) ∞-category of cdga into $\mathbf{dSch}$. Here the ∞-category of cdga can be presented concretely as the category whose objects are quasi-free cdga together with the standard simplicial sets of morphisms $Map(A, B)$[2]. A general derived scheme X is locally equivalent to $\mathbf{Spec}\, A$ for some cdga A, and many of the notions defined for cdga can be extended to arbitrary derived scheme by sheafification. This is for instance the case for the notions of smooth, flat and étale maps, as well as for the notion of cotangent complexes of derived schemes, etc.

3.3.2. Derived algebraic stacks. The reader should have already guessed that derived schemes are not quite enough for our purpose and that we will need the notion of derived algebraic stacks (including derived higher stacks in some cases). These are defined in a similar fashion as algebraic stacks and higher algebraic stacks (see [32, 37] for details). In a nutshell a derived algebraic stack is given by a quotient of a derived scheme X by an action of a smooth groupoid. Concretely a derived algebraic stack is associated to a simplicial object

[2]Whose set of n-simplicies is $Hom(A, B \otimes_k DR(\Delta^n))$, where $DR(\Delta^n)$ is the algebraic de Rham complex of the algebraic n-dimensional simplex $\{\sum x_i = 1\} \subset \mathbb{A}^{n+1}$.

X_* made of derived schemes satisfying some smooth Kan lifting conditions (see [24]). A typical example is the action of an algebraic group G on a derived scheme Y, for which the simplicial objects is the standard nerve of the action $([n] \in \Delta) \mapsto Y \times G^n$, where the face maps are defined by the action of G on Y and the multiplication on G, and the degeneracies are induced by the unit in G. The derived algebraic stack obtained this way will be denoted by $[Y/G]$, and it should be noted that already some interesting derived algebraic stacks are of this form (but these are not enough to represent all moduli problems).

Derived algebraic stacks are good objects to do algebraic geometry with, and many of the standard notions and results known for underived algebraic stacks can be extended to the derived setting. One construction of fundamental importance for us is the (dg-)category (see [15]) of quasi-coherent complexes on a given derived algebraic stack. For an affine derived scheme $X = \mathbf{Spec}\, A$, the quasi-coherent complexes over X are declared to be the A-dg-modules. The A-dg-modules form a nice k-linear dg-category $L(A)$, for which one explicit model consists of the dg-category of quasi-free A-dg-modules. For a general derived algebraic stack X the dg-category of quasi-coherent complexes is defined by approximating X by affine derived schemes

$$L(X) = L_{qcoh}(X) := \lim_{\mathbf{Spec}\, A \to X} L(A),$$

where the limit is taken inside a suitable ∞-category of dg-categories (see [34]), or equivalently is understood as a homotopy limit inside the homotopy theory of dg-categories of [29].

Another important notion we will use is the cotangent complex. Any derived algebraic stack possesses a canonically defined object $\mathbb{L}_X \in L(X)$, which is the derived version of the sheaf of Kalher 1-forms. When X is a smooth scheme then $\mathbb{L}_X$ is the vector bundle Ω^1_X considered as an object in the quasi-coherent derived category of X. When $X = \mathbf{Spec}\, A$ is an affine derived scheme, $\mathbb{L}_A$ is the A-dg-module representing the so-called André-Quillen homology, and can be defined as the left derived functor of $A \mapsto \Omega^1_A$. For a scheme X, $\mathbb{L}_X$ coincides with Illusie's cotangent complex. For a general derived algebraic stack X the cotangent complex $\mathbb{L}_X \in L_{qcoh}(X)$ is obtained by gluing the cotangent complexes of each stage in a simplicial presentation, but can also be characterized by a universal property involving square zero extensions (see [32, §3.1]). The dual object $\mathbb{T}_X := \underline{Hom}_{\mathcal{O}_X}(\mathbb{L}_X, \mathcal{O}_X) \in L_{qcoh}(X)$ is called the tangent complex of X and is a derived version of the sheaf of derivations.

To finish with general facts about derived algebraic stacks, we would like to mention a specific class of objects which are particularly simple to describe in algebraic terms, and which already contains several non-trivial examples. This class consists of derived algebraic stacks of the form $X = [Y/G]$, where Y is an affine derived scheme and G a linear algebraic group acting on Y. The derived affine scheme Y is the spectrum of a commutative dg-algebra A, which up to a quasi-isomorphism can be chosen to be a cdga inside the category $Rep(G)$ of linear representations of G (and A can even be assumed to be free as a commutative graded algebra). The cdga A, together with its strict G-action, can be used in order to describe $L_{qcoh}(X)$ as well as $\mathbb{L}_X \in L_{qcoh}(X)$. A model for the dg-category $L_{qcoh}(X)$ is the dg-category of cofibrant and fibrant A-dg-modules inside $Rep(G)$, where here fibrant refers to a model category structure on the category of complexes of representations of G (and fibrant means $\mathcal{K}$-injective complex of representations). In particular the homs in the dg-category $L_{qcoh}(X)$ compute G-equivariant ext-groups of A-dg-modules. The object $\mathbb{L}_X \in L_{qcoh}(X)$ can be described as follows. The G-action on A induces a morphism of A-dg-modules

$\mathbb{L}_A \longrightarrow \mathfrak{g}^\vee \otimes_k A$, where $\mathfrak{g}$ is the Lie algebra of G, which is a morphism of A-dg-modules inside $Rep(G)$. The cone of this morphism, or more precisely a cofibrant and fibrant model for this cone, is a model for $\mathbb{L}_X$ as an object in $L_{qcoh}(X)$.

3.4. Representability of derived mapping stacks. As for the case of derived schemes, derived algebraic stacks form an ∞-category denoted by $\mathbf{dArSt}$ (where "Ar" stands for "Artin"). This ∞-category is itself a full sub-∞-category of $\mathbf{dSt}$, the ∞-category of (possibly non-algebraic) derived stacks. The objects of $\mathbf{dSt}$ are ∞-functors $F : \mathbf{cdga} \longrightarrow \mathbb{S}$ satisfying étale descent conditions, and are also called *derived moduli problems*. The derived moduli problems can sometimes be represented by schemes, by derived schemes, or by derived algebraic stacks, in the sense that there exists a derived algebraic stack X together with functorial equivalences $F(A) \simeq Map_{\mathbf{dArSt}}(\mathbf{Spec}\, A, X)$. Proving that a given derived moduli problem is representable is in general not a trivial task, and the following theorem provides a way to construct new derived algebraic stacks.

Theorem 3.2 ([37, Thm. 2.2.6.11]). *Let X be a smooth and proper scheme and Y a derived algebraic stack which is locally of finite presentation over the base field k. Then, the derived moduli problem $A \mapsto Map_{\mathbf{dSt}}(X \times \mathbf{Spec}\, A, Y)$ is representable by a derived algebraic stack denoted by $\mathbf{Map}(X, Y)$.*

One important aspect of the theorem above lies in the fact that the (co-)tangent complexes of the derived mapping stacks $\mathbf{Map}(X, Y)$ are easy to compute: there is a diagram of derived algebraic stacks $Y \xleftarrow{ev} X \times \mathbf{Map}(X, Y) \xrightarrow{p} \mathbf{Map}(X, Y)$, where ev is the evaluation map and p is the natural projection, and we have

$$\mathbb{T}_{\mathbf{Map}(X,Y)} \simeq p_* ev^*(\mathbb{T}_Y) \in L_{qcoh}(\mathbf{Map}(X, Y)). \tag{3.1}$$

At a given point $f \in \mathbf{Map}(X, Y)$, corresponding to a morphism $f : X \longrightarrow Y$, the formula states that the tangent complex at f is $C^*(X, f^*(\mathbb{T}_Y))$, the cochain complex of cohomology of X with coefficients in the pull-back of $\mathbb{T}_Y$ by f. This last formula is moreover compatible with the dg-Lie structures: the formal completion of $\mathbf{Map}(X, Y)$ corresponds, via the DDT correspondence, to the dg-Lie algebra $C^*(X, f^*(\mathbb{T}_Y))[-1]$ (here $\mathbb{T}_Y[-1]$ is equipped with its natural dg-Lie structure, see [12]). This provides a nice and efficient way to understand the derived moduli space $\mathbf{Map}(X, Y)$ at the formal level.

The above theorem also possesses several possible variations, which often can be reduced to the statement 3.2 itself. For instance X can be replaced by a finite homotopy type (e.g. a compact smooth manifold), or by formal groupoids such as X_{DR} or X_{Dol} (see [28, §9]). This provides existence of derived moduli stacks for maps $X \longrightarrow Y$ understood within different settings (e.g. locally constant maps, maps endowed with flat connections or with a Higgs field etc ...). The formula for the tangent complex remains correct for these variants as well, with a suitable definition of the functors p_* and ev^* involved. Pointwise this is reflected in the fact that $C^*(X, f^*(\mathbb{T}_Y))$ is replaced with the appropriate cohomology theory (cohomology with local coefficients when X is a finite homotopy type, algebraic de Rham cohomology for X_{DR} etc ...).

The main examples of applications of theorem 3.2 will be for us when $Y = BG$, the classifying stack of G-bundles, in which case $\mathbf{Map}(X, BG)$ is by definition the derived moduli stack of G-bundles on X. Other interesting instances of applications arise when $Y = \mathbf{Perf}$ is the derived stack of perfect complexes (in which case $\mathbf{Map}(X, Y)$ is the

derived moduli stack of perfect complexes on X), or when Y is the total space of a shifted total cotangent bundle (see below).

4. Symplectic and Poisson structures in the derived setting

In the previous section we saw why and how moduli problems can be represented by derived schemes and derived algebraic stacks. In the sequel we will be further interested in the representability of derived mapping stacks provided by our theorem 3.2, as well as the formula for their tangent complexes. This formula is the key for the construction of symplectic structures on derived mapping stacks, by using cup products in cohomology in order to define pairing on tangent complexes. This brings us to the notions of *shifted symplectic structures*, and of *shifted poisson structures*, of major importance in order to achieve the goal proposed by our principle 2.1.

4.1. Algebraic de Rham theory of derived algebraic stacks. Let X be a derived algebraic stack locally of finite presentation over our ground field k. We have seen that X possesses a cotangent complex $\mathbb{L}_X$, which is a quasi-coherent complex on X. In our situation $\mathbb{L}_X$ is moreover a perfect $\mathcal{O}_X$-module (see [37, Def. 1.2.4.6]) because of the locally finite presentation condition, and is thus a dualizable object in $L_{qcoh}(X)$. Its dual (dual here as an $\mathcal{O}_X$-module), is the tangent complex $\mathbb{T}_X := \mathbb{L}_X^\vee$. A *p-form on* X is simply defined as an element in $H^0(X, \wedge^p_{\mathcal{O}_X}\mathbb{L}_X)$, or equivalently as a homotopy class of morphisms $w : \wedge^p_{\mathcal{O}_X}\mathbb{T}_X \longrightarrow \mathcal{O}_X$ in $L_{qcoh}(X)$. More generally, if $n \in \mathbb{Z}$, a *p-form of degree* n *on* X is an element in $H^n(X, \wedge^p_{\mathcal{O}_X}\mathbb{L}_X)$, or equivalently a homotopy class of morphisms $w : \wedge^p_{\mathcal{O}_X}\mathbb{T}_X \longrightarrow \mathcal{O}_X[n]$ in $L_{qcoh}(X)$. For p fixed, p-forms of various degrees form a complex of k-vector spaces $\mathcal{A}^p(X) = \Gamma(X, \wedge^p_{\mathcal{O}_X}\mathbb{L}_X)$[3], whose n-th cohomology space is the space of p-forms of degree n.

The total complex of differential forms on X is defined as an infinite product

$$\mathcal{A}(X) := \prod_{i\geq 0} \mathcal{A}^i(X)[-i],$$

and except in the very special case where X is a smooth scheme, this infinite product does not restrict to a finite product in general (i.e. $\mathcal{A}^p(X) \neq 0$ for arbitrary large p's in general). The complex $\mathcal{A}(X)$ can be shown to carry an extra differential called the *de Rham differential* and denoted by dR[4]. The differential dR commutes with the cohomological differential, and the complex $\mathcal{A}(X)$ will be always considered endowed with the corresponding total differential. When X is a smooth scheme $\mathcal{A}(X)$ is simply the algebraic de Rham complex of X. When X is a singular scheme $\mathcal{A}(X)$ is the derived de Rham complex of X, which is known to compute the algebraic de Rham cohomology of X (see [3]). When X is a derived algebraic stack the complex $\mathcal{A}(X)$ is by definition the (algebraic and derived) de Rham complex of X, and it can be shown to compute the algebraic de Rham cohomology of the underlying algebraic stack, and thus the Betti cohomology of its geometric realization when $k = \mathbb{C}$.

The de Rham complex comes equipped with a standard Hodge filtration, which is a decreasing sequence of sub-complexes $F^p\mathcal{A}(X) \subset F^{p-1}\mathcal{A}(X) \subset \mathcal{A}(X)$, where $F^p\mathcal{A}(X)$

[3]Here and in the sequel $\Gamma(X, -)$ stands for the ∞-functor of global sections, and thus computes hypercohomology on X.

[4]The existence of this differential is not a trivial fact because of the stackyness of X, see [22].

consists of the sub-complex $\prod_{i\geq p}\mathcal{A}^i(X)[-i]\subset\prod_{i\geq 0}\mathcal{A}^i(X)$. The complex $F^p\mathcal{A}(X)[p]$ is also denoted by $\mathcal{A}^{p,cl}(X)$ and is by definition the complex of closed p-forms on X. We note here that an n-cocycle in $\mathcal{A}^{p,cl}(X)$ consists of a formal series $\sum_{i\geq 0}\omega_i\cdot t^i$, where ω_i an element of degree $n-i$ in $\mathcal{A}^{p+i}(X)$, and satisfies the infinite number of equations

$$dR(\omega_{i-1})+d(\omega_i)=0 \qquad \forall\, i\geq 0,$$

where dR is the de Rham differential, d is the cohomological differential and ω_i is declared to be 0 when $i<0$. With this notation, ω_0 is the underlying p-form and the higher forms ω_i are the *closeness structures*, reflecting that ω_0 *is closed up to homotopy*.

By definition a closed p-form of degree n on X is an element in $H^n(\mathcal{A}^{p,cl}(X))$. Any closed p-form $\sum_{i\geq 0}\omega_i\cdot t^i$ of degree n has an underlying p-form ω_0 of degree n, and thus defines a morphism $\wedge^p_{\mathcal{O}_X}\mathbb{T}_X\longrightarrow\mathcal{O}_X[n]$ in $L_{qcoh}(X)$. We note here that a given p-form of degree n can come from many different closed p-forms of degree n, or in other words that the projection map $\mathcal{A}^{p,cl}(X)\longrightarrow\mathcal{A}^p(X)$, sending $\sum_{i\geq 0}\omega_i\cdot t^i$ to ω_0, needs not be injective in cohomology. This aspect presents a major difference with the setting of differential forms on smooth schemes, for which a given p-forms is either closed or not closed. This aspect can also be understood in the setting of cyclic homology, as differential forms on X can be interpreted as elements in Hochschild homology of X (suitably defined to encode the eventual stackyness of X), and closed forms as elements in negative cyclic homology.

4.2. Shifted symplectic structures.

Definition 4.1 ([22, Def. 1.18]). *An* n-shifted symplectic structure on X *consists of a closed* 2*-form of degree* n *whose underlying morphism*

$$\wedge^2_{\mathcal{O}_X}\mathbb{T}_X\longrightarrow\mathcal{O}_X[n]$$

is non-degenerate: the adjoint map $\mathbb{T}_X\longrightarrow\mathbb{L}_X[n]$ *is an equivalence of quasi-coherent complexes on* X.

There are some basic examples of n-shifted symplectic structures which are the building blocks of more evolved examples. A 0-shifted symplectic structure on a smooth scheme is simply a symplectic structure understood in the usual sense. For a reductive algebraic group G, the 2-shifted symplectic structures on the stack BG are in one-to-one correspondence with non-degenerate and G-invariant scalar products on the Lie algebra $\mathfrak{g}$ of G. Such a structure always exists and is even unique up to a constant when G is a simple reductive group. When $G=Gl_n$, there is a canonical choice for a 2-shifted symplectic structure on BG by considering the standard invariant scalar product on the space of matrices given by $(A,B)\mapsto Tr(A.B)$.

Another source of examples is provided by shifted cotangent bundles. For X a derived algebraic stack and n an arbitrary integer we define the n-shifted total cotangent derived stack of X by

$$T^*X[n]:=\mathbb{V}(\mathbb{L}_X[n])=\mathbf{Spec}\,(Sym_{\mathcal{O}_X}(\mathbb{T}_X[-n])),$$

as the linear derived algebraic stack over X determined by the perfect complex $\mathbb{L}_X[n]$. The derived algebraic stack $T^*X[n]$ comes equipped with a standard Liouville 1-form of degree n[5], whose de Rham differential provides an n-shifted symplectic structure on X. This is

[5]This form represents the universal 1-form of degree n on X.

already interesting for X a smooth scheme as it provides instances of n-shifted symplectic structures for arbitrary values of n. Note here that when X is a smooth scheme, then $T^*X[n]$ is either a smooth (and thus non-derived) algebraic n-stack if $n \geq 0$, or a derived scheme when $n < 0$. Another interesting and useful example is when $X = BG$ and $n = 1$, as $T^*X[1]$ is then identified with the quotient stack $[\mathfrak{g}^*/G]$, for the co-adjoint action of G. The quotient stack $[\mathfrak{g}^*/G]$ is thus equipped with a canonical 1-shifted symplectic structure, which sheds new light on symplectic reduction (we refer to [9, §2.2] for more on the subject). A third important example is the derived algebraic stack of perfect complexes $\mathbf{Perf}$ (see [22, §2.3]), which is a generalization of the stack BGl_n (BGl_n sits as an open in $\mathbf{Perf}$).

More evolved examples of shifted symplectic structures can be constructed by means of the following existence theorem. This result can be seen as a geometrico-algebraic counter part of the so-called AKSZ formalism.

Theorem 4.2 ([22, Thm. 2.5]). *Let X be either a connected compact oriented topological manifold of dimension d, or a connected smooth and proper scheme of dimension d equipped with a nowhere vanishing top form $s \in \Omega^d_X$. Let Y be a derived algebraic stack endowed with an n-shifted symplectic structure. Then the derived algebraic stack $\mathbf{Map}(X, Y)$ is equipped with a canonical $(n-d)$-shifted symplectic structure.*

An important special case is when $Y = BG$ for G a reductive algebraic group, equipped with the 2-shifted symplectic structure corresponding to a non-degenerate element in $Sym^2(\mathfrak{g}^*)^G$. We find this way that the derived moduli stack of G-bundles on X, $Bun_G(X) := \mathbf{Map}(X, BG)$ carries a canonical $(2-d)$-shifted symplectic structure, which is a first step towards a mathematical formulation of our principle 2.1.

Corollary 4.3. *With the above notations, $Bun_G(X)$ carries a canonical $(2-d)$-shifted symplectic structure.*

When $d = 2$ the above corollary recovers the well known symplectic structures on moduli spaces of G-local systems on a compact Riemann surface and of G-bundles on K3 and abelian surface. However, even in this case, the corollary is new and contains more as the 0-shifted symplectic structure exists on the whole derived moduli stack, not only on the nice part of this moduli stack which is a smooth scheme (see for instance our comments in §6.1).

In dimension 3 the corollary states that $Bun_G(X)$ is equipped with a natural (-1)-shifted symplectic structure. The underlying 2-form of degree -1 is an equivalence of perfect complexes $\mathbb{T}_{Bun_G(X)} \simeq \mathbb{L}_{Bun_G(X)}[-1]$. When restricted to the underived part of the moduli stack this equivalence recovers the symmetric obstruction theory used in Donaldson-Thomas theory (see [8, Def. 1.1]). However, here again the full data of the (-1)-shifted symplectic structure contains strictly more than the underlying symmetric obstruction theory, essentially because of the fact that a shifted symplectic structure is not uniquely determined by its underlying 2-form (see [21]).

Finally, when the dimension d is different from 2 and 3 the content of the corollary seems completely new, thought in dimension 1 it essentially states that $[G/G]$ is 1-shifted symplectic, which can be used in order to provide a new understanding of quasi-hamiltonian actions (see [9, §2.2]).

The idea of the proof of theorem 4.2 is rather simple, and at least the underlying 2-form can be described explicitly in terms of the formula for the tangent complexes (formula (3.1) of §3.4). We define a pairing of degree $(n-d)$ on this complex by the composition of the

natural maps and the pairing of degree n on $\mathbb{T}_Y$

$$\wedge^2 p_*(ev^*(\mathbb{T}_Y)) \longrightarrow p_*(ev^*(\wedge^2\mathbb{T}_Y)) \longrightarrow p_*(\mathcal{O})[n] \longrightarrow \mathcal{O}[n-d],$$

where the last map comes from the fundamental class in $H^d(X,\mathcal{O}_X) \simeq k$. This defines a non-degenerate 2-form on $\mathbf{Map}(X,Y)$, and the main content of the theorem 4.2 is that this 2-form comes from a canonically defined closed 2-form of degree $(n-d)$.

For variants and generalizations of theorem 4.2 we refer to [9, 22, 32] in which the reader will find non-commutative generalizations as well as versions with boundary conditions, but also several other possible admissible sources.

4.3. Derived critical loci. To finish the part on n-shifted symplectic structures let us mention critical loci and their possible generalizations. We have already seen that for a given derived algebraic stack X the shifted cotangent $T^*X[n]$ carries a canonical n-shifted symplectic structure. Moreover, the zero section $X \longrightarrow T^*X[n]$ has a natural Lagrangian structure (see [22, Def. 2.8]). More generally, if $f \in H^n(X,\mathcal{O}_X)$ is a function of degree n on X, its de Rham differential $dR(f)$ defines a morphism of derived algebraic stacks $dR(f) : X \longrightarrow T^*X[n]$ which is also equipped with a natural Lagrangian structure. Therefore, the intersection of the zero section with the section $dR(f)$ defines a natural $(n-1)$-shifted derived algebraic stack (see [22, Thm. 2.9]) denoted by $\mathbb{R}Crit(f)$ and called the derived critical locus of f. When $f = 0$ the derived critical locus $\mathbb{R}Crit(f)$ is simply $T^*X[n-1]$ together with its natural $(n-1)$-shifted symplectic structure. When X is a smooth scheme and f is a function of degree 0 (i.e. simply a function $X \longrightarrow \mathbb{A}^1$), then the symplectic geometry of $\mathbb{R}Crit(f)$ is closely related to the singularity theory of the function f. From a general point of view derived critical loci provide a nice source of examples of n-shifted symplectic derived algebraic stacks, which contain already examples of geometric interests. It is shown in [5, 6] that every (-1)-shifted symplectic derived scheme is locally the derived critical locus of a function defined on a smooth scheme.

Derived critical loci are important because they are easy to describe and their quantizations can be understood explicitly. Moreover, derived critical loci and their generalizations can be used to provide local models for n-shifted symplectic structures by means of a formal Darboux lemma we will not reproduce here (see for instance [2, 5, 6]).

4.4. Shifted polyvector fields and poisson structures. The notion of shifted Poisson structure is the dual notion of that of shifted symplectic structure we have discussed so far. The general theory of shifted Poisson structures has not been fully settled down yet and we will here present the basic definitions as well as its, still hypothetical, relations with shifted symplectic structures. They are however a key notion in the existence of quantization that will be presented in the next section.

4.4.1. Shifted polyvectors on derived algebraic stacks. A derived algebraic stack X (as usual assumed locally of finite presentation over the ground field k) has a tangent complex $\mathbb{T}_X$, which is the $\mathcal{O}_X$-linear dual to the cotangent complex. The complex of n-shifted polyvector fields on X is defined by

$$\mathcal{P}ol(X,n) := \bigoplus_i \Gamma(X, Sym^i_{\mathcal{O}_X}(\mathbb{T}_X[-1-n])).$$

The complex $\mathcal{P}ol(X,n)$ has a natural structure of a graded commutative dg-algebra, for which the piece of weight i is $\Gamma(X, Sym^i_{\mathcal{O}_X}(\mathbb{T}_X[-1-n]))$ and the multiplication is induced by the canonical multiplication on the symmetric algebra. We note here that depending of the parity of n we either have $Sym^i_{\mathcal{O}_X}(\mathbb{T}_X[-1-n]) \simeq (\wedge^i \mathbb{T}_X)[-i-ni]$ (if n is even), or $Sym^i_{\mathcal{O}_X}(\mathbb{T}_X[-1-n]) \simeq (Sym^i \mathbb{T}_X)[-i-ni]$ (if n is odd). When X is a smooth scheme and $n=0$, $\mathcal{P}ol(X,0) = \oplus_i \Gamma(X, \wedge^i \mathbb{T}_X)[-i]$ is the standard complex of polyvector fields of X. When $n=1$, and still X a smooth scheme, $\mathcal{P}ol(X,1)$ coincides with $\Gamma(T^*X, \mathcal{O}_{T^*X})$, the cohomology of the total cotangent space of X with coefficients in $\mathcal{O}$. In general, $\mathcal{P}ol(X,n)$ can be interpreted as the graded cdga of cohomology of the shifted cotangent derived stack $T^*X[n+1]$ with coefficients in $\mathcal{O}$ (that is "functions" on $T^*X[n+1]$).

When X is a smooth scheme, $\mathbb{T}_X$ is a sheaf (say on the small étale site of X) of k-linear Lie algebras with the bracket of vector fields. This extends easily to the case where X is a derived Deligne-Mumford stack, $\mathbb{T}_X$ can be made into a sheaf of k-linear dg-Lie algebras for the bracket of dg-derivations. Therefore, polyvector fields $\mathcal{P}ol(X,n)$ can also be endowed with a k-linear dg-Lie bracket of cohomological degree $-1-n$, making it into a graded Poisson dg-algebra where the bracket has cohomological degree $(-1-n)$ and weight (-1). In particular $\mathcal{P}ol(X,n)[n+1]$ always comes equipped with a structure of a graded dg-Lie algebra over k. It is expected that this fact remains valid for a general derived algebraic stack X, but there is no precise construction at the moment. One complication when considering general algebraic stacks comes from the fact that vector fields can not be pulled-back along smooth morphisms (as opposed to étale maps), making the construction of the Lie bracket on $\mathcal{P}ol(X,n)$ much more complicated than for the case of a scheme. For a derived algebraic stack of the form $[\mathbf{Spec}\, A/G]$, for G linear, there are however two possible constructions. A first very indirect construction uses natural operations on the derived moduli stacks of branes (see [33]). A more direct construction can be done as follows. We can take A to be a cofibrant and fibrant cgda inside the category of representations $Rep(G)$. We let $\mathbb{T}_A$ be the A-dg-module of dg-derivations from A to itself. The action of G on A induces a morphism of dg-Lie algebras $\mathfrak{g} \otimes_k A \longrightarrow \mathbb{T}_A$ representing the infinitesimal action of G on A. We consider the co-cône $\mathbb{T}$ of the morphism $\mathfrak{g} \otimes_k A \longrightarrow \mathbb{T}_A$. The complex $\mathbb{T}$ is obviously a k-linear Lie algebra for the bracket induced from the brackets on $\mathbb{T}_A$ and on $\mathfrak{g}$, but this lie structure is *not* compatible with the cohomological differential and thus is not a dg-Lie algebra. However, its fixed points by G (assume G reductive for simplicity) is a dg-Lie algebra over k, which is a model for $\Gamma(X, \mathbb{T}_X)$ where $X = [\mathbf{Spec}\, A/G]$. This construction can be also applied to the G-invariant of the various symmetric powers of shifts of $\mathbb{T}$ in order to get the desired dg-Lie structure on $\mathcal{P}ol(X,n)[n+1]$ in this special case.

4.4.2. Shifted Poisson structures. Let X be a derived algebraic stack and fix an integer $n \in \mathbb{Z}$. We can define n-shifted Poisson structures as follows. We let $\mathcal{P}ol(X,n)[n+1]$ be the shifted polyvector fields on X, endowed with the structure of a graded dg-Lie algebra just mentioned. We let $k(2)[-1]$ be the graded dg-Lie algebra which is k in cohomological degree 1, with zero bracket and k is pure of weight 2. An n-shifted Poisson structure on X is then defined to be a morphism of graded dg-Lie algebras

$$p : k(2)[-1] \longrightarrow \mathcal{P}ol(X,n)[n+1].$$

Here, a morphism of graded dg-Lie algebras truly means a morphism inside the ∞-category of graded dg-Lie algebras, or a morphism in an appropriate homotopy category. Using the dictionary between dg-Lie algebras and formal moduli problems (see [19]), such a morphism

p is determined by a Mauer-Cartan element in $\mathcal{P}ol(X,n)[n+1]\otimes tk[[t]]$, which is of weight 2 with respect to the grading on $\mathcal{P}ol(X,n)$. Such an element can be described explicitly as a formal power series $\sum_{i\geq 1} p_i\cdot t^i$, where p_i is an element of cohomological degree $n+2$ in $\Gamma(X, Sym^{i+1}(\mathbb{T}_X[-1-n]))$, and satisfies the equations

$$d(p_i)+\frac{1}{2}\cdot\sum_{a+b=i}[p_a,p_b]=0 \qquad \forall i\geq 1.$$

As we already mentioned, shifted Poisson structures can be developed along the same lines as shifted symplectic structures (e.g. there is a notion of co-isotropic structures on a map with an n-shifted Poisson target, and a Poisson version of the existence theorem 4.2), but at the moment this work has not been carried out in full details. It is believed that for a given X and $n\in\mathbb{Z}$, there is a one-to-one correspondence between n-shifted symplectic structures on X and n-shifted Poisson structures on X which are non-degenerate in an obvious sense. However, this correspondence has not been established yet, except in some special cases, and remains at the moment an open question for further research (see §6.2).

5. Deformation quantization of n-shifted Poisson structures

In this section we finally discuss the existence of quantization of n-shifted Poisson structures, a far reaching generalization of the existence of deformation quantization of Poisson manifolds due to Kontsevich. For this we first briefly discuss the output of the quantization, namely the notion of deformation of categories and iterated monoidal categories, which already contains some non-trivial aspects. We then present the formality conjecture, which is now a theorem except in some very particular cases, and whose main corollary is the fact that every n-shifted Poisson structure defines a canonical formal deformation of the E_n-monoidal category of quasi-coherent complexes. We also discuss the case $n<0$ by presenting the *red shift trick* consisting of working with a formal parameter $\hbar$ living in some non-trivial cohomological degree.

5.1. The deformation theory of monoidal dg-categories. As we have seen in §2, a derived algebraic stack X has a dg-category of quasi-coherent complexes $L(X)$. It is a k-linear dg-category which admits arbitrary colimits. We will assume in this section that $L(X)$ is a compactly generated dg-category, or equivalently that it can be realized as the category of dg-modules over a small dg-category. More generally we will assume that X is a *perfect* derived algebraic stack, in the sense that perfect complexes on X are compact generators of $L(X)$. This is known to be the case under the assumption that X can be written as a quotient $[\mathbf{Spec}\, A/G]$ for a linear algebraic G acting on a cdga A.

5.1.1. Deformations of dg-categories. We let $T_0:=L(X)$ and we would like to study the deformation theory of T_0. For this, we define a first naive deformation functor

$$Def^{naive}(T_0):\mathbf{dg}-\mathbf{art}^*\longrightarrow\mathbb{S},$$

from the ∞-category of augmented local artinian cdga to the ∞-category of spaces as follows. To $A\in\mathbf{dg}-\mathbf{art}^*$ we assign the ∞-category $\mathbb{D}g^c(A)$, of cocomplete and compactly generated A-linear dg-categories and A-linear colimit preserving dg-functors (see

[34, §3.1]). For a morphism of dg-artinian rings $A \to B$, we have a base change ∞-functor $-\widehat{\otimes}_A B : \mathbb{D}g^c(A) \longrightarrow \mathbb{D}g^c(B)$. We then set

$$Def^{naive}(T_0)(A) := \mathbb{D}g^c(A) \times_{\mathbb{D}g^c(k)} \{T_0\}.$$

Here $\mathbb{D}g^c(A) \times_{\mathbb{D}g^c(k)} \{T_0\}$ is the fiber taken at the point T_0 of the ∞-functor $-\widehat{\otimes}_A k$ induced by the augmentation $A \to k$. As is, $Def^{naive}(T_0)(A)$ is an ∞-category, from which we extract a space by taking the geometric realization of its sub-∞-category of equivalences (i.e. taking the nerve of the maximal sub-∞-groupoid). Intuitively, $Def^{naive}(T_0)(A)$ is the classifying space of pairs (T, u), with T a compactly generated A-linear dg-category and u a k-linear equivalence $u : T\widehat{\otimes}_A k \simeq T_0$.

As already observed in [14] the ∞-functor $Def^{naive}(T_0)$ is not a formal moduli problem, it does not satisfies the Schlessinger conditions of [19], and thus can not be equivalent to the functor of Mauer-Cartan elements in a dg-Lie algebra. This bad behavior of the ∞-functor $Def^{naive}(T_0)$ has been a longstanding major obstacle preventing the understanding of the deformation theory of dg-categories. There have been several tentative modifications of $Def^{naive}(T_0)$ attempting to overcome this problem, for instance by allowing curved dg-categories as possible deformations, however none of these were successful. We propose here a new solution to this problem which provides the only complete understanding of deformations of dg-categories that we are aware of. For this, we introduce $Def(T_0) : \mathbf{dg} - \mathbf{art}^* \longrightarrow \mathbb{S}$, which is the universal ∞-functor constructed out of $Def^{naive}(T_0)$ and satisfying the Schlessinger conditions of [19] (in other words it is the best possible approximation of $Def^{naive}(T_0)$ by an ∞-functor associated to a dg-Lie algebra). By construction there is a natural transformation $l : Def^{naive}(T_0) \longrightarrow Def(T_0)$, as well as a dg-Lie algebra L such that $Def(T_0)$ is given by $A \mapsto \underline{MC}_*(L \otimes m_A)$ (where as usual $m_A \subset A$ is the augmentation dg-ideal in A, and $\underline{MC}_*$ denotes the space of Mauer-Cartan elements). Moreover the natural transformation l is universal for these properties, and in particular the dg-Lie L is uniquely determined and only depends on $Def^{naive}(T_0)$.

The following theorem is folklore and known to experts. It appears for instance in a disguised form in [23].

Theorem 5.1. *Let T_0 be a compactly generated dg-category.*

(1) *The dg-Lie algebra associated to the formal moduli problem $Def(T_0)$ is $HH(T_0)[1]$, the Hochschild cochains on T_0 endowed with its usual Gerstenhaber bracket (see e.g. [15, §5.4]).*

(2) *The space $Def(T_0)(k[[t]])$ is naturally equivalent to the classifying space of $k[\beta]$-linear structures on T_0, where $k[\beta]$ is the polynomial dg-algebra over k with one generator β in degree 2.*

The above theorem subsumes the two main properties of the formal moduli problem $Def(T_0)$, but much more can be said. The formla for the $k[[t]]$-points of $Def(T_0)$ can be generalized to any (pro-)artinian augmented dg-algebra A, by using B_A-linear structures on T_0, where now B_A is the E_2-Koszul dual of A (see [19], and the E_2-Koszul dual of $k[[t]]$ is of course $k[\beta]$). By construction we have a map of spaces $Def^{naive}(T_0)(A) \longrightarrow Def(T_0)(A)$. This map is not an equivalence but can be shown to have 0-truncated fibers (so it induces isomorphisms on π_i for $i > 1$ and is injective on π_1). It is interesting to note here that not only $Def(T_0)$ contains more objects than $Def^{naive}(T_0)$ but also contains more morphisms. There are natural conditions one can impose on T_0 in order to make $Def^{naive}(T_0)$ closer to

$Def(T_0)$. It is for instance believed that they coincide when T_0 is a smooth and proper dg-category, as well as for dg-categories of complexes in Grothendieck abelian categories. In our situation, $T_0 = L(X)$, with X a derived algebraic stack which is not smooth in general, it is not reasonable to expect any nice assumptions on T_0, and the above theorem is probably the best available result in order to understand formal deformations of $L(X)$.

5.1.2. Deformations of monoidal dg-categories. Theorem 5.1 also possesses monoidal and iterated monoidal versions as follows. First of all the ∞-category $\mathbb{D}g^c(A)$ of compactly generated A-linear dg-categories is equipped with a tensor product $\widehat{\otimes}_A$, making it into a symmetric monoidal ∞-category. It is therefore possible to use the notion of an E_n-monoid in $\mathbb{D}g^c(A)$ of [20], in order to define E_n-monoidal A-linear dg-categories (also called n-fold monoidal A-linear dg-categories). In a nutshell, an E_n-monoidal A-linear dg-category consists of a compactly generated A-linear dg-category T together with morphisms

$$\mu_k : E_n(k) \otimes T^{\widehat{\otimes}_A k} \longrightarrow T,$$

where the tensor by the space $E_n(k)$ and the tensor products are taken in the symmetric monoidal ∞-category $\mathbb{D}g^c(A)$, and together with compatibilty conditions/structures. For our derived algebraic stack X, the dg-category $L(X)$ is equipped with a symmetric monoidal structure and thus is naturally an E_n-monoidal dg-category for all $n \geq 0$, where by convention an E_0-monoidal dg-category simply is a dg-category.

For a cdga A, we set $\mathbb{D}g^c_{E_n}(A)$ for the ∞-category of compactly generated E_n-monoidal A-linear dg-categories. Here compactly generated also means that the compact objects are stable by the monoidal structure, so objects in $\mathbb{D}g^c_{E_n}(A)$ can also be described as dg-categories of dg-modules over small A-linear E_n-monoidal dg-categories. Morphisms in $\mathbb{D}g^c_{E_n}(A)$ must be defined with some care as they involve higher dimensional versions of Morita morphisms between algebras. For two E_n-monoidal dg-categories T and T' in $\mathbb{D}g^c_{E_n}(A)$, the dg-category of A-linear colimit preserving dg-functors can be written as $T^\vee \widehat{\otimes}_A T'$, where $T^\vee$ is the dual of T (i.e. we take the opposite of the sub-dg-category of compact generators). The A-linear dg-category $T^\vee \widehat{\otimes}_A T'$ is a new object in $\mathbb{D}g^c_{E_n}(A)$, and in particular it makes sense to consider E_n-algebras inside the dg-category $T^\vee \widehat{\otimes}_A T'$. From the point of view of dg-functors these correspond to E_n-lax monoidal A-linear colimit preserving dg-functors $T \to T'$. For two E_n-algebras M and N inside $T^\vee \widehat{\otimes}_A T'$, we can form a new E_n-algebra $M^{op} \otimes N$. The A-linear dg-category of $M^{op} \otimes N$-modules inside $T^\vee \widehat{\otimes}_A T'$ is then E_{n-1}-monoidal, so the process can be iterated. We can consider two E_{n-1}-algebras inside $M^{op} \otimes N$-modules, say M' and N', as well as their tensor product $M'^{op} \otimes N'$ and the A-linear dg-category of $M'^{op} \otimes N'$-modules, which is itself E_{n-2}-monoidal ... and so on and so forth. We are describing here $\mathbb{D}g^c_{E_n}(A)$ as an $(\infty, n+1)$-category (see [26]), whose objects are E_n-monoidal compactly generated A-linear dg-categories, whose 1-morphism from T to T' are E_{n-1}-algebras inside $T^\vee \widehat{\otimes}_A T'$, whose 2-morphisms between M' and N' are E_{n-2}-algebras inside $M'^{op} \otimes N'$-modules, etc .... The $(\infty, n+1)$-category $\mathbb{D}g^c_{E_n}(A)$ produces a space by considering the geometric realization of its maximal sub-∞-groupoid (i.e. realizing the sub-∞-category of equivalences).

For $T_0 = L(X)$, assuming that $L(X)$ is compactly generated and that its compact objects are the perfect complexes, we define a naive deformation functor $Def^{naive}_{E_n}(T_0)$, of T_0 considered as an E_n-dg-category, by sending an augmented dg-artinian ring $A \in \mathbf{dg} - \mathbf{art}^*$

to the fiber at T_0 of the restriction map

$$-\widehat{\otimes}_A k : \mathbb{D}g^c_{E_n}(A) \longrightarrow \mathbb{D}g^c_{E_n}(k).$$

The space $Def^{naive}_{E_n}(T_0)$ is the space of pairs (T, u), where T is an E_n-monoidal compactly generated A-linear dg-category, and $u : T\widehat{\otimes}_A k \simeq T_0$ an equivalence in $\mathbb{D}g^c_{E_n}(k)$. Similar to the case $n = 0$ we already discussed, the ∞-functor $Def^{naive}_{E_n}(T_0)$ does not satisfy the Schlessinger's conditions, and the bigger n is, the more this fails. We denote by $Def_{E_n}(T_0)$ the formal moduli problem generated by $Def^{naive}_{E_n}(T_0)$. The following theorem is the generalization of 5.1 to the iterated monoidal setting.

Theorem 5.2. *Let T_0 be a compactly generated E_n-monoidal dg-category.*

(1) *The dg-Lie algebra associated to the formal moduli problem $Def_{E_n}(T_0)$ is $HH^{E_{n+1}}(T_0)\,[n+1]$, the E_{n+1}-Hochschild cochains on T_0 of [11].*

(2) *The space $Def_{E_n}(T_0)(k[[t]])$ is naturally equivalent to the classifying space of $k[\beta_n]$-linear structures on T_0, where $k[\beta_n]$ is the commutative polynomial dg-algebra over k with one generator β_n in degree $2 + n$.*

Theorems 5.1 and 5.2 provides a way to understand the relations between (higher) Hochschild cohomology and formal deformations of dg-categories and iterated monoidal dg-categories. They state in particular that the correct manner to define a formal deformation of a given dg-category T_0, parametrized by $k[[t]]$, is by considering $k[\beta]$-linear structures on T_0, and similarly for the iterated monoidal setting with $k[\beta_n]$-linear structures. In the sequel, we will freely use the expression "formal deformation of the dg-category $L(X)$ considered as an E_n-monoidal dg-category", by which we mean an element in $Def_{E_n}(L(X))(k[[t]])$, and thus a $k[\beta_n]$-linear structure on T_0. We however continue to think of these deformations as actual deformations of $L(X)$ over $k[[\hbar]]$ for a formal parameter $\hbar$, even thought they are not quite as naive objects.

5.2. The higher formality conjecture. We have just seen that formal deformations of a given E_n-monoidal compactly generated dg-category T_0 is controlled by its higher Hochschild cochain complex $HH^{E_{n+1}}(T_0)[n+1]$, endowed with its natural structure of a dg-Lie algebra. We now turn to the specific case where $T_0 = L(X)$, the quasi-coherent dg-category of a derived algebraic stack X. We continue to assume that X is nice enough (e.g. of the form $[\mathbf{Spec}\, A/G]$) so that $L(X)$ is compactly generated by the perfect complexes). The higher Hochschild cohomology of T_0 can then be described in geometric terms as follows. We let $n \geq 0$, and let $S^n = \partial B^{n+1}$ be the topological n-sphere considered as a constant derived stack. We consider the derived mapping stack $\mathcal{L}^{(n)}(X) := \mathbf{Map}(S^n, X)$, also called the *$n$-dimensional derived loop stack of X*. There is a constant map morphism $j : X \longrightarrow \mathcal{L}^{(n)}(X)$, and thus a quasi-coherent complex $j_*(\mathcal{O}_X) \in L(\mathcal{L}^{(n)}(X))$. The E_{n+1}-Hochschild cohomology of the dg-category $L(X)$ can be identified with

$$HH^{E_{n+1}}(X) \simeq End_{L(\mathcal{L}^{(n)}(X))}(j_*(\mathcal{O}_X)).$$

Note that when $n = 0$ and X is a scheme this recovers the description of the Hochschild complex of X as the self extension of the diagonal. Because of the stackyness of X this definition can be modified by replacing $\mathcal{L}^{(n)}(X)$ by its formal completion $\widehat{\mathcal{L}^{(n)}(X)}$ along the map $X \longrightarrow \mathcal{L}^{(n)}(X)$, which is called the formal n-dimensional derived loop space

(when X is a derived scheme the formal and non-formal versions of the derived loop stacks coincide). We then have the formal version of Hochschild complex

$$\widehat{HH}^{E_{n+1}}(X) \simeq End_{L(\widehat{\mathcal{L}^{(n)}(X)})}(j_*(\mathcal{O}_X)).$$

Note that we have a natural morphism $\widehat{HH}^{E_{n+1}}(X) \longrightarrow HH^{E_{n+1}}(X)$.

The complexe $\widehat{HH}^{E_{n+1}}(X)$ has a structure of an E_{n+2}-algebra, as predicted by the so-called Deligne's conjecture which is now a theorem (see [11, 20]). In particular $\widehat{HH}^{E_{n+1}}(X)[n+1]$ is a dg-Lie algebra. The formality conjecture asserts that the dg-Lie algebra $\widehat{HH}^{E_{n+1}}(X)[n+1]$ can be described in simple terms involving shifted polyvector fields.

Conjecture 5.3 (Higher formality). *For a nice enough derived algebraic stack X, and $n \geq 0$, the dg-Lie algebra $\widehat{HH}^{E_{n+1}}(X)[n+1]$ is quasi-isomorphic to $\mathcal{P}ol(X,n)[n+1]$. The quasi-isomorphism is canonical up to a universal choice of a Drinfeld associator.*

Note that when X is a smooth scheme and $n = 0$ the conjecture 5.3 is the so-called *Kontsevich's formality theorem*. The conjecture has been proven in already many cases.

Theorem 5.4 ([33, Cor. 5.4]). *The above higher formality conjecture is true for all $n > 0$ and for all derived algebraic stacks X of the form $[\mathbf{Spec}\, A/G]$ for G a linear algebraic group acting on the cdga A. When X is a derived Deligne-Mumford stack it is also true for $n = 0$.*

The theorem above provides many cases in which conjecture 5.3 is satified. We believe it is also true in the remaining case when $n = 0$ and for non Deligne-Mumford stacks. We also believe that the restriction for X being of the form $[\mathbf{Spec}\, A/G]$ in the theorem 5.4 is not necessary, and that the theorem should be true for a large class of derived higher algebraic stacks as well.

5.3. Existence of deformation quantization. We finally arrive at the existence of quantization of derived algebraic stacks X endowed with n-shifted Poisson structures, and its consequence: the mathematical incarnation of our principle 2.1. Let X be a derived algebraic stack, and $n \geq 0$ to start with (the case of negative values will be treated below). We assume that X is nice enough and that the conjecture 5.4 is satisfied (e.g. under the hypothesis of theorem 5.4). Let p be an n-shifted Poisson structure on X. By definition it provides a morphism of dg-Lie algebras $p : k[-1] \longrightarrow Pol(X,n)[n+1]$. Using the conjecture 5.4 we find a morphism of dg-Lie algebras $p : k[-1] \longrightarrow \widehat{HH}^{E_{n+1}}(X)[n+1]$, which composed with the natural morphism $\widehat{HH}^{E_{n+1}}(X) \longrightarrow HH^{E_{n+1}}(X)$ provides a morphism of dg-Lie algebras

$$p : k[-1] \longrightarrow HH^{E_{n+1}}(X)[n+1].$$

The derived deformation theory (see [19]) and theorem 5.2 tell us that the morphism p provides a formal deformation of $L(X)$ as an E_n-monoidal dg-category, denoted by $L(X,p)$. This is the deformation quantization of the pair (X,p).

Assume now that $n < 0$ and that X is equipped with an n-shifted Poisson structure p such that the conjecture 5.4 is satisfied for X and $-n$. The n-shifted Poisson structure p is a morphism of graded dg-Lie algebras $k(2)[-1] \longrightarrow \mathcal{P}ol(X,n)[n+1]$ where $k(2)[-1]$ is the

abelian dg-Lie algebra which is k in cohomological degree 1 and pure weight 2. The category of $\mathbb{Z}$-graded complexes has a tensor auto-equivalence, sending a complex E pure of weight i to $E[-2i]$ again pure of weight i. This auto-equivalence induces an auto-equivalence of the ∞-category of graded dg-Lie algebras, and sends $\mathcal{P}ol(X,n)[n+1]$ to $\mathcal{P}ol(X,n+2)[n+3]$ and $k(2)[-1]$ to $k(2)[-3]$. Iterated n times, the morphism p goes to a new morphism of dg-Lie algebras $p' : k(2)[-2n-1] \longrightarrow \mathcal{P}ol(X,-n)[-n+1]$, which by conjecture 5.3 induces a morphism of dg-Lie algebras

$$p' : k[-2n-1] \longrightarrow HH^{E_{-n+1}}(X)[-n+1].$$

The abelian dg-Lie algebra $k[-2n-1]$ corresponds to the formal derived scheme $\mathbf{Spec}\, k[[\hbar_{2n}]]$, where now $\hbar_{2n}$ has cohomological degree $2n$. By the general DDT and theorem 5.2 we do find a formal deformation of $L(X)$, considered as an E_{-n}-monoidal dg-category, over $k[[\hbar_{2n}]]$. This deformation will be denoted by $L(X,p)$. This trick to deal with cases where $n<0$ is called the *red shift trick*. It is not new, and already appears in the conjecture [13, Page 14] where $\mathbb{Z}/2$-graded derived categories are considered instead of $\mathbb{Z}$-graded derived categories, and canceling out the red shift.

Definition 5.5. *The formal deformation $L(X,p)$ constructed above is the* deformation quantization of (X,p). *It is a formal deformation of $L(X)$ considered as an E_n-monoidal dg-category if $n \geq 0$, and a formal deformation of $L(X)$ considered as an E_{-n}-dg-category over $k[[\hbar_{2n}]]$ if $n<0$.*

Definition 5.5 applies in particular to the case $X = Bun_G(Y)$, making our principle 2.1 into a mathematical statement.

6. Examples and open questions

We present here some examples as well as some further questions.

6.1. Three examples. We start by coming back to the three situations we mentioned in §2.

Quantum groups. We let $X = BG$, for G reductive. We have seen that X has a 2-shifted symplectic structure given by the choice of non-degenerate G-invariant scalar product on $\mathfrak{g}$. The dg-category $L(X)$ here is the dg-category of complexes of representations of G. Our quantization is then a formal deformation of $L(X)$ as an E_2-monoidal dg-category, and is simply realized by taking the dg-category of complexes of representations of the quantum group.

Skein dg-algebras. We now let $X = Bun_G(\Sigma)$ be the derived moduli stack of G-bundles on a compact oriented surface Σ. We know that X carries a natural 0-shifted symplectic structure (depending on a choice of a non-degenerate G-invariant scalar product on $\mathfrak{g}$), whose quantization $L(X,p)$ in our sense is a deformation of the dg-category $L(X)$[6]. The dg-category $L(X,p)$ is an interesting refinement of the skein algebra of Σ which, as far as the author is aware, has not been considered before. The structure sheaf $\mathcal{O}_X \in L(X)$

[6]Note however that here $n=0$ and the formality conjecture 5.3 is not established yet, so this situation is still conjectural at the moment.

deforms to a uniquely defined object $\widetilde{\mathcal{O}_X} \in L(X,p)$, whose endormophisms form a dg-algebra $B_\hbar = End(\widetilde{\mathcal{O}_X})$ over $k[[\hbar]]$, which is a deformation of $\mathcal{O}_X(X)$ the dg-algebra of functions on X. The skein algebra is recovered as $H^0(B_\hbar)$, but $B_\hbar$ is not cohomologically concentrated in degree 0 in general and contains strictly more than $K_\hbar(\Sigma)$. The higher cohomology groups of $B_\hbar$ are directly related to the non-trivial derived structure of X, which is concentrated around the singular points corresponding to G-bundles with many automorphisms. Outside these bad points the dg-category $L(X,p)$ is essentially given by complexes of $K_\hbar(\Sigma)$-modules. Formally around a given singular point $\rho \in X$, the dg-category $L(X,p)$ has a rather simple description as follows. The formal completion of X at ρ is controlled by the formal dg-Lie algebra $L_\rho := H^*(\Sigma, \mathfrak{g}_\rho)$, where $\mathfrak{g}_\rho$ is the local system of Lie algebras associated to the G-bundle ρ. The dg-Lie algebra L_ρ is endowed with a non-degenerate pairing of degree 2 induced by the choice of a G-invariant scalar product on $\mathfrak{g}$ which defines a non-degenerate pairing $p : L_\rho^\vee[-1] \wedge L_\rho^\vee[-1] \longrightarrow k$. The pairing p defines itself a Poisson structure on the completed Chevalley complex $\widehat{\mathcal{O}_{X,\rho}} \simeq \widehat{Sym}_k(L_\rho^\vee[-1])$, which is the cdga of formal functions on X around ρ. The quantization of this Poisson cdga, which can be described in simple terms as the Weyl dg-algebra associated to L_ρ with the pairing p, is the quantization of X around ρ and can be used to describe the full sub-dg-category of $L(X,p)$ generated by objects supported at ρ.

Donaldson-Thomas theory. We now turn to the case where $X = Bun_G(Y)$ for a Calabi-Yau 3-fold Y, which is endowed with a (-1)-shifted symplectic form. Our quantization $L(X,p)$ here is a formal deformation of $L(X)$ as a monoidal dg-category with a formal parameter $\hbar_{-2}$ of degree -2. To simplify a bit we can consider this as a formal deformation of $L^{\mathbb{Z}/2}(X)$, the 2-periodic dg-category of quasi-coherent complexes on X, considered as a monoidal dg-category and with a formal parameter $\hbar$ sitting now in degree 0. Locally, X is essentially given as the critical locus of a function f, whose category of matrix factorizations $MF(f)$ provides a natural $L(X,p)$-module (i.e. $MF(f)$ is enriched over the monoidal dg-category $L(X,p)$). In a precise sense, $MF(f)$ can be viewed as an object $\mathcal{M}$ in the quantization of X[7]. The object $\mathcal{M}$ only exists locally, but when X is endowed with orientation data we can expect more and maybe an existence globally on X (for instance, the class of $\mathcal{M}$ in a suitable Grothendieck group has been constructed in [17]). This suggests a possible relation with the perverse sheaf $\mathcal{E}$ we mentioned in §2, as $\mathcal{E}$ should be somehow the Betti realization of the sheaf of dg-categories $\mathcal{M}$. Our quantization should thus refine and reinterpret some already known constructions in Donaldson-Thomas theory.

6.2. Further questions. We finish by a sample of further possible research directions.

Symplectic to Poisson and formality for n=0. As already mentioned in the text the precise way to obtain an n-shifted Poisson structure out of an n-shifted symplectic structure is not clear at the moment, except in some special case (e.g. for derived scheme for which a version of the Darboux lemma holds and can be used, see [5, 6]). Also recall that our conjecture 5.3 remains open for non Deligne-Mumford derived algebraic stacks.

Quantization of Lagrangian morphisms. For a morphism between derived algebraic stacks, the correct analog of a shifted symplectic structure is that of a Lagrangian structure

[7]This is so when monoidal dg-categories are considered through their $(\infty,2)$-category of modules.

(see [22]). These are the maps that are candidates to survive after the deformation quantization. For this a version of the formality conjecture 5.3 must be stated and proved (if at all true). The basic idea here is that a Lagrangian morphim $f : X \longrightarrow Y$, with Y n-shifted symplectic ($n > 0$), should deform $L(X)$ as an E_{n-1}-monoidal dg-category enriched over the deformation quantization of $L(Y)$. According to [9], fully extended TQFT should be obtained this way, by quantization of fully extended TQFT with values in a certain category of n-shifted symplectic derived algebraic stacks and Lagrangian correspondences between them.

Quantization for $n = -1, -2$. When $n = -1$, and $n = -2$ the output of our quantization is respectively a monoidal dg-category and braided monoidal dg-category. There are other possible interpretations of the quantization in these two specific cases, as the expression "E_{-1}-monoidal dg-category" can be understood as "an object in a dg-category", and "E_{-2}-monoidal dg-category" as "an endomorphism of an object in a dg-category". In particular, the quantization of a derived algebraic stack X endowed with a (-1)-shifted (resp. (-2)-shifted) Poisson structure could also be interpreted as the construction of a deformation of an object in $L(X)$ (resp. the deformation of an endomorphism in $L(X)$). For $n = -1$ this is the point of view taken by Joyce and his coauthors (see [2, 7, 8]). Note that in this setting the existence of quantization is predicated on the existence of orientation data which may not exist. The precise relations with the quantization of 5.5 remains to be investigated, and at the moment there is no precise explanations of the construction of the constructible sheaf of [2, 7, 8] in term of derived deformation theory.

Motivic aspects. Deformation quantization possesses an interesing interaction with the motivic world. This is particularly clear when $n = -1$ (e.g. in the setting of Donaldson-Thomas theory): DT are made "motivic" in [17], and the constructible sheaf $\mathcal{E}$ we mentioned above is expected to be the Betti realization of a certain "motive" over $Bun_G(X)$. Because of deformation quantization these motives most probably are instances of "non-commutative motives" over non-commutative schemes ("E_2-schemes" in the setting of DT theory). For commutative base schemes non-commutative motives have been studied in [25], for which the constructions of [4] provides a possible Betti realization functor. From a general point of view, the specific example of Donaldson-Thomas theory suggests the notion of *E_n-motives*, related to our deformation quantization for arbitrary values of n, as well as E_n-motives over a base E_{n-1}-scheme (or stack), which is worth studying along the same lines as [25, 30]

Geometric quantization. Only deformation quantization has been considered in this text. However, derived algebraic geometry can also interact nicely with geometric quantization, a direction currently investigated in [39].

Acknowledgements. The content of this manuscript represents works still in progress in collaboration with T. Pantev, M. Vaquié and G. Vezzosi. The author is thankful to the three of them for numerous conversations that have helped a lot in the writting of this manuscript. An important part of the present text has been prepared during a, snowy but enjoyable, visit at Yale Mathematics department. A particular thanks to S. Goncharov and M. Kapranov for many enlightening conversations on related subjects.

References

[1] Artin, M., *Versal deformations and algebraic stacks,* Invent. Math. **27** (1974), 165–189.

[2] O. Ben-Bassat, C. Brav, V. Bussi, and D. Joyce., *A 'Darboux Theorem' for shifted symplectic structures on derived Artin stacks, with applications*. Preprint arXiv:1312.0090.

[3] Bhatt, B., *Completions and derived de Rham cohomology*. Preprint arXiv:1207.6193.

[4] Blanc, A., *Invariants topologiques des Espaces non commutatifs*. Thesis, Montpellier July 2013, available as arXiv:1307.6430.

[5] Bouaziz, E. and Grojnowski I., *A d-shifted Darboux theorem*. Preprint arXiv:1309. 2197.

[6] C. Brav and V. Bussi, D. Joyce, *A 'Darboux theorem' for derived schemes with shifted symplectic structure*. Preprint arXiv:1305.6302.

[7] C. Brav, V. Bussi, D. Dupont, D. Joyce, and B. Szendroi, *Symmetries and stabilization for sheaves of vanishing cycles*. Preprint arXiv:1211.3259.

[8] Bussi, V., *Generalized Donaldson-Thomas theory over fields $K \neq \mathbb{C}$*, Thesis, available as arXiv:1403.2403.

[9] Calaque, D., *Lagrangian structures on mapping stacks and semi-classical TFTs*. Preprint arXiv:1306.3235.

[10] Drinfeld, V. G., *Quantum groups*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986), 798–820, Amer. Math. Soc., Providence, RI, 1987.

[11] Francis, J., *The tangent complex and Hochschild cohomology of E_n-rings*, Compos. Math. **149** (2013), no. 3, 430–480.

[12] Hennion, B., *Tangent Lie algebra of derived Artin stacks*. Preprint arXiv:1312.3167.

[13] Kapustin, A., *Topological field theory, higher categories, and their applications*, Proceedings of the International Congress of Mathematicians. Volume III, 2021–2043, Hindustan Book Agency, New Delhi, 2010.

[14] Keller, B. and Lowen, W., *On Hochschild cohomology and Morita deformations*, Int. Math. Res. Not. IMRN 2009, no. **17**, 3221–3235.

[15] Keller, B., *On differential graded categories*, International Congress of Mathematicians. Vol. II, 151-190, Eur. Math. Soc., Zürich, 2006.

[16] Kontsevich, M., *Enumeration of rational curves via torus actions*, The moduli space of curves (Texel Island, 1994), 335-368, Progr. Math., **129**, Birkhäuser Boston, Boston, MA, 1995.

[17] Kontsevich, M. and Soibelman, Y. *Motivic Donaldson-Thomas invariants: summary of results,* Mirror symmetry and tropical geometry, 55–89, Contemp. Math., 527, Amer. Math. Soc., Providence, RI, 2010.

[18] Lurie, J., *The "DAG" series*, Available at the author home page http://www.math.harvard.edu/~lurie/.

[19] ———, *Moduli problems for ring spectra,* Proceedings of the International Congress of Mathematicians. Volume II, 1099–1125, Hindustan Book Agency, New Delhi, 2010.

[20] ———, *Higher Algebra*, Available at the author home page http://www.math.harvard.edu/~lurie/.

[21] Pandharipande, R. and Thomas, R. P., *Almost closed 1-forms*, Glasg. Math. J. **56** (2014), no. 1, 169–182.

[22] Pantev, T., Toën, B., Vaquié, M., and Vezzosi, G., *Shifted symplectic structures*, Publ. Math. Inst. Hautes Études Sci. **117** (2013), 271–328.

[23] Preygel, A., *Thom-Sebastiani and Duality for Matrix Factorizations, and Results on the Higher Structures of the Hochschild Invariants*, Thesis (Ph.D.) Massachusetts Institute of Technology. 2012. Preprint arXiv:1101.5834.

[24] Pridham, J. P., *Representability of derived stacks*, J. K-Theory **10** (2012), no. 2, 413–453.

[25] Robalo, M., *Noncommutative Motives I, II*, Preprints arXiv:1206.3645, arXiv:1306.3795.

[26] Simpson, C., *Homotopy theory of higher categories*, New Mathematical Monographs, 19. Cambridge University Press, Cambridge, 2012. xviii+634 pp.

[27] ———, *Algebraic (geometric) n-stacks*. Preprint arXiv:alg-geom/9609014.

[28] ———, *Algebraic aspects of higher nonabelian Hodge theory*, Motives, polylogarithms and Hodge theory, Part II (Irvine, CA, 1998), 417–604, Int. Press Lect. Ser., 3, II, Int. Press, Somerville, MA, 2002.

[29] Tabuada, G., *Une structure de catégorie de modèles de Quillen sur la catégorie des dg-catégories*, C. R. Math. Acad. Sci. Paris **340** (2005), no. 1, 15–19.

[30] ———, *A guided tour through the garden of noncommutative motives* , Topics in noncommutative geometry, 259–276, Clay Math. Proc., 16, Amer. Math. Soc., Providence, RI, 2012.

[31] Toën, B., *Higher and derived stacks: a global overview*, Algebraic geometry Seattle 2005. Part 1, 435–487, Proc. Sympos. Pure Math., 80, Part 1, Amer. Math. Soc., Providence, RI, 2009.

[32] ———, *Derived Algebraic Geometry*. Preprint arXiv:1401.1044.

[33] ———, *Operations on derived moduli spaces of branes*. Preprint arXiv:1307.0405.

[34] ———, *Derived Azumaya algebras and generators for twisted derived categories*, Invent. Math. **189** (2012), no. 3, 581–652.

[35] Toën, B. and Vaquié, M., *Moduli of objects in dg-categories*, Ann. Sci. École Norm. Sup. (4) **40** (2007), no. 3, 387–444.

[36] Toën, B. and Vezzosi, G., *Homotopical algebraic geometry. I. Topos theory*, Adv. Math. **193** (2005), no. 2, 257–372.

[37] ———, *Homotopical algebraic geometry. II. Geometric stacks and applications*, Mem. Amer. Math. Soc. **193** (2008), no. 902, x+224 pp.

[38] Turaev, V. G., *Skein quantization of Poisson algebras of loops on surfaces*, Ann. Sci. École Norm. Sup. (4) **24** (1991), no. 6, 635–704.

[39] Wallbridge, J., *Derived analytic stacks and prequantum categories*, In preparation.

Université de Montpellier 2, Case courrier 051, Bât 9, Place Eugène Bataillon, Montpellier Cedex 5, France

E-mail: bertrand.toen@um2.fr

Teichmüller spaces, ergodic theory and global Torelli theorem

Misha Verbitsky

Abstract. A Teichmüller space Teich is a quotient of the space of all complex structures on a given manifold M by the connected components of the group of diffeomorphisms. The mapping class group Γ of M is the group of connected components of the diffeomorphism group. The moduli problems can be understood as statements about the Γ-action on Teich. I will describe the mapping class group and the Teichmüller space for a hyperkähler manifold. It turns out that this action is ergodic. We use the ergodicity to show that a hyperkähler manifold is never Kobayashi hyperbolic.

Mathematics Subject Classification (2010). Primary 32G13; Secondary 53C26.

Keywords. Torelli theorem, hyperkähler manifold, moduli space, mapping class group, Teichmüller space, ergodicity.

This talk is based on two papers, [46] and [47]. In these papers one can find details, examples, and rigorous proofs omitted here.

1. Teichmüller spaces

1.1. Teichmüller spaces and period maps. The notion of Teichmüller spaces has a long history since its discovery by Teichmüller in 1944 ([38]) and further development by Ahlfors, Bers and others. However, it is rarely applied to complex manifolds of dimension > 1. It turns out that this notion is interesting and useful for many purposes of complex geometry in any dimension.

Definition 1.1. Let M be a smooth manifold. An **almost complex structure** is an operator $I:\ TM \longrightarrow TM$ which satisfies $I^2 = -\operatorname{Id}_{TM}$. An almost complex structure is **integrable** if $\forall X, Y \in T^{1,0}M$, one has $[X,Y] \in T^{1,0}M$. In this case I is called **a complex structure operator**. A manifold with an integrable almost complex structure is called **a complex manifold**.

Definition 1.2. The space of almost complex structures is an infinite-dimensional Fréchet manifold X_M of all tensors $I \in \operatorname{End}(TM)$ satisfying $I^2 = -\operatorname{Id}_{TM}$. Similarly, one considers **the group of diffeomorphisms** as a Fréchet Lie group.

Remark 1.3. Definition of Fréchet manifolds and Fréchet spaces and many results on the geometry of infinite-dimensional manifolds can be found in [18].

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Definition 1.4. Let M be a compact complex manifold, and $\operatorname{Diff}_0(M)$ a connected component of its diffeomorphism group (**the group of isotopies**). Denote by Comp the space of complex structures on M, considered with the topology induced from a Fréchet manifold of almost complex structures, and let $\operatorname{Teich} := \operatorname{Comp} / \operatorname{Diff}_0(M)$ be the quotient space with the quotient topology. We call it **the Teichmüller space.**

Remark 1.5. When the complex manifold M admits a certain geometric structure, such as Kähler, or hyperkähler structure, it is natural to consider the Teichmüller space of complex structure compatible with (say) Kähler structure. Consider the open subset $\operatorname{Comp}_K \subset \operatorname{Comp}$ of all complex structures I such that (M, I) admits a Kähler structure. The corresponding Teichmüller space is $\operatorname{Teich}_K := \operatorname{Comp}_K / \operatorname{Diff}_0$. When working with the Teichmüller space of hyperkähler manifolds, or a torus, we shall always restrict ourselves to Comp_K and Teich_K.

Results of Kuranishi about local structure of deformation spaces can be summarized as a statement about local structure of Comp as follows ([14, 24, 25]).

Theorem 1.6. *Let M be a compact complex manifold and $I \in \operatorname{Comp}$. Then there exists an open neighbourhood $U \ni I$ in Comp and a neighbourhood R of unit in Diff_0 satisfying the following. Consider the quotient U/R of U by an equivalence relation generated by $x \sim wy$, for all $x, y \in U$ and $w \in R$. Then U/R is a complex analytic stack, equipped with a natural holomorphic embedding to the stacky quotient $H^1(TM)/\operatorname{Aut}(M)$. In particular, when $\operatorname{Aut}(M)$ is finite, the space U/R is a complex variety, embeddable to an orbifold $H^1(TM)/\operatorname{Aut}(M)$.*

Remark 1.7. The quotient space U/R obtained by Kuranishi is called **the Kuranishi space**. Let $\operatorname{Teich}(U)$ be an image of U in the Teichmüller space. Clearly, the Kuranishi space admits a surjective, continuous map to $\operatorname{Teich}(U)$. It is not entirely clear whether this map is always a homeomorphism. However, if it is always a homeomorphism, for a given M, the space Teich acquires a structure of a complex variety.

As shown by F. Catanese ([11, Proposition 15]), for Kähler manifolds with trivial canonical bundle, e.g. for the hyperkähler manifolds, the Teichmüller space is locally isomorphic to the Kuranishi moduli space, hence it is a complex variety. In this case it is actually a complex manifold, by Bogomolov-Tian-Todorov theorem ([7, 39, 42]).

It is not clear if this is true for a general complex manifold; in the present work we deal with hyperkähler manifolds, which are Calabi-Yau.

Question 1.8. *Consider a compact complex manifold M, and let Teich be its Teichmüller space. Can we equip Teich with a structure of a complex variety (possibly non-Hausdorff), in a way which is compatible with the local charts obtained from the Kuranishi theorem?*

When M is a torus, or a hyperkähler manifold, Teich is a complex manifold which can be described explicitly (Theorem 2.1, Theorem 4.5). However, even for a hyperkähler manifold, Teich is not Hausdorff.

Claim 1.9. *Assume that M is Kähler, and Teich the Teichmüller space of all complex structures of Kähler type on M. For a given $I \in \operatorname{Teich}$, choose a representative $\tilde{I} \in \operatorname{Comp}$. Then the Hodge decomposition $H^*(M) = \bigoplus H^{p,q}(M, \tilde{I})$ is independent from the choice of $\tilde{I}$.*

Proof. The ambiguity of a choice of $\tilde{I}$ lies in Diff_0. However, Diff_0 acts trivially on $H^*(M)$. □

This elementary claim allows one to define **the period map**.

Definition 1.10. Let M be a Kähler manifold, Teich its Teichmüller space, and Per the map associating to I the Hodge decomposition $H^*(M) = \bigoplus H^{p,q}(M, I)$. Then Per is called **the period map** of M.

Remark 1.11. Consider the product $\mathrm{Comp} \times M$ trivially fibered over Comp. The fibers of $\pi : M \times \mathrm{Comp} \longrightarrow \mathrm{Comp}$ can be considered as complex manifolds, with complex structure at $I \in \mathrm{Comp}$ given by I. This complex structure is clearly Diff_0-invariant, giving a complex structure on the fibers of the quotient fibration $(M \times \mathrm{Comp})/\mathrm{Diff}_0 \longrightarrow \mathrm{Teich}$. At each $I \in \mathrm{Teich}$, the fiber of this fibration (called **the universal fibration**) is isomorphic to (M, I).

1.2. Marked moduli spaces. A more conventional approach to the moduli problem goes as follows. Given a complex manifold M, one defines **the deformation functor** from marked complex spaces to sets as a functor mapping a complex space (B, x) to the set of equivalence classes of deformations $\pi : \mathcal{X} \longrightarrow B$ of M over B with M identified with the fiber of π at x.

If the deformation functor is representable by a complex space, this space is called **the fine moduli space** of deformations of M.

Usually, the fine moduli space does not exist. In this case, one considers the category of natural transformations from the deformation functor to representable functors. The initial object in this category is called **the coarse moduli space**. The points of coarse moduli are identified with equivalence classes of deformations of M.

In this setup, an analogue of Teichmüller space can be defined as follows. Fix an abelian group which is isomorphic to $H^*(M, \mathbb{Z})$, and define **a marked manifold** as a pair $(M, \varphi : V \tilde{\longrightarrow} H^*(M, \mathbb{Z}))$, where M is a complex manifold, and φ a group isomorphism. In the same way as above, one defines a coarse moduli space of deformations of marked manifolds. To compare this space with Teichmüller space, consider a subgroup group Γ_0 of mapping class group which acts trivially on cohomology. Clearly, the points of Teich/Γ_0 are in bijective correspondence with the equivalence classes of marked complex structures on M.

Given a coarse marked moduli space W, one obtains the tautological map $W \longrightarrow \mathrm{Teich}/\Gamma_0$, by construction continuous. For hyperkähler manifolds (or compact tori), this map is a diffeomorphism on each connected component ([46, Corollary 4.31]).

2. Torelli theorem

2.1. Torelli theorem: an introduction. Torelli theorems are a broad class of results which describe the Teichmüller spaces in terms of the period maps (Definition 1.10).

The name originates with Ruggiero Torelli, who has shown that it is possible to reconstruct a Riemann surface from its Jacobian [43]. The term "Torelli theorems" is due to André Weil [50], who gave a modern proof of this classical result, and explained its possible generalizations.

One may distinguish between the "local Torelli theorem", where a local structure of deformation space is described in terms of periods, and "global Torelli", where the Teichmüller space is described globally.

Weil, who was the first to define and study K3 surfaces, spent much time trying to prove

the Torelli theorem for K3 surfaces, but it was notoriously difficult. Its local version is due to Tjurina, Piatetski-Shapiro and Shafarevich [35, 40]. The local Torelli was generalized by Bogomolov to hyperkähler manifolds [6] and by Bogomolov-Tian-Todorov to Calabi-Yau manifolds [7, 39, 42], building foundation for the theory of Mirror Symmetry.

In dimension > 1, the global Torelli theorem was known only for compact tori (where it is essentially trivial) and the K3 surfaces, where it was proven by Kulikov in 1977 [23], and then improved many times during the 1980-ies [3, 15, 27, 41, 44].

2.2. Birational Teichmüller space. In what follows, **a hyperkähler manifold** is a compact complex manifold admitting a Kähler structure and a holomorphic symplectic form.

In generalizing global Torelli to more general hyperkähler manifolds, two problems were apparent. First of all, bimeromorphic hyperkähler manifolds have the same periods, hence the period map cannot distinguish between them. However, for $\dim_{\mathbb{C}} M > 0$, birational holomorphically symplectic manifolds can be non-isomorphic [12].

Another (mostly psychological) difficulty is based on attachment to moduli spaces, as opposed to marked moduli or Teichmüller spaces. For a K3, one can reconstruct a K3 from its Hodge structure, and this gives an identification between the moduli and the space of Hodge structures. In bigger dimension, one has to use the Teichmüller space. Indeed, for some classes of hyperkähler manifolds, the group $O(H^2(M, \mathbb{Z})$ of Hodge isometries of cohomology is strictly bigger than the image of the mapping class group. This gives elements $\gamma \in O(H^2(M, \mathbb{Z})$ acting non-trivially on the Teichmüller space in such a way that the complex manifolds (M, I) and $(M, \gamma(I))$ are not birationally equivalent ([30, 33]). However, their Hodge structures *are* equivalent, by construction. This example explains the necessity of using the Teichmüller spaces (or marked moduli) to state the Torelli theorem: its Hodge-theoretic version is often false.

For Teichmüller spaces, the Torelli theorem is a statement about the period map (Definition 1.10). Ideally, we want the period map to give a diffeomporphism between Teich and the corresponding space of Hodge structures. This is what happens for a compact torus.

Theorem 2.1. *Let M be a compact torus, $\dim_{\mathbb{R}} M = 2n$, and Teich the Teichmüller space of all complex structures of Kähler type on M. Denote by $\mathbb{P}\mathsf{er}$ the space $SL(2n, \mathbb{R})/SL(n, \mathbb{C})$ of all Hodge structures of weight one on $H^1(M, \mathbb{C})$, that is, the space of all complex operators on $H^1(M, \mathbb{R})$ compatible with the orientation. Then the period map $\mathsf{Per} : \operatorname{Teich} \longrightarrow \mathbb{P}\mathsf{er}$ is a diffeomorphism on each connected component of $\mathbb{P}\mathsf{er}$.*

Unfortunately, this ideal situation is almost never realized. Even in the simplest cases (such as for hyperkähler manifolds), the Teichmüller space is no longer Hausdorff. However, in some situations it is still possible to deal with non-Hausdorff points.

Remark 2.2. A non-Hausdorff manifold is a topological space locally diffeomorphic to $\mathbb{R}^n$ (but not necessarily Hausdorff).

Definition 2.3. Let X be a topological space, and $X \stackrel{\varphi}{\longrightarrow} X_0$ a continuous surjection. The space X_0 is called **a Hausdorff reduction** of X if any continuous map $X \longrightarrow X'$ to a Hausdorff space is factorized through φ.

Definition 2.4. Let M be a topological space. We say that $x, y \in M$ are **non-separable** (denoted by $x \sim y$) if for any open sets $V \ni x, U \ni y$, one has $U \cap V \neq \emptyset$.

Remark 2.5. Suppose that $\sim$ is an equivalence relation, and the quotient $M/\sim$ is Hausdorff. Then $M/\sim$ is a Hausdorff reduction of M.

Unfortunately, this notion cannot be applied universally. Firstly, $\sim$ is not always an equivalence relation; and secondly, even if $\sim$ is equivalence, the $M/\sim$ is not always Hausdorff. Fortunately, for Teichmüller space of a hyperkähler manifold, Hausdorff reduction can be defined, using the following theorem due to D. Huybrechts ([19]).

Theorem 2.6. *If I_1, $I_2 \in \operatorname{Teich}$ are non-separate points, then (M, I_1) is birationally equivalent to (M, I_2).*

Using this result and geometry of the period map (Bogomolov's local Torelli theorem), it is elementary to show that the quotient $\operatorname{Teich}_b := \operatorname{Teich} / \sim$ is a Hausdorff manifold. This quotient is called **the birational Teichmüller space** of a hyperkähler manifold.

Global Torelli theorem implies that for hyperkähler manifolds the period map induces a diffeomorphism between the Hausdorff reduction of the Teichmüller space and the appropriate period domain.

3. Hyperkähler manifolds and Bogomolov-Beauville-Fujiki form

3.1. Hyperkähler manifolds: definition and examples. The standard definition of hyperkähler manifolds is rather differential geometric. It is, indeed, synonymous with "holomorphic symplectic", but this synonymity follows from Calabi-Yau theorem. For more details about hyperkähler manifolds, please see [2] or [4].

Definition 3.1. A **hyperkähler structure** on a manifold M is a Riemannian structure g and a triple of complex structures I, J, K, satisfying quaternionic relations $I \circ J = -J \circ I = K$, such that g is Kähler for I, J, K.

Remark 3.2. This is equivalent to $\nabla I = \nabla J = \nabla K = 0$: the parallel translation along the connection preserves I, J, K.

Remark 3.3. A hyperkähler manifold has three symplectic forms: $\omega_I := g(I\cdot,\cdot)$, $\omega_J := g(J\cdot,\cdot)$, $\omega_K := g(K\cdot,\cdot)$.

Definition 3.4. Let M be a Riemannian manifold, $x \in M$ a point. The subgroup of $GL(T_xM)$ generated by parallel translations (along all paths) is called **the holonomy group** of M.

Remark 3.5. A hyperkähler manifold can be defined as a manifold which has holonomy in $Sp(n)$ (the group of all endomorphisms preserving I, J, K).

Definition 3.6. A holomorphically symplectic manifold is a complex manifold equipped with non-degenerate, holomorphic $(2,0)$-form.

Remark 3.7. Hyperkähler manifolds are holomorphically symplectic. Indeed, $\Omega := \omega_J + \sqrt{1}\omega_K$ is a holomorphic symplectic form on (M, I).

Theorem 3.8 (Calabi-Yau). *A compact, Kähler, holomorphically symplectic manifold admits a unique hyperkähler metric in any Kähler class.*

Remark 3.9. For the rest of this talk, a hyperkähler manifold means a compact complex manifold admitting a Kähler structure and a holomorphically symplectic structure.

Definition 3.10. A hyperkähler manifold M is called **simple**, or **IHS**, if $\pi_1(M) = 0$, $H^{2,0}(M) = \mathbb{C}$.

The rationale for this terminology comes from Bogomolov's decomposition theorem.

Theorem 3.11 (Bogomolov, [5]). *Any hyperkähler manifold admits a finite covering which is a product of a torus and several simple hyperkähler manifolds.*

Further on, all hyperkähler manifolds are assumed to be simple.

Remark 3.12. A hyperkähler manifold is simple if and only if its holonomy group is $Sp(n)$, and not a proper subgroup of $Sp(n)$ [4].

Example 3.13. Take a 2-dimensional complex torus T, then the singular locus of $T/\pm 1$ is 16 points locally of form $\mathbb{C}^2/\pm 1$. Its resolution by blow-up is called **a Kummer surface**. It is not hard to see that it is holomorphically symplectic.

Definition 3.14. A K3 surface is a hyperkähler manifold which is diffeomorphic to a Kummer surface.

In real dimension 4, the only compact hyperkähler manifolds are tori and K3 surfaces, as follows from the Kodaira-Enriques classification.

Definition 3.15. A **Hilbert scheme** $M^{[n]}$ of a complex surface M is a classifying space of all ideal sheaves $I \subset \mathcal{O}_M$ for which the quotient $\mathcal{O}_M/I$ has dimension n over $\mathbb{C}$.

Remark 3.16. A Hilbert scheme is obtained as a resolution of singularities of the symmetric power $\operatorname{Sym}^n M$.

Theorem 3.17 (Fujiki, Beauville). *A Hilbert scheme of a hyperkähler manifold of real dimension 2 is hyperkähler.*

Example 3.18. Let T be a torus. Then T acts on its Hilbert scheme freely and properly by translations. For $n = 2$, the quotient $T^{[n]}/T$ is a Kummer K3-surface. For $n > 2$, a universal covering of $T^{[n]}/T$ is called **a generalized Kummer variety**.

Remark 3.19. There are 2 more "sporadic" examples of compact hyperkähler manifolds, constructed by K. O'Grady ([34]). All known simple hyperkaehler manifolds are these 2 and the two series: Hilbert schemes of K3 and generalized Kummer.

3.2. Bogomolov-Beauville-Fujiki form and the mapping class group.

Theorem 3.20 (Fujiki, [16]). *Let $\eta \in H^2(M)$, and $\dim M = 2n$, where M is hyperkähler. Then $\int_M \eta^{2n} = cq(\eta,\eta)^n$, for some primitive integer quadratic form q on $H^2(M,\mathbb{Z})$, and $c > 0$ a positive rational number, called* **Fujiki constant**.

Definition 3.21. This form is called **Bogomolov-Beauville-Fujiki form**. It is defined by the Fujiki's relation uniquely, up to a sign. The sign is determined from the following formula (Bogomolov, Beauville)

$$\begin{aligned}\lambda q(\eta,\eta) = &\int_X \eta\wedge\eta\wedge\Omega^{n-1}\wedge\bar\Omega^{n-1} - \\ &- \frac{n-1}{n}\left(\int_X \eta\wedge\Omega^{n-1}\wedge\bar\Omega^n\right)\left(\int_X \eta\wedge\Omega^n\wedge\bar\Omega^{n-1}\right)\end{aligned}$$

where Ω is the holomorphic symplectic form, and $\lambda > 0$.

Remark 3.22. The BBF form q has signature $(b_2 - 3, 3)$. It is negative definite on primitive forms, and positive definite on $\langle \Omega, \bar{\Omega}, \omega \rangle$, where ω is a Kähler form.

Using the BBF form, it is possible to describe the automorphism group of cohomology in a very convenient way.

Theorem 3.23. *Let M be a simple hyperkähler manifold, and $G \subset GL(H^*(M))$ a group of automorphisms of its cohomology algebra preserving the Pontryagin classes. Then G acts on $H^2(M)$ preserving the BBF form. Moreover, the map $G \longrightarrow O(H^2(M, \mathbb{R}), q)$ is surjective on a connected component, and has compact kernel.*

Proof.

Step 1 : Fujiki formula $v^{2n} = q(v, v)^n$ implies that Γ_0 preserves the Bogomolov-Beauville-Fujiki up to a sign. The sign is fixed, if n is odd.

Step 2 : For even n, the sign is also fixed. Indeed, G preserves $p_1(M)$, and (as Fujiki has shown in [16]), $v^{2n-2} \wedge p_1(M) = q(v, v)^{n-1} c$, for some $c \in \mathbb{R}$. The constant c is positive, because the degree of $c_2(B)$ is positive for any non-trivial Yang-Mills bundle with $c_1(B) = 0$.

Step 3 : $\mathfrak{o}(H^2(M, \mathbb{R}), q)$ acts on $H^*(M, \mathbb{R})$ by derivations preserving Pontryagin classes ([45]). Therefore $\operatorname{Lie}(G)$ surjects to $\mathfrak{o}(H^2(M, \mathbb{R}), q)$.

Step 4 : The kernel K of the map $G \longrightarrow G\big|_{H^2(M,\mathbb{R})}$ is compact, because it commutes with the Hodge decomposition and Lefschetz $\mathfrak{sl}(2)$-action, hence preserves the Riemann-Hodge form, which is positive definite. □

Using this result, the mapping class group can also be computed. We use a theorem of D. Sullivan, who expressed the mapping group in terms of the rational homotopy theory, and expressed the rational homotopy in terms of the algebraic structure of the de Rham algebra.

Theorem 3.24 (Sullivan, [37, Theorem 10.3, Theorem 12.1, Theorem 13.3]). *Let M be a compact, simply connected Kähler manifold, $\dim_{\mathbb{C}} M \geqslant 3$. Denote by Γ_0 the group of automorphisms of an algebra $H^*(M, \mathbb{Z})$ preserving the Pontryagin classes $p_i(M)$. Then the natural map $\operatorname{Diff}(M)/\operatorname{Diff}_0 \longrightarrow \Gamma_0$ has finite kernel, and its image has finite index in Γ_0.*

As a corollary of this theorem, we obtain a similar result about hyperkähler manifolds.

Theorem 3.25. *Let M be a simple hyperkähler manifold, and Γ_0 the group of automorphisms of an algebra $H^*(M, \mathbb{Z})$ preserving the Pontryagin classes $p_i(M)$. Then*

(i) *$\Gamma_0\big|_{H^2(M,\mathbb{Z})}$ is a finite index subgroup of $O(H^2(M, \mathbb{Z}), q)$.*

(ii) *The map $\Gamma_0 \longrightarrow O(H^2(M, \mathbb{Z}), q)$ has finite kernel.*

We obtained that the mapping group is **arithmetic** (commensurable to a subgroup of integer points in a rational Lie group).

As follows from [20, Theorem 2.1], there are only finitely many connected components of Teich. Let Γ^I be the group of elements of mapping class group preserving a connected component of Teichmüller space containing $I \in \operatorname{Teich}$. Then Γ^I is also arithmetic. Indeed, it has finite index in Γ.

Definition 3.26. The image of Γ^I in $GL(H^2(M,\mathbb{Z}))$ is called **monodromy group of a manifold**.

Remark 3.27. The monodromy group can also be obtained as a group generated by monodromy of all Gauss-Manin local system for all deformations of M ([46, Theorem 7.2]). This explains the term. This notion was defined and computed in many special cases by E. Markman [29, 30].

4. Global Torelli theorem

4.1. Period map. To study the moduli problem, one should understand the mapping class group (described above) and the Teichmüller space. It turns out that the birational Teichmüller space has a very simple description in terms of the period map, inducing a diffeomorphism

$$\operatorname{Teich}_b \longrightarrow \frac{SO(b_2-3,3)}{SO(b_2-3,1)\times SO(2)}$$

on each connected component of Teich_b.

Definition 4.1. Let $\mathsf{Per}: \operatorname{Teich} \longrightarrow \mathbb{P}H^2(M,\mathbb{C})$ map J to a line $H^{2,0}(M,J) \in \mathbb{P}H^2(M,\mathbb{C})$. The map $\mathsf{Per}: \operatorname{Teich} \longrightarrow \mathbb{P}H^2(M,\mathbb{C})$ is called **the period map**.

Remark 4.2. Per maps Teich into an open subset of a quadric, defined by

$$\mathbb{P}\mathsf{er} := \{l \in \mathbb{P}H^2(M,\mathbb{C}) \mid q(l,l)=0, q(l,\bar{l})>0\}.$$

The manifold $\mathbb{P}\mathsf{er}$ is called **the period space** of M.

As follows from Proposition 4.8 below, $\mathbb{P}\mathsf{er} = \frac{SO(b_2-3,3)}{SO(b_2-3,1)\times SO(2)}$.

Theorem 4.3 (Bogomolov, [6]). *Let M be a simple hyperkähler manifold, and* Teich *its Teichmüller space. Then the period map* $\mathsf{Per}: \operatorname{Teich} \longrightarrow \mathsf{Per}$ *is etale (has invertible differential everywhere).*

Remark 4.4. Bogomolov's theorem implies that Teich is smooth. It is non-Hausdorff even in the simplest examples.

Now the global Torelli theorem can be stated as follows. Recall that the birational Teichmüller space Teich_b is a Hausdorff reduction of the Teichmüller space of the holomorphic symplectic manifolds of Kähler type.

Theorem 4.5. *Let M be a simple hyperkähler manifold, and* $\mathsf{Per}: Teich_b \longrightarrow \mathbb{P}\mathsf{er}$ *the period map. Then* Per *is a diffeomorphism on each connected component.*

The following proposition is proven in a straghtforward manner using 1950-ies style arguments of geometric topology.

Proposition 4.6 (The Covering Criterion). *Let $X \stackrel{\varphi}{\longrightarrow} Y$ be an etale map of smooth manifolds. Suppose that each $y \in Y$ has a neighbourhood $B \ni y$ diffeomorphic to a closed ball, such that for each connected component $B' \subset \varphi^{-1}(B)$, B' projects to B surjectively. Then φ is a covering.*

Now, the Global Torelli implied by the following result, which is proven in Subsection 4.2 using hyperkähler structures.

Proposition 4.7. *In assumptions of Theorem 4.5, the period map satisfies the conditions of the Covering Criterion.*

4.2. Moduli of hyperkähler structures and twistor curves.

Proposition 4.8. *The period space*

$$\mathbb{P}\mathsf{er} := \{l \in \mathbb{P}H^2(M,\mathbb{C}) \mid q(l,l)=0, q(l,\bar{l})>0\}$$

is identified with $\frac{SO(b_2-3,3)}{SO(2)\times SO(b_2-3,1)}$*, which is a Grassmannian of positive oriented 2-planes in* $H^2(M,\mathbb{R})$.

Proof.

Step 1 : Given $l \in \mathbb{P}H^2(M,\mathbb{C})$, the space generated by $\operatorname{Im} l, \operatorname{Re} l$ is 2-dimensional, because $q(l,l)=0, q(l,\bar{l})$ implies that $l \cap H^2(M,\mathbb{R})=0$.

Step 2 : This 2-dimensional plane is positive, because $q(\operatorname{Re} l, \operatorname{Re} l) = q(l+\bar{l}, l+\bar{l}) = 2q(l,\bar{l})>0$.

Step 3 : Conversely, for any 2-dimensional positive plane $V \in H^2(M,\mathbb{R})$, the quadric $\{l \in V\otimes_{\mathbb{R}}\mathbb{C} \mid q(l,l)=0\}$ consists of two lines; a choice of a line is determined by orientation. □

Remark 4.9. Two hyperkähler structures (M,I,J,K,g) and (M,I',J',K',g) are called **equivalent** if there exists a unitary quaternion h such that $I'=hIh^-1$, $J'=hJh^-1$, $K'=hKh^-1$. From the holonomy characterization of simple hyperkähler manifolds (Remark 3.12) it follows that two hyperkähler structures are isometric if and only if they are equivalent.

Definition 4.10. Let (M,I,J,K,g) be a hyperkähler manifold. **A hyperkähler 3-plane** in $H^2(M,\mathbb{R})$ is a positive oriented 3-dimensional subspace W, generated by three Kähler forms $\omega_I,\omega_J,\omega_K$.

Definition 4.11. Similarly to the Teichmüller space and period map of complex structures, one can define the period space of hyperkähler metrics. Denote it by Teich_H. The corresponding period map is

$$\mathsf{Per}: \operatorname{Teich}_H \longrightarrow \mathbb{P}\mathsf{er}_H,$$

where $\mathbb{P}\mathsf{er}_H = \frac{SO(b_2-3,3)}{SO(3)\times SO(b_2-3)}$ is the space of positive, oriented 3-planes, and Per maps a hyperkähler structure to the corresponding hyperkähler 3-plane.

Remark 4.12. There is one significant difference between Teich and the hyperkähler Teichmüller space Teich_H: the latter is Hausdorff, and, in fact, metrizable. Indeed, we could equip the space Teich_H of hyperkähler metrics with the Gromov-Hausdorff metric.

Let $I \in \operatorname{Teich}$ be a complex structure, and $\mathcal{K}(I)$ its Kähler cone. The set of hyperkähler metrics compatible with I is parametrized by $\mathcal{K}(I)$, by Calabi-Yau theorem. The corresponding 3-dimensional subspaces are generated by $\mathsf{Per}(I)+\omega$, where $\omega \in \mathcal{K}(I)$. The local Torelli

theorem implies that locally $I \in \mathrm{Teich}$ is uniquely determined by the 2-plane generated by ω_J and ω_K; Calabi-Yau theorem implies that the hyperkähler metric is uniquely determined by the complex structure and the Kähler structure. This gives the following hyperkähler version of the local Torelli theorem.

Theorem 4.13. *Let M be a simple hyperkähler manifold, and Teich_H its hyperkähler Teichmüller space. Then the period map $\mathsf{Per} : \mathrm{Teich} \longrightarrow \mathsf{Per}_H$ mapping an equivalence class of hyperkähler structures to is its 3-plane is etale (has invertible differential everywhere).*

Remark 4.14. Let $W \subset H^2(M, \mathbb{R})$ be a positive 3-dimensional plane. The set $S_W \subset \mathbb{P}\mathsf{er}$ of oriented 2-dimensional planes in W is identified with $S^2 = \mathbb{C}P^1$. When W is a hyperkähler 3-plane, S_W is called **the twistor family** of a hyperkähler structure. A point in the twistor family corresponds to a complex structure $aI + bJ + cK \in \mathbb{H}$, with $a^2 + b^2 + c^2 = 1$. We call the corresponding rational curves $\mathbb{C}P^1 \subset \mathrm{Teich}$ **the twistor lines**. It is not hard to see that the twistor lines are holomorphic.

Definition 4.15. Let $W \in \mathbb{P}\mathsf{er}_H$ be a positive 3-plane, $S_W \subset \mathbb{P}\mathsf{er}$ the corresponding rational curve, and $x \in S_W$ be a point. It is called **liftable** if for any point $y \in \mathsf{Per}^{-1}(x) \subset \mathrm{Teich}$ there exists $\mathcal{H} \in \mathrm{Teich}_H$ such that the corresponding twistor line contains y.

When W is generic, the corresponding line S_W is liftable, as indicated below.

Definition 4.16. The Neron-Severi lattice $NS(I)$ of a hyperkähler manifold (M, I) is

$$H^{1,1}(M, I) \cap H^2(M, \mathbb{Z}).$$

The following theorem, based on results of [13], was proven by D. Huybrechts.

Theorem 4.17 ([19]). *Let M be a hyperkaehler manifold with $\mathsf{NS}(M) = 0$. Then its Kaehler cone is one of two components of the set*

$$\{\nu \in H^{1,1}(M, \mathbb{R}) \mid q(\nu, \nu) \geqslant 0\}.$$

Definition 4.18. Let $S \subset \mathrm{Teich}$ be a $\mathbb{C}P^1$ associated with a twistor family. It is called **generic** if it passes through a point $I \in \mathrm{Teich}$ with $\mathsf{NS}(M, I) = 0$. Clearly, a hyperkähler 3-plane $W \subset H^2(M, \mathbb{R})$ corresponds to a generic twistor family if and only if its orthogonal complement $W^\perp \subset H^2(M, \mathbb{R})$ does not contain rational vectors. A 3-plane $W \in \mathbb{P}\mathsf{er}_H$ is called **generic** if $W^\perp \subset H^2(M, \mathbb{R})$ does not contain rational vectors. The corresponding rational curve $S_W \subset \mathbb{P}\mathsf{er}$ is called **a GHK line**. GHK lines are liftable, which is very useful for many purposes, including the proof of Torelli theorem (see also [1], where GHK lines were used to study Kähler cones of hyperkähler manifolds).

The following theorem immediately follows from the Calabi-Yau theorem and the description of the Kähler cone given in Theorem 4.17.

Theorem 4.19. *Let $W \in \mathbb{P}\mathsf{er}_H$ be a generic plane, $S_W \subset \mathbb{P}\mathsf{er}$ the corresponding rational curve, and $x \in S_W$ a generic point. Then (S_W, x) is liftable.*

Assumptions of the covering criterion (Proposition 4.7) immediately follow from Theorem 4.19. Indeed, it is not hard to see that any two points on a closed ball $B \subset \mathbb{P}\mathsf{er}$ can be connected inside B by a sequence of GHK curves intersecting in generic points of B. Since these curves are liftable, any connected component of $\mathsf{Per}^{-1}(B)$ is mapped to B surjectively.

5. Teichmüller spaces and ergodic theory

5.1. Ergodic complex structures. After the Teichmüller space and the mapping class group are understood, it is natural to consider the quotient space $\mathrm{Comp} / \mathrm{Diff} = \mathrm{Teich} / \Gamma$ of the Teichmüller space by the mapping class group $\Gamma := \mathrm{Diff} / \mathrm{Diff}_0$.

Claim 5.1. *Let M be a simple hyperkähler manifold, Γ its mapping class group, and Teich_b the birational Teichmüller space. Then the quotient $\mathrm{Teich}_b / \Gamma$ parametrizes the birational classes of deformations of M.*

One could call the quotient $\mathrm{Teich}_b / \Gamma$ "the moduli space", but, unfortunately, this is not a space in any reasonable sense. Indeed, as we shall see, non-trivial closed subsets of $\mathrm{Teich}_b / \Gamma$ are at most countable, making $\mathrm{Teich}_b / \Gamma$ terribly non-Hausdorff. This means that the concept of "moduli space" has no meaning, and all interesting information about moduli problems is hidden in dynamics of Γ-action on Teich.

Let $I \in \mathrm{Teich}$ be a point, and $\mathrm{Teich}^I \subset \mathrm{Teich}$ its connected component. Since Teich has finitely many components, a subgroup mapping class group fixing Teich has finite index. Its image in $\mathrm{Aut}(\mathrm{Teich}^I)$ is called **monodromy group** and denoted Γ^I (Definition 3.26). It is a finite index subgroup in $SO(H^2(M, \mathbb{Z}))$.

All that said, we find that the moduli problem for hyperkähler manifold is essentially reduced to the dynamics of the Γ^I-action on the space $\mathbb{P}\mathrm{er}$, which is understood as a Grassmannian of positive, oriented 2-planes in $H^2(M, \mathbb{R})$.

It is natural to study the dynamics of a group action from the point of view of ergodic theory, ignoring measure zero subsets. However, the quotient map $\mathrm{Teich} \longrightarrow \mathrm{Teich}_b$ is bijective outside of a union of countably many divisors, corresponding to complex structures I with $NS(M, I)$ non-zero. This set has measure 0. Therefore, the quotient map $\mathrm{Teich} \longrightarrow \mathrm{Teich}_b$ induces an equivalence of measured spaces. For the purposes of ergodic theory, we shall identify Teich^I with the corresponding homogeneous space $\mathbb{P}\mathrm{er}$.

Ths first observation, based on a theorem of C. Moore, implies that the monodromy action on $\mathbb{P}\mathrm{er}$ is ergodic.

Definition 5.2. Let (M, μ) be a space with measure, and G a group acting on M. This action is **ergodic** if all G-invariant measurable subsets $M' \subset M$ satisfy $\mu(M') = 0$ or $\mu(M \backslash M') = 0$.

Claim 5.3. *Let M be a manifold, μ a Lebesgue measure, and G a group acting on (M, μ) ergodically. Then the set of non-dense orbits has measure 0.*

Proof. Consider a non-empty open subset $U \subset M$. Then $\mu(U) > 0$, hence $M' := G \cdot U$ satisfies $\mu(M \backslash M') = 0$. For any orbit $G \cdot x$ not intersecting U, one has $x \in M \backslash M'$. Therefore, the set of such orbits has measure 0. □

Definition 5.4. Let $I \in \mathrm{Comp}$ be a complex structure on a manifold. It is called **ergodic** if its Diff-orbit is dense in its connected component of Comp.

Remark 5.5. This is equivalent to density of Γ-orbit of I in its Teichmüller component.

5.2. Ergodicity of the monodromy group action.

Definition 5.6. Let G be a Lie group, and $\Gamma \subset G$ a discrete subgroup. Consider the pushforward of the Haar measure to G/Γ. We say that Γ **has finite covolume** if the Haar measure of G/Γ is finite. In this case Γ is called **a lattice subgroup**.

Remark 5.7. Borel and Harish-Chandra proved that an arithmetic subgroup of a reductive group G is a lattice whenever G has no non-trivial characters over $\mathbb{Q}$ (see e.g. [48]). In particular, all arithmetic subgroups of a semi-simple group are lattices.

Theorem 5.8 (Calvin C. Moore, [31, Theorem 7]). *Let Γ be an arithmetic subgroup in a non-compact simple Lie group G with finite center, and $H \subset G$ a non-compact subgroup. Then the left action of Γ on G/H is ergodic.*

Theorem 5.9. *Let* Teich *be a connected component of a Teichmüller space, and Γ^I its monodromy group. Then the set of all non-ergodic points of* Teich *has measure 0.*

Proof. Global Torelli theorem identifies Teich (as a measured space) and G/H, where $G = SO(b_2 - 3, 3)$, $H = SO(2) \times SO(b_2 - 3, 1)$. Since Γ^I is an arithmetic lattice, Γ^I-action on G/H is ergodic, by Moore's theorem. □

Moore's theorem implies that outside of a measure zero set, all complex structures on Teich are ergodic. If we want to determine which exactly complex structures are ergodic, we have to use Ratner's theorem, giving precise description of a closure of a Γ^I-orbit in a homogeneous space.

Now I will state some basic results of Ratner theory. For more details, please see [22] and [32].

Definition 5.10. Let G be a Lie group, and $g \in G$ any element. We say that g is **unipotent** if $g = e^h$ for a nilpotent element h in its Lie algebra. A group G is **generated by unipotents** if G is multiplicatively generated by unipotent elements.

Theorem 5.11 ([32, 1.1.15 (2)]). *Let $H \subset G$ be a Lie subroup generated by unipotents, and $\Gamma \subset G$ a lattice. Then a closure of any H-orbit in G/Γ is an orbit of a closed, connected subgroup $S \subset G$, such that $S \cap \Gamma \subset S$ is a lattice.*

When this lattice is arithmetic, one could describe the group S very explicitly.

Claim 5.12 ([22, Proposition 3.3.7] or [36, Proposition 3.2]). *Let $x \in G/H$ be a point in a homogeneous space, and $\Gamma \cdot x$ its Γ-orbit, where Γ is an arithmetic lattice. Then its closure is an orbit of a group S containing stabilizer of x. Moreover, S is a smallest group defined over rationals and stabilizing x.*

For the present purposes, we are interested in a pair $SO(3, k) \supset SO(1, k) \times SO(2) \subset G$ (or, rather, their connected components $G = SO^+(3, k)$ and $H = SO(1, k) \times SO(2) \subset G$). In this case, there are no intermediate subgroups.

Claim 5.13. *Let $G = SO^+(3, k)$, and $H \cong SO^+(1, k) \times SO(2) \subset G$. Then any closed connected Lie subgroup $S \subset G$ containing H coincides with G or with H.*

Corollary 5.14. *Let $J \in \mathbb{P}\mathsf{er} = G/H$. Then either J is ergodic, or its Γ-orbit is closed in* $\mathbb{P}\mathsf{er}$.

By Ratner's theorem, in the latter case the H-orbit of J has finite volume in G/Γ. Therefore, its intersection with Γ is a lattice in H. This brings

Corollary 5.15. *Let $J \in \mathbb{P}\mathsf{er}$ be a point such that its Γ-orbit is closed in* $\mathbb{P}\mathsf{er}$. *Consider its stabilizer $\mathrm{St}(J) \cong H \subset G$. Then $\mathrm{St}(J) \cap \Gamma$ is a lattice in $\mathrm{St}(J)$.*

Corollary 5.16. *Let J be a non-ergodic complex structure on a hyperkähler manifold, and $W \subset H^2(M, \mathbb{R})$ be a plane generated by $\operatorname{Re}\Omega, \operatorname{Im}\Omega$. Then W is rational. Equivalently, this means that $Pic(M)$ has maximal possible dimension.*

Similar results are true for a torus of dimension > 1; we refer the reader to [47] for precise statements and details of the proof.

6. Applications of ergodicity

6.1. Ergodic complex structures, Gromov-Hausdorff closures, and semicontinuity. The ergodicity theorem (Subsection 5.2) has some striking and even paradoxical implications. For instance, consider a Kähler cone Kah of a hyperkähler manifold (or a torus of dimension > 1) equipped with an ergodic complex structure. By Calabi-Yau theorem, each point of Kah corresponds to a Ricci-flat metric on M. If we restrict ourselves to those metrics which satisfy $\mathsf{diam}(M, g) \leqslant d$ (with bounded diameter), then, by Gromov's compactness theorem ([17]), the set X_d of such metrics is precompact in the Gromov's space of all metric spaces, equipped with the Gromov-Hausdorff metric. It is instructive to see what kind of metric spaces occur on its boundary (that is, on $\bar{X}_d \backslash X_d$). To see this, let ν_i be a sequence of diffeomorphisms satisfying $\lim_i \nu_i(I) = I'$. By Kodaira stability theorem, the Kähler cone of (M, I) is lower continuous on I. Therefore, there exists a family of Kähler classes ω_i on $(M, \nu_i(I))$ which converge to a given Kähler class ω' on (M, I'). This implies convergence of the corresponding Ricci-flat metrics. We obtain that any Ricci-flat metric on (M, I') (for any I' in the same deformation class as I) is obtained as a limit of Ricci-flat metrics on (M, I).

This gives the following truly bizzarre theorem.

Theorem 6.1. *Let (M, I) be an ergodic complex structure on a hyperkähler manifold, $X \cong \mathrm{Kah}$ the set of all Ricci-flat Kähler metrics on (M, I), and g' another Ricci-flat metric on M in the same deformation class. Then g' lies in the closure of X with respect to the Gromov topology on the space of all metrics.*

This result is very strange, because Kah is a smooth manifold of dimension $b_2(M) - 2$. By Theorem 4.13, the space Teich_H of all hyperkähler metrics is a smooth manifold of dimension $\frac{b_2(b_2-1)(b_2-2)}{6}$, clearly much bigger than $\dim \mathrm{Kah}$. Obviously, the boundary of X is highly irregular and chaotic.

For another application, consider some numerical quantity $\mu(I)$ associated with an equivalence class of complex structures. Suppose that μ is continuous or semi-continuous on Teich. Then μ is constant on ergodic complex structures. To see this, suppose that μ is upper semicontinuous, giving

$$\mu(\lim_k I_k) \geqslant \lim_k (\mu(I_k)). \tag{6.1}$$

Given an ergodic complex structire I, find a sequence $I_k = \nu_k(I)$ converging to a complex structure I'. Then (6.1) gives $\mu(I) \leqslant \mu(I')$. This implies that any ergodic complex structure satisfies $\mu(I) = \inf_{I' \in \mathrm{Teich}} \mu(I')$.

This observation can be applied to Kobayashi pseudometric and Kobayashi hyperbolicity.

6.2. Kobayashi non-hyperbolicity of hyperkähler manifolds.

Definition 6.2. Pseudometric on M is a function $d: M \times M \longrightarrow \mathbb{R}^{\geqslant 0}$ which is symmetric: $d(x,y) = d(y,x)$ and satisfies the triangle inequality $d(x,y) + d(y,z) \geqslant d(x,z)$.

Remark 6.3. Let $\mathfrak{D}$ be a set of pseudometrics. Then $d_{\mathsf{max}}(x,y) := \sup_{d \in \mathfrak{D}} d(x,y)$ is also a pseudometric.

Definition 6.4. The Kobayashi pseudometric on a complex manifold M is d_{max} for the set $\mathfrak{D}$ of all pseudometrics such that any holomorphic map from the Poincaré disk to M is distance-non-increasing.

In other words, a Kobayashi pseudo-distance between two points x, y is an infimum of distance from x to y in Poincare metric for any sequence of holomorphic disks connecting x to y.

The following observation is not difficult to see.

Claim 6.5. *Let $\pi: \mathcal{M} \longrightarrow X$ be a smooth holomorphic family, which is trivialized as a smooth manifold: $\mathcal{M} = M \times X$, and d_x the Kobayashi metric on $\pi^{-1}(x)$. Then $d_x(m, m')$ is upper continuous on x.*

Corollary 6.6. *Denote the diameter of the Kobayashi pseudometric by*

$$\mathsf{diam}(d_x) := \sup_{m,m'} d_x(m, m').$$

Then the Kobayashi diameter of a fiber of π is an upper continuous function: $\mathsf{diam}: X \longrightarrow \mathbb{R}^{\geqslant 0}$.

For a projective K3 surface, the Kobayashi pseudometric vanishes ([49, Lemma 1.51]). However, all non-projective K3 surfaces are ergodic (Corollary 5.16). This proves the vanishing of Kobayashi pseudodistance for all K3 surfaces. A more general version of this result is due to due to Kamenova-Lu-Verbitsky.

Theorem 6.7 ([21]). *Let M be a Hilbert scheme of K3. Then the Kobayashi pseudometric on M vanishes.*

Definition 6.8. A complex manifold is called **Kobayashi hyperbolic** if the Kobayashi pseudometric is a metric.

Definition 6.9. An entire curve is a non-constant map $\mathbb{C} \longrightarrow M$.

Brody has shown that a compact manifold is Kobayashi hyperbolic if and only if it admits no entire curves. The same argument also proves semicontinuity.

Theorem 6.10 ([8]). *Let I_i be a sequence of complex structures on M which are not hyperbolic, and I its limit. Then (M, I) is also not hyperbolic.*

With ergodicity, this can be used to prove that all hyperkähler manifolds are non-hyperbolic. Recall that **a twistor family** of complex structures on a hyperkähler manifold (M, I, J, K) is a family of complex structures of form $S^2 \cong \{L := aI + bJ + cK, \quad a^2 + b^2 + c^2 = 1\}$. F. Campana has obtained a remarkable partial result towards non-hyperbolicity.

Theorem 6.11 ([10]). *Let M be a hyperkähler manifold, and $S \subset$ Teich a twistor family. Then there exists an entire curve in some $I \in S$.*

Claim 6.12. *There exists a twistor family which has only ergodic fibers.*

Proof. There are only countably many complex structures which are not ergodic; however, twistor curves move freely through the Teichmüller space of a hyperkähler manifold, as seen from Theorem 4.19. □

Applying Campana's theorem to the family constructed in Claim 6.12, we obtain an ergodic complex structure which is non-hyperbolic. Then the Brody's theorem implies that all complex structures in the same deformation class are non-hyperbolic.

Theorem 6.13. *All hyperkähler manifolds are non-hyperbolic.*

6.3. Symplectic packing and ergodicity. I will finish this talk with a list of open problems of hyperkähler and holomorphically symplectic geometry which might be solvable with ergodic methods.

Question 6.14. *Let M be a hyperkähler manifold, and* Teich *its Teichmüller space. Consider the universal fibration $\mathcal{X} \longrightarrow$ Teich (Remark 1.11). The mapping class group Γ acts on $\mathcal{X}$ in a natural way. Is this action ergodic?*

This question (suggested by Claire Voisin) seems to be related to the following conjecture.

Conjecture 6.15. *Let M be a K3 surface. Then for each $x \in M$ and $v \in T_xM$ there exists an entire curve $C \ni x$ with $T_xC \ni v$.*

The symplectic packing problem is a classical subject of symplectic geometry ([28]). However, its holomorphically symplectic version seems to be completely unexplored.

Definition 6.16. A holomorphic symplectic ball B_r of radius r is a complex holomorphically symplectic manifold admitting a holomorphic symplectomorphism to an open ball in $\mathbb{C}^{2n}$ of radius r with the standard holomorphic symplectic form $\sum_{i=1}^{n} dz_{2i-1} \wedge dz_{2i}$.

Notice that by a holomorphic symplectic version of Darboux theorem, any holomorphically symplectic manifold is locally symplectomorphic to a holomorphic symplectic ball.

Definition 6.17. Let M be a holomorphically symplectic manifold. **Symplectic packing** of radii $r_1, ..., r_k$ of M is a set of holomorphic symplectomorphisms $\varphi_i : \ B_{r_i} \hookrightarrow M$ with images of φ_i not intersecting.

Obviously, in these assumptions, $\sum \mathrm{Vol}(B_{r_i}) \leqslant \mathrm{Vol}_M$, where Vol denotes the symplectic volume of a holomorphic symplectic manifold (M, Ω_M):

$$\mathrm{Vol}(M) = \int_M (\Omega_M \wedge \bar{\Omega}_M), \ \ 2n = \dim_{\mathbb{C}} M.$$

The volume inequality puts certain restrictions on the possible symplectic packing. Are there any other restrictions?

For the usual (smooth) symplectic packing, some additional restrictions are obtained from the Gromov's symplectic capacity theorem and from the study of pseudoholomorphic

curves. However, it seems that in holomorphic symplectic situation these restrictions are also trivial. For a general compact torus of real dimension 4, volume is known to be the only restriction to existence of symplectic packing ([26]). It seems that a similar result about the smooth symplectic packings is true for K3 surfaces as well, and, possibly, for any hyperkähler manifold.

The arguments used to treat the usual (smooth) symplectic packings don't work for the holomorphic symplectic case. However, the set of possible radii for symplectic packing is obviously semicontinous, hence it can be studied by ergodic methods, in the same way as one studies the Kobayashi pseudometric.

The following classical question was treated Buzzard and Lu in [9].

Definition 6.18. A complex manifold M of dimension n is called **dominated by** $\mathbb{C}^n$ if there exists a holomorphic map $\varphi:\ \mathbb{C}^n \longrightarrow M$ which has non-degenerate differential in generic point.

Buzzard and Lu proved that Kummer K3 surfaces are dominated by $\mathbb{C}^2$. So far, there is not a single example of a hyperkähler manifold M for which it is proven that M is not dominated. This leads to the following conjecture

Conjecture 6.19. *Any compact hyperkähler manifold is dominated by* $\mathbb{C}^n$.

There is no semicontinuity in dominance, because Brody lemma fails to produce dominating maps $\mathbb{C}^n \longrightarrow M$ for $n > 1$ as limits of sequences of dominating maps. In the proof of Brody's lemma (showing that a limit of a sequence of entire curves contains an entire curve) one takes a reparametrizations of each of the curves in the sequence. Starting from a sequence of dominating maps, one could apply the same argument, but each subsequent reparametrization can lead to smaller Jacobian of the differential, and the differential of the limit could be zero.

It seems that more of the Brody's argument can be retained if we restrict ourselves to symplectomorphisms.

Question 6.20. *Consider a flat holomorphically symplectic structure on* $\mathbb{C}^2$. *Is there a holomorphic map* $\mathbb{C}^2 \longrightarrow M$ *to a K3 surface which is compatible with the holomorhic symplectic form?*

Probably not. However, a quantitative version of this question makes sense. Let M be a hyperkähler manifold, and $K(M)$ the supremum of all r such that there exists a symplectic immersion from a symplectic ball of radius r to M. It is not hard to see that $K(M)$ is semicontinuous in families, hence constant on ergodic complex structures.

Question 6.21. *For a given hyperkähler manifold, find* $K(M)$.

It is not clear if $K(M)$ is finite or infinite, even for a K3 surface (it is clearly infinite for a torus).

Acknowledgements. The author is partially supported by RFBR grants 12-01-00944-Ðř and AG Laboratory NRU-HSE, RF government grant, ag. 11.G34.31.0023. Many thanks to Laurent Meersseman for his comments to an early version of this manuscript.

References

[1] Ekaterina Amerik and Misha Verbitsky, *Rational curves on hyperkahler manifolds*, arXiv:1401.0479, 34 pages.

[2] Beauville, A., *Varietes Kähleriennes dont la première classe de Chern est nulle*, J. Diff. Geom. **18** (1983), 755–782.

[3] Beauville, Arnaud, *Le theoreme de Torelli pour les surfaces $K3$: fin de la demonstration*, Geometry of $K3$ surfaces: moduli and periods (Palaiseau, 1981/1982). Asterisque No. 126 (1985), 111–121.

[4] Besse, A., *Einstein Manifolds*, Springer-Verlag, New York, 1987.

[5] Bogomolov, F. A., *On the decomposition of Kähler manifolds with trivial canonical class*, Math. USSR-Sb. **22** (1974), 580–583.

[6] _____, *Hamiltonian Kählerian manifolds*, Dokl. Akad. Nauk SSSR 243 (1978), no. 5, 1101–1104.

[7] _____, *Kähler manifolds with trivial canonical class,* Preprint, Institut des Hautes Etudes Scientifiques (1981), 1–32.

[8] Brody, R., *Compact manifolds and hyperbolicity*, Trans. Amer. Math. Soc. **235** (1978), 213–219.

[9] Buzzard, G. and Lu, S. S.-Y., *Algebraic surfaces holomorphically dominable by* $\mathbb{C}^2$, to appear in Invent. Math.

[10] F. Campana, *An application of twistor theory to the nonhyperbolicity of certain compact symplectic Kähler manifolds*, J. Reine Angew. Math., 425:1–7, 1992.

[11] F. Catanese, *A Superficial Working Guide to Deformations and Moduli*, arXiv:1106.1368, 56 pages.

[12] Debarre, O., *Un contre-exemple au théorème de Torelli pour les variétés symplectiques irréductibles*, C. R. Acad. Sci. Paris Sér. I Math. **299** (1984), no. 14, 681–684.

[13] Demailly, J.-P. and Paun, M., *Numerical characterization of the Kähler cone of a compact Kähler manifold*, Annals of Mathematics, **159** (2004), 1247–1274, math.AG/0105176.

[14] Douady, A., *Le probleme des modules pour les varietes analytiques complexes*, Seminaire Bourbaki, 1964/1965, No 277.

[15] Friedman, Robert, *A new proof of the global Torelli theorem for $K3$ surfaces*, Ann. of Math. (2) **120** (1984), no. 2, 237–269.

[16] Fujiki, A., *On the de Rham Cohomology Group of a Compact Kähler Symplectic Manifold*, Adv. Stud. Pure Math. **10** (1987), 105–165.

[17] Gromov, Misha, *Metric structures for Riemannian and non-Riemannian spaces*, Based on the 1981 French original. With appendices by M. Katz, P. Pansu and S. Semmes. Translated from the French by Sean Michael Bates. Progress in Mathematics, 152. Birkhäuser Boston, Inc., Boston, MA, 1999. xx+585 pp.

[18] Hamilton, Richard S., *The inverse function theorem of Nash and Moser*, Bull. Amer. Math. Soc. (N.S.) **7** (1982), no. 1, 65–222.

[19] Huybrechts, D., *The Kähler cone of a compact hyperkähler manifold*, Math. Ann. **326** (2003), no. 3, 499–513, arXiv:math/9909109.

[20] _____, *Finiteness results for hyperkähler manifolds*, J. Reine Angew. Math. **558** (2003), 15–22, arXiv:math/0109024.

[21] Ljudmila Kamenova, Steven Lu, and Misha Verbitsky, *Kobayashi pseudometric on hyperkahler manifolds*, arXiv:1308.5667, 21 pages.

[22] Kleinbock, Dmitry, Shah, Nimish, and Starkov, Alexander, *Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory*, Handbook of dynamical systems, Vol. 1A, 813-930, North-Holland, Amsterdam, 2002.

[23] Kulikov, Vik. S., *Degenerations of K3 surfaces and Enriques surfaces*, Mathematics of the USSR-Izvestiya, 11:5 (1977), 957–989.

[24] Kuranishi, Masatake, *New proof for the existence of locally complete families of complex structures*, 1965 Proc. Conf. Complex Analysis (Minneapolis, 1964) pp. 142-154 Springer, Berlin.

[25] _____, *A note on families of complex structures* 1969 Global Analysis (Papers in Honor of K. Kodaira) pp. 309–313 Univ. Tokyo Press, Tokyo.

[26] Janko Latschev, Dusa McDuff, and Felix Schlenk, *The Gromov width of 4-dimensional tori*, arXiv:1111.6566.

[27] E. Loojenga, *A Torelli theorem for Kähler-Einstein K3 surfaces*, Lecture Notes in Mathematics, 1981, Volume 894/1981, 107–112.

[28] D. McDuff and L. Polterovich, *Symplectic packings and algebraic geometry,* Invent. Math. **115** (1994), 405–434.

[29] Markman, E., *On the monodromy of moduli spaces of sheaves on $K3$ surfaces*, J. Algebraic Geom. **17** (2008), no. 1, 29–99, arXiv:math/0305042.

[30] _____, *Integral constraints on the monodromy group of the hyperkahler resolution of a symmetric product of a K3 surface,* International Journal of Mathematics Vol. 21, No. 2 (2010) 169–223, arXiv:math/0601304.

[31] Calvin C. Moore, *Ergodicity of Flows on Homogeneous Spaces*, American Journal of Mathematics Vol. 88, No. 1 (Jan., 1966), pp. 154–178

[32] Morris, Dave Witte, *Ratner's Theorems on Unipotent Flows,* Chicago Lectures in Mathematics, University of Chicago Press, 2005.

[33] Namikawa, Y., *Counter-example to global Torelli problem for irreducible symplectic manifolds*, Math. Ann. **324** (2002), no. 4, 841–845.

[34] O'Grady, Kieran G., *A new six-dimensional irreducible symplectic variety*, J. Algebraic Geom. **12** (2003), no. 3, 435–505.

[35] Piatecki-Shapiro, I.I. and Shafarevich I.R., *Torelli's theorem for algebraic surfaces of type K3*, Izv. Akad. Nauk SSSR Ser. Mat. **35** (1971), 530–572.

[36] N. A. Shah, *Uniformly distributed orbits of certain flows on homogeneous spaces*, Math. Ann. **289** (2) (1991), 315–33.

[37] Sullivan, D., *Infinitesimal computations in topology*, Publications Mathématiques de l'IHÉS, 47 (1977), p. 269–331

[38] Teichmüller, Oswald, *Veränderliche Riemannsche Flächen*, Abh. Preuss. Deutsche Mathematik **7** (1944), 344–359.

[39] G. Tian, *Smoothness of the universal deformation space of compact Calabi-Yau manifolds and its Petersson-Weil metric*, in Math. Aspects of String Theory, S.-T. Yau, ed., Worlds Scientific, 1987, 629–646.

[40] Tjurina, G. N., *The space of moduli of a complex surface with q = 0 and K = 0*, In: "Algebraic Surfaces", Seminar Shafarevich, Proc. Steklov Inst. **75** (1965).

[41] Todorov, A. N., *Applications of the Kähler-Einstein-Calabi-Yau metric to moduli of K3 surfaces*, Inventiones Math. 6-1 (1980), 251–265.

[42] ———, *The Weil-Petersson geometry of the moduli space of* $SU(n \geqslant 3)$ *(Calabi-Yau) manifolds*, Comm., Math. Phys. **126** (1989), 325–346.

[43] Ruggiero Torelli, *Sulle varietà di Jacobi*, Rend. della R. Acc. Nazionale dei Lincei, (5) **22** (1913), 98–103.

[44] Siu, Y.-T., *Every K3 surface is Kähler*, Invent. Math. **73** (1983), 139–150.

[45] Verbitsky, M., *Cohomology of compact hyperkähler manifolds and its applications*, GAFA vol. 6 (4) (1996), 601–612.

[46] ———, *A global Torelli theorem for hyperkähler manifolds*, Duke Math. J. Volume 162, Number 15 (2013), 2929–2986.

[47] ———, *Ergodic complex structures on hyperkahler manifolds*, arXiv:1306.1498, 22 pages.

[48] Vinberg, E. B., Gorbatsevich, V. V., Shvartsman, O. V., *Discrete Subgroups of Lie Groups*, in "Lie Groups and Lie Algebras II", Springer-Verlag, 2000.

[49] Claire Voisin, *On some problems of Kobayashi and Lang; algebraic approaches*, Current Developments in Mathematics 2003, no. 1 (2003), 53–125.

[50] A. Weil., *Zum Beweis des Torellischen Satzes*, Nachr. Akad. Wiss. Göttingen, Math.-Phys. Kl. IIa: 32–53 (1957).

7 Vavilova Str., Moscow, Russia, 117312

E-mail: verbit@verbit.ru

5. Geometry

Family Floer cohomology and mirror symmetry

Mohammed Abouzaid

Abstract. Ideas of Fukaya and Kontsevich-Soibelman suggest that one can use Strominger-Yau-Zaslow's geometric approach to mirror symmetry as a torus duality to construct the mirror of a symplectic manifold equipped with a Lagrangian torus fibration as a moduli space of those simple objects of the Fukaya category which are supported on the fibres. In the absence of singular fibres, the construction of the mirror is explained in this framework, and, given a Lagrangian submanifold, a (twisted) coherent sheaf on the mirror is constructed.

Mathematics Subject Classification (2010). Primary 53D40; Secondary 14J33.

Keywords. Lagrangian Floer cohomology, homological mirror symmetry.

1. Introduction

1.1. Overview. Mirror symmetry is a prediction from string theory identifying invariants associated to the complex geometry of a family of Calabi-Yau manifolds with invariants associated to the Kähler geometry of a possibly different Calabi-Yau manifold that is called the mirror. In our setting, we take as definition of a Calabi-Yau manifold a complex manifold equipped with a nowhere vanishing holomorphic volume form. The original focus in mathematics was on the dualities of Hodge diamonds which gave a straightforward though non-trivial check, and on the enumerative predictions for the number of curves of a given genus and degree that went beyond the computations which could be performed using rigourous methods. While it is not reasonable to expect the existence of a mirror partner to every Calabi-Yau manifold, there are large classes of examples (e.g. toric complete intersections [6, 17]) for which various original forms of the conjecture have been verified.

In [22], Kontsevich introduced a homological version of the conjecture: the invariants to be related would be the derived category of coherent sheaves on the complex side and the Fukaya category of Lagrangian submanifolds on the symplectic side. Strominger, Yau, and Zaslow [28] later introduced a geometric version of the conjecture: mirror pairs should arise as dual torus fibrations over the same base; these are often called SYZ fibrations. The degenerating family can then be understood as arising from rescaling the fibres.

It is easier to state precise versions of the SYZ conjecture (which still hold in a large class of examples) if one analyses the two sides of mirror symmetry separately, i.e. fixing a Kähler form on a Calabi-Yau manifold X whose symplectic topology will be related to the complex geometry of a Calabi-Yau variety Y over the ring of power series $\mathbb{C}[[T]]$ or the analogous rings appearing naturally in symplectic topology in which real exponents are

Proceedings of International Congress of Mathematicians, 2014, Seoul

allowed; weakening one side to a formal family is related to a convergence problem in Floer theory.

In this context, the starting point of a reformulation of the SYZ conjecture is the existence of a Lagrangian fibration $\pi\colon X \to Q$ with singularities (there is still no good approach to the class of allowable singularities). The space Y should then be constructed from the Fukaya category $\mathrm{Fuk}(X)$ as a moduli space of objects supported on fibres of this map [11]. While the moduli space of such objects can be described locally over the base as the dual torus fibration, the fact that we consider them as objects of the Fukaya category introduces non-trivial identifications of the local charts given by the sum of a classical term with *instanton corrections* that arise from the moduli space of holomorphic discs bounded by such fibres. These corrections are expected to be expressible in terms of the geometry of the base via *wall-crossing formulae* [23, 19].

In this setting, the homological mirror conjecture asserts the existence of a derived equivalence $D\,\mathrm{Fuk}(X) \cong D\,\mathrm{Coh}(Y)$, where both categories are linear over the appropriate version of power series rings. Much effort has gone in verifying this conjecture in certain examples [26, 1, 27], and extracting some of the classical statements of mirror symmetry from it [8, 20].

Unfortunately, all the current proofs of mirror symmetry rely on *ad hoc* methods to construct the mirror functor, neglecting the construction of the mirror as a moduli space of objects of the Fukaya category. To address this issue, Fukaya has introduced family Floer cohomology [10, 11, 12, 13] as a strategy for directly assigning a sheaf on Y to a Lagrangian L in X; as noted in [11], the main difficulty arises from the caustics of L, i.e. the singularities of its projection to the base. In Section 4, we use the invariance properties of Floer cohomology under continuation maps to bypass the difficulties arising from caustics (as in [10, Section 6]), and prove convergence in the rigid analytic sense.

1.2. A twisted example. Since the problem is sufficiently complicated in the absence of singular fibres, we shall henceforth only consider symplectic manifolds that admit smooth Lagrangian torus fibrations. This class includes, for example, a codimension $\frac{n(n-1)}{2}$ subspace of the $n(2n-1)$-dimensional space of linear symplectic structures on $\mathbb{R}^{2n}/\mathbb{Z}^{2n}$, corresponding to those structures that can be represented as the quotient of $\mathbb{R}^{2n}$ by a lattice which intersects some Lagrangian plane in a rank n subgroup. It also includes the following twisted example due to Thurston [30]:

Equip $\mathbb{R}^4$ with coordinates (x_1, x_2, x_3, x_4), and symplectic structure

$$dx_1 \wedge dx_2 + dx_3 \wedge dx_4. \tag{1.1}$$

This form is invariant under the transformation

$$(x_1, x_2, x_3, x_4) \to (x_1 + 1, x_2, x_3, x_4 + x_3) \tag{1.2}$$

as well as under translation by integral vectors of the form $(0, x_2, x_3, x_4)$. The Thurston manifold is the quotient of $\mathbb{R}^4$ by the group generated by these transformations.

Thurston considered the symplectic fibration obtained by forgetting the (x_3, x_4) coordinates, which gives a description of this space as a twisted (flat) torus bundle over the torus. In joint work with Auroux, Katzarkov, and Orlov [2] we noticed the existence of two (inequivalent) Lagrangian fibrations on this space:

(1) The fibration obtained by forgetting the x_2 and x_4 coordinates is a principal bundle on

which the (x_2, x_4)-torus acts. Since the total space is not trivial, this fibration admits no continuous section, and hence, in particular, no Lagrangian section.

(2) The fibration obtained by forgetting the x_1 and x_4 coordinates admits a Lagrangian section $(0, x_2, x_3, 0)$.

The above two examples show that one can have Lagrangian fibrations with completely different behaviour on the same symplectic manifold. The first fibration is mirror to a gerbe on an abelian variety, while the second is mirror to a Kodaira surface. To see this, write the first fibration as the quotient of $[0,1]^2 \times (\mathbb{R}/\mathbb{Z})^2$ by the equivalence relation

$$(0, x_3, x_2, x_4) \sim (1, x_3, x_2, x_4 + x_3) \tag{1.3}$$

$$(x_1, 0, x_2, x_4) \sim (x_1, 1, x_2, x_4). \tag{1.4}$$

Note that the gluing maps act trivially on the homology of the fibre, so the action on the space of local systems is trivial. Since the fibres bound no holomorphic disc, the mirror is the moduli space of such local systems; it agrees with the mirror of the symplectic manifold obtained via the trivial identifications. This symplectic manifold is simply the product of two tori of area 2π, hence the mirror is a product of two (families of) elliptic curves as is well-understood by now [1].

At this stage, it is clear that additional data are needed to implement mirror symmetry from this point of view. At the most basic level, the Lagrangian fibres are null-homologous, which implies that Floer cohomology has vanishing Euler characteristic whenever one of the two inputs is a fibre. Under mirror symmetry the fibres map to skyscraper sheaves of point, and there are therefore many coherent sheaves (e.g. vector bundles of non-vanishing rank) whose mirrors would be expected to have Floer cohomology groups with a fibre whose Euler characteristic does not vanish.

The additional data arise naturally on both sides: by obstruction theory, the failure of this torus fibration to be trivial is detected by a second cohomology group of the base as one can easily construct a Lagrangian section in the complement of a point. This obstruction class is constructed in Section 2.1 using a Čech cover, and a simple exponentiation procedure in Section 2.4 then yields an $\mathcal{O}^*$-valued second cohomology class on the mirror space; i.e. a gerbe. The correct statement of mirror symmetry involves sheaves twisted by this gerbe.

For completeness, we elaborate on the description of the mirror of the second fibration: a convenient starting point is the vector bundle $(\mathbb{R}/\mathbb{Z})^2 \times \mathbb{R}^2$ over the torus in the (x_2, x_3) coordinates whose fibre is the plane spanned by (x_1, x_4). The key observation is that the Thurston manifold is obtained by taking the quotient of the fibre over (x_2, x_3) by the lattice spanned by $(1, x_3)$ and $(0, 1)$. The corresponding family of lattices in $\mathbb{R}^2$ has non-trivial monodromy (given by an elementary transvection) around a loop in the x_3 direction. Constructing the mirror (complex) manifold by dualising the fibres, we conclude that the underlying smooth manifold is also a torus fibration over the torus with total Betti number 3, hence a primary Kodaira surface [21, p. 787-788]. With a bit more care, one can avoid appealing to the classification of surfaces and identify the (complex) mirror as the degenerating family of quotients of $\mathbb{C}^* \times \mathbb{C}^*$ by the groups of automorphisms

$$(z_1, z_2) \mapsto (z_1, T \cdot z_2) \tag{1.5}$$

$$(z_1, z_2) \mapsto (T \cdot z_1, z_1 z_2) \tag{1.6}$$

parametrised by a variable T. To obtain a precise statement of Homological mirror symmetry

in this setting, one then interprets the above as giving a rigid analytic primary Kodaira surface [33, p. 788].

One interesting outcome of this observation is that the two homological mirror symmetry statements imply the existence of a derived equivalence between twisted coherent sheaves on the mirror abelian variety and coherent sheaves on the mirror primary Kodaira surface. Such a result can be proved independently of mirror symmetry by exhibiting a Fourier-Mukai kernel [2].

1.3. Statement of the results. Let (X, ω) be a compact symplectic manifold, equipped with a fibration $\pi\colon X \to Q$ over a smooth manifold Q with Lagrangian fibres; the triple (X, ω, π) is called a *Lagrangian fibration*; write F_q for the fibre at a point q, and assume that $\pi_2(Q) = 0$.

Remark 1.1. The vanishing of the second homotopy group of Q implies that $\pi_2(X, F_q)$ vanishes, and hence that F_q bounds no holomorphic disc, which implies that there are no instanton corrections, i.e. that the mirror should be the space of rank-1 unitary local systems on the fibres. There seem to be no known examples where this condition fails (though this can be arranged if X is not assumed to be compact, or if singular fibres are allowed).

The symplectic topology of such fibrations is reviewed in Section 2.1, but for now, recall that the tangent space of Q at a point q is naturally isomorphic to $H^1(F_q, \mathbb{R}) \equiv H^1(F_q; \mathbb{Z}) \otimes \mathbb{R}$. Arnol'd's Liouville theorem implies that the corresponding lattice in TQ arises from an integral affine structure on Q. In particular, Q can be obtained by gluing polytopes in $\mathbb{R}^n$ by transformations whose differentials lie in $SL(n, \mathbb{Z})$.

To this integral affine structure, one associates a rigid analytic space (in the sense of Tate), which we denote Y (this is the same construction used in [23, 15, 32]): fix a field $\mathbf{k}$, and consider the Novikov field

$$\Lambda = \Big\{ \sum_{i=1}^{\infty} a_i t^{\lambda_i} | a_i \in \mathbf{k},\ \lambda_i \in \mathbb{R},\ \lambda_i \to +\infty \Big\}. \tag{1.7}$$

This is a non-archimedian field with valuation $\mathrm{val}\colon \Lambda - \{0\} \to \mathbb{R}$

$$\sum_{i=1}^{\infty} a_i t^{\lambda_i} \mapsto \min(\lambda_i | a_i \neq 0). \tag{1.8}$$

Denote by U_Λ the units of Λ, i.e. those elements with 0-valuation.

As a set, the space Y is the union $\coprod_{q \in Q} H^1(F_q; U_\Lambda)$. This description makes explicit the fact that Y parametrises Lagrangian fibres together with the datum of a *rank*-1 Λ-*local system* with monodromy in U_Λ. The analytic structure on Y arises from the natural identifications of the first cohomology groups of nearby fibres; the explicit construction appears in Section 2.2.

Given a Lagrangian $L \subset X$, satisfying the technical conditions required to make Floer cohomology well-defined, one obtains Floer cohomology groups $HF^k((F_q, b), L; \Lambda)$ for each $q \in Q$, and $b \in H^1(F_q; U_\Lambda)$.

These groups are locally the fibres of coherent sheaves on Y, as can be seen by adapting an argument of Fukaya [12] who studied the case of self-Floer cohomology. In order to encode the full data of the global Floer theory of X on the Y-side, an *analytic gerbe* α_X on Y is introduced in Section 2.4.

Theorem 1.2. *There is an α_X-twisted coherent sheaf on Y whose dual fibre at the point of Y corresponding to a pair (q, b) is $HF^k((F_q, b), L; \Lambda)$.*

Remark 1.3. One can interpret this theorem as an attempt to make rigourous the strategy introduced in Fukaya's announcement [10], which assigns to a Lagrangian a complex analytic sheaf *assuming convergence*. In the special case of Lagrangian surfaces constructed by Hyperkähler rotation, analytic continuation may provide an alternative approach for bypassing the problem arising from caustics, as noted in [11]. As part of an ongoing project to study mirror symmetry from the point of view of family Floer cohomology [31, 32], Junwu Tu has an independent argument to prove a similar result.

1.4. Difficulties lying ahead. While the above result points in the right direction, it is unfortunately not adequate for serious applications. The last section of this paper outlines the construction of an object $\mathcal{L}$ in a category of *α_X-twisted sheaves of perfect complexes* that is a differential graded enhancement of the derived category of α_X-twisted coherent sheaves on Y. The twisted sheaf in Theorem 1.2 is obtained from $\mathcal{L}$ by passing to cohomology.

The following conjecture makes clear why $\mathcal{L}$, rather than its cohomology sheaves, is the right object to study:

Conjecture 1.4. *If L_1 and L_2 are Lagrangians in X with corresponding twisted sheaves of perfect complexes $\mathcal{L}_1$ and $\mathcal{L}_2$, there is an isomorphism*

$$HF^*(L_1, L_2) \cong H^*(\mathrm{Hom}_*(\mathcal{L}_1, \mathcal{L}_2)). \tag{1.9}$$

There are two essential difficulties that arise in applying these techniques to the SYZ fibrations $X \to Q$ that are expected to exist, for example, on Calabi-Yau hypersurfaces in toric varieties:

(1) In general, some smooth fibres may bound non-constant holomorphic discs. Assuming such fibres are *unobstructed*, which essentially means that the counts of holomorphic discs with boundary on a single fibre algebraically vanish, Tu constructed the candidate (open subset) of the mirror in [31]. The basic idea, following Fukaya [11] and Kontsevich and Soibelman [23], is that we should obtain the mirror space by gluing affinoids as in the uncorrected case, but the gluing maps take into account moduli spaces of holomorphic discs. One can interpret part of the program of Gross and Siebert [19] as providing such a construction for fibrations arising from toric degenerations, though it is not yet known how to prove that their construction agrees with the one intrinsic to symplectic topology. Given the appropriate technical tools (i.e. a theory of virtual fundamental chains on moduli spaces of holomorphic discs, as in [14]), the extension of our results to this setting should be straightforward.

(2) The construction of the mirror space from an SYZ perspective also requires understanding the Floer theory of singular fibres. Whenever the fibre is *immersed*, this Floer theory is well-understood [4]. In the simplest situation, such a fibre is an immersed Lagrangian 2-sphere with a single double point in a 4-manifold, and the nearby Lagrangian (torus) fibres are obtained by Lagrangian surgery [24]. Using the relation between moduli spaces of holomorphic discs before and after surgery [14], Fukaya [13] has announced the construction of the mirror space in this setting. In particular, for a symplectic structure on a $K3$-surface admitting a Lagrangian torus fibration,

Fukaya's method provides a construction of the mirror space using family Floer cohomology. Unfortunately, in higher dimensions, SYZ fibrations are expected to have singular fibres which cannot be described as Lagrangian immersions, putting them beyond the reach of current techniques.

2. Background

2.1. Lagrangian torus fibrations. Let Q be the base of a Lagrangian fibration as in Section 1.3. The Arnol'd-Liouville theorem implies that we have canonical identifications

$$T_q Q = H^1(F_q; \mathbb{R}) \text{ and } T_q^* Q = H_1(F_q; \mathbb{R}). \tag{2.1}$$

Write $T_q^{\mathbb{Z}} Q$ for the image of $H^1(F_q; \mathbb{Z})$ in $T_q Q$ under the first isomorphism above, $T^{\mathbb{Z}} Q$ for the corresponding rank-n local system on Q, and $T_{\mathbb{Z}}^* Q$ for the dual. The key property satisfied by this sublattice of $T^* Q$ is that its (local) flat sections are closed, hence exact. On a subset P of Q, a choice of n functions whose differentials span $T_{\mathbb{Z}}^* P$ at every point defines an immersion into $\mathbb{R}^n$ mapping the fibres of $T^{\mathbb{Z}} P$ to the standard lattice $\mathbb{Z}^n$ in the tangent space of $\mathbb{R}^n$. By choosing P sufficiently small, we obtain a coordinate patch for the *integral affine structure* on Q induced by $T^{\mathbb{Z}} Q$.

The inverse image X_P of such a subset under π is fibrewise symplectomorphic to $T^* P / T_{\mathbb{Z}}^* P$. Given sets $P, P' \subset Q$ which intersect, the restrictions of the two symplectomorphisms induce a fibrewise symplectomorphism

$$T^*\left(P \cap P'\right) / T_{\mathbb{Z}}^*\left(P \cap P'\right) \to X_{P \cap P'} \to T^*\left(P \cap P'\right) / T_{\mathbb{Z}}^*\left(P \cap P'\right). \tag{2.2}$$

Such a symplectomorphism can be written as fibrewise addition by a closed 1-form which is uniquely defined up to an element of $T_{\mathbb{Z}}^*(P \cap P')$.

In order to classify symplectic fibrations which induce a given integral affine structure, choose a finite partially ordered set $\mathcal{A}$ labelling a simplicial triangulation of Q, i.e. there are vertices labelled by elements of $\mathcal{A}$, and every cell is the span of a unique ordered subset I of $\mathcal{A}$. Assume that this triangulation is sufficiently fine that there is a cover of Q by codimension 0 submanifolds with boundary $\{P_i\}_{i \in \mathcal{A}}$ so that P_i contains the open star of the vertex i and all iterated intersections are contractible. Let

$$P_I = \bigcap_{i \in I} P_i \tag{2.3}$$

and note that P_I contains the open star of the cell corresponding to I.

Choose a trivialisation τ_i of the inverse image X_{P_i}, i.e. a fibrewise symplectic identification with $T^* P_i / T_{\mathbb{Z}}^* P_i$. If $i < j$, the restrictions of τ_i and τ_j to $X_{P_{ij}}$ differ by fibrewise addition of a closed 1-form; choose a primitive for this 1-form

$$f_{ij} \colon P_{ij} \to \mathbb{R}. \tag{2.4}$$

If $i < j < k$, the cyclic sum

$$\alpha_X(ijk) = f_{ij} + f_{jk} - f_{ik} \tag{2.5}$$

is function on P_{ijk} whose differential lies in $T_{\mathbb{Z}}^* P_{ijk}$; i.e. an integral affine function. Such functions define a sheaf on Q that will be denoted Aff, and α_X yields a cocycle in $\check{C}^2(Q; \mathrm{Aff})$.

Write $[\alpha_X]$ for the corresponding cohomology class. Here, the Čech complex associated to a sheaf $\mathcal{F}$ on Q is given by

$$\check{C}^*_{\mathcal{A}}(Q;\mathcal{F}) = \bigoplus_{\substack{I=(i_0,\dots,i_r)\subset\mathcal{A}\\ i_0<\dots<i_r}} \mathcal{F}(P_I)[-r] \tag{2.6}$$

with differential given by restriction.

The importance of the sheaf Aff was noted by Gross and Siebert [18], who related it to classical invariants of affine structures. In that spirit, the following is a reformulation of a result of Duistermaat [9, Equation (2.6)]:

Proposition 2.1 (c.f. p. 476 of [5]). *X is determined up to fibrewise symplectomorphism by the triple $(Q, T^*_{\mathbb{Z}}Q, [\alpha_X])$.*

Proof. The fibrewise symplectic automorphisms of X_P are given by $\Omega^1_c(P)/T^*_{\mathbb{Z}}P$, where Ω^1_c is the sheaf of closed 1-forms. As noted by Duistermaat, this implies that Lagrangian fibrations which induce the integral affine structure $T^*_{\mathbb{Z}}Q$ are classified up to fibrewise symplectomorphism by $H^1(Q, \Omega^1_c/T^*_{\mathbb{Z}}Q)$.

To obtain the desired result, note that the identification of Ω^1_c with $C^\infty/\mathbb{R}$ induces an isomorphism of sheaves

$$C^\infty/\operatorname{Aff} \cong \Omega^1_c/T^*_{\mathbb{Z}}Q. \tag{2.7}$$

Since C^∞ is a soft sheaf, this implies the existence of a canonical isomorphism

$$H^1(Q, \Omega^1_c/T^*_{\mathbb{Z}}Q) \cong H^2(Q, \operatorname{Aff}). \tag{2.8}$$

□

Remark 2.2. The differentiable type of X is determined by a Chern class with values in $H^2(Q, T^*_{\mathbb{Z}}Q)$. The short exact sequence $\mathbb{R} \to \operatorname{Aff} \to T^*_{\mathbb{Z}}Q$ induces a long exact sequence

$$\cdots \to H^2(Q,\mathbb{R}) \to H^2(Q, \operatorname{Aff}) \to H^2(Q, T^*_{\mathbb{Z}}Q) \to \cdots \tag{2.9}$$

which has the following interpretation: once a smooth torus fibration over Q which is compatible with its affine structure is fixed, the set of symplectic structures on the total space for which this fibration is Lagrangian is either empty or an affine space over $H^2(Q,\mathbb{R})$, where the action is given by adding the pullback of a 2-form on Q.

2.2. Construction of the mirror space. The next step is to associate a rigid analytic space Y to the integral affine structure $T^{\mathbb{Z}}Q$ on Q. As a set, Y is simply the flat bundle over Q

$$Y = T^{\mathbb{Z}}Q \otimes_{\mathbb{Z}} U_\Lambda \tag{2.10}$$

where U_Λ is the multiplicative subgroup of the units in the Novikov ring as in Section 1.3. Write

$$\operatorname{val}\colon Y \to Q \tag{2.11}$$

for the projection. The fibre of val at a point $q \in Q$ is $H^1(F_q; U_\Lambda)$.

To construct an analytic structure on Y, start by considering the valuation

$$H^1(F_q;\Lambda^*) \to H^1(F_q;\mathbb{R}), \tag{2.12}$$

whose fibre is $H^1(F_q; U_\Lambda)$. Splitting the above map by taking a real number λ to T^λ, yields an isomorphism

$$H^1(F_q; U_\Lambda) \times H^1(F_q; \mathbb{R}) \cong H^1(F_q; \Lambda^*). \tag{2.13}$$

If P is a sufficiently small neighbourhood of $q \in Q$, it can be identified using parallel transport with respect to $T^{\mathbb{Z}}Q$ with a neighbourhood of the origin in T_qQ. This gives rise to a natural embedding

$$Y_P \equiv \mathrm{val}^{-1}(P) = \coprod_{p \in P} H^1(F_p; U_\Lambda) \subset H^1(F_q; \Lambda^*). \tag{2.14}$$

Assume now that $P \subset H^1(F_q; \mathbb{R})$ is a polytope defined by integral affine equations, i.e. that there exist integral homology classes $\{\alpha_i\}_{i=1}^d$, and real numbers $\{\lambda_i\}_{i=1}^d$ such that

$$P = \{v \in H^1(F_q; \mathbb{R}) | \langle v, \alpha_i \rangle \leq \lambda_i \text{ for } 1 \leq i \leq d\}. \tag{2.15}$$

If P is such a polytope, Y_P is a *special affine subset* in the sense of Tate [29, Definiton 7.1]; these are now usually studied as examples of the more general class of *affinoid domains* [7]. The affinoid ring $\mathcal{O}_P$ corresponding to Y_P in this case consists of formal series

$$\sum_{A \in H_1(F_q, \mathbb{Z})} f_A z_q^A, \quad f_A \in \Lambda \tag{2.16}$$

which T-adically converge at every point of Y_P, i.e. so that

$$\lim_{|A| \to +\infty} \mathrm{val}(f_A) + \langle v, A \rangle = +\infty \tag{2.17}$$

whenever v lies in P.

Remark 2.3. Despite the fact that Equation (2.16) refers to the basepoint q, the ring $\mathcal{O}_P$ does not depend on it. One way to see this is to construct a natural isomorphism of the rings associated to different choices of basepoints. Say that $p \in Q$ is obtained by exponentiating $v' \in H^1(F_q; \mathbb{R}) = T_qQ$. Using parallel transport to identify the tangent space of q with that of p, associate to P the polytope

$$P - v' \subset H^1(F_p; \mathbb{R}) = H^1(F_q; \mathbb{R}). \tag{2.18}$$

Note that the transformation

$$z_q^A \mapsto t^{\langle v', A \rangle} z_p^A \tag{2.19}$$

maps series in z_q coordinates which are convergent in P bijectively to series in z_p coordinates which are convergent in $P - v'$.

We now assume that the cover of Q chosen in Section 2.1 has the property that

$$\text{for each ordered subset } I \subset \mathcal{A}, P_I \text{ is an integral affine polytope.} \tag{2.20}$$

Covers satisfying this property exist for the following reason: every point in Q has a neighbourhood which is an integral affine polytope, and two such polytopes intersect along an integral affine polytope whenever they are sufficiently small. Using such a cover, we see that Y is obtained by gluing affinoid sets; it is therefore an affinoid variety.

2.3. Sheaves as functors. Consider the category $\mathcal{O}_{\mathcal{A}}$ whose objects are the ordered subsets of $\mathcal{A}$. The morphisms in $\mathcal{O}_{\mathcal{A}}$ are given by

$$\mathcal{O}(I,J) = \begin{cases} \mathcal{O}_J & \text{if } I \subset J \\ 0 & \text{otherwise.} \end{cases} \tag{2.21}$$

where $\mathcal{O}_I$ is the ring of functions on Y_I (and Y_I denotes Y_{P_I}). Composition is defined as

$$\mathcal{O}(J,K) \otimes \mathcal{O}(I,J) \cong \mathcal{O}_K \otimes \mathcal{O}_J \to \mathcal{O}_K \otimes \mathcal{O}_K \to \mathcal{O}_K \cong \mathcal{O}(I,K), \tag{2.22}$$

where the middle two arrows are respectively given by restriction (from Y_J to Y_K), and by multiplication (in $\mathcal{O}_K$).

Definition 2.4. A pre-sheaf of $\mathcal{O}$-modules on Y is a functor from $\mathcal{O}_{\mathcal{A}}$ to the category of Λ-vector spaces.

To see that this definition is reasonable, recall that a functor $\mathcal{F}$ assigns to each set I a Λ-vector space we denote $\mathcal{F}(I)$. Since the endomorphisms of the object I in $\mathcal{O}_{\mathcal{A}}$ is the ring of functions on Y_I, $\mathcal{F}(I)$ is equipped with an $\mathcal{O}_I$ module structure. Since $\mathcal{O}(I,J) = \mathcal{O}_J$ whenever $I \subset J$, we obtain a map

$$\mathcal{O}_J \otimes_\Lambda \mathcal{F}(I) \to \mathcal{F}(J). \tag{2.23}$$

The associativity equation implies that this is a map of $\mathcal{O}_J$ modules, and that it descends to a map

$$\mathcal{O}_J \otimes_{\mathcal{O}_I} \mathcal{F}(I) \to \mathcal{F}(J), \tag{2.24}$$

which exactly implies that we have pre-sheaf of $\mathcal{O}$-modules in the usual sense.

Definition 2.5. A *sheaf of* $\mathcal{O}$*-modules* on Y is a presheaf such that the structure maps in Equation (2.24) are isomorphisms.

The key point here is that the category of modules over the ring $\mathcal{O}_I$ is equivalent to the category of sheaves of $\mathcal{O}$-modules on the affinoid space Y_I (see, e.g. [7, Section 9.4.3]). The datum of a sheaf on Y_I can therefore be replaced by that of a single module $\mathcal{F}(I)$. As in the usual description, a sheaf is therefore a presheaf satisfying an additional property.

Remark 2.6. Since the ring of regular functions on an affinoid domain is Noetherian [7, p. 222], the notions of coherence and finite generation agree. So we may define a sheaf of coherent modules on Y to be a sheaf of $\mathcal{O}$-modules such that each module $\mathcal{F}(I)$ is finitely generated; i.e. admits a surjection from a finite rank free module. A standard argument implies that the cohomology modules of finite rank free cochain complexes over $\mathcal{O}_I$ are coherent modules; it is in this way that coherent sheaves on Y will arise from the mirror.

2.4. Rigid analytic gerbes and twisted sheaves. There is a natural map

$$\exp\colon \ \mathrm{Aff}(P) \to \mathcal{O}^*(Y_P) \tag{2.25}$$

$$F \mapsto t^{F(q)} z_q^{dF}, \tag{2.26}$$

which induces a map

$$H^2(Q, \mathrm{Aff}) \to H^2(Y, \mathcal{O}^*). \tag{2.27}$$

This map assigns to each Lagrangian fibration over Q an (analytic) gerbe on Y.

Remark 2.7. The above map is not surjective, but it is reasonably to expect surjectivity by considering deformations of Floer theory in X by the pullback of classes in $H^2(Q, U_\Lambda)$. For the subgroups $H^2(Q, \mathbb{Z}_2)$ and $H^2(Q, 1+\Lambda_+)$, such deformations were considered separately in [14] as *background class* and *bulk deformation*.

To define a twisted module over this gerbe, one needs a model for sheaf cohomology on Y: choose a cocycle α_X in $\check{C}^2_{\mathcal{A}}(Q, \mathrm{Aff})$ as in Equation (2.5); this consists of an assignment $\alpha_X(ijk) \in \mathrm{Aff}(P_{ijk})$ for every triple $i < j < k$, satisfying the cocycle condition. Given a triple $I \subset J \subset K$ of ordered subsets of $\mathcal{A}$ with final elements (i, j, k), define

$$\mathrm{Aff}(P_K) \ni \alpha_X(IJK) = \begin{cases} \alpha_X(ijk)|P_K & \text{if } i < j < k \\ 1 & \text{otherwise.} \end{cases} \tag{2.28}$$

Associated to this cocycle, we define a new category $\mathcal{O}^{\alpha_X}_{\mathcal{A}}$ with the same objects and morphisms as $\mathcal{O}_{\mathcal{A}}$. The composition is given by

$$\mathcal{O}(J, K) \otimes \mathcal{O}(I, J) \to \mathcal{O}(I, K) \tag{2.29}$$

$$f_K \otimes f_J \mapsto \exp(\alpha_X(ijk)) \cdot f_J|_{Y_K} \cdot f_K. \tag{2.30}$$

The cocycle property of α_X implies that composition is associative. As in the untwisted case, a functor $\mathcal{F}$ from $\mathcal{O}^{\alpha_X}_{\mathcal{A}}$ to Vect_Λ induces a map of $\mathcal{O}_J$ modules

$$\mathcal{O}_J \otimes_{\mathcal{O}_I} \mathcal{F}(I) \to \mathcal{F}(J). \tag{2.31}$$

Definition 2.8. An α_X-twisted $\mathcal{O}$-module is a functor from $\mathcal{O}^{\alpha_X}_{\mathcal{A}}$ to Λ-vector spaces such that the map in Equation (2.31) is an isomorphism for every pair $I \subset J$. □

The above definition unwinds into something more familiar: an α_X-twisted $\mathcal{O}$-module over Y is a collection $\mathcal{F}(I)$ of $\mathcal{O}_I$ modules, together with isomorphisms of $\mathcal{O}_J$ modules

$$\mathcal{F}_{IJ} \colon \mathcal{O}_J \otimes_{\mathcal{O}_I} \mathcal{F}(I) \to \mathcal{F}(J), \tag{2.32}$$

defined whenever $I \subset J$, such that

$$\mathcal{F}_{JK} \circ \mathcal{F}_{IJ}|_{Y_K} = \exp(\alpha_X(IJK)) \cdot \mathcal{F}_{IK} \tag{2.33}$$

for an ordered triple $I \subset J \subset K$. Here, $\mathcal{F}_{IJ}|_{Y_K}$ denotes the map induced by $\mathcal{F}_{IJ}$:

$$\mathcal{O}_K \otimes_{\mathcal{O}_I} \mathcal{F}(I) \xrightarrow{=} \mathcal{O}_K \otimes_{\mathcal{O}_J} \mathcal{O}_J \otimes_{\mathcal{O}_I} \mathcal{F}(I) \xrightarrow{\mathcal{F}_{IJ}} \mathcal{O}_K \otimes_{\mathcal{O}_I} \mathcal{F}(J). \tag{2.34}$$

3. Local constructions

3.1. Basics of Floer theory. Assume we are given a Lagrangian L so that

$$\begin{aligned} &L \text{ is } \textit{tautologically unobstructed,} \text{ i.e. there exists a tame almost complex structure } J_L \text{ on } X \text{ so that } L \\ &\text{bounds no } J_L\text{-holomorphic disc.} \end{aligned} \tag{3.1}$$

This is a technical condition, which will allow us to avoid discussing virtual fundamental chains, and should be replaced by the condition that L is unobstructed in the sense of [14].

Given a Hamiltonian diffeomorphism ϕ mapping L to a Lagrangian transverse to F_q, there is an ungraded Floer complex

$$CF^*(F_q, \phi(L)), \tag{3.2}$$

generated over $\Lambda_{\mathbb{F}_2}$ by the intersection points of $\phi(L)$ and F_q. To define the differential, choose a generic family of almost complex structures $\{J_t \in \mathcal{J}\}_{t\in[0,1]}$ so that $J_0 = \phi^*(J_L)$. For each pair (x, y) of intersection points between $\phi(L)$ and F_q, the space of J_t-holomorphic maps from the strip $B = \mathbb{R} \times [0, 1]$ to X satisfying the following boundary and asymptotic conditions

$$u(s, 1) \in \phi(L) \qquad u(s, 0) \in F_q \tag{3.3}$$

$$\lim_{s\to-\infty} u(s, t) = x \qquad \lim_{s\to+\infty} u(s, t) = y \tag{3.4}$$

admits a natural $\mathbb{R}$-action by translation in the s-coordinate. The quotient by this action of the space of such maps is the moduli space of strips $\mathcal{M}^q(x, y)$, and the matrix coefficient of x in dy is the count of rigid elements of this moduli space. The key point here is that this moduli space is regular for generic almost complex structures, so the count of such isolated elements gives a differential by standard methods.

In order for the mirror of L to be an object of the bounded derived category and to be defined away from fields of characteristic 2, this construction must be refined to a chain complex of free abelian groups which is $\mathbb{Z}$ graded. Combining the discussions of orientations in [14] and [25], assume that

$$w_2(L) = \pi^*(w_2(Q)). \tag{3.5}$$

Under this assumption, one could make an arbitrary choice of Pin^+ structure on the bundles $TF_q \oplus \pi^*(T^*Q \otimes |Q|^{\oplus 3})$ and $TL \oplus \pi^*(T^*Q \otimes |Q|^{\oplus 3})$ to define the Floer cohomology of L and F_q. It will be important to make a *global* choice, i.e. one obtained by restriction from X. To this end, identify the restriction of $\pi^*(T^*Q)$ to F_q with its tangent space via the Arnol'd-Liouville theorem. In particular, a Pin^+ structure on

$$T^*Q \oplus T^*Q \oplus |Q|^{\oplus 3} \tag{3.6}$$

will induce one by the pullback to all fibres. The above bundle has vanishing second Stiefel-Whitney class, which is the obstruction to such a structure.

Upon fixing Pin^+ structures on $TL \oplus \pi^*(T^*Q \otimes |Q|^{\oplus 3})$ and in Equation (3.6), index theory assigns a rank 1 free abelian group δ_x to each intersection point $x \in \phi(L) \cap F_q$, with the property that every rigid element of $\mathcal{M}^q(x, y)$ induces a canonical map

$$d_u \colon \delta_y \to \delta_x, \tag{3.7}$$

which should be thought of as the signed contribution of u to the differential.

It remains to lift the grading of the Floer complex to a $\mathbb{Z}$-grading. Equipped with any almost complex structure for which the fibres are totally real, there is a natural isomorphism of vector bundles $TX \cong \pi^*(TQ) \otimes_{\mathbb{R}} \mathbb{C}$. This implies that a density on Q induces an almost complex quadratic volume form on X. Evaluating such a form on a basis for the tangent space of a Lagrangian defines the phase function

$$\eta_\Omega \colon L \to S^1. \tag{3.8}$$

By assuming that the phase function on L is null homotopic and fixing a lift to $\mathbb{R}$, index theory assigns a degree $\deg(x) \in \mathbb{Z}$ to every intersection point $x \in \phi(L) \pitchfork F_q$ with the property that the moduli space $\mathcal{M}^q(x,y)$ has pure dimension

$$\dim(\mathcal{M}^q(x,y)) = \deg(x) - \deg(y) - 1. \tag{3.9}$$

The differential defined in Equation (3.7) then raises degree by 1 on the Floer complex

$$CF^d(F_q, \phi(L)) = \bigoplus_{\deg(x)=d} \delta_x \otimes \Lambda. \tag{3.10}$$

3.2. Convergence of the Floer differential and restriction. Let P be a polytope in Q containing q and $\mathcal{O}_P$ denote the affinoid ring of Y_P. This section adapts an argument of Fukaya showing that, whenever P is sufficiently small, the Floer complex in Equation (3.10) is the fibre of a complex of vector bundles on Y_P

$$\bigoplus_{x \in F_q \cap L} \mathcal{O}_P \otimes_{\mathbb{Z}} \delta_x. \tag{3.11}$$

More precisely, choosing P small enough, there is a differential on Equation (3.11) which specialises to the Floer complex

$$CF^*((F_p, b), \phi(L)) \tag{3.12}$$

for every point $(p,b) \in Y_P$.

In order to define the differential using this moduli space, it is useful to think of the intersection point $x \in F_q \cap \phi(L)$ as a sheet of $\phi(L)$ over P. Fixing a trivialisation

$$\tau_P \colon X_P \cong T^*P/T^*_{\mathbb{Z}}P, \tag{3.13}$$

this can be written as the differential of a function $g_x \colon P \to \mathbb{R}$ which is well-defined up to an integral affine function.

Definition 3.1. A collection of *Floer data* D_P consists of the choices $(\tau_P, \phi, J, \{g_x\})$. They *are tame in* P if there is a (smooth) map $\psi \colon P \to \mathrm{Diff}(X)$ which maps q to the identity, such that ψ_p maps F_q to F_p and preserves $\phi(L)$ and the tameness of the almost complex structures $\{J_t\}_{t\in[0,1]}$.

Choosing the functions $\{g_x\}$ yields for each strip u with sides mapping to L and F_q a class

$$[\partial u] \in H_1(F_q, \mathbb{Z}). \tag{3.14}$$

In order to define this class explicitly, note that the choice of trivialisation of X_P determines a 0-section of X_P, and hence a basepoint on F_q. The linear path tdg_x has endpoints the basepoint at $t=0$ and x at $t=1$. Define $[\partial u]$ to be the homology class of the loop obtained by concatenating the paths associated to dg_x, dg_y, and the restriction of u to the boundary.

Letting $z^{[\partial u]}$ be the exponential of the unique linear function on Q which vanishes at q and whose differential is given by $[\partial u]$ under the identification of Equation (2.1), define

$$d|\delta_y = \bigoplus_x \sum_{u \in \mathcal{M}^q(x,y)} T^{\mathcal{E}(u)} z^{[\partial u]} \otimes d_u, \tag{3.15}$$

where $\mathcal{M}^q(x,y)$ is the moduli space of strips defining the Floer differential, d_u is the map induced on determinant lines by u, and $\mathcal{E}(u)$ is the energy of u:

$$\mathcal{E}(u) = \int u^*\omega. \tag{3.16}$$

We shall presently use an idea of Fukaya to show that the infinite sum in the expression of the differential lies in $\mathcal{O}_P$.

Recall from Equation (2.14) that every element $z' \in Y_P$ can be written as a pair (p, b'), with $p \in P$ and $b' \in H^1(F_p; U_\Lambda)$. There is a natural isomorphism of Λ-modules

$$\bigoplus_x \mathcal{O}_P \underset{z=z'}{\otimes} \Lambda \otimes_{\mathbb{Z}} \delta_x = \bigoplus_x \Lambda \otimes_{\mathbb{Z}} \delta_x = CF^*((F_p, b'), \phi(L)). \tag{3.17}$$

If $q = p$, this map commutes with the differential defined using J, which in particular proves that the series in Equation (3.15) are convergent at such points. By defining the right hand side of Equation (3.17) using the almost complex structure $(\psi_p^{-1})^* J$ (c.f. [12]) so that composition with ψ_p gives a bijection between holomorphic strips with boundary on F_q and $\phi(L)$, and those with boundary on F_p and $\phi(L)$, convergence is achieved when $q \neq p$. This is where Definition 3.1 is used. To state the result, define $p - q$ as the point in T_qQ which exponentiates to p.

Lemma 3.2. *If u lies in $\mathcal{M}^q(x,y)$, the energy of $\psi_p \circ u$ is*

$$\mathcal{E}(\psi_p \circ u) = \mathcal{E}(u) + \langle p - q, [\partial u] \rangle + g_x(q) - g_y(q) + g_y(p) - g_x(p). \tag{3.18}$$

Proof. Consider the linear path in Q from q to p. The term $\langle p - q, [\partial u] \rangle$ is the area of a cylinder in X, lying over this path, and which intersects each fibre in a circle of homotopy class $[\partial u]$. The terms $g_x(p) - g_x(q)$ and $g_y(p) - g_y(q)$ are respectively the areas of strips over this path whose intersections with each fibre are the paths from the basepoint to the intersection of each fibre with the local sheets of $\phi(L)$ labelled x and y. The right hand side is therefore the sum of the area of u with that of a strip in X, which intersects each fibre along the segment from q to p in a path from the intersection with x to the intersection with y, lying in the homotopy class of $u|\mathbb{R} \times \{0\}$.

The result of gluing these two strips is homotopic to $\psi_p \circ u$, and Equation (3.18) follows from the invariance of the topological energy under homotopies with fixed Lagrangian boundary conditions. □

By the previous result, the contributions of a curve u to the differentials on the two sides of Equation (3.17) differ by multiplication by

$$T^{g_x(q) - g_y(q) + g_y(p) - g_x(p)}. \tag{3.19}$$

This readily implies that the pre-composition of the isomorphism in Equation (3.17) with multiplication by

$$T^{g_x(q) - g_x(p)} \tag{3.20}$$

on the $\Lambda \otimes_{\mathbb{Z}} \delta_x$ factor is a cochain isomorphism (the disappearance of $\langle p-q, [\partial u] \rangle$ is explained by Remark 2.3). Gromov compactness applied for all fibres F_p over the polygon P implies:

Proposition 3.3 (c.f. [12]). *For every pair of intersections (x, y) the series*

$$\sum_{u \in \mathcal{M}^q(x,y)} T^{\mathcal{E}(u)} z^{[\partial u]} \tag{3.21}$$

is convergent in Y_P. □

As a consequence, Equation (3.15) defines a differential on the complex

$$\mathcal{L}(Y_P; D_P) \equiv \bigoplus_{x \in F_q \cap \phi(L)} \mathcal{O}_P \otimes_{\mathbb{Z}} \delta_x. \tag{3.22}$$

It will be convenient to drop the Floer data from the notation whenever they are unambiguously given.

Given an inclusion of polytopes $P' \to P$, with basepoints q on P and q' on P', there are restricted data

$$D_P|_{P'} \equiv (\tau_P|_{X_{P'}}, \phi, (\psi_{q'}^{-1})^* J, \{g_x\}). \tag{3.23}$$

Multiplication of each summand by Equation (3.20) defines a cochain map

$$\mathcal{L}(Y_P; D_P) \to \mathcal{L}(Y_{P'}; D_P|_{P'}). \tag{3.24}$$

3.3. Change of trivialisation. Any pair of fibrewise identifications

$$\tau_i \colon X_P \cong T^*P / T^*_{\mathbb{Z}} P,\ i \in \{1, 2\}. \tag{3.25}$$

differ by the differential of a function $f \colon P \to \mathbb{R}$. With respect to two such trivialisations, choose functions g_x^1 and g_x^2 defining every local section of $\phi(L)$, and consider the two sets of data

$$D^i = (\tau_i, \phi, J, \{g_x^i\}),\ i \in \{1, 2\}. \tag{3.26}$$

Define the *change of trivialisation* cochain map

$$\mathcal{L}(Y_P; D^1) \to \mathcal{L}(Y_P; D^2) \tag{3.27}$$

as a diagonal map given on the factor δ_x by multiplication with $T^{f(q)} z_q^{df - dg_x^1 + dg_x^2}$.

Since this map does not entail counting any holomorphic curves, it is easy to check that given an inclusion $P' \subset P$, we have a commutative diagram

$$\begin{array}{ccc} \mathcal{L}(Y_P; D^1) & \longrightarrow & \mathcal{L}(Y_P; D^2) \\ \downarrow & & \downarrow \\ \mathcal{L}(Y_{P'}; D^1|_{P'}) & \longrightarrow & \mathcal{L}(Y_{P'}; D^2|_{P'}). \end{array} \tag{3.28}$$

3.4. Continuation maps. Let D^+ and D^- denote Floer data $(\tau, \phi^{\pm}, J^{\pm}, \{g_{x_\pm}\})$, which share a common trivialisation. This section recalls the construction of the continuation map

$$CF^*(F_q, \phi^+(L)) \to CF^*(F_q, \phi^-(L)) \tag{3.29}$$

as a count of pseudo-holomorphic sections of the projection $X \times B \to B$.

Pick a path of Hamiltonian diffeomorphisms ϕ^s such that

$$\phi^s = \begin{cases} \phi^+ \text{ if } 0 \ll s \\ \phi^- \text{ if } s \ll 0. \end{cases} \tag{3.30}$$

Recall that there is a unique function H on $X \times \mathbb{R}$ which generates this flow such that

$$\int_X H_s \omega^n = 0, \tag{3.31}$$

and pick a compactly supported function

$$G\colon\ X \times B \to \mathbb{R}, \tag{3.32}$$

which agrees with H on $X \times \mathbb{R} \times \{1\}$.

If α is a symplectic form on B of finite area, the 2-form

$$\omega_X - dG \wedge ds + C\alpha \tag{3.33}$$

defines a symplectic structure on $X \times B$ whenever C is a sufficiently large constant. We denote by $\tilde{\mathcal{J}}$ the space of almost complex structures $\tilde{J}$ on $B \times X$ which are of the form

$$\tilde{J} = \begin{pmatrix} J & K \\ 0 & j \end{pmatrix} \tag{3.34}$$

with $J \in \mathcal{J}$, and K vanishes outside a compact set in B. Any such almost complex structure will be tamed by the symplectic structure in Equation (3.33) whenever C is sufficiently large.

We choose an almost complex structure $\tilde{J} \in \tilde{\mathcal{J}}$ whose restrictions to $0 \ll s$ and $s \ll 0$ agree with

$$\tilde{J}^{\pm} = \begin{pmatrix} J^{\pm} & 0 \\ 0 & j \end{pmatrix}. \tag{3.35}$$

Definition 3.4. An *elementary continuation datum* D^{+-} from D^+ to D^- is a choice of the pair of data $(\{\phi^s\}, \tilde{J})$ above.

For each pair (x_-, x_+) of intersections points, we then define $\mathcal{M}^q_\kappa(x_-, x_+)$ with respect to the data D^{+-} to be the moduli space of maps $v\colon\ B \to X$ with $\tilde{J}$-holomorphic graph $\tilde{v}\colon\ B \to B \times X$ such that

$$v(s,0) \in F_q \tag{3.36}$$

$$v(s,1) \in \phi^s L \tag{3.37}$$

$$\lim_{s \to \pm\infty} v(s,t) = x_\pm. \tag{3.38}$$

For a generic almost complex structure $\tilde{J}$, these moduli spaces are regular. Computing the linearisation of the index of a solution shows that

$$\dim(\mathcal{M}^q_\kappa(x_-, x_+)) = \deg(x_-) - \deg(x_+). \tag{3.39}$$

The (topological) energy of a solution to the continuation equation is

$$\mathcal{E}(v) = \int_B v^*(\omega) - \int_{\mathbb{R}} H_s(v(s,1))ds = \int_B \tilde{v}^* \left(\omega - dG \wedge ds\right) \tag{3.40}$$

Lemma 3.5. *If* $\deg(x_-) = \deg(x_+)$*, then for any real number* E*, there are only finitely many elements* v *of* $\mathcal{M}_\kappa^q(x_-, x_+)$ *such that*

$$\mathcal{E}(v) \leq E. \tag{3.41}$$

Proof. By adding to $\mathcal{E}$ the total area of the form $C\alpha$ in Equation (3.33), we obtain the area of $\tilde{v}$ with respect to a symplectic structure on $X \times B$ for which $\tilde{J}$ is tame. A standard application of Gromov compactness therefore implies that the energy is proper on the Gromov-Floer compactification.

Having excluded bubbling via Assumption (3.1), the only broken curves in the limit arise when some energy escapes along the ends, which gives rise to components of hypothetical broken curves that are graphs of Floer trajectories. Since the moduli space of Floer trajectories is assumed to be regular, this is impossible whenever the moduli space we are considering has vanishing virtual dimension. □

As in the setting of Floer trajectories, an element $v \in \mathcal{M}_\kappa^q(x_-, x_+)$ induces a canonical map

$$\kappa_v \colon \delta_{x_+} \to \delta_{x_-} \tag{3.42}$$

whenever $\deg(x_+) = \deg(x_-)$. Taking the sum over all elements of such rigid moduli spaces defines the continuation map

$$\kappa \colon CF^*(F_q, \phi^+(L)) \to CF^*(F_q, \phi^-(L)) \tag{3.43}$$

$$\kappa|\delta_{x_+} = \sum_{\substack{\deg(x_+)=\deg(x_-) \\ v \in \mathcal{M}_\kappa^q(x_-, x_+)}} T^{\mathcal{E}(v)} \kappa_v. \tag{3.44}$$

3.5. Convergence of continuation maps. Let P be a polytope based at q, with the property that $\mathcal{L}(Y_P; D^\pm)$ are well defined. Pick diffeomorphisms $\psi^\pm$ as in Definition 3.1, and extend them to a family

$$\Psi \colon P \times B \to \mathrm{Diff}(X) \tag{3.45}$$

$$\{q\} \times B \mapsto \mathrm{Id} \tag{3.46}$$

such that the following properties hold for all $p \in P$ (see Figure 3.1):

$$\Psi_{p,s,t} = \psi_p^+ \text{ if } 0 \ll s \tag{3.47}$$

$$\Psi_{p,s,t} = \psi_p^- \text{ if } s \ll 0 \tag{3.48}$$

$$\Psi_{p,s,0}(F_q) = F_p \tag{3.49}$$

$$\Psi_{p,s,1} = \mathrm{Id} \text{ if } H_s \not\equiv 0. \tag{3.50}$$

For each $p \in P$, denote by $\tilde{\Psi}_p$ the fibrewise diffeomorphism of $B \times X$

$$(s, t, x) \mapsto (s, t, \Psi_{p,s,t}(x)). \tag{3.51}$$

Define $\tilde{J}_p = \left(\tilde{\Psi}_p^{-1}\right)^* \tilde{J}$. This is an almost complex structure on $B \times X$ which has the upper triangular form required in Equation (3.35).

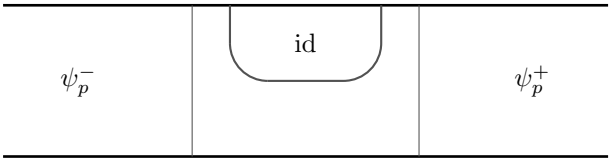


Figure 3.1.

Definition 3.6. The continuation data D^{+-} *are tame in* P if $\tilde{J}_p$ lies in $\tilde{\mathcal{J}}$ for all $p \in P$.

Of course, tameness depends on the choice of the fibrewise diffeomorphism Ψ of $B \times X$, but as this will be clear from the context, it is omitted. Condition (3.50) ensures that the off-diagonal term in $\tilde{J}_p$ vanishes outside a compact set. Openness of the taming condition implies:

Lemma 3.7. *If* P *is sufficiently small the data* D^{+-} *are tame in* P. □

Define $\mathcal{M}_\kappa^p(x_-, x_+)$ to be the space of maps from a strip to X, with boundary conditions F_p and $\phi^s(L)$, converging to $x_\pm$ at the respective ends, whose graphs are $\tilde{J}_p$-holomorphic.

Lemma 3.8. *Composition with* $\tilde{\Psi}_p$ *defines a bijection*

$$\mathcal{M}_\kappa^q(x_-, x_+) \cong \mathcal{M}_\kappa^p(x_-, x_+). \tag{3.52}$$

Proof. The key point is that the diffeomorphism $\tilde{\Psi}_p$ is compatible with the Lagrangian boundary conditions. In particular, given $v \in \mathcal{M}_\kappa^q(x_-, x_+)$, Equation (3.50) ensures that the boundary condition of $\tilde{\Psi}_p \circ \tilde{v}$ along $\mathbb{R} \times \{1\}$ is the path $\phi^s(L)$, and Equation (3.49) ensures that the boundary condition along $\mathbb{R} \times \{0\}$ is F_p. □

The advantage of introducing both moduli spaces is that we have independent energy estimates:

Lemma 3.9. *If* $\deg(x_-) = \deg(x_+)$, *the topological energy defines a proper map on* $\mathcal{M}_\kappa^p(x_-, x_+)$.

Proof. Condition (3.50) ensures that the off-diagonal term in $\tilde{J}_p$ vanishes outside a compact set, so escape of energy along the ends gives rise to Floer trajectories as in the proof of Lemma 3.5. The remainder of that proof applies mutatis mutandis. □

These results imply the existence of a cochain map

$$\kappa\colon\ \mathcal{L}(Y_P; D^+) \to \mathcal{L}(Y_P; D^-) \tag{3.53}$$

by counting solutions to continuation maps as follows: recall that the construction of the differential on the two $\mathcal{O}_P$ modules relied on choosing primitives defining the sheets of $\phi^+(L)$ and $\phi^-(L)$ over P. These primitives define a class $[\partial v] \in H_1(F_q, \mathbb{Z})$ associated to the boundary of $v \in \mathcal{M}_\kappa^q(x_-, x_+)$.

The continuation map is defined on each factor by the formula

$$\kappa|\delta_{x_+} = \bigoplus_{x_-} \sum_{v \in \mathcal{M}_\kappa(x_-, x_+)} T^{\mathcal{E}(v)} z^{[\partial v]} \otimes \kappa_v \tag{3.54}$$

The argument proving the convergence of the differential applies verbatim in this case.

Remark 3.10. It is important to note that κ is not defined over Λ_0 since the energy $\mathcal{E}(v)$ for a solution to the continuation equation is not necessarily positive. This raises obstacles to using the methods developed here to detect quantitative information about Lagrangians in X, e.g. displacement energy as in [16].

3.6. Families of continuation maps. Let Δ be a compact manifold with boundary parametrising a family $\{(\{\phi_\delta^s\}, \tilde{J}_\delta)\}_{\delta \in \Delta}$ of continuation data as in Definition 3.4.

For each pair (x_-, x_+) of intersections points, let $\mathcal{M}^q_{\kappa^\delta}(x_-, x_+)$ denote the moduli space of continuation maps corresponding to $\delta \in \Delta$, and consider the parametrised space

$$\mathcal{M}^q_{\kappa^\Delta}(x_-, x_+) \equiv \coprod_{\delta \in \Delta} \mathcal{M}^q_{\kappa^\delta}(x_-, x_+). \tag{3.55}$$

Assuming the data are chosen generically, this is a manifold with boundary of dimension

$$\deg(x_-) - \deg(x_+) + \dim(\Delta). \tag{3.56}$$

In particular, if this moduli space is rigid, we may consider the series:

$$\sum_{v \in \mathcal{M}^q_{\kappa^\Delta}(x_-, x_+)} T^{\mathcal{E}(v)} z^{[\partial v]}, \tag{3.57}$$

where the class $[\partial v] \in H_1(F_q, \mathbb{Z})$ is defined as in Equation (3.14).

Proposition 3.11. *There is a polytope $P_\Delta \subset Q$ so that Equation* (3.57) *defines an element of $\mathcal{O}_{P_\Delta}$.*

Proof. Condition (3.30) and the compactness of Δ imply that there is a constant S_Δ so that for all $\delta \in \Delta$, H^s_δ agrees with H^+ if $S_\Delta \leq s$, and with H^- if $s \leq -S_\Delta$. Consider a smooth family $\Psi_{\delta,p,s,t}$ of diffeomorphisms of X parametrised by $(\delta, p, s, t) \in \Delta \times P \times B$ satisfying, for fixed $\delta \in \Delta$, Conditions (3.46)-(3.50). The assumption that these diffeomorphisms are the identity for $p = q$ and the compactness of Δ imply that the corresponding fibrewise diffeomorphisms $\tilde{\Psi}_{\delta,p}$ of $B \times X$ preserve the tameness of the almost complex structure whenever p lies in a sufficiently small neighbourhood of q. This yields the analogue of Lemma 3.9, and hence convergence in this neighbourhood. □

3.7. Composition of continuation map. Let (ϕ^+, J^+), (ϕ^0, J^0) and (ϕ^-, J^-) denote three choices of Hamiltonian diffeomorphisms and almost complex structures, and pick (regular) continuation data D^{+0}, D^{0-} and D^{+-} which define cochain maps

$$\begin{array}{ccc} \mathcal{L}(Y_P; D^+) & \xrightarrow{\kappa_{+-}} & \mathcal{L}(Y_P; D^-) \\ & {\scriptstyle \kappa_{+0}} \searrow \quad \nearrow {\scriptstyle \kappa_{0-}} & \\ & \mathcal{L}(Y_P; D^0), & \end{array} \tag{3.58}$$

whenever P is sufficiently small.

Gluing the data D^{+0} and D^{0-}, defines a continuation map from $\mathcal{L}(Y_P; D^+)$ to $\mathcal{L}(Y_P; D^-)$ which agrees with the composition $\kappa_{0-} \circ \kappa_{+0}$. Choosing a homotopy between D^{+-} and the

glued data yields a family parametrised by an interval. Possibly upon shrinking P, Equation (3.57) and Proposition 3.11 produce a map

$$\kappa^1 \colon \mathcal{L}(Y_P; D^+) \to \mathcal{L}(Y_P; D^-)[1]. \tag{3.59}$$

For each pair of intersections (x_-, x_+) between $\phi^{\pm}(L)$ and F so that $\deg(x_+) = \deg(x_-)$, the moduli space $\mathcal{M}^q_{\kappa^1}(x_-, x_+)$ had dimension 1, and its boundary consists of the strata

$$\mathcal{M}^q_{\kappa}(x_-, x_+) \tag{3.60}$$

$$\coprod_{x_0 \in \phi^0(L) \cap F} \mathcal{M}^q_{\kappa}(x_-, x_0) \times \mathcal{M}^q_{\kappa}(x_0, x_+) \tag{3.61}$$

$$\coprod_{x'_+ \in \phi^+(L) \cap F} \mathcal{M}^q_{\kappa}(x_-, x'_+) \times \mathcal{M}^q(x'_+, x_+) \tag{3.62}$$

$$\coprod_{x'_- \in \phi^-(L) \cap F} \mathcal{M}^q(x_-, x'_-) \times \mathcal{M}^q_{\kappa}(x'_-, x_+). \tag{3.63}$$

Counting elements of the first two moduli spaces corresponds to the two compositions in Diagram (3.58). The second two moduli spaces respectively define the composition of κ^1 with the differentials in $\mathcal{L}(Y_P; D^+)$ and $\mathcal{L}(Y_P; D^-)$.

Proposition 3.12. *If P is sufficiently small, κ^1 defines a homotopy between the two compositions in Diagram* (3.58). □

Applying the above construction to the null-homotopy for the concatenation of a path and its inverse implies:

Corollary 3.13. *If P is sufficiently small, κ is a chain equivalence.* □

3.8. Compatibility of restriction, continuation, and change of trivilisations. Choose data $(\phi^{\pm}, J^{\pm})$ as in Section 3.4, and trivalisations $\{\tau_i\}_{i=1,2}$. These yield four sets of Floer data

$$D_i^{\pm} \equiv (\tau_i, \phi^{\pm}, J^{\pm}, \{g^i_{x_{\pm}}\}). \tag{3.64}$$

Using the same continuation equation to map the Floer complexes defined from the data D_i^+ to those defined from the data D_i^-, we have:

Lemma 3.14. *The following diagram, in which the vertical arrows are continuation maps and the horizontal ones are changes of coordinates, commutes*

$$\begin{array}{ccc} \mathcal{L}(Y_P; D_1^+) & \longrightarrow & \mathcal{L}(Y_P; D_2^+) \\ \downarrow & & \downarrow \\ \mathcal{L}(Y_P; D_1^-) & \longrightarrow & \mathcal{L}(Y_P; D_2^-). \end{array} \tag{3.65}$$

Proof. The class in $H_1(F_q, \mathbb{Z})$ associated to a continuation map v depends on the choices of local primitives. We write $[\partial v]^i$ for the choice associated to i. With this in mind, the commutativity of the diagram reduces to the equality

$$[\partial v]^1 + dg^1_{x_+} - dg^1_{x_-} - df = [\partial v]^2 + dg^2_{x_+} - dg^2_{x_-} - df \in H_1(F_q, \mathbb{Z}) \subset T^*_q P. \tag{3.66}$$

□

Similarly, given a subpolytope $P' \subset P$ with basepoint q', define $\kappa|P'$ to be the continuation map associated to the data

$$(\{\phi^s\}, \tilde{J}_{q'}). \tag{3.67}$$

As in Lemma 3.14, there is a commutative diagram

$$\begin{array}{ccc} \mathcal{L}(Y_P; D^+) & \xrightarrow{\kappa} & \mathcal{L}(Y_P; D^-) \\ \downarrow & & \downarrow \\ \mathcal{L}(Y_{P'}; D^+|_{P'}) & \xrightarrow{\kappa|P'} & \mathcal{L}(Y_{P'}; D^-|_{P'}). \end{array} \tag{3.68}$$

4. From Lagrangians to twisted sheaves

4.1. Homological patching. In this section, Floer cohomology groups are used to define an α_X-twisted sheaf $H^*\mathcal{L}$ associated to a Lagrangian L.

Start by choosing a simplicial triangulation as in Section 2.1, with associated cover P_i satisfying Condition (2.20). Denoting by $\mathcal{O}_I$ the ring of functions on the inverse image Y_I of P_I, this cover should be sufficiently fine that:

1. for each vertex $i \in \mathcal{A}$, there are Floer data D_i defining complexes of $\mathcal{O}_i$ modules $\mathcal{L}(i) \equiv \mathcal{L}(Y_i; D_i)$.
2. for each pair of vertices $i < j$, there are continuation data D_{ij} defining chain equivalences
$$\mathcal{L}_{ij} \colon \mathcal{L}(Y_{ij}; D_i|P_{ij}) \to \mathcal{L}(Y_{ij}; D_j|P_{ij}). \tag{4.1}$$
3. for each triple of vertices $i < j < k$, there are homotopies D_{ijk} of continuation data between D_{ik} and the gluing of D_{ij} and D_{jk} defining a chain homotopy $\mathcal{L}_{ijk}$ in the diagram
$$\begin{array}{ccc} \mathcal{L}(Y_{ijk}; D_i|P_{ijk}) & \longrightarrow & \mathcal{L}(Y_{ijk}; D_k|P_{ijk}) \\ & \searrow \quad \nearrow & \\ & \mathcal{L}(Y_{ijk}; D_j|P_{ijk}). & \end{array} \tag{4.2}$$

Lemma 4.1. *If the triangulation of Q is sufficiently fine, there are choices of data satisfying the above properties.*

Proof. Start with a finite cover $\mathcal{U}_1$ by polytopes equipped with tame Floer data. Then choose a second cover $\mathcal{U}_2$, subordinate to $\mathcal{U}_1$, so that, for each element of $\mathcal{U}_2$ contained in a pair of elements of $\mathcal{U}_1$, there are convergent continuation data between the two Floer data obtained by restriction, and fix a choice of such data for pairs. Finally, we pick a cover $\mathcal{U}_3$, subordinate to $\mathcal{U}_2$, so that there are convergent chain homotopies between all compositions of the continuation data chosen for $\mathcal{U}_2$.

Now, assume that the triangulation of Q labelled by $\mathcal{A}$ is subordinate to $\mathcal{U}_3$ (i.e. so that all open stars of all vertices are contained in an element of the cover), and choose the polytope P_i for each element $i \in \mathcal{A}$ to also be subordinate to this cover and contain the open star of i.

Pick the data D_i arbitrarily among the Floer data associated to elements of $\mathfrak{U}_3$ which contain P_i. The above choices determine continuation maps and homotopies satisfying the desired properties. □

Lemma 4.2. *The modules $H^*\mathcal{L}(i)$ and the structure maps $\mathcal{L}_{ij}$ define an α_X-twisted sheaf on Y.*

Proof. It suffices to show that the restriction maps commute up to multiplication by $\exp(\alpha_X(ijk))$. Lemma 3.14 reduces the proof to the case in which the Floer data (ϕ_j, J_j) and (ϕ_k, J_k) are obtained by restricting common data (ϕ_i, J_i). Let (f_{ij}, f_{jk}, f_{ik}) denote the transition functions between the three different trivialisations fixed on P_i, P_j and P_k. The composition $\mathcal{L}_{jk} \circ \mathcal{L}_{ij}$ is given on δ_x by multiplication with

$$T^{f_{jk}+f_{ij}} z^{df_{jk}+df_{ij}-dg_x^k+dg_x^i} = \exp(f_{jk}+f_{ij}) T^{dg_x^i-dg_x^k}, \tag{4.3}$$

while the restriction of $\mathcal{L}_{ik}$ to δ_x agrees with

$$T^{f_{ik}} z^{df_{ik}-dg_x^k+dg_x^i} = \exp(f_{ik}) T^{dg_x^i-dg_x^k}. \tag{4.4}$$

The result now follows immediately from Equation (2.5). □

4.2. Towards the mirror functor. To each ordered subset $I = (i_0, \ldots, i_r) \subset \mathcal{A}$ corresponding to an r-dimensional simplex in Q, Adams' construction [3] associates an $r-1$-dimensional cube σ_I of paths in Q from the initial to the final vertex. Paths parametrised by the boundary of this cube are given by (i) the family of paths associated to codimension 1 subsimplices, and (ii) the product of the cubes associated to a pair of complementary simplices in I. The homotopy constructed in Section 3.7 for a triple arose from a family of continuation maps associated to such a 1-dimensional cube in the case $r = 2$.

By gluing and induction on dimension, one obtains a family D_I of continuation maps from D_{i_0} to D_{i_r} parametrised by σ_I, whose restriction to the boundary strata of σ_I are given either by the continuation maps associated to a subsimplex, or the concatenation of continuation maps associated to complementary simplices. We are in the setting of Section 3.6 so, assuming the parametrised data are chosen generically and the triangulation is sufficiently fine, the count of rigid elements of such a moduli space defines a map

$$\mathcal{L}_I \colon \mathcal{L}(Y_{i_0}; D_{i_0}) \to \mathcal{L}(Y_I; D_{i_r}|P_I) \tag{4.5}$$

of degree $-r$. Adopting the convention that

$$\mathcal{L}(I) \equiv \mathcal{L}(Y_I; D_{i_r}|P_I), \tag{4.6}$$

the maps in Equation (4.5) naturally extend to an A_∞ module over $\mathcal{O}_{\mathcal{A}}^{\alpha_X}$, i.e an A_∞ functor

$$\mathcal{L} \colon \mathcal{O}_{\mathcal{A}}^{\alpha_X} \to \mathrm{Vect}_\Lambda\,. \tag{4.7}$$

This data is exactly that of an α_X-twisted A_∞-presheaf of $\mathcal{O}$-complexes on Y. Keeping in mind the fact that $\mathcal{L}(I)$ is a finite rank free $\mathcal{O}_I$ module, and that the map associated to an inclusion is a quasi-isomorphism (after restriction), $\mathcal{L}$ in fact defines an object of the A_∞-category of α_X-twisted sheaves of perfect complexes on Y, with respect to the cover $\mathcal{A}$. In this sense, the assignment $L \to \mathcal{L}$ gives, at the level of objects, the mirror functor between the derived Fukaya category of X and the derived category of α_X-twisted coherent sheaves on Y.

Remark 4.3. It is easy in this setting to implement one of the standard equivalences between A_∞ and dg-modules, and replace $\mathcal{L}$ by a quasi-equivalent dg-module over $\mathcal{O}^{\alpha x}_{\mathcal{A}}$, see e.g. [10, Theorem 6.15]. Since the Fukaya category is an A_∞ category, such a replacement does not seem to particularly simplify this approach to Homological mirror symmetry.

Acknowledgments. The author was supported by NSF grant DMS-1308179. It is my pleasure to thank my collaborators, on past and ongoing projects, for teaching me much of what I know about mirror symmetry and symplectic topology. Comments from Ivan Smith and Nick Sheridan were useful in clarifying the exposition and removing confusing conventions. In addition, I am grateful to Kenji Fukaya for explaining to me Tate's acyclicity theorem and its relevance to mirror symmetry, and to Paul Seidel for having instilled in me the lesson that constructions involving the Fukaya category should first be explained at the cohomological level.

References

[1] Mohammed Abouzaid and Ivan Smith, *Homological mirror symmetry for the 4-torus*, Duke Math. J. **152** (2010), no. 3, 373–440.

[2] Mohammed Abouzaid, Denis Auroux, Dmitri Orlov, and L. Katzarkov, *Homological mirror symmetry for Kodaira-Thurston manifolds*, in preparation.

[3] J. F. Adams, *On the cobar construction*, Proc. Nat. Acad. Sci. U.S.A. **42** (1956), 409–412.

[4] Manabu Akaho and Dominic Joyce, *Immersed Lagrangian Floer theory*, J. Differential Geom. **86** (2010), no. 3, 381–500.

[5] Paul S. Aspinwall, Tom Bridgeland, Alastair Craw, Michael R. Douglas, Mark Gross, Anton Kapustin, Gregory W. Moore, Graeme Segal, Balázs Szendrői, and P. M. H. Wilson, *Dirichlet branes and mirror symmetry*, Clay Mathematics Monographs, vol. 4, American Mathematical Society, Providence, RI, 2009.

[6] Victor V. Batyrev, *Mirror symmetry and toric geometry*, Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998), 1998, pp. 239–248 (electronic).

[7] S. Bosch, U. Güntzer, and R. Remmert, *Non-Archimedean analysis*, Grundlehren der Mathematischen Wissenschaften, vol. 261, Springer-Verlag, Berlin, 1984.

[8] Kevin Costello, *The partition function of a topological field theory*, J. Topol. **2** (2009), no. 4, 779–822.

[9] J. J. Duistermaat, *On global action-angle coordinates*, Comm. Pure Appl. Math. **33** (1980), no. 6, 687–706.

[10] Kenji Fukaya, *Floer homology for families—a progress report*, Integrable systems,

topology, and physics (Tokyo, 2000), Contemp. Math., vol. 309, Amer. Math. Soc., Providence, RI, 2002, pp. 33–68.

[11] ———, *Multivalued Morse theory, asymptotic analysis and mirror symmetry*, Graphs and patterns in mathematics and theoretical physics, Proc. Sympos. Pure Math., vol. 73, Amer. Math. Soc., Providence, RI, 2005, pp. 205–278.

[12] ———, *Cyclic symmetry and adic convergence in Lagrangian Floer theory*, Kyoto J. Math. **50** (2010), no. 3, 521–590.

[13] ———, *Lagrangian surgery and ridig analytic family of Floer homologies* (2009), available at http://www.math.kyoto-u.ac.jp/~fukaya/fukaya.html.

[14] Kenji Fukaya, Yong-Geun Oh, Hiroshi Ohta, and Kaoru Ono, *Lagrangian intersection Floer theory: anomaly and obstruction. Part I*, AMS/IP Studies in Advanced Mathematics, vol. 46, American Mathematical Society, Providence, RI, 2009.

[15] ———, *Lagrangian Floer theory on compact toric manifolds. I*, Duke Math. J. **151** (2010), no. 1, 23–174.

[16] ———, *Displacement of polydisks and Lagrangian Floer theory*, J. Symplectic Geom. **11** (2013), no. 2, 231–268.

[17] Alexander Givental, *A mirror theorem for toric complete intersections*, Topological field theory, primitive forms and related topics (Kyoto, 1996), Progr. Math., vol. 160, Birkhäuser Boston, Boston, MA, 1998, pp. 141–175.

[18] Mark Gross and Bernd Siebert, *Mirror symmetry via logarithmic degeneration data. I*, J. Differential Geom. **72** (2006), no. 2, 169–338.

[19] ———, *From real affine geometry to complex geometry*, Ann. of Math. (2) **174** (2011), no. 3, 1301–1428.

[20] L. Katzarkov, M. Kontsevich, and T. Pantev, *Hodge theoretic aspects of mirror symmetry*, From Hodge theory to integrability and TQFT tt*-geometry, Proc. Sympos. Pure Math., vol. 78, Amer. Math. Soc., Providence, RI, 2008, pp. 87–174.

[21] K. Kodaira, *On the structure of compact complex analytic surfaces. II*, Amer. J. Math. **88** (1966), 682–721.

[22] Maxim Kontsevich, *Homological algebra of mirror symmetry*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), Birkhäuser, Basel, 1995, pp. 120–139.

[23] Maxim Kontsevich and Yan Soibelman, *Affine structures and non-Archimedean analytic spaces*, The unity of mathematics, Progr. Math., vol. 244, Birkhäuser Boston, Boston, MA, 2006, pp. 321–385.

[24] L. Polterovich, *The surgery of Lagrange submanifolds*, Geom. Funct. Anal. **1** (1991), no. 2, 198–210.

[25] Paul Seidel, *Fukaya categories and Picard-Lefschetz theory*, Zurich Lectures in Advanced Mathematics, European Mathematical Society (EMS), Zürich, 2008.

[26] _____, *Homological mirror symmetry for the quartic surface*, available at ArXiv: 0310414.

[27] Nick Sheridan, *Homological Mirror Symmetry for Calabi-Yau hypersurfaces in projective space*, available at ArXiv:1111.0632.

[28] Andrew Strominger, Shing-Tung Yau, and Eric Zaslow, *Mirror symmetry is T-duality*, Nuclear Phys. B **479** (1996), no. 1-2, 243–259.

[29] John Tate, *Rigid analytic spaces*, Invent. Math. **12** (1971), 257–289.

[30] W. P. Thurston, *Some simple examples of symplectic manifolds*, Proc. Amer. Math. Soc. **55** (1976), no. 2, 467–468.

[31] Junwu Tu, *On the reconstruction problem in mirror symmetry*, Adv. Math. **256** (2014), 449–478, DOI 10.1016/j.aim.2014.02.005. MR3177298

[32] _____, *Homological Mirror Symmetry and "Fourier-Mukai" Transform*, available at ArXiv:1208.5912. To appear in Int. Math. Res. Not.

[33] Kenji Ueno, *Compact rigid analytic spaces with special regard to surfaces*, Algebraic geometry, Sendai, 1985, Adv. Stud. Pure Math., vol. 10, North-Holland, Amsterdam, 1987, pp. 765–794.

Department of Mathematics, Columbia University, 2990 Broadway New York, NY 10027, USA
E-mail: abouzaid@math.columbia.edu

Hyperbolic orbifolds of small volume

Mikhail Belolipetsky

Abstract. Volume is a natural measure of complexity of a Riemannian manifold. In this survey, we discuss the results and conjectures concerning n-dimensional hyperbolic manifolds and orbifolds of small volume.

Mathematics Subject Classification (2010). Primary 22E40; Secondary 11E57, 20G30, 51M25.

Keywords. Volume, Euler characteristic, hyperbolic manifold, hyperbolic orbifold, arithmetic group.

1. Volume in hyperbolic geometry

A hyperbolic manifold is an n-dimensional manifold equipped with a complete Riemannian metric of constant sectional curvature -1. Any such manifold $\mathcal{M}$ can be obtained as the quotient of the hyperbolic n-space $\mathbf{H}^n$ by a torsion-free discrete group Γ of isometries of $\mathbf{H}^n$:

$$\mathcal{M} = \mathbf{H}^n/\Gamma.$$

If we allow more generally the discrete group to have elements of finite order, then the resulting quotient space $\mathcal{O} = \mathbf{H}^n/\Gamma$ is called a *hyperbolic n-orbifold*.

We can descend the volume form from $\mathbf{H}^n$ to $\mathcal{O}$ and integrate it over the quotient space. This defines the hyperbolic volume of $\mathcal{O}$. The generalization of the Gauss-Bonnet theorem says that in even dimensions the volume is proportional to the Euler characteristic. More precisely, we have for n even:

$$\mathrm{Vol}(\mathcal{M}) = \frac{\mathrm{Vol}(\mathbf{S}^n)}{2} \cdot (-1)^{n/2}\chi(\mathcal{M}), \tag{1.1}$$

where $\mathrm{Vol}(\mathbf{S}^n)$ is the Euclidean volume of the n-dimensional unit sphere and $\chi(\mathcal{M})$ denotes the Euler characteristic. This formula generalizes to hyperbolic n-orbifolds with the orbifold Euler characteristic in place of χ. Conceptually it says that the hyperbolic volume is a topological invariant and, like for the Euler characteristic, its value is a measure of complexity of the space. In odd dimensions the Euler characteristic vanishes but the volume is still a non-trivial topological invariant that measures the complexity of $\mathcal{M}$. Indeed, the Mostow–Prasad rigidity theorem implies that every geometric invariant of a finite volume hyperbolic n-manifold (or orbifold) of dimension $n \geq 3$ is a topological invariant. One particular example of an application of the volume of the hyperbolic 3-manifolds as a measure of complexity appears in knot theory — see [15] and [16] where all knots up to a certain complexity are enumerated. Note that although many knots in the tables of Callahan–Dean–Weeks and

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Champanerkar–Kofman–Patterson have the same number of simplexes in the minimal triangulations, only very few of them share the same volume. For large volume the picture is different (see [41] for the recent results in dimension 3 and [21] for higher dimensions and other symmetric spaces), but nevertheless, in practice, hyperbolic volume has proven to be very effective in distinguishing manifolds.

The main purpose of this report is to discuss what is currently known about the simplest (i.e. *minimal volume*) hyperbolic n-manifolds and orbifolds. More information about this topic with a particular emphasis on a connection with hyperbolic reflection groups can be found in a recent survey paper by Kellerhals [32].

The minimal volume problem for hyperbolic n-orbifolds goes back to the paper of Siegel [50] where the general setup is described and the solution to the problem for $n = 2$ is given. In fact, the solution of the 2-dimensional problem can be traced back to the earlier work of Hurwitz [29], which is briefly mentioned in Siegel's paper. The qualitative solution to Siegel's problem in general was obtained by Kazhdan and Margulis in [31] (the title of [31] refers to a conjecture of Selberg about the existence of unipotent elements in non-uniform lattices which was also resolved in the same 5-page paper). We are going to come back to the discussion of the Kazhdan–Margulis theorem in Section 5, but before that we shall consider the sharp lower bounds for the volumes of *arithmetic orbifolds*.

2. Arithmeticity and volume

The group of isometries of the hyperbolic n-space is isomorphic to the real Lie group $\mathrm{PO}(n,1)$. Its subgroup of orientation preserving isometries corresponds to the identity component $H = \mathrm{PO}(n,1)^\circ$, which can be further identified with the matrix group $\mathrm{SO}_0(n,1)$ – the subgroup of $\mathrm{SO}(n,1)$ that preserves the upper half space. We shall mainly consider *orientable finite volume hyperbolic n-orbifolds*

$$\mathcal{O} = \mathbf{H}^n/\Gamma, \quad \Gamma \text{ is a lattice in } H.$$

Let $\mathbf{G}$ be an algebraic group defined over a number field k which admits an epimorphism $\phi : \mathbf{G}(k \otimes_{\mathbb{Q}} \mathbb{R})^\circ \to H$ whose kernel is compact. Then, by the Borel–Harish-Chandra theorem [13], $\phi(\mathbf{G}(\mathcal{O}_k))$ is a finite covolume discrete subgroup of H (here and further on $\mathcal{O}_k$ denotes the ring of integers of k). Such subgroups and all the subgroups of H which are commensurable with them are called *arithmetic lattices* (or *arithmetic subgroups*), and the field k is called their *field of definition*.

It can be shown that to define all arithmetic subgroups of H it is sufficient to consider only simply connected, absolutely simple k-groups $\mathbf{G}$ of absolute type $\mathrm{B}_{n/2}$, if n is even, or $\mathrm{D}_{(n+1)/2}$, if n is odd. In this case $\mathbf{G}(k \otimes_{\mathbb{Q}} \mathbb{R}) \cong \widetilde{H} \times K$, where $\widetilde{H} = \mathrm{Spin}(n,1)$ is the simply connected covering of H and K is a compact Lie group. We shall call such groups $\mathbf{G}$ and corresponding fields k *admissible*. The Godement compactness criterion implies that for $n \geq 4$ the quotient $\mathbf{H}^n/\Gamma$ is noncompact if and only if it is defined over $k = \mathbb{Q}$.

From the classification of semisimple algebraic groups [55] it follows that if n is even then $\mathbf{G}$ has to be the spinor group of a quadratic form of signature $(n,1)$ defined over a totally real field k, i.e. in even dimensions the arithmetic subgroups are commensurable with the groups of units of the quadratic forms. For odd n there is another family of arithmetic subgroups corresponding to the groups of units of appropriate Hermitian forms over quater-

nion algebras. Moreover, if $n = 7$ there is also the third type of arithmetic subgroups of H which are associated to the Cayley algebra.

The number theoretic local-global principle gives us a way to construct arithmetic lattices that is particularly suitable for the volume computations. Let $\mathrm{P} = (\mathrm{P}_v)_{v \in V_\mathrm{f}}$ be a collection of parahoric subgroups $\mathrm{P}_v \subset \mathbf{G}(k_v)$, where v runs through all finite places of k and k_v denotes the non-archimedean completion of the field (see e.g. [43, Sec. 0.5] for the definition of parahoric subgroups). The family P is called *coherent* if $\prod_{v \in V_\mathrm{f}} \mathrm{P}_v$ is an open subgroup of the finite adèle group $\mathbf{G}(\mathbb{A}_f(k))$. Following [43], the group

$$\Lambda = \mathbf{G}(k) \cap \prod_{v \in V_\mathrm{f}} \mathrm{P}_v$$

is called the *principal arithmetic subgroup* of $\mathbf{G}(k)$ associated to P. We shall also call $\Lambda' = \phi(\Lambda)$ a principal arithmetic subgroup of H. This construction is motivated by a simple observation that the integers $\mathbb{Z} = \mathbb{Q} \cap \prod_{p \text{ prime}} \mathbb{Z}_p$, where $\mathbb{Q}$ is embedded diagonally into the product of p-adic fields $\prod_{p \text{ prime}} \mathbb{Q}_p$, and can be understood as its generalization to the algebraic groups defined over number fields. We refer to the books by Platonov–Rapinchuk [42] and Witte Morris [59] for more material about arithmetic subgroups and their properties.

The Lie group H carries a Haar measure μ that is defined uniquely up to a scalar factor. We can normalize μ so that the hyperbolic volume satisfies

$$\mathrm{Vol}(\mathbf{H}^n/\Gamma) = \mu(H/\Gamma).$$

The details of this normalization procedure are explained, for instance, in Section 2.1 of [8]. If Γ is a principal arithmetic subgroup, its covolume can be effectively computed. The first computations of this kind can be traced back to the work of Smith, Minkowski and Siegel on masses of lattices in quadratic spaces. After the work of Kneser, Tamagawa and Weil these computations were brought into the framework of algebraic groups and number theory. More precisely, if Γ is an arithmetic subgroup of $\mathbf{G}$ defined over k, then its covolume can be expressed through the volume of $\mathbf{G}(\mathbb{A}_k)/\mathbf{G}(k)$ with respect to a volume form ω associated naturally to Γ, and one can relate ω to the Tamagawa measure of $\mathbf{G}(\mathbb{A}_k)$ by virtue of certain local densities. Assuming that the Tamagawa number of $\mathbf{G}$ is known, the computation of the covolume of Γ is thus reduced to the computation of these local densities. The precise expressions of this form are known as the *volume formulas*, among which we would like to mention the Gauss–Bonnet formula of Harder [26], Borel's volume formula [12], Prasad's formula [43], and its motivic extension by Gross [25]. In his paper, Harder worked out an explicit formula for the split groups $\mathbf{G}$ but in our case, if $n > 3$, the corresponding algebraic groups are never split. In Borel's influential paper the case of the semisimple groups of type A_1 is covered in full generality. This corresponds to the hyperbolic spaces of dimensions 2 and 3 and their products. Our primary interest lies in higher dimensions, where the computations can be carried out via Prasad's volume formula.

Let us recall *Prasad's formula* adapted to our setup. Let Λ be a principal arithmetic subgroup of an admissible group $\mathbf{G}/k$ associated to a coherent collection of parahoric subgroups P. Following [8, Section 2.1], assuming Λ does not contain the center of $\mathbf{G}$, we have

$$\mu(H/\Lambda') = \mathrm{Vol}(\mathbf{S}^n) \cdot \mathcal{D}_k^{\frac{1}{2}\dim(\mathbf{G})} \left(\frac{\mathcal{D}_\ell}{\mathcal{D}_k^{[\ell:k]}} \right)^{\frac{1}{2}s} \left(\prod_{i=1}^{r} \frac{m_i!}{(2\pi)^{m_i+1}} \right)^{[k:\mathbb{Q}]} \tau_k(\mathbf{G})\, \mathcal{E}(\mathrm{P}), \quad (2.1)$$

where

(i) $\mathrm{Vol}(\mathbf{S}^n) = \dfrac{2\pi^{\frac{n+1}{2}}}{\Gamma(\frac{n+1}{2})}$ is the volume of the unit sphere in $\mathbb{R}^{n+1}$;

(ii) $\mathcal{D}_K$ denotes the absolute value of the discriminant of the number field K;

(iii) ℓ is a Galois extension of k defined as in [43, 0.2] (if $\mathbf{G}$ is not a k-form of type ${}^6\mathrm{D}_4$, then ℓ is the splitting field of the quasi-split inner k-form of $\mathbf{G}$, and if $\mathbf{G}$ is of type ${}^6\mathrm{D}_4$, then ℓ is a fixed cubic extension of k contained in the corresponding splitting field; in all cases $[\ell : k] \leq 3$);

(iv) $\dim(\mathbf{G})$, r and m_i denote the dimension, rank and Lie exponents of $\mathbf{G}$:

- if n is even, then $r = \frac{n}{2}$, $\dim(\mathbf{G}) = 2r^2 + r$, and $m_i = 2i - 1$ $(i = 1, \ldots, r)$;
- if n is odd, then $r = \frac{1}{2}(n+1)$, $\dim(\mathbf{G}) = 2r^2 - r$, and $m_i = 2i - 1$ $(i = 1, \ldots, r-1)$, $m_r = r - 1$;

(v) $s = 0$ if n is even and $s = 2r - 1$ for odd dimensions (cf. [43, 0.4]);

(vi) $\tau_k(\mathbf{G})$ is the Tamagawa number of $\mathbf{G}$ over k (since $\mathbf{G}$ is simply connected and k is a number field, $\tau_k(\mathbf{G}) = 1$); and

(vii) $\mathcal{E}(\mathrm{P}) = \prod_{v \in V_f} e_v$ is an Euler product of the local densities $e_v = e(\mathrm{P}_v)$ which can be explicitly computed using Bruhat–Tits theory.

When Λ contains the center of $\mathbf{G}$ its covolume is twice the above value.

In even dimensions the right-hand side of the volume formula is related to the generalized Euler characteristic of the quotient (cf. [14, Section 4.2]) and we obtain a variant of the classical Gauss–Bonnet theorem.

If $\mathcal{O} = \mathbf{H}^n/\Gamma$ is a minimal volume hyperbolic orbifold then Γ is a *maximal lattice* in H. It is known that any maximal arithmetic subgroup Γ can be obtained as the normalizer in H of some principal arithmetic subgroup Λ, and that the index $[\Gamma : \Lambda]$ can be evaluated or estimated using Galois cohomology. We refer for more details and some related computations to the corresponding sections of [5], [6] and [8]. The upshot is that this technique allows us to study the minimal volume arithmetic hyperbolic n-orbifolds using volume formulas.

3. Minimal volume arithmetic hyperbolic orbifolds

The minimal volume 2-orbifold corresponds to the Hurwitz triangle group $\Delta(2, 3, 7)$ (cf. [50]). This group is arithmetic and defined over the cubic field $k = \mathbb{Q}[\cos(\frac{\pi}{7})]$, which follows from Takeuchi's classification of arithmetic triangle groups [53]. The smallest noncompact 2-orbifold corresponds to the modular group $\mathrm{PSL}(2, \mathbb{Z})$. In dimension 3 the minimal covolume arithmetic subgroup was found by Chinburg and Friedman [17], who used Borel's volume formula [12]. Much later Gehring, Martin and Marshall showed that this group solves Siegel's minimal covolume problem in dimension 3 [24, 38]. The noncompact hyperbolic 3-orbifold of minimal volume was found by Meyerhoff [40], it corresponds to the arithmetic Bianchi group $\mathrm{PSL}(2, \mathcal{O}_3)$, with $\mathcal{O}_3$ the ring of integers in $\mathbb{Q}[\sqrt{-3}]$. For even dimensions $n \geq 4$ the minimal volume problem for arithmetic hyperbolic n-orbifolds was solved in my paper [5] (with addendum [6]). The odd dimensional case of this problem for $n \geq 5$ was studied by Emery in his thesis [20] and appeared in our joint paper [8]. These results complete the solution of Siegel's problem for arithmetic hyperbolic n-orbifolds. We shall now review our work and discuss some corollaries.

The main results of [5, 6, 8, 20] can be summarized in two theorems:

Theorem 3.1. *For every dimension $n \geq 4$, there exists a unique orientable compact arithmetic hyperbolic n-orbifold $\mathcal{O}_0^n$ of the smallest volume. It is defined over the field $k_0 = \mathbb{Q}[\sqrt{5}]$ and has $\mathrm{Vol}(\mathcal{O}_0^n) = \omega_c(n)$.*

Theorem 3.2. *For every dimension $n \geq 4$, there exists a unique orientable noncompact arithmetic hyperbolic n-orbifold $\mathcal{O}_1^n$ of the smallest volume. It is defined over the field $k_1 = \mathbb{Q}$ and has $\mathrm{Vol}(\mathcal{O}_1^n) = \omega_{nc}(n)$.*

The values of the minimal volume are as follows:

$$\omega_c(n) = \begin{cases} \frac{4\cdot 5^{r^2+r/2}\cdot(2\pi)^r}{(2r-1)!!}\prod_{i=1}^{r}\frac{(2i-1)!^2}{(2\pi)^{4i}}\zeta_{k_0}(2i), & \text{if } n=2r,\ r \text{ even;} \\ \frac{2\cdot 5^{r^2+r/2}\cdot(2\pi)^r\cdot(4r-1)}{(2r-1)!!}\prod_{i=1}^{r}\frac{(2i-1)!^2}{(2\pi)^{4i}}\zeta_{k_0}(2i), & \text{if } n=2r,\ r \text{ odd;} \\ \frac{5^{r^2-r/2}\cdot 11^{r-1/2}\cdot(r-1)!}{2^{2r-1}\pi^r}\,L_{\ell_0|k_0}(r)\,\prod_{i=1}^{r-1}\frac{(2i-1)!^2}{(2\pi)^{4i}}\zeta_{k_0}(2i), & \text{if } n=2r-1; \end{cases} \tag{3.1}$$

$$\omega_{nc}(n) = \begin{cases} \frac{4\cdot(2\pi)^r}{(2r-1)!!}\prod_{i=1}^{r}\frac{(2i-1)!}{(2\pi)^{2i}}\zeta(2i), & \text{if } n=2r,\ r\equiv 0,\ 1(\mathrm{mod}\ 4); \\ \frac{2\cdot(2^r-1)\cdot(2\pi)^r}{(2r-1)!!}\prod_{i=1}^{r}\frac{(2i-1)!}{(2\pi)^{2i}}\zeta(2i), & \text{if } n=2r,\ r\equiv 2,\ 3(\mathrm{mod}\ 4); \\ \frac{3^{r-1/2}}{2^{r-1}}\,L_{\ell_1|\mathbb{Q}}(r)\,\prod_{i=1}^{r-1}\frac{(2i-1)!}{(2\pi)^{2i}}\zeta(2i), & \text{if } n=2r-1,\ r \text{ even;} \\ \frac{1}{2^{r-2}}\,\zeta(r)\,\prod_{i=1}^{r-1}\frac{(2i-1)!}{(2\pi)^{2i}}\zeta(2i), & \text{if } n=2r-1,\ r\equiv 1(\mathrm{mod}\ 4); \\ \frac{(2^r-1)(2^{r-1}-1)}{3\cdot 2^{r-1}}\,\zeta(r)\,\prod_{i=1}^{r-1}\frac{(2i-1)!}{(2\pi)^{2i}}\zeta(2i), & \text{if } n=2r-1,\ r\equiv 3(\mathrm{mod}\ 4). \end{cases} \tag{3.2}$$

(The fields ℓ_0 and ℓ_1 are defined by $\ell_1 = \mathbb{Q}[\sqrt{-3}]$ and ℓ_0 is the quartic field with a defining polynomial $x^4 - x^3 + 2x - 1$. The functions $\zeta_K(s)$, $L_{L|K}(s)$ and $\zeta(s)$ denote the Dedekind zeta function of a field K, the Dirichlet L-function associated to a quadratic field extension L/K, and the Riemann zeta function, respectively.)

The groups Γ_0^n and Γ_1^n can be described as the normalizers of stabilizers of integral lattices in quadratic spaces (V, f),

$$f = dx_0^2 + x_1^2 + \ldots x_n^2,$$

where for Γ_0^n we take $d = -\frac{1}{2}(1+\sqrt{5})$ if n is even and $d = (-1)^r 3 - 2\sqrt{5}$ if $n = 2r-1$ is odd, and for Γ_1^n we have $d = -1$ except for the case when $n = 2r-1$ is odd and r is even where $d = -3$. In even dimensions the stabilizers of the lattices under consideration appear to be maximal discrete subgroups so the index $[\Gamma_i^{2r} : \Lambda_i^{2r}] = 1$ and Γ_i^{2r} are principal arithmetic subgroups ($i = 0, 1$). In the odd dimensional case the index is equal to 2 except $[\Gamma_1^{2r-1} : \Lambda_1^{2r-1}] = 1$ when $r = 2m+1$ with m even. In the noncompact case the corresponding covolumes (without proof of minimality) were previously computed by Ratcliffe and Tschantz (cf. [45, 47]), who achieved this by explicitly evaluating the limit in the classical Siegel's volume formula [48, 49].

The formulas (3.1)–(3.2) look scary but they are explicit and can be applied for computation and estimation of the volumes. As an example of such computation we can study the growth of minimal volume depending on the dimension of the space. Figure 3.1 presents a graph of the logarithm of the minimal volume of compact/noncompact arithmetic orbifolds in dimensions $n < 30$. The logarithmic graph of the Euler characteristic in even dimensions has a similar shape. By analyzing this image we come up with several interesting corollaries which can be then confirmed analytically (as was done in [5] and [8]):

Corollary 3.3. *The minimal volume decreases with n till $n = 8$ (resp. $n = 17$) in the compact (resp. noncompact) case. After this it starts to grow eventually reaching a very fast super-exponential growth. For every dimension $n \geq 5$, the minimal volume of a noncompact arithmetic hyperbolic n-orbifold is smaller than the volume of any compact arithmetic hyperbolic n-orbifold. Moreover, the ratio between the minimal volumes $\mathrm{Vol}(\mathcal{O}_0^n)/\mathrm{Vol}(\mathcal{O}_1^n)$ grows super-exponentially with n.*

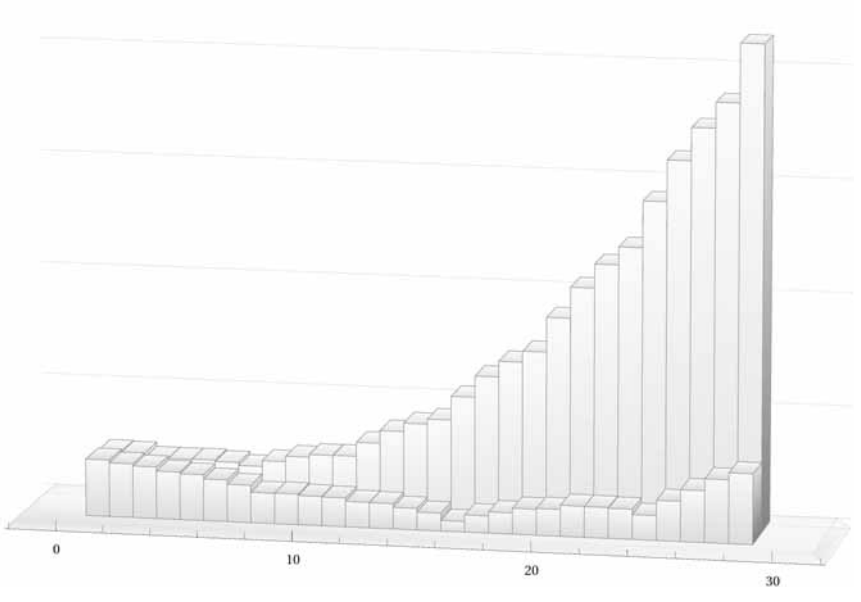


Figure 3.1. The logarithm of the minimal volume of noncompact (front) and compact (back) arithmetic hyperbolic n-orbifolds for $n = 2, 3, \ldots, 29$.

Another interesting corollary of the growth of minimal volume was obtained by Emery:

Theorem 3.4 (Emery, [19]). *For $n > 4$ there is no compact arithmetic hyperbolic n-manifold $\mathcal{M}$ with $|\chi(\mathcal{M})| = 2$.*

In particular, there do not exist arithmetically defined hyperbolic rational homology n-spheres with n even and bigger than 4. We can remark that for $n > 10$ this theorem follows from the results in [5] pertaining to Corollary 3.3, but smaller dimensions are harder and require more careful analysis of the Euler characteristic of arithmetic subgroups.

It is conjectured that all results in this section are true without assuming arithmeticity. We shall discuss this conjecture more carefully in Section 5.

4. Minimal volume manifolds and cusps

An interesting and somewhat surprising corollary of the results about the minimal volume arithmetic hyperbolic n-orbifolds is that for $n \geq 5$ the minimum is attained on *noncompact* hyperbolic n-orbifolds. In a joint work with Emery we observed that the picture becomes even more interesting when we restrict our attention to manifolds. As a result we came up with a conjecture [7]:

Conjecture 4.1. *Let $\mathcal{M}$ be a compact hyperbolic manifold of dimension $n \neq 3$. Then there exists a noncompact hyperbolic n-manifold $\mathcal{M}_1$ whose volume is smaller than the volume of $\mathcal{M}$.*

Dimension $n = 3$ is special because it is the only dimension in which we can perform hyperbolic Dehn fillings on the cusps of a finite volume noncompact manifold $\mathcal{M}_1$ to obtain compact hyperbolic manifolds of smaller volume. This follows from Thurston's Dehn surgery theorem (cf. [54, Sections 5 and 6]). Our conjecture essentially says that this is the only way to produce very small compact hyperbolic n-manifolds.

The conjecture is known to be true in dimensions $n \leq 4$ and 6. More precisely, for these n there exist noncompact hyperbolic n-manifolds $\mathcal{M}$ with $|\chi(\mathcal{M})| = 1$ [22, 46], whereas it is a general fact that the Euler characteristic of a compact hyperbolic manifold is even (cf. [44, Theorem 4.4]). The main result of [7] is a proof of the conjecture for arithmetic hyperbolic n-manifolds of dimension $n \geq 30$. In the next section we shall discuss the minimal volume conjecture (MVC), which together with this result would imply Conjecture 4.1 for $n \geq 30$, however, the above conjecture is weaker than the MVC and we hope that it might be possible to attack it directly.

The proof of the theorem in [7] is based on the results about minimal volume arithmetic hyperbolic n-orbifolds discussed in the previous section combined with a certain control over their manifold covers. For the latter we use explicit arithmetic constructions providing the upper bounds and the orbifold Euler characteristic for the lower bounds. One of our findings was that even in odd dimensions the Euler characteristic could provide an effective tool for bounding the degree of the smooth covers. Indeed, if an orbifold $\mathcal{O}$ under consideration has an even dimensional totally geodesic suborbifold $\mathcal{S}$, then the denominator of $\chi(\mathcal{S})$ gives a lower bound for the degrees of the manifold covers of $\mathcal{O}$. It appears that small volume hyperbolic n-orbifolds tend to contain many totally geodesic codimension-one suborbifolds whose Euler characteristics we can use.

The main feature of noncompact finite volume hyperbolic orbifolds is that they have infinite ends that are called *cusps*. Any such cusp is diffeomorphic to $\mathcal{N} \times [0, +\infty)$ for some closed connected flat $(n-1)$-orbifold $\mathcal{N}$. Geometry of the cusps plays a major role in the study of noncompact hyperbolic orbifolds and their volumes. For example, the cusp volume was used by Meyerhoff in his work on the noncompact minimal volume hyperbolic 3-orbifold [40], and later by Hild in the proof of minimality for hyperbolic n-orbifolds in dimensions $n \leq 9$ [27].

Long and Reid showed that any closed flat $(n-1)$-manifold $\mathcal{N}$ is diffeomorphic to a cusp cross-section of a finite volume hyperbolic n-orbifold $\mathcal{M}$ [37]. It was later proved by McReynolds that the same can be achieved with $\mathcal{M}$ being a manifold [39]. In both constructions the resulting n-orbifold or manifold can be chosen to be arithmetic. By Margulis lemma, any finite volume hyperbolic n-orbifold has a finite number of cusps (cf. [54, Proposition 5.11.1]). In the same paper Long and Reid raised a question about existence of 1-cusped hyperbolic n-manifolds for $n \geq 4$. In dimension 4 this problem was recently solved

by Kolpakov and Martelli [35], who constructed infinite families of hyperbolic 4-manifolds with any given number $k \geq 1$ of cusps. The method of Kolpakov–Martelli is specific for $n = 4$ and not applicable in higher dimensions. In particular, it is not known if there exists a 5-dimensional 1-cusped hyperbolic manifold. On the other hand, Stover [52] has shown that in dimensions $n \geq 30$ there are no 1-cusped arithmetic hyperbolic n-orbifolds (or manifolds). Following our yoga, this suggests that there should be no any 1-cusped hyperbolic n-orbifolds in high dimensions. To conclude the discussion, let us mention that (arithmetic) 1-cusped hyperbolic n-orbifolds in dimensions $n \leq 9$ appear, for example, in Hild's paper [27], and for $n = 10$ and 11 were constructed by Stover in [52]. The existence of 1-cusped orbifolds in dimensions $12 \leq n \leq 29$ is not known and there is not even a conjecture about it.

A careful reader would notice that dimension bound 30 appeared in two independent results discussed in this section. It is also the dimension bound in the celebrated Vinberg's theorem that says that there no arithmetic hyperbolic reflection groups in dimensions $n \geq 30$ [56]. There is no reason, however, to expect that any of these bounds is sharp. The coincidence can be explain by the fact that arithmetic methods work very well in higher dimensions and 30 is about the place where this starts to be noticeable. It might be possible to push down the bounds using the same methods but it would require a considerable effort and obtaining a sharp bound for any of these problems would most likely require some totally new ideas.

5. Minimal volume without arithmeticity

By the Kazhdan–Margulis theorem [31] and the subsequent work of Wang [58], we know that for $n \geq 4$ there exists a minimal volume hyperbolic n-orbifold. The classical results on uniformization of Riemann surfaces and classification of Fuchsian groups imply that the same holds true for $n = 2$, while the work of Jørgensen–Thurston [54, § 5.12] implies the same for $n = 3$. These papers also imply that there are smallest representatives in the restricted classes of compact/noncompact orbifolds or manifolds. Thus for each dimension n we have four positive numbers representing the minimal values in the volume spectra.

A *folklore conjecture*, which we call the *MVC*, says that the minimal volume is always attained on arithmetic quotient spaces. This conjecture was known for a long time for $n = 2$. (Note that the smallest volume for compact or noncompact manifolds in dimension 2 is attained also on nonarithmetic surfaces, and conjecturally this is the only dimension when it happens.) The MVC has now been completely confirmed for $n = 3$ — see [1, 23, 24, 38, 40] for the results covering each of the four cases. The smallest noncompact hyperbolic n-orbifolds for $n \leq 9$ were determined by Hild in his thesis [28] (see also [27]) and they are all arithmetic. For $n = 4$ and $n = 6$ there are examples of noncompact arithmetic hyperbolic n-manifolds $\mathcal{M}$ with $|\chi(\mathcal{M})| = 1$ which is the smallest possible [22, 46]. The smallest known compact orientable hyperbolic 4-manifolds have $\chi = 16$. They were constructed independently by Conder–Maclachlan [18] and Long [36] and can be described as finite-sheeted covers of the smallest compact arithmetic hyperbolic 4-orbifold from [5], but it is not known if there exist any smaller examples. In fact, the problem of finding a compact hyperbolic 4-manifold of minimal volume was one of the main motivations for [5]. Most of the small dimensional examples discussed here are ultimately related to hyperbolic reflection groups and we refer to the survey paper by Kellerhals for more about this connection [32].

So far, these are the only known results supporting the conjecture.

In this section we are going to discuss known general lower bounds for the volume that do not require arithmeticity. These bounds come either from a quantitative analysis of the proof of the Kazhdan–Margulis theorem or from *the Margulis lemma* and related estimates.

Let us recall the Margulis lemma for the case of hyperbolic spaces (cf. [54, Lemma 5.10.1]):

Lemma 5.1. *For every dimension n there is a constant $\mu = \mu_n > 0$ such that for every discrete group $\Gamma < \mathrm{Isom}(\mathbf{H}^n)$ and every $x \in \mathbf{H}^n$, the group $\Gamma_\mu(x) = \langle \gamma \in \Gamma \mid \mathrm{dist}(x, \gamma(x)) \le \mu \rangle$ has an abelian subgroup of finite index.*

For a given discrete group Γ, the maximal value of μ such that $\Gamma_\mu(x)$ is virtually abelian is called the *Margulis number* of $\mathbf{H}^n/\Gamma$, and the constant μ_n from the lemma is called the *Margulis constant* of $G = \mathrm{Isom}(\mathbf{H}^n)$.

If $\mathcal{M} = \mathbf{H}^n/\Gamma$ is a manifold, this result allows us to define its decomposition $\mathcal{M} = \mathcal{M}_{(0,\mu]} \cup \mathcal{M}_{[\mu,\infty)}$ into a thin and thick parts, and then to estimate from below the volume of $\mathcal{M}$ by the volume $v(\epsilon)$ of a hyperbolic ball of radius $\epsilon = \mu/2$ which embeds into the thick part $\mathcal{M}_{[\mu,\infty)}$. The case of orbifolds is much more delicate but still it is possible to use the Margulis lemma to give a lower bound for the volume. This was shown by Gelander in [9]. The resulting bound for the volume of $\mathcal{O}^n = \mathbf{H}^n/\Gamma$ is

$$\mathrm{Vol}(\mathcal{O}^n) \ge \frac{2v(0.25\epsilon)^2}{v(1.25\epsilon)}, \quad \epsilon = \min\{\frac{\mu_n}{10}, 1\}. \tag{5.1}$$

The problem with this bound is that in higher dimensions we do not have a good estimate for the value of μ_n. To my best knowledge the only appropriate general estimate can be found in [4, p. 107]. It gives

$$\mu_n \ge \frac{0.49}{16\left(1 + 2\left(\frac{4\pi}{0.49}\right)^{n(n-1)/2}\right)}. \tag{5.2}$$

In [34], Kellerhals gave a much better bound for the Margulis constant of hyperbolic n-manifolds but it is not clear if her result should extend to orbifolds.

In connection with these results it would be interesting to understand how the Margulis constant depends on the dimension of the hyperbolic space. All known bounds for μ_n decrease to zero exponentially fast when n goes to infinity, but does μ_n actually tend to zero? Note that if we define the arithmetic Margulis constant μ_n^a as the minimal value of the Margulis numbers of arithmetic hyperbolic n-orbifolds, then a positive solution to Lehmer's problem about the Mahler measure of algebraic numbers (cf. [51]) would imply that there is a uniform lower bound for μ_n^a. So conjecturally μ_n^a is bounded. The situation with μ_n is different as it is shown by the following result, which I learnt from M. Kapovich:

Proposition 5.2. *There exists a constant $C > 0$ such that $\mu_n \le \frac{C}{\sqrt{n}}$.*

Proof. The argument is based on ideas from [30].

Let us fix $\epsilon > 0$. We want to construct a discrete group of isometries $\Gamma < \mathrm{Isom}(\mathbf{H}^n)$ for which the Margulis number of $\mathbf{H}^n/\Gamma$ is less than ϵ. Let $\Gamma = F_2 = \langle f, g \rangle$, a two-generator free group. We would like to define a Γ-invariant quasi-isometric embedding of the Cayley graph T of Γ into $\mathbf{H}^n$ such that the generators of Γ act by isometries with small displacement. The graph T is a regular tree of degree four whose edges can be labeled by

the generators f, g and their inverses (starting from the root of T). Let us map the root to $p_0 \in \mathbf{H}^n$ and embed the edge corresponding to f as a geodesic segment $[p_0, p_1]$ of length ϵ. Now choose a geodesic through p_1 orthogonal to $[p_0, p_1]$ and map the g-edge adjacent to p_1 as an ϵ-segment $[p_1, p_2]$ of this geodesic. We can continue this process inductively each time choosing a geodesic which is orthogonal to the subspace containing the previously embedded edges. The process terminates when we get to p_n as n is the dimension of the space. Along the way we have defined the action of the generators f, g on the points $p_0, p_1, \dots, p_n$ which can be extended to isometry of $\mathbf{H}^n$. Now use these isometries to embed the rest of the tree. The construction gives an embedding $\rho : T \to \mathbf{H}^n$ and an isometric action of Γ on $\mathbf{H}^n$ leaving invariant the image $\rho(T)$. It remains to check that ρ is a quasi-isometric embedding.

We need to estimate the distance in $\mathbf{H}^n$ between the images of different vertices $x, y \in T$ and check that it satisfies the quasi-isometric property with respect to $\mathrm{dist}_T(x, y)$. Let $x = p_0$ and $y = p_m$ for some $m > n$ — this is a typical case and all the other easily reduce to it. Let $b_i = \mathrm{dist}_{\mathbf{H}^n}(p_0, p_i)$, so $b_0 = \epsilon$ and $b_n = \mathrm{dist}_{\mathbf{H}^n}(p_0, p_n)$. By the hyperbolic Pythagorean theorem we have

$$\cosh(b_{i+1}) = \cosh(b_i)\cosh(\epsilon),$$

therefore, $\cosh(b_n) = (\cosh(\epsilon))^n$. We now can use the disjoint bisectors test (cf. [30, Section 3]). If the length b_n is bigger than a certain constant (which can be taken $= 2.303$ as in the proof of Lemma 3.2 [loc. cit.]), then the bisectors of $[p_0, p_n]$ and $[p_n, p_{2n}]$ do not intersect and hence are separated by the distance $\delta = \delta(\rho) > 0$. Hence we have $\mathrm{dist}_{\mathbf{H}^n}(p_0, p_m) \geq \delta[m/n]$, as the geodesic joining p_0 and p_m will have to intersect all the intermediate bisectors. It follows that ρ is a quasi-isometry, provided

$$b_n \geq 2.303.$$

It remains to apply [30, Lemma 2.2], which shows that the isometric action of Γ on $\mathbf{H}^n$ is discrete and hence

$$\mu_n \leq \mu(\mathbf{H}^n/\Gamma) = \epsilon.$$

We conclude with an estimate for the constants:

$$\cosh(b_n) = (\cosh(\epsilon))^n \geq \cosh(2.303).$$

When $\epsilon \to 0$, we have $\cosh(\epsilon) \approx 1 + \frac{\epsilon^2}{2}$, so there exists $C > 0$ (can take

$$C = \sqrt{2\log(\cosh(2.303))} = 1.799\dots$$

if n is sufficiently large) such that if $\epsilon \geq \frac{C}{\sqrt{n}}$ then ρ is a quasi-isometric embedding. □

The groups in Proposition 5.2 have infinite covolume. It is tempting to try a similar argument on non-arithmetic lattices with small systole. Such lattices can be obtained by the inbreeding construction found by Agol for $n = 4$ [3] and generalized to higher dimensions in [10] and [11]. However, as it stands for now, it is not clear how to make this work and the problem remains open. The other important open problem is to find a better lower bound for μ_n. We can speculate that some kind of polynomially decreasing lower bound should exist.

The other approach is to bound the volume via quantitative version of the Kazhdan–Margulis theorem. It goes back to the paper [57] by Wang, who found an explicit lower

bound for the radius of a ball embedded into a Zassenhaus neighborhood of a Lie group G (with $G \cong \mathrm{PO}(n,1)^\circ$ in our case). Adeboye and Wei combined this bound with a bound for the sectional curvature of G to obtain an explicit lower bound for the volume [2]. Their main result is

$$\mathrm{Vol}(\mathcal{O}^n) \geq \frac{2^{[\frac{6-n}{4}]}\pi^{[\frac{n}{4}]}(n-2)!(n-4)!\cdots 1}{(2+9n)^{[\frac{n^2+n}{4}]}\Gamma(\frac{n^2+n}{4})} \int_0^{\min[0.08\sqrt{2+9n},\pi]} \sin^{\frac{n^2+n-2}{2}} \rho \; d\rho. \quad (5.3)$$

This lower bound decreases super-exponentially when n goes to infinity and it is currently the best known general lower bound for the volume.

For the noncompact hyperbolic orbifolds and manifolds there is also another approach to bounding the volume. It is based on estimating the density of Euclidean sphere packings associated to cusps. It was used in the papers of Meyerhoff, Adams, and Hild that were mentioned above. For the case of arbitrarily large dimension n, Kellerhals applied this method to obtain the best available lower bound for the volume of non-compact hyperbolic n-manifolds [33]:

$$\mathrm{Vol}(\mathcal{M}_1^n) \geq m\frac{2n}{n(n+1)}\nu_n, \quad (5.4)$$

where m is the number of cusps of the manifold $\mathcal{M}_1^n$ and ν_n denotes the volume of the ideal regular simplex in $\mathbf{H}^n$. Note that by Milnor's formula for large n we have $\nu_n \approx \frac{e\sqrt{n}}{n!}$ and hence again we have a lower bound that decreases super-exponentially with n.

In conclusion, we see that for large dimensions there is a very large gap between the known (*super-exponentially decreasing*) bound and the conjectural (*super-exponentially increasing*) values of the minimal volume. This gap highlights our limited understanding of the complexity and structure of high-dimensional hyperbolic manifolds and orbifolds, especially the non-arithmetic ones. I hope that the future research will shed more light onto this problem.

Acknowledgements. The author is supported by a CNPq research grant. I would like to thank Vincent Emery for the fruitful collaboration throughout the years which gave rise to an essential part of the results that are considered in this report. I thank Misha Kapovich for allowing me to include Proposition 5.2. I also thank Matthew Stover for the comments on a preliminary version of the paper.

References

[1] Colin C. Adams, *The noncompact hyperbolic* 3*-manifold of minimal volume*, Proc. Amer. Math. Soc. **100** (1987), no. 4, 601–606. MR 894423 (88m:57018)

[2] Ilesanmi Adeboye and Guofang Wei, *On volumes of hyperbolic orbifolds*, Algebr. Geom. Topol. **12** (2012), no. 1, 215–233. MR 2916274

[3] Ian Agol, *Systoles of hyperbolic 4-manifolds*, preprint, 2006, arXiv:math/0612290.

[4] Werner Ballmann, Mikhael Gromov, and Viktor Schroeder, *Manifolds of nonpositive curvature*, Progress in Mathematics, vol. 61, Birkhäuser Boston Inc., Boston, MA, 1985. MR 823981 (87h:53050)

[5] Mikhail Belolipetsky, *On volumes of arithmetic quotients of* $\mathrm{SO}(1,n)$, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **3** (2004), no. 4, 749–770. MR 2124587 (2005k:11080)

[6] ———, *Addendum to: "On volumes of arithmetic quotients of* $\mathrm{SO}(1,n)$*"* [Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **3** (2004), no. 4, 749–770; mr2124587], Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **6** (2007), no. 2, 263–268. MR 2352518 (2008f:11048)

[7] Mikhail Belolipetsky and Vincent Emery, *Hyperbolic manifolds of small volume*, preprint arXiv:1310.2270 [math.MG].

[8] ———, *On volumes of arithmetic quotients of* $\mathrm{PO}(n,1)^\circ$*,* n *odd*, Proc. Lond. Math. Soc. (3) **105** (2012), no. 3, 541–570. MR 2974199

[9] Mikhail Belolipetsky, Tsachik Gelander, Alexander Lubotzky, and Aner Shalev, *Counting arithmetic lattices and surfaces*, Ann. of Math. (2) **172** (2010), no. 3, 2197–2221. MR 2726109 (2011i:11150)

[10] Mikhail V. Belolipetsky and Scott A. Thomson, *Systoles of hyperbolic manifolds*, Algebr. Geom. Topol. **11** (2011), no. 3, 1455–1469. MR 2821431 (2012k:53072)

[11] Nicolas Bergeron, Frédéric Haglund, and Daniel T. Wise, *Hyperplane sections in arithmetic hyperbolic manifolds*, J. Lond. Math. Soc. (2) **83** (2011), no. 2, 431–448. MR 2776645 (2012f:57037)

[12] A. Borel, *Commensurability classes and volumes of hyperbolic* 3*-manifolds*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **8** (1981), no. 1, 1–33. MR 616899 (82j:22008)

[13] Armand Borel and Harish-Chandra, *Arithmetic subgroups of algebraic groups*, Ann. of Math. (2) **75** (1962), 485–535. MR 0147566 (26 #5081)

[14] Armand Borel and Gopal Prasad, *Finiteness theorems for discrete subgroups of bounded covolume in semi-simple groups*, Inst. Hautes Études Sci. Publ. Math. (1989), no. 69, 119–171. MR 1019963 (91c:22021)

[15] Patrick J. Callahan, John C. Dean, and Jeffrey R. Weeks, *The simplest hyperbolic knots*, J. Knot Theory Ramifications **8** (1999), no. 3, 279–297. MR 1691433 (2000c:57005)

[16] Abhijit Champanerkar, Ilya Kofman, and Eric Patterson, *The next simplest hyperbolic knots*, J. Knot Theory Ramifications **13** (2004), no. 7, 965–987. MR 2101238 (2005k:57010)

[17] Ted Chinburg and Eduardo Friedman, *The smallest arithmetic hyperbolic three-orbifold*, Invent. Math. **86** (1986), no. 3, 507–527. MR 860679 (88a:22022)

[18] Marston Conder and Colin Maclachlan, *Compact hyperbolic 4-manifolds of small volume*, Proc. Amer. Math. Soc. **133** (2005), no. 8, 2469–2476 (electronic). MR 2138890 (2006d:57025)

[19] Vincent Emery, *On compact hyperbolic manifolds of Euler characteristic two*, Algebr. Geom. Topol., to appear, preprint arXiv:1304.3509 [math.GT].

[20] ———, *Du volume des quotients arithmétiques de l'espace hyperbolique*, Ph.D. thesis, University of Fribourg, 2009.

[21] ———, *Arbitrarily large families of spaces of the same volume*, Geom. Dedicata **160** (2012), 313–320. MR 2970057

[22] Brent Everitt, John G. Ratcliffe, and Steven T. Tschantz, *Right-angled Coxeter polytopes, hyperbolic six-manifolds, and a problem of Siegel*, Math. Ann. **354** (2012), no. 3, 871–905. MR 2983072

[23] David Gabai, Robert Meyerhoff, and Peter Milley, *Minimum volume cusped hyperbolic three-manifolds*, J. Amer. Math. Soc. **22** (2009), no. 4, 1157–1215. MR 2525782 (2011a:57031)

[24] Frederick W. Gehring and Gaven J. Martin, *Minimal co-volume hyperbolic lattices. I. The spherical points of a Kleinian group*, Ann. of Math. (2) **170** (2009), no. 1, 123–161. MR 2521113 (2010h:57029)

[25] Benedict H. Gross, *On the motive of a reductive group*, Invent. Math. **130** (1997), no. 2, 287–313. MR 1474159 (98m:20060)

[26] G. Harder, *A Gauss-Bonnet formula for discrete arithmetically defined groups*, Ann. Sci. École Norm. Sup. (4) **4** (1971), 409–455. MR 0309145 (46 #8255)

[27] Thierry Hild, *The cusped hyperbolic orbifolds of minimal volume in dimensions less than ten*, J. Algebra **313** (2007), no. 1, 208–222. MR 2326144 (2008g:57038)

[28] ———, *Cusped hyperbolic orbifolds of minimla volume in dimesnions less than 11*, Ph.D. thesis, University of Fribourg, 2007.

[29] A. Hurwitz, *Ueber algebraische Gebilde mit eindeutigen Transformationen in sich*, Math. Ann. **41** (1893), no. 3, 403–442. MR 1510753

[30] Michael Kapovich, *Representations of polygons of finite groups*, Geom. Topol. **9** (2005), 1915–1951 (electronic). MR 2175160 (2006g:20069)

[31] D. A. Každan and G. A. Margulis, *A proof of Selberg's hypothesis*, Mat. Sb. (N.S.) **75 (117)** (1968), 163–168. MR 0223487 (36 #6535)

[32] Ruth Kellerhals, *Hyperbolic orbifolds of minimal volume*, preprint 2013, to appear in the F. Gehring memorial volume, CMFT.

[33] ———, *Volumes of cusped hyperbolic manifolds*, Topology **37** (1998), no. 4, 719–734. MR 1607720 (99c:57039)

[34] ———, *On the structure of hyperbolic manifolds*, Israel J. Math. **143** (2004), 361–379. MR 2106991 (2005i:53038)

[35] Alexander Kolpakov and Bruno Martelli, *Hyperbolic four-manifolds with one cusp*, Geom. Funct. Anal. **23** (2013), no. 6, 1903–1933. MR 3132905

[36] Cormac Long, *Small volume closed hyperbolic 4-manifolds*, Bull. Lond. Math. Soc. **40** (2008), no. 5, 913–916. MR 2439657 (2009f:57030)

[37] D. D. Long and A. W. Reid, *All flat manifolds are cusps of hyperbolic orbifolds*, Algebr. Geom. Topol. **2** (2002), 285–296 (electronic). MR 1917053 (2003e:57029)

[38] T. H. Marshall and G. J. Martin, *Minimal co-volume hyperbolic lattices, II: Simple torsion in a Kleinian group*, Ann. of Math. (2) **176** (2012), no. 1, 261–301. MR 2925384

[39] D. B. McReynolds, *Controlling manifold covers of orbifolds*, Math. Res. Lett. **16** (2009), no. 4, 651–662. MR 2525031 (2010i:57042)

[40] Robert Meyerhoff, *The cusped hyperbolic* 3*-orbifold of minimum volume*, Bull. Amer. Math. Soc. (N.S.) **13** (1985), no. 2, 154–156. MR 799800 (87b:22022)

[41] Christian Millichap, *Factorial growth rates for the number of hyperbolic 3-manifolds of a given volume*, Proc. Amer. Math. Soc., to appear, preprint arXiv:1209.1042 [math.GT].

[42] Vladimir Platonov and Andrei Rapinchuk, *Algebraic groups and number theory*, Pure and Applied Mathematics, vol. 139, Academic Press Inc., Boston, MA, 1994, Translated from the 1991 Russian original by Rachel Rowen. MR 1278263 (95b:11039)

[43] Gopal Prasad, *Volumes of* S*-arithmetic quotients of semi-simple groups*, Inst. Hautes Études Sci. Publ. Math. (1989), no. 69, 91–117, With an appendix by Moshe Jarden and the author. MR 1019962 (91c:22023)

[44] John G. Ratcliffe, *Chapter 17 - hyperbolic manifolds*, Handbook of Geometric Topology (R.J. Daverman and R.B. Sher, eds.), North-Holland, Amsterdam, 2001, pp. 899–920.

[45] John G. Ratcliffe and Steven T. Tschantz, *Volumes of integral congruence hyperbolic manifolds*, J. Reine Angew. Math. **488** (1997), 55–78. MR 1465367 (99b:11076)

[46] ———, *The volume spectrum of hyperbolic 4-manifolds*, Experiment. Math. **9** (2000), no. 1, 101–125. MR 1758804 (2001b:57048)

[47] ———, *On volumes of hyperbolic Coxeter polytopes and quadratic forms*, Geom. Dedicata **163** (2013), 285–299. MR 3032695

[48] Carl Ludwig Siegel, *Über die analytische Theorie der quadratischen Formen*, Ann. of Math. (2) **36** (1935), no. 3, 527–606. MR 1503238

[49] ———, *Über die analytische Theorie der quadratischen Formen. II*, Ann. of Math. (2) **37** (1936), no. 1, 230–263. MR 1503276

[50] ———, *Some remarks on discontinuous groups*, Ann. of Math. (2) **46** (1945), 708–718. MR 0014088 (7,239c)

[51] Chris Smyth, *The Mahler measure of algebraic numbers: a survey*, Number theory and polynomials, London Math. Soc. Lecture Note Ser., vol. 352, Cambridge Univ. Press, Cambridge, 2008, pp. 322–349. MR 2428530 (2009j:11172)

[52] Matthew Stover, *On the number of ends of rank one locally symmetric spaces*, Geom. Topol. **17** (2013), no. 2, 905–924. MR 3070517

[53] Kisao Takeuchi, *Arithmetic triangle groups*, J. Math. Soc. Japan **29** (1977), no. 1, 91–106. MR 0429744 (55 #2754)

[54] William P. Thurston, *The geometry and topology of 3-manifolds*, 1979, Princeton Univ., online version at: http://www.msri.org/publications/books/gt3m/.

[55] J. Tits, *Classification of algebraic semisimple groups*, Algebraic Groups and Discontinuous Subgroups (Proc. Sympos. Pure Math., Boulder, Colo., 1965) (Providence, R.I., 1966), Amer. Math. Soc., 1966, pp. 33–62. MR 0224710 (37 #309)

[56] È. B. Vinberg, *The nonexistence of crystallographic reflection groups in Lobachevskiĭ spaces of large dimension*, Funktsional. Anal. i Prilozhen. **15** (1981), no. 2, 67–68. MR 617472 (83d:51026)

[57] Hsien-Chung Wang, *Discrete nilpotent subgroups of Lie groups*, J. Differential Geometry **3** (1969), 481–492. MR 0260930 (41 #5550)

[58] Hsien Chung Wang, *Topics on totally discontinuous groups*, Symmetric spaces (Short Courses, Washington Univ., St. Louis, Mo., 1969–1970), Dekker, New York, 1972, pp. 459–487. Pure and Appl. Math., Vol. 8. MR 0414787 (54 #2879)

[59] Dave Witte Morris, *Introduction to arithmetic groups*, preliminary version v5, arXiv: math/0106063 [math.DG].

IMPA Estrada Dona Castorina, 110, 22460-320 Rio de Janeiro, Brazil
E-mail: mbel@impa.br

Einstein 4-manifolds and singularities

Olivier Biquard

Abstract. In this note we report on recent progress on the desingularization of real Einstein 4-manifolds. A new type of obstruction is introduced, with applications to the compactification of the moduli space of Einstein metrics, and to the correspondence between conformal metrics in dimension d and asymptotically hyperbolic Einstein metrics in dimension $d+1$.

Mathematics Subject Classification (2010). Primary 53C25; Secondary 53A30.

Keywords. Einstein metric, conformal metric, gravitational instantons, AdS/CFT correspondance.

1. The Einstein equation

The Einstein equation reads

$$\mathrm{Ric}(g) = \Lambda g,$$

where g is a metric on a manifold M^n (in local coordinates $g = \sum g_{ij} dx^i dx^j$), Λ is a real number called the cosmological constant, and $\mathrm{Ric}(g) = \sum R_{ij} dx^i dx^j$ is the Ricci tensor of the metric g.

Of course the equation comes from general relativity, in which case the manifold is 4-dimensional and the metric g is Lorentzian, that is of signature (1,3). Here, we consider the case a Riemannian Einstein metric (positive signature), which has deep connections with geometry and topology, as is illustrated by the situation in low dimension that we now review briefly.

In dimension 2, a metric is Einstein if it has constant curvature equal to λ; there always exists an Einstein metric on a compact Riemann surface (this is equivalent to the uniformization theorem), and one has the dichotomy

- $\Lambda > 0$: M is a sphere;
- $\Lambda = 0$: M is a torus;
- $\Lambda < 0$: M is a surface of genus $g \geq 2$.

In dimension 3, again a metric is Einstein if it has constant curvature equal to 2Λ, so we have a similar dichotomy between spherical, flat or hyperbolic geometry according to the sign of Λ. The question of understanding which compact 3-manifolds carry an Einstein metric is completely understood, and deeply connected with the topology: the case $\Lambda > 0$ is that of a 3-sphere (and its finite quotients), and it is related to the Poincaré conjecture (proved by Perelman) saying that a compact simply connected 3-manifold is a 3-sphere—this can be

Proceedings of the International Congress of Mathematicians, Seoul, 2014

phrased by saying that it carries a constant curvature metric, and the Einstein equation plays an important role in the proof through its heat flow, the Ricci flow

$$\frac{dg}{dt} = -2\,\mathrm{Ric}(g).$$

The Ricci flow has been more generally used by Perelman to prove Thurston's geometrization conjecture, according to which any compact 3-manifold is decomposed into pieces, each of which carries one of eight homogeneous geometries (including the three constant curvature ones).

In higher dimension, it is no more true that an Einstein metric has constant curvature: the Ricci tensor is just a part of the Riemannian curvature, which contains another component (the Weyl curvature). The questions of existence and uniqueness are far from being solved. In dimension at least 5, there is no known obstruction to the existence of Einstein metrics. In dimension 4, the situation is more interesting: there is a strong relation between Einstein metrics and topology: this can be illustrated by the Hitchin-Thorpe inequality between the Euler characteristic χ and the signature τ of a compact Einstein 4-manifold:

$$2\chi(M) \geq 3|\tau(M)|. \tag{1.1}$$

This gives a topological restriction on the manifold M. More subtle obstructions are based on Gromov's idea of minimal volume (Besson-Courtois-Gallot) or on Seiberg-Witten theory (LeBrun), see the nice survey of LeBrun [16] and the references there.

The Riemannian Einstein equation is a nonlinear elliptic equation (transversely to the action of the group of diffeomorphisms), and the linearization L is a selfadjoint operator, see for example [5]. This means that one cannot extract much information on the deformations of a solution:

- either $\ker L = 0$, then the solution is rigid;
- either $\ker L \neq 0$, then there are infinitesimal deformations, but there is a space $\mathrm{coker}\, L = \ker L$ of the same dimension of obstructions, so one cannot say anything in general on the local structure of the deformation space. For example there is no known bound on the dimension of the moduli space of Einstein metrics on a given manifold.

Except in the case of special structures (Kähler or other special holonomies like quaternion-Kähler, hyper-Kähler, etc.) there is no general method to produce Einstein metrics (running the Ricci flow is of course a method, but it remains very difficult to analyze in higher dimension). Things are better for Einstein metrics on manifolds with boundary: it turns out that there exists a natural boundary problem for Einstein metrics, on which some general features of Einstein metrics can be tested, and which has its own geometric interest, in relation with conformal geometry. We now explain these ideas which originated in the work of Fefferman and Graham [11].

So let now (M, g) be a manifold with boundary $\partial M = X$, and choose on M a defining function x of X, so that $x > 0$ in the interior of M, and vanishes at first order over X. Given a metric γ on X, we consider metrics g in the interior of M such that, when $x \to 0$,

$$g \sim \frac{dx^2 + \gamma}{x^2}. \tag{1.2}$$

This behaviour depends only on the conformal class $[\gamma]$ of g: indeed, if γ is transformed into $\varphi^2\gamma$, then for $\tilde{x} = \varphi x$ one has $\frac{dx^2+\gamma}{x^2} = \frac{d\tilde{x}^2+\varphi^2\gamma}{\tilde{x}^2} +$ l.o.t. The conformal metric $[\gamma]$ is called the conformal infinity of g.

For example, if g is the hyperbolic metric on the ball, then the conformal infinity is the standard conformal metric on the boundary sphere. More generally, the behaviour (1.2) implies that the sectional curvature of g goes to -1 when $x \to 0$, hence the name of these metrics, which are called asymptotically hyperbolic (AH).

Dirichlet problem at infinity. Given a conformal class $[\gamma]$ on X, find an AH Einstein metric g in M such that the conformal infinity of g is $[\gamma]$.

The motivation of the original work of Fefferman and Graham is the study of conformal geometry through the corresponding Einstein metrics. The idea is that the formal development of g near the boundary captures invariant conformal properties of γ. This perspective was very fruitful, see [6, 12]. The correspondance received a lot of attention because it underlies a physical correspondance, the AdS/CFT correspondance [17, 24].

The global problem is well-behaved: when the metric g is non degenerate (meaning that the linearization of the problem has trivial L^2 kernel, which often happens), then one can fill a small deformation of $[\gamma]$ by a small deformation of g. This was first observed by Graham and Lee [13]. Important ideas to solve the problem were introduced by Anderson (see later in the text), but the main difficulty remains to analyze the compactness problem: is the map which associates to the Einstein metric g its conformal infinity $[\gamma]$ proper ? it is clear that such a property, together with the nice local deformation property, enables to solve the Dirichlet problem by a continuity method.

2. Compactness

So we now pass to compactness problems. We specialize to dimension 4. There is a very good compactness result on Einstein metrics, which was obtained by Anderson [1] and by Bando, Kasue and Nakajima [3].

Theorem 2.1. *Suppose (M_i, g_i) is a sequence of compact Einstein 4-manifolds, with cosmological constant ± 1 or 0, satisfying the following hypothesis:*

(1) *the diameter of (M_i, g_i) is bounded above;*

(2) *the volume of (M_i, g_i) is bounded from below;*

(3) *the L^2 norm of the curvature, $\int_{M_i} |R(g_i)|^2 dvol(g_i)$, is bounded above.*

Then a subsequence (M_i, g_i) converges for the Gromov-Hausdorff distance to a 4-orbifold (M_0, g_0) with isolated orbifold singularities. The convergence is C^∞ outside the singularities.

Moreover, for each singularity, there is a rescaling $\frac{g_i}{t_i}$ with $t_i \to 0$ such that $(M_i, \frac{g_i}{t_i})$ converges to a noncompact Ricci flat 4-manifold which is Asymptotically Locally Euclidean (ALE), that is it has one end and this end is asymptotic to the flat metric on $\mathbb{R}^4/\Gamma$ for some finite subgroup $\Gamma \subset SO_4$.

There has been a lot of progress recently to understand the limits of Einstein manifolds in higher dimension, see the article by Naber in the same volume [18].

The first hypothesis of the theorem guarantees that there is no cusp formation; the second hypothesis that there is no collapsing on a lower dimensional space; the third hypothesis is topological, because for an Einstein metric g on a compact 4-manifold M, one has $\frac{1}{8\pi^2} \int_M |R(g)|^2 d\,\mathrm{vol}(g) = \chi(M)$.

The ALE spaces which appear at the limit are the "bubbles" of the problem. This notion of bubble appears similarly in a lot of geometric problems (pseudo-holomorphic disks, instantons, harmonic maps, etc.). Similarly to these problems, another bubble can appear where a singularity forms, and one gets a tree of bubbles: the smooth ALE space mentioned in the statement is the deepest bubble.

A basic problem in understanding the possible limits of Einstein 4-manifolds is to classify the possible bubbles, that is the Ricci flat ALE 4-manifolds. There is a well-known family of hyper-Kähler (hence Ricci flat) ALE 4-manifolds (also called gravitational instantons), constructed by Kronheimer [14], who also classified all hyper-Kähler ALE 4-manifolds [15]. The finite subgroups of SO_4 which appear are the finite subgroups of SU_2. Also some cyclic subgroups of SO_4 which are not contained into a SU_2 appear as finite quotients of Kronheimer's ALE spaces. It is an old open important question whether all simply connected Ricci flat ALE 4-manifolds are hyper-Kähler (and therefore one of Kronheimer's spaces). Nakajima [19] proved that if one adds the condition that the manifold is spin for a spin structure which is also ALE in some sense, then the answer is yes.

The simplest example of a Ricci flat ALE space is the Eguchi-Hanson space [10]: topologically it is T^*S^2. The Eguchi-Hanson metric g_{EH} is asymptotic to the flat metric on $\mathbb{R}^4/\mathbb{Z}_2$. Actually T^*S^2 with the zero section removed is diffeomorphic to $(\mathbb{R}^4 \setminus \{0\})/\mathbb{Z}_2$; from the complex geometry point of view it is a desingularization of the A_1 singularity $\mathbb{C}^2/\mathbb{Z}_2$; all Kronheimer's spaces are deformations of desingularizations of the Kleinian singularities, that is of $\mathbb{C}^2/\Gamma$ for Γ a finite subgroup of SU_2. In this way one gets a short list of singularities (A_k, D_k, E_6, E_7 and E_8).

The kind of degeneration described in theorem 2.1 does occur. Actually Kähler geometry provides lots of examples. The first one [20, 23] was the singular Kummer surface (M_0, g_0), with

$$M_0 = \mathbb{T}^4/\mathbb{Z}_2,$$

a quotient of the 4-torus by an involution with 16 singular points of type $\mathbb{C}^2/\mathbb{Z}_2$, and g_0 is the flat metric. Then there is a family (M_t, g_t) of smooth K3 surfaces with their Ricci flat metrics g_t (coming from Yau's solution of the Calabi conjecture), which degenerate to (M_0, g_0) exactly in the way described by the theorem. Moreover, one can describe quite concretely the behaviour of g_t when $t \to 0$. Consider the two following regions in M_0 and in the Eguchi-Hanson space:

1. near a singular point $p_0 \in M_0$, note r the radius from p_0, the region

$$A_t = \{t^{\frac{1}{4}} \leq r \leq 2t^{\frac{1}{4}}\};$$

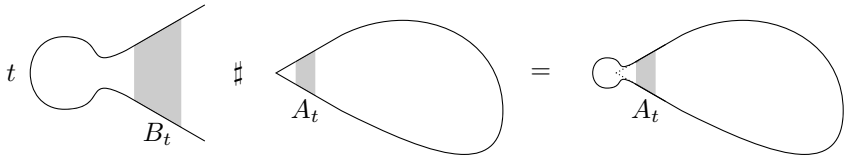
2. at the end of the metric g_{EH}, which is asymptotic to the cone $\mathbb{C}^2/\mathbb{Z}_2$, the region (where R is the radius near infinity)

$$B_t = \{t^{-\frac{1}{4}} \leq R \leq 2t^{-\frac{1}{4}}\}.$$

(Actually g_{EH} is close to the flat cone metric by a factor $O(R^{-4})$).

The homothety h_t of factor $\sqrt{t}$ identifies B_t with A_t, and sends the metric tg_{EH} to a metric which is very close to g_0 when $t \to 0$. So we can construct a new manifold M with a new metric $g_0 \sharp t g_{\mathrm{EH}}$ by gluing at each singular point the region $(\{r \geq t^{\frac{1}{4}}\}, g_0)$ in M_0 with the region $(\{R \leq 2t^{-\frac{1}{4}}\}, tg_{\mathrm{EH}})$ in the Eguchi-Hanson space, identifying A_t and B_t by h_t and

interpolating between the (very close) metrics g_0 and tg_{EH} on A_t. The process is illustrated by the figure below.



The metric $g_0 \sharp t g_{\mathrm{EH}}$ does not satisfy any more the Einstein equation in the damage area, but one can prove that it is indeed a very good approximation of g_t. In particular it illustrates well the behaviour of g_t when $t \to 0$: on one hand, $g_0 \sharp t g_{\mathrm{EH}} \to g_0$, on the other hand $\frac{1}{t}(g_0 \sharp t g_{\mathrm{EH}}) \to g_{\mathrm{EH}}$.

The compactness theorem 2.1 says basically that all the limits arise in this way, but, as mentioned before, there is no classification of the possible Ricci flat ALE spaces at the limit. In the sequel, we will see that it is not true that any 4-orbifold with a singular Einstein metric can be approximated by smooth Einstein metrics in a similar way. This leads to new restrictions on the compactification of the moduli space of Einstein metrics.

3. Desingularization

It is a fundamental algebraic fact that the 2-forms in dimension 4 decompose into selfdual and antiselfdual 2-forms:

$$\Omega^2 = \Omega_+ \oplus \Omega_-. \tag{3.1}$$

The Riemannian curvature tensor can be seen as a symmetric endomorphism of Ω^2. Therefore it decomposes on (3.1), and the various components are

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_+ & \mathrm{Ric}_0 \\ \mathrm{Ric}_0 & \mathbf{R}_- \end{pmatrix}. \tag{3.2}$$

Moreover, $\mathbf{R}_\pm$ decompose into a scalar part and a trace free part, which can be identified with the Weyl tensor W:

$$\mathbf{R}_\pm = \frac{\mathrm{Scal}}{12} \pm W_\pm. \tag{3.3}$$

We now start from an Einstein 4-orbifold (M_0, g_0), which can be compact or AH. We consider only the case of the simplest singularity $\mathbb{R}^4/\mathbb{Z}_2$. Let p_0 be a singular point (so of type $\mathbb{R}^4/\mathbb{Z}_2$). To simplify the statements, we assume that there is only one point, but the results are unchanged if there are several ones.

The following says that there is an obstruction to the existence of a sequence of metrics which desingularize g_0:

Theorem 3.1 ([8]). *Suppose that a sequence of Einstein manifolds (M_i, g_i) converges to a non degenerate (M_0, g_0), in such a way that g_i is close to $g_0 \sharp t_i g_{\mathrm{EH}}$ for a sequence of real numbers $t_i \to 0$. Then*

$$\det \mathbf{R}_+^{g_0}(p_0) = 0. \tag{3.4}$$

Here 'close' in the theorem refers to some weighted $C^{1,\alpha}$ Hölder norm.

In particular, the spaces of constant curvature $\mathbf{R}_\pm = \frac{\mathrm{Scal}}{12}$, so if the curvature is nonzero then (3.4) cannot be satisfied. It follows that spherical or hyperbolic orbifolds cannot be limits of Einstein manifolds as in the theorem:

Corollary 3.2 ([8]). *Suppose (M_0^4, g_0) satisfies the same hypothesis and has constant curvature ± 1. Then (M_0, g_0) is not the limit of a sequence of Einstein manifolds (M_i, g_i) as in theorem 3.1.*

For example, the corollary applies to the round metric on $S^4/\mathbb{Z}_2$, where the action of $\mathbb{Z}_2$ has two fixed points (the two poles), or to $\mathbb{B}^4/\mathbb{Z}_2$, the quotient of the hyperbolic 4-ball by $\mathbb{Z}_2$; it was already known that there is no U_2-invariant desingularization (U_2-invariant Einstein metrics in dimension 4 are explicitly understood).

Of course the corollary is still a partial result: a stronger result would be: if $(M_i, g_i) \to (M_0, g_0)$ in the Gromov-Hausdorff sense, and at each singularity a rescaling $(M_i, \frac{g_i}{t_i})$ converges Gromov-Hausdorff to the Eguchi-Hanson space, then the obstruction (3.4) is satisfied, and in particular the limit cannot be spherical or hyperbolic. This statement requires to strenghten the convergence of the metric to get the hypothesis of the theorem. Nevertheless we believe that theorem 3.1 already exhibits a new type of restriction on the Einstein metrics which can appear in the compactification of the moduli space of Einstein metrics.

One may also ask the question for the other singularities: for the other Kleinian singularities and their finite quotients, the answer is that the obstruction (3.4) still holds, together with other obstructions: actually the number of scalar obstructions equals the b_2 of the corresponding ALE space (work in preparation). So the corollary should remain true for these singularities. The case of other singularities depend on the question mentioned above of the classification of all Ricci flat ALE spaces.

Now pass to some more precise remarks about theorem 3.1. First, note that if the Eguchi-Hanson space is glued with the opposite orientation (which results in a different topological space) then the condition (3.4) becomes $\det \mathbf{R}_-^{g_0}(p_0) = 0$ (this is clear since the Einstein equation does not depend on the orientation).

Also note that in the Kähler case, choosing a basis $(\omega_1, \omega_2, \omega_3)$ of Ω_+ such that ω_1 is the Kähler form, one has

$$\mathbf{R}_+ = \begin{pmatrix} \frac{\mathrm{Scal}}{4} & & \\ & 0 & \\ & & 0 \end{pmatrix}, \tag{3.5}$$

so the condition (3.4) is automatically satisfied. Indeed it is well known that there is no such obstruction in the Kähler case.

When one considers the gluing $g_0 \sharp t g_{\mathrm{EH}}$, there is an ambiguity which gives a gauge parameter: indeed one can apply an element $u \in SO_4/\mathbb{Z}_2$ when identifying the parts A_t and B_t of the cone $\mathbb{R}^4/\mathbb{Z}_2$ (applying an orientation reversing element of O_4 amounts to changing the orientation of the Eguchi-Hanson space and was considered just above). It turns out that the isometry group of g_{EH} is $\mathrm{Isom}(g_{\mathrm{EH}}) = (U_2/\pm 1) \rtimes \mathbb{Z}_2$ (where $U_2 \subset SO_4$ is the standard unitary subgroup, and the $\mathbb{Z}_2$ is generated by $(z_1, z_2) \mapsto (-\bar{z}_2, \bar{z}_1)$, inducing the antipodal map on S^2). Taking u in $\mathrm{Isom}(g_{\mathrm{EH}})$ does not change $g_0 \sharp t g_{\mathrm{EH}}$, so the remaining parameter is in $SO_4/(U_2 \rtimes \mathbb{Z}_2) = \mathbf{P}\Omega_+(\mathbb{R}^4)$.

This means that the ambiguity u can be interpreted as a real line in $\Omega_+(\mathbb{R}^4)$. This is related to the obstruction (3.4): note u_i the gauge parameter used for $g_0 \sharp t_i g_{\mathrm{EH}}$, then one can add to the statement of the theorem the fact that the directions in $\Omega_+(\mathbb{R}^4)$ corresponding to the limits of the gauge parameters u_i must be in the kernel of $\mathbf{R}_+^{g_0}(p_0)$. (This also fits with

the Kähler picture, since when $\mathrm{Scal} \neq 0$, this condition implies that the complex structure of the orbifold must be glued with a complex structure of the Eguchi-Hanson space which is orthogonal to that of $T^*\mathbb{C}P^1$, and in particular does not admit a holomorphic sphere; but indeed a Kähler-Einstein metric with $\mathrm{Scal} \neq 0$ can not admit a holomorphic sphere of self intersection -2).

In particular, if $\mathrm{rk}\, \mathbf{R}_+^{g_0}(p_0) = 1$, then the kernel of $\mathbf{R}_+^{g_0}(p_0)$ gives a direction in $\Omega_+(\mathbb{R}^4)$, so a gauge parameter $u = \lim u_i$. This is of importance in the reverse construction, that we now describe.

To see if the condition (3.4) is the only local obstruction to the desingularization, it is important to produce an Einstein desingularization (M, g_t) from the singular (M, g_0). It turns out that this is not possible in general on a compact manifold, because, as mentioned earlier, there are always global obstructions to deformation which make the problem untractable. Fortunately, the problem becomes much better in the AH setting:

Theorem 3.3 ([8]). *Suppose that (M_0, g_0) is a non degenerate AH Einstein orbifold, with a singularity of type $\mathbb{R}^4/\mathbb{Z}_2$ at the point p_0. If g_0 satisfies the condition (3.4), then there exists a family of AH Einstein metrics g_t on a topological desingularization M such that (M_0, g_0) is the limit of (M, g_t) when $t \to 0$.*

Again the theorem is still valid if there are several singular points: the topological desingularization M is obtained by replacing each singular point by a sphere of self intersection -2.

An important fact to note in the theorem is that the conformal infinity γ_t induced by g_t on ∂M depends on t, and converges to the conformal infinity γ_0 of g_0 on $\partial M_0 = \partial M$: it is this flexibility which enables to solve the problem in the AH case.

There is an explicit family [4, 21], called the AdS-Taub-Bolt family, of U_2 invariant metrics on T^*S^2, which converge to an orbifold metric on $\mathbb{B}^4/\mathbb{Z}_2$. The limit is not the hyperbolic metric (this is impossible by corollary 3.2), but a $\mathbb{Z}_2$ quotient of a selfdual Einstein metric on $\mathbb{B}^4$, which is a member of a 1-parameter family found by Pedersen [22]; more precisely, it is the unique member of this family which satisfies the obstruction (3.4).

4. Degree theory and wall crossing

We now consider the AH setting, and study the consequences of theorem 3.3 on the Dirichlet problem at infinity stated in section 1.

Let (M_0, g_0) be an AH Einstein 4-orbifold, with conformal infinity $[\gamma_0]$ on the boundary $\partial M_0 = X$. Again for simplicity, suppose that we have only one singular point. We still restrict to the simplest singularity A_1, and we ask g_0 to be non degenerate (remind this means that the L^2 kernel of the linearization vanishes). This implies that, given a small deformation γ of γ_0, there exists a deformation g_0^γ of g_0, which is an AH Einstein orbifold with conformal infinity γ.

We also suppose that condition (3.4) holds for g_0. Then, inside the space $\mathscr{C}$ of all conformal metrics on X, we can consider, at least near γ_0, the space of conformal metrics on X such that the corresponding orbifold Einstein metric also satisfies (3.4):

$$\mathscr{C}_0 = \{\gamma \in \mathscr{C}, \det \mathbf{R}_+^{g_0^\gamma}(p_0) = 0\}. \tag{4.1}$$

Therefore, all the metrics g_0^γ with $\gamma \in \mathscr{C}_0$ can be desingularized by theorem 3.3, leading to AH Einstein metrics g_t^γ ($t > 0$) on the topological desingularization M of M_0.

Theorem 4.1 ([8, 9]). *Suppose that* $\mathrm{rk}\, \mathbf{R}_+^{g_0}(p_0) = 2$ *(this is a way to say that the vanishing of* $\det \mathbf{R}_+^{g_0}(p_0)$ *is non degenerate). Then*

(1) *The set* $\mathscr{C}_0$ *is a smooth hypersurface of* $\mathscr{C}$ *near* γ_0.

(2) *For* γ *near* $\mathscr{C}_0$*, all the desingularized Einstein metrics have their conformal infinity on the side of* $\mathscr{C}_0$ *determined by*

$$\det \mathbf{R}_+^{g_0^\gamma}(p_0) > 0. \tag{4.2}$$

This result means that $\mathscr{C}_0$ is a 'wall' for the Dirichlet problem at infinity on M: for a conformal infinity on the side (4.2) of the wall, there is an AH Einstein metric with this conformal infinity (one of the metrics g_t^γ); when the conformal infinity goes to the wall, the Einstein metric degenerates, and disappears on the other side.

This is better understood in the setting of the degree theory proposed by Anderson [2] for the Dirichlet problem at infinity. The idea is the following: let $\mathscr{M}$ be the space of all AH Einstein metrics on M, and consider the map

$$\Phi : \mathscr{M} \longrightarrow \mathscr{C} \tag{4.3}$$

defined by: $\Phi(g)$ is the conformal infinity of g. Anderson proved that, in a suitable Banach topology, if $\pi_1(M, X) = 0$, the map Φ is Fredholm of index 0. If there exists some open set $U \subset \mathscr{C}$ over which the map Φ is proper, then Sard-Smale theory gives a well-defined notion of degree of Φ which counts the number of preimages of an element of U. A priori, the degree is only defined in $\mathbb{Z}_2$, but there is a way to count the solutions with sign (the sign is the number of negative eigenvalues of the linearization) and to define a degree with values in $\mathbb{Z}$. In some cases, one may hope to calculate the degree at some special points of U, and if it does not vanish, this implies that the map Φ is surjective over U.

It turns out that the properness of the map (4.3) is a difficult problem, which is far from being solved in general. The paper [2] is written under the following assumptions:

1. $\dim M = 4$: this is to be able to use the strong compactness results for Einstein metrics in dimension 4;
2. $U = \{\gamma \text{ on } X, \mathrm{Scal}^\gamma > 0\}$: this is used to avoid cusp formation in the limits, and is also natural from the point of view of the physicists; it replaces the hypothesis on the volume in theorem 2.1; there are also counterexamples to to properness with flat conformal infinities;
3. the map $H_2(X, \Bbbk) \to H_2(M, \Bbbk)$ is surjective for any field $\Bbbk$: this is to avoid the degeneration to an Einstein orbifold, because in that case some 2-homology must exist in the interior of M (for example, the 2-sphere in the case of the degeneration to a $\mathbb{R}^4/\mathbb{Z}_2$ singularity).

The general case to consider in dimension 4 is when one relaxes the third hypothesis. Here, theorem 4.1 gives insight on what to expect. We are far from being able to prove something here, but the following speculations may help to understand the meaning of the theorem.

It is clear that in this general case, the map Φ is not proper: indeed we have explicit examples of orbifold degenerations of AH Einstein metrics. But we at least understand what

is happening when there is a degeneration of M to an orbifold M_0 with an A_1 singularity, obtained by contracting a 2-sphere of self intersection -2: the number of preimages of Φ changes when one goes through the wall $\mathscr{C}_0$ defined by (4.1), in the precise way given by theorem 4.1. In this way, theorem 4.1 can be interpreted as a wall crossing formula calculating the jump of the degree on M when one goes across $\mathscr{C}_0$.

Of course in general there are several (-2) spheres which can be contracted, so they give rise to several walls in $\mathscr{C}$: one can hope to have Φ proper in the regions delimited by these walls, and jumping across the walls like in theorem 4.1.

Now, all this is for A_1 singularities, so what is happening for the other singularities? the other Kleinian singularities are obtained by contracting a number of (-2) spheres, say k, and indeed one expects to obtain k obstructions to desingularization: so it seems that the generic case is that of A_1 singularities, the other Kleinian singularities being obtained when k walls intersect in a certain way; so the wall crossing formula for the A_1 case might be sufficient. The case of finite quotients of Kleinian singularities is also similar.

To transform these speculations into a proof, one would need to prove the properness of the map Φ outside the walls obtained from the various possible orbifolds obtained from M: in particular, this requires the classification of the Ricci flat ALE spaces, mentioned in section 2.1, and a better understanding of the behaviour of a degenerating Einstein metric.

5. Some ideas of the proofs

The beginning of the proof builds on usual ideas in 'gluing problems' appearing in geometric analysis. For small t we have a metric $g_0 \sharp t g_{\mathrm{EH}}$ which is an approximate solution of the Einstein equation, which is better and better when $t \to 0$, and one wants to deform it into a true solution if t is small enough. In general, this is possible if the two pieces (here g_0 and g_{EH}) are not obstructed for the deformation theory of the Einstein problem. The point is that this is never true for g_{EH} (or more generally for any ALE space), because g_{EH} comes in a 1-parameter family given by scaling. More precisely, the linearization of the Einstein equation on the Eguchi-Hanson space (or more generally any hyper-Kähler space) is

$$L = d_-^* d_- : \Omega_- \Omega_+ \longrightarrow \Omega_- \Omega_+, \tag{5.1}$$

where one uses the identification $\Omega_-(\mathbb{R}^4)\Omega_+(\mathbb{R}^4) = \mathrm{Sym}_0^2(\mathbb{R}^4)$ given by $u \otimes v \mapsto u \circ v$. (The operator on the trace part is just the usual Laplacian). On a hyper-Kähler manifold, the bundle Ω_+ is a flat trivial bundle: $\Omega_+ = \mathbb{R}^3$. So the operator L is identified with the Laplacian $d_-^* d_-$ acting on $\Omega_- \otimes \mathbb{R}^3$, and its L^2-kernel is therefore the L^2 cohomology of the Eguchi-Hanson space:

$$\ker_{L^2} L = L^2 H^2 \otimes \mathbb{R}^3 \simeq \mathbb{R}^3. \tag{5.2}$$

Indeed the L^2 cohomology of Eguchi-Hanson is generated by the Poincaré dual of the 2-sphere. Let choose a basis (o_1, o_2, o_3) of this obstruction space. Then usual techniques enable to deform $g_0 \sharp t g_{\mathrm{EH}}$ into a (basically unique) solution of the Einstein equation modulo these obstructions:

$$\mathrm{Ric}(g_t) - \Lambda g_t = \sum_1^3 \lambda_i(t) o_i. \tag{5.3}$$

(This is not the exact equation to be solved because one must respect the Bianchi identity, but it gives the idea). The problem becomes to analyse the functions $\lambda_i(t)$ and their possible vanishing.

The way to do this is to refine the approximate metric $g_0 \sharp t g_{\mathrm{EH}}$: if one has an approximation to a better order of a solution of (5.3), then g_t will be closer to this new approximation and this can give the first terms of the development of $\lambda_i(t)$.

The idea here is to refine the ALE metric g_{EH} into a metric h_t before gluing it to g_0: the metric h_t is a perturbation of g_{EH} which should satisfy the equation

$$\mathrm{Ric}(h_t) = t\Lambda h_t \tag{5.4}$$

instead of $\mathrm{Ric}(g_{\mathrm{EH}}) = 0$ (so that $\mathrm{Ric}(th_t) = \Lambda(th_t)$); and it should match better $\frac{g_0}{t}$ near infinity: denote euc the standard Euclidean metric, then near p_0, in normal coordinates, one has

$$g_0 = \mathrm{euc} + g_2 + O(r^4), \tag{5.5}$$

where g_2 is an order 2 term:

$$g_2 = \sum a_{ijkl} x^i x^j dx^k dx^l. \tag{5.6}$$

We can ask h_t to match these order 2 terms in the following way: when we perform the homothety h_t, we transfer the coordinates x^i near 0 into the coordinates $X^i = t^{-\frac{1}{2}} x^i$ near infinity on Eguchi-Hanson, so

$$\frac{g_2}{t} = t \sum a_{ijkl} X^i X^j dX^k dX^l. \tag{5.7}$$

So it is natural to look for a first order deformation $h_t = g_{\mathrm{EH}} + th$ which satisfies at infinity

$$h \sim \sum a_{ijkl} X^i X^j dX^k dX^l \tag{5.8}$$

while (5.4) becomes

$$Lh = \Lambda g_{\mathrm{EH}}. \tag{5.9}$$

The first order deformation $g_{\mathrm{EH}} + th$ is not a metric on the whole Eguchi-Hanson space, since the perturbation h blows up at infinity. Nevertheless it will define a metric on the region which is considered in the gluing, that is $R \leq 2t^{-\frac{1}{4}}$.

Now it turns out that the system (5.8) (5.9) is obstructed and has no solution in general, because of the cokernel of L (which equals its kernel). Actually, instead of (5.9), one can only solve

$$Lh = \Lambda g_{\mathrm{EH}} + \sum_1^3 \lambda_i o_i, \tag{5.10}$$

where the real numbers λ_i are also unknown.

At the end, the system (5.8) (5.10) has a solution (h, λ_i), and the λ_i depends only on the second order terms g_2 of g_0 at p_0, that is on the curvature of g_0 at p_0. There are some arguments using in particular the invariance of the system to calculate precisely the λ_i and one finds (up to a constant)

$$\lambda_i = \langle \mathbf{R}_+^{g_0}(p_0)\omega_1, \omega_i \rangle, \tag{5.11}$$

where (ω_i) is an orthonormal basis of Ω_+. Then, using the approximate metric $g_0 \sharp t(g_{\mathrm{EH}} + th)$, one can show that the coefficient $\lambda_i(t)$ appearing in (5.3) has the expansion

$$\lambda_i(t) = t\lambda_i + O(t^2). \tag{5.12}$$

In particular, the vanishing of $\lambda_i(t)$ forces $\lambda_i = 0$, which by (5.11) means

$$\mathbf{R}_+^{g_0}(p_0)\omega_1 = 0. \tag{5.13}$$

Therefore $\mathbf{R}_+^{g_0}(p_0)$ has a kernel; using the gauge freedom, one can reduce this condition to $\det \mathbf{R}_+^{g_0}(p_0) = 0$, which proves theorem 3.1.

For the desingularization itself (theorem 3.3), the work is far from being finished, since from the hypothesis $\mathbf{R}_+^{g_0}(p_0)\omega_1 = 0$ we have killed only the first term in the development of $\lambda_i(t)$. Here one uses the fact that (M_0, g_0) is an AH Einstein manifold, which gives the flexibility to vary g_0 varying its conformal infinity γ_0. In particular, one considers the map $F = (F_1, F_2, F_3) : \mathscr{C} \to \mathbb{R}^3$ defined by

$$\gamma \longmapsto (\lambda_1(g_0^\gamma), \lambda_2(g_0^\gamma), \lambda_3(g_0^\gamma)), \tag{5.14}$$

where g_0^γ is the Einstein orbifold metric on M_0 with conformal infinity γ, and the λ_i are defined by (5.11). Then one proves that the map F is submersive at γ_0: despite the fact that the space $\mathscr{C}$ is infinite dimensional, this is not an obvious fact, and the proof relies in particular on a unique continuation theorem proved in [7]. This means that there exist directions γ_i in the space of conformal structures, such that

$$d_{\gamma_0} F_i(\gamma_j) = \delta_{ij}. \tag{5.15}$$

Consider now the metric g_t and the functions $\lambda_i(t)$ in (5.3) as depending also of the conformal infinity γ, and note this dependence as $g_t(\gamma)$, $\lambda_i(t, \gamma)$. From equations (5.12) and (5.15) it is now immediate that there exist functions $a_i(t) = O(t)$ such that

$$\lambda_i(t, \gamma_0 + \sum_1^3 a_j(t)\gamma_j) = 0, \tag{5.16}$$

which means that the metric $g_t(\gamma_0 + \sum_1^3 a_j(t)\gamma_j)$ is the expected solution of the Einstein equation.

Proving theorem 4.1 requires substantially new arguments. The first step is to refine the previous arguments.

If $\mathrm{rk}\, \mathbf{R}_+^{g_0}(p_0) = 2$, one can show that actually $\lambda_2(t)$ and $\lambda_3(t)$ can be killed just by varying the gauge parameter, so there is no need to deform the conformal infinity in the directions γ_2 and γ_3, the direction γ_1 is sufficient. So one can obtain a solution $g_t(\gamma_0 + a_1(t)\gamma_1)$.

Moreover one can construct a more refined deformation of g_{EH} which matches even better g_0 at infinity before gluing, by obtaining the coincidence not only of the terms of order 2, but also the terms of order 4; the whole construction is then refined to obtain a better expansion of $\lambda_1(t)$:

$$\lambda_1(t) = t\lambda_1 + t^2\mu_1 + O(t^3), \tag{5.17}$$

where μ_1 is some a priori non explicit number, obtained when finding the second order terms of the solution modulo obstructions of the equation (5.4). Then it is clear that the function $a_1(t)$ such that $g_t(\gamma_0 + a_1(t)\gamma_1)$ is the expected Einstein metric satisfies

$$a_1(t) \sim -\mu_1 t \tag{5.18}$$

when $t \to 0$. If μ_1 has a sign, then all the solutions are exactly on the side of $\mathscr{C}_0$ determined by the direction $-\mu_1\gamma_1$ at γ_0.

Calculating μ_1 is difficult. Up to now, the analysis used essentially the linearization the Einstein equation (and some global properties). But calculating μ_1 involves understanding the second order terms of the equation, in order to find the second order terms of the solution of (5.4). We will not give any detail here, see [9], except to say that the hyper-Kähler nature of g_{EH} helps a lot to get insight on these second order terms and on the calculation of μ_1. From this theorem 4.1 is deduced.

Finally let us say that the proofs of both theorems do not rely on the precise form of the Eguchi-Hanson metric, but more on the fact that the Eguchi-Hanson space has a one dimensional L^2-cohomology and has a Hamiltonian circle action which rotates the other complex structures. There are lots of other spaces with the same geometric properties, if one allows orbifold singularities inside. For example, the A_k singularity $\mathbb{C}^2/\mathbb{Z}_{k+1}$ has a partial desingularization satisfying the same properties, but with an orbifold point with a A_{k-1} singularity. Using the same techniques as above, one can calculate an expansion for $\det \mathbf{R}_+$ at the singular point and find an obstruction to continuing the desingularization. So it seems that an inductive process can be started, leading to k obstructions to desingularization. Unfortunately this process can not be carried out so easily, because the non degeneracy hypothesis seems difficult to prove for the partial desingularizations. Nevertheless the author believes he is able to overcome this technical problem using some refined analysis.

References

[1] M. T. Anderson., *Ricci curvature bounds and Einstein metrics on compact manifolds*, J. Amer. Math. Soc., **2**(3) (1989), 455–490.

[2] ———, *Einstein metrics with prescribed conformal infinity on 4-manifolds*, Geom. Funct. Anal., **18**(2) (2008), 305–366.

[3] S. Bando, A. Kasue, and H. Nakajima., *On a construction of coordinates at infinity on manifolds with fast curvature decay and maximal volume growth*, Invent. Math., **97**(2) (1989), 313–349.

[4] L. Bérard-Bergery., *Sur de nouvelles variétés riemanniennes d'Einstein*, Inst. Élie Cartan, **6** (1982), 1–60.

[5] A. L. Besse., *Einstein manifolds*, Springer-Verlag, Berlin, 1987.

[6] O. Biquard, ed., *AdS/CFT correspondence: Einstein metrics and their conformal boundaries*, IRMA Lectures in Mathematics and Theoretical Physics, 8. European Mathematical Society, Zürich, 2005.

[7] O. Biquard., *Continuation unique à partir de l'infini conforme pour les métriques d'Einstein*, Math. Res. Lett., **15**(6) (2008), 1091–1099.

[8] ———, *Désingularisation de métriques d'Einstein. I*, Inventiones Math., **192**(1) (2013), 197–252.

[9] ———, *Désingularisation de métriques d'Einstein. II.*, arXiv:1311.0956 [math.DG].

[10] T. Eguchi and A. J. Hanson, *Asymptotically flat self-dual solutions to euclidean gravity*, Phys. Lett. B, **74**(3) (1978), 249–251.

[11] C. Fefferman and C. R. Graham, *Conformal invariants*, Astérisque, (hors série), 95–116, 1985, The mathematical heritage of Élie Cartan (Lyon, 1984).

[12] ———, *The ambient metric*, Princeton, NJ: Princeton University Press, 2012.

[13] C. R. Graham and J. M. Lee, *Einstein metrics with prescribed conformal infinity on the ball*, Adv. Math., **87**(2) (1991), 186–225.

[14] P. B. Kronheimer, *The construction of ALE spaces as hyper-Kähler quotients*, J. Differential Geom., **29**(3) (1989), 665–683.

[15] ———, *A Torelli-type theorem for gravitational instantons*, J. Differential Geom., **29**(3) (1989), 685–697.

[16] C. LeBrun, *Four-dimensional Einstein manifolds, and beyond*, In Surveys in differential geometry. Vol. VI: Essays on Einstein manifolds. Lectures on geometry and topology, sponsored by Lehigh University's Journal of Differential Geometry, pp. 247–285. Cambridge, MA: International Press, 1999.

[17] J. Maldacena, *The large N limit of superconformal field theories and supergravity*, Adv. Theor. Math. Phys. **2**(2) (1998), 231–252.

[18] A. Naber, *The geometry of Ricci curvature*, Proceedings of ICM 2014.

[19] H. Nakajima, *Self-duality of ALE Ricci-flat 4-manifolds and positive mass theorem*, In Recent topics in differential and analytic geometry, volume 18 of Adv. Stud. Pure Math., pp. 385–396. Academic Press, Boston, MA, 1990.

[20] D. N. Page, *A physical picture of the K3 gravitational instanton*, Phys. Lett. B, **80**(1-2) (1978), 55–57.

[21] D. N. Page and C. Pope, *Inhomogeneous Einstein metrics on complex line bundles*, Classical Quantum Gravity, **4** (1987), 213–225.

[22] H. Pedersen, *Einstein metrics, spinning top motions and monopoles*, Math. Ann. **274**(1) (1986), 35–59.

[23] P. Topiwala, *A new proof of the existence of Kähler-Einstein metrics on K3. I. and II*, Invent. Math. **89** (1987), 425–448 and 449–454.

[24] E. Witten, *AdS/CFT correspondence and topological field theory*, J. High Energy Phys. **12** (1998), Paper 12, 31 p. (electronic).

UPMC Université Paris 6 et École Normale Supérieure

E-mail: olivier.biquard@ens.fr

Non-negatively curved manifolds and Tits geometry

Fuquan Fang

Abstract. We explain a surprising passage from non-negatively curved manifolds with polar actions to Tits geometries, which is the basic tool for the rigidity theorem on positively curved polar manifolds established in [13], as well as for works in progress for further rigidity theorems on non-negatively curved hyperpolar manifolds. The latter possibly leads to a new characterization of riemannian symmetric spaces.

Mathematics Subject Classification (2010). Primary 53C24; Secondary 53C35.

Keywords. Curvature, polar actions, Chamber system, Tits building.

1. From non-negative curvature to topology, a brief review

The interest in positively/non-negatively curved manifolds goes back to the 19-th century. From the Gauss-Bonnet theorem, a compact surface with a positively curved metric is either diffeomorphic to $\mathbb{S}^2$ or $\mathbb{RP}^2$. But it was not known untill the early 1980's, after the fundamental work of Hamilton, that all compact 3-dimensional manifolds with positively curved metrics are diffeomorphic to spherical space forms. In dimensions larger than 24, there are no known examples of compact simply connected manifolds with positive sectional curvature other than the rank one symmetric spaces. Surprisingly, infinitely many positively (pinched) curved simply connected manifolds were found only in dimensions 7 and 13, among others, e.g.,

- the Aloff-Wallach spaces $\mathsf{SU}(3)/i_{p,q}(\mathsf{S}^1)$, where $i_{p,q}(\mathsf{S}^1) \subset \mathsf{T}^2$ is a family of circles in the maximal torus T^2 parameterized by pairs of coprime integers p, q;
- the Bazaikin spaces $i_{q_1,\cdots,q_6}(\mathsf{S}^1) \setminus \mathsf{SU}(6)/\mathsf{Sp}(3)$, where $i_{q_1,\cdots,q_6}(\mathsf{S}^1) \subset \mathsf{SU}(6)$ is a diagonal circle subgroup parameterized by six odd integers $q_1, \cdots, q_6$ such that the biquotients are manifolds.

The structure of positively/non-negatively curved manifolds has not yet been well-understood, except for a few well-known general topological constraints, such as the Bonnet-Meyer theorem, the Synge theorem, Gromov's Betti number theorem, and the vanishing of the $\hat{A}$-genus (α-invariant) for Spin manifolds with positive scalar curvature.

For a compact riemannian manifold, one can always normalize the metric so that the maximum of the sectional curvature is 1. The curvature is said to be δ-*pinched* if $\delta < \sec \leq 1$ everywhere. A very challenging conjecture is (cf. Berger [Be]):

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Conjecture (Optimal pinching constant). *For any given integer $n > 1$, there is a positive constant $\delta(n)$ depending only on n, such that, every compact positively curved n-manifold admits a $\delta(n)$-pinched metric.*

For positively pinched manifolds, the following finiteness theorem of Cheeger (Weinstein obtained homotopy type finiteness) explains how small the class is:

Theorem 1.1 (Cheeger, Weinstein). *For any given positive constant $\delta < 1$, up to diffeomorphism, there are only finitely many $2n$-dimensional riemannian manifolds with δ-pinched curvature.*

The same statement can not be true in any odd dimensions, because there are infinitely many spherical space forms in every odd dimension $2n + 1$. Furthermore, in dimension 7 (also 13), the Aloff-Wallach (also Bazaikin) spaces indeed contain infinitely many curvature uniformly pinched manifolds with pairwisely different topological types (cf. Puttmann [Pu]), hence the same statement fails even for simply connected positively δ-pinched curved manifolds. However, one has the following analogue of the above Cheeger-Weinstein theorem:

Theorem 1.2 (Fang-Rong [14, 15], Petrunin-Tuschmann [25]). *For any given positive constant $\delta < 1$, up to diffeomorphism, there are only finitely many $(2n + 1)$-dimensional positively δ-pinched curved, simply connected riemannian manifolds with trivial second rational Betti numbers.*

In contrast, there are many more examples of manifolds with non-negatively curved metrics. Products of non-negatively curved manifolds are still non-negatively curved. All compact Lie groups, and in fact all symmetric spaces of compact type have non-negative sectional curvature. By Cheeger, the connected sums of two complex/quaternionic projective spaces admit non-negatively curved riemannian metrics. Starting from cohomogeneity one actions, Grove-Ziller found infinitely many new examples of non-negatively curved manifolds, which includes in particular 10 of the 14 unoriented exotic 7-spheres (cf. [GZ]). It is natural to ask whether this sort of construction can be extended to higher cohomogeneity to create new families of examples. As we will see in later sections, it is much more rigid for polar actions in higher cohomogeneity.

We finally mention a few important conjectures:

- (Hopf) There exists no metric with positive sectional curvature on $\mathbb{S}^2 \times \mathbb{S}^2$. More generally, there are no positively curved metrics on the product of two compact manifolds, or on a symmetric space of rank at least two.
- (Hopf) A compact manifold with sec ≥ 0 has non-negative Euler characteristic. An even dimensional manifold with positive curvature has positive Euler characteristic.
- (Bott-Grove-Halperin) A compact simply connected manifold M with sec ≥ 0 is rational elliptic, i.e., the sequence of rational Betti numbers of the loop space of M grows at most polynomially.
- (Klingenberg-Sakai-Yau) There are only finitely many diffeomorphism classes of positively curved manifolds in a given homotopy type.

2. Reflection groups in Non-negatively curved manifolds

The theory of discrete groups of motions generated by reflections originates in the study of plane regular polygons and space polyhedra. An extensive account of the history of the theory of reflection groups in euclidean and spherical spaces appeared in Bourbaki, Groupes et Algèbres de Lie, Chapters IV-VI [5]. According to this account the modern theory originates from the works of geometers A. Möbius and L. Schläfli in the middle of the 19th century, then was extended and applied to the theory of Lie algebras in the works of E. Cartan and W. Killing at the end of the same century, and culminated in the works of H. S. M. Coxeter [8].

Reflection groups in the hyperbolic plane were described at least back in 1882 by Poincaré in a memoir on Fuchsian groups [28], and by von Dyck [11]. A complete classification of reflection groups in hyperbolic 3-space was achieved by Andreev [2] (cf. also [30]). Hyperbolic reflection groups in higher dimensions are very rich and far from being classified. A surprising theorem of Vinberg [37] asserts there are no co-compact hyperbolic reflection groups in dimensions ≥ 30. An extended non-existence theorem was established by Khovanskii [20] for co-finite volume hyperbolic reflection groups in dimension at least 995.

In this section we present a joint work with Karsten Grove [12].

Assume Σ is an n-dimensional nonnegatively curved complete manifold with a discrete reflection group W, such that the orbit space Σ/W is compact (equivalent to *finite covolume*). Here a reflection refers to an isometric involution with a codimension 1 fixed point set, and a mirror refers to a codimension 1 component. In this generality, a *mirror* may not separate the manifold. We will give a complete description not only of the reflection group, but also of the equivariant structure of the manifold.

Note, if Σ is non compact but Σ/W is compact, it follows from Cheeger-Gromoll soul theorem that Σ splits into a metric product of a flat $\mathbb{R}^k$ with the soul, a compact totally geodesic submanifold. Therefore, Σ/W splits into a metric product of eucledian simplices and the orbit space of a reflection group on the soul.

The mirrors of all reflections in W form a configuration in Σ. A Dirichlet domain will be called an open *chamber* which is a locally convex set. We point out that, it often happens that there are more faces in the closure of the chamber than the minimal number of generators of W. For instance, A_2 acts on the tiling of a flat torus $\mathbb{T}^2$ by six equilateral flat triangles, A_2 is generated by any two of the three reflections.

To explain the appearance of building blocks, we say that the action $\mathsf{W} \times \Sigma \to \Sigma$ is *decomposable* if the orbitspace Σ/W metrically is a finite quotient of a product, and *indecomposable* otherwise. With this terminology one of our main results is the following *Rigidity Theorem*

Theorem 2.1 (Fang-Grove [12]). *A nonnegatively curved manifold Σ^n with an indecomposable cocompact action by a reflection group W is isometric to either $\mathbb{R}^n$, or $\mathbb{T}^n$, or equivariantly diffeomorphic to either $\mathbb{S}^n$, or $\mathbb{RP}^n$ with a linear action, unless all mirrors in Σ meet.*

Here the spherical case relies on showing that the orbit space is a simplex, whereas the part where the universal cover of Σ is non-compact also relies on Cheeger- Gromoll splitting results for cocompact actions and for compact manifolds with infinite fundamental group, as well as on Bieberbach's celebrated Theorem (cf. [12]). Recall, that by the latter, any compact flat manifold is finitely covered by a flat torus, i.e., $\Sigma = \mathbb{T}^n/\,\mathsf{G}$, where $\mathsf{G} \subset \mathsf{O}(n)$ is the holonomy. In particular, Theorem 2.1 shows that the holonomy group G must be trivial

when the action is indecomposible. More generally, we point out that, if the orbit space splits as a metric product of eucledian simplices, then $\mathbb{T}^n/\,\mathsf{G}$ must be an iterated torus bundle, with holonomy group G a very special elementary abelian 2-group in $\mathsf{GL}(\mathbb{Z}, n)$. The Klein bottle serves as the simplest example.

To describe the structure that arises when all mirrors meet let us introduce the fibre product

$$\prod(\nu_1 \oplus \varepsilon^{k_1}, \cdots, \nu_\ell \oplus \varepsilon^{k_\ell})$$

of the sphere bundles $\mathbb{S}(\nu_1 \oplus \varepsilon^{k_1}), \cdots, \mathbb{S}(\nu_\ell \oplus \varepsilon^{k_\ell})$ over a compact manifold X: the subset of the product $\mathbb{S}(\nu_1 \oplus \varepsilon^{k_1}) \times \cdots \times \mathbb{S}(\nu_\ell \oplus \varepsilon^{k_\ell})$ which projects to the diagonal Δ in X^ℓ, where $\nu_1, \cdots, \nu_\ell$ are vector bundles and ε is a trivial line bundle. Notice, $\prod(\nu_1 \oplus \varepsilon^{k_1}, \cdots, \nu_\ell \oplus \varepsilon^{k_\ell})$, admits a fiberwise product linear action by the product of spherical Coxeter groups $\mathsf{W}_1 \times \cdots \times \mathsf{W}_\ell$.

Theorem 2.2 (Fang-Grove [12]). *Given a spherical reflection group W, a nonnegatively curved compact simply connected W-manifold whose all mirrors meet is equivariantly diffeomorphic to*

$$\prod(\nu_1 \oplus \varepsilon^{k_1}, \cdots, \nu_\ell \oplus \varepsilon^{k_\ell})$$

where $\nu_1, \cdots, \nu_\ell$ are vector bundles with non-negative sectional curvature over a soul X.

When passing to the universal cover, the above results in particular lead to the following general *Splitting Theorem*

Theorem 2.3 (Fang-Grove [12]). *Let Σ be a complete non negatively curved manifold with co-compact reflection group W. Then the lifted reflection group $\hat{\mathsf{W}}$ on the universal cover $\tilde{\Sigma}$ is a product of Coxeter groups,*

$$\hat{\mathsf{W}} = \hat{\mathsf{W}}_0 \times \hat{\mathsf{W}}_1 \times \cdots \times \hat{\mathsf{W}}_{\ell-1} \times\ \hat{\mathsf{W}}_\ell,$$

where $\hat{\mathsf{W}}_0$ is affine, and the remaining factors are spherical. Correspondingly, $\tilde{\Sigma}$ admits a $\hat{\mathsf{W}}$ invariant metric splitting,

$$\tilde{\Sigma} = \mathbb{R}^k \times \mathbb{S}^{k_1} \times \cdots \times \mathbb{S}^{k_{\ell-1}} \times \Theta_\ell \times N,$$

where N can be any simply connected compact manifold of nonnegative curvature on which all $\hat{\mathsf{W}}_i$ act trivially, $\mathbb{S}^{k_i}$ is a non negatively curved standard sphere with a linear $\hat{\mathsf{W}}_i$ action, and Θ_ℓ is a compact simply connected non-negatively curved manifold as in Theorem B.

As a consequence we derive the following *Group Structure Theorem*,

Corollary 2.4 (Fang-Grove [12]). *A group W is a co-compact reflection group of a complete non negatively curved manifold if and only if*

$$\mathsf{W} \cong \hat{\mathsf{W}}_0 \times \cdots \times \hat{\mathsf{W}}_\ell/\mathsf{N},$$

where $\hat{\mathsf{W}}_0$ is an affine Coxeter group, $\hat{\mathsf{W}}_i$, $1 \le i \le \ell$, is a spherical Coxeter group, and $\mathsf{N} \lhd \hat{\mathsf{W}}$ a normal subgroup isomorphic to a product of a torsion free lattice and an elementary abelian 2-group.

The overall strategy in our approach is based on the fact that follows from the work of Wörner [39] that the chamber C for a Coxeter action is a product $C = C_0 \times C_1 \times C_2 \times \ldots \times C_\ell$ where C_0 is a manifold without boundary (typically a point), and each C_i, $i \ge 1$ is a smooth non negatively curved convex manifold with corners, and either

(1) C_i has more than $n_i = \dim C_i$ faces, but any n_i faces of C_i meet,

or

(2) C_i has $k_i \leq n_i$ faces and they all meet.

Based on critical point theory of distance functions we can prove that, if there is only one factor and it is of type (1) then C is a simplex, which is a key to prove theorem 2.1.

3. Polar actions

Introduced by Szenthe [Sz] and independently Palais-Terng [PTe], an isometric action of a Lie group G on a compact Riemannian manifold M is said to be *polar* if there is a complete isometrically immersed submanifold $\Sigma \looparrowright M$, called a *section*, that meets all orbits of G and all intersections of Σ with orbits of G are perpendicular. It is clear that the dimension of Σ is equal to the cohomogeneity of the action. Observe that

- a section is always totally geodesic.

A polar action is called *hyperpolar* if the section Σ is flat.
Simple examples of polar Actions are

- *The rotation of* $\mathsf{SO}(2)$ *on* $\mathbb{R}^2$. A line passing through the origin is a section.
- *Every isometric action of cohomogeneity* 1. A normal geodesic that starts perpendicularly to a principal orbit is a section.
- *The linear conjugation action of* $\mathsf{SO}(n)$ *on the vector space* V *of real symmetric* $n \times n$ *matrices with trace zero.* The subspace of diagonal matrices in V serves as a section.
- *The conjugation action of a compact Lie group* G *on itself with a bi-invariant Riemannian metric.* The maximal tori are the sections.
- *The left action of* K *on the symmetric space* $\mathsf{G}\,/\,\mathsf{H}$*, and vice-versa left* H*-action on* $\mathsf{G}\,/\,\mathsf{K}$ where (G, K) and (G, H) are symmetric pairs. Such actions are called *Hermann actions.*
- *The isotropy representation of* K *of a symmetric space* $\mathsf{G}\,/\,\mathsf{K}$.

We recall that isometric actions of Lie groups G_1 and G_2 on Riemannian manifolds M_1 and M_2 respectively are said to be *orbit equivalent* if there is an isometry between M_1 and M_2 under which the orbits of G_1 and G_2 correspond.

Theorem 3.1 (Dadok [9]). *Let* K *be a connected compact Lie group and* $\rho : \mathsf{K} \to \mathsf{SO}(n)$ *a polar representation. Then there is a symmetric space* M *such that* ρ *is orbit equivalent to the isotropy representation of* M.

An easy but basic lemma due to Palais and Terng is

Lemma 3.2 (Palais-Terng [24]). *For any polar action of a Lie group* G *on a riemannian manifold, all of its slice representations are polar.*

Therefore, given a polar G action, by Dadok's theorem there is a symmetric space associated to the slice of every singular orbit. Moreover, given a section Σ through a singular point p, the tangent space $T_p\Sigma$ is the section of the slice representation of the isotropy group G_p. Let W_p be the Weyl group of the symmetric space, a linear reflection group on $T_p\Sigma$.

Definition 3.3 (Reflection Group). The reflection group W_Σ associated to a section Σ is the group generated by W_p for all singular orbit $p \in \Sigma$.

For a polar G action it is natural to make a

Definition 3.4 (Generalized Weyl group). Let G be a compact connected Lie group acting on a riemannian manifold M in a polar fashion. For a section Σ, the generalized Weyl group $\Pi_\Sigma(\mathsf{G}) = \mathsf{N}_\Sigma(\mathsf{G})/Z_\Sigma(\mathsf{G})$, where $N_\Sigma(\mathsf{G}) = \{g \in \mathsf{G} : g(\Sigma) = \Sigma\}$ is the stabilizer of Σ, and $Z_\Sigma(\mathsf{G})$ is the kernel of $N_\Sigma(\mathsf{G})$ action on Σ.

It is clear that the generalized Weyl group does not depend on the choice of a particular section. In the case the polar action is the conjugation action of a Lie group on itself, the generalized Weyl group is exactly the Weyl group of the Lie group.

Lemma 3.5 (Alexandrino [1]). *If M is simply connected, then $\Pi_\Sigma = \mathsf{W}_\Sigma$.*

• *Coisotropic actions and polar actions*. Recall that a submanifold N of a symplectic manifold $(M;\omega)$ is called *coisotropic* if

$$(T_pN)^{\perp_\omega} \subset T_pN$$

for all $p \in N$, where $(T_pN)^{\perp_\omega}$ denotes the subspace of T_pM that is ω-orthogonal to T_pN. In the special case that (M,ω) is a Kähler manifold it is easy to see that a submanifold N of M is coisotropic if and only if $J(\nu_pN) \subset T_pN$ for all $p \in N$, where J denotes the complex structure of M, and ν_pN the normal space of N in p.

A symplectic G-action on (M,ω) is called *Poisson* if there is a Lie algebra homomorphism $\lambda : \mathfrak{g} \to C^\infty(M)$ such that the hamiltonian vector field $X_{\lambda(\xi)}$ agrees with the infinitesimal action of ξ on M. The moment map of a Poisson action is defined as

$$\Phi : M \to \mathfrak{g}^*, \Phi(p)(\xi) = \lambda(\xi)(p)$$

For a compact Kähler manifold (M,ω), an isometric G action is called *multiplicity-free* ([18]) or *coisotropic* ([19]) if the principal G-orbits are coisotropic with respect to ω.

Let $\mathfrak{t}^*_+$ be the Weyl chamber of the dual abelian Lie algebra $\mathfrak{t}^*$. General convexity theorem established in [18, 21] claims that $\Phi(M)\cap\mathfrak{t}^*$ is a convex polytope. The following result, essentially due to [27], identifies a chamber of a polar action with the image of the moment map.

Theorem 3.6 (Podesta-Thorbergsson [27]). *Let $M \subset \mathbb{P}^n(\mathbb{C})$ be an irreducible smooth projective variety with the Fubini-Study metric. If G is a compact connected Lie group acting isometrically on M in a polar fashion. Then the orbit space M/G is homeomorphic to the convex polytope $\Phi(M) \cap \mathfrak{t}^*_+$.*

4. Polar actions and chamber systems

Let M be a compact, simply connected Riemannian manifold with a polar action by a compact Lie group G. Let Σ be a section. By 3.5 the generalized Weyl group Π (:=Π_Σ) coincides with the reflection group W, and hence acts on Σ by reflections. Let C be a closed chamber. Notice that C is isometric to $M/\mathsf{G} = \Sigma/\mathsf{W}$. Moreover, W may not be a Coxeter group in this generality, but is generated by finitely many reflections $s_1, s_2, \cdots, s_k$ satisfying

- $(s_i s_j)^{m_{ij}} = 1$ where $m_{ii} = 1$, and $m_{ij} = m_{ji} \in \mathbb{N}_{\geq 2} \cup \infty$.

Let $\mathsf{M} = (m_{ij})$ denote the symmetric $k \times k$ matrix. In this special case where W is a Coxeter group, M is the Coxeter matrix. In general W may have additional relations. The codimension 1 faces of C are named as types in $I = \{1, 2, \cdots, k\}$, corresponding to $s_1, \cdots, s_k$.

One can associate a (homogeneous) *chamber system* $\mathcal{C}(M; \mathsf{G})$ of type I by setting

$$\mathcal{C}(M; \mathsf{G}) = \bigcup\nolimits_{g \in \mathsf{G}} gC$$

with adjacency relations as follows: chambers $g_1 C$ and $g_2 C$ are called i-adjacent, denoted by $g_1 C \sim_i g_2 C$, if they share a face of type i. It is clear that every point of M is contained in some chamber gC.

We now recall some basic concepts from Tits geometry (cf. [30, 36]).

A chamber system $\mathcal{C}$ is called *connected* if any two chambers can be joint by a (finite) sequence of adjacent chambers, a *gallery*. In our circumstance, if $\mathcal{C}(M; \mathsf{G})$ is a connected chamber system, we can have a length metric d_H on M, namely, $d_H(x, y)$ being the length of shortest horizontal curves connecting x and y.

- *J-Residue*. For a subset $J \subset I$, a J-residue in a chamber subsystem of $\mathcal{C}$ is a set of chambers maximal with respect to being connected by J-galleries, i.e, galleries of types $i_1 \cdots i_\ell$ where all $i_j \in J$, it is, in particular, a connected chamber system over J.
- *Generalized m-gon*. For any integer $m \geq 1$, or for $m = \infty$, a generalized m-gon is a connected, bipartite graph of diameter m and girth $2m$, in which each vertex lies on at least two edges. Recall that a graph is bipartite if its set of vertices can be partitioned into two disjoint subsets such that no two vertices in the same subset lie on a common edge; the diameter is the maximum distance between two vertices, and the girth is the length of a shortest circuit. If $m = \infty$, this is simply a tree with no end points.

A generalized m-gon can be considered as a chamber system by taking the edges as chambers, and adjacency to mean having a common vertex, of one of the two appropriate types.

- *Geometry of Type* M. By definition a chamber system $\mathcal{C}$ is of type M, if for every $i \neq j$, the residue of type $\{i, j\}$ is a generalized m_{ij}-gon.
- *Buildings of Type* M. Let W be a Coxeter group with Coxeter matrix M. For a gallery of type $f = i_1 \cdots i_\ell$ there is an associated element $s_{i_1} \cdots s_{i_\ell} \in \mathsf{W}$. A *building* of type M is a chamber system $\mathcal{C}$ over I such that each codim. one face lies on at least two chambers, and having a W-distance function

$$\delta : \mathcal{C} \times \mathcal{C} \to \mathsf{W}$$

such that if f is a reduced word (i.e., minimal), then $\delta(C_1, C_2) = s_f$ if and only if C_1 and C_2 can be joined by a gallery of type f. In particular, any two chambers can be joined by a minimal gallery.

Example 4.1 (Polar Representations). The chamber system, $\mathcal{C}(\mathbb{S}^n; \mathsf{K})$ associated to the restriction of a polar representation of a compact Lie group K to the unit sphere $\mathbb{S}^n$ (without fixed points) is a fundamental example of a (spherical) Tits building (see [9, 35]).

Example 4.2 (Thin building). For a spherical Coxeter group W generated by reflections $s_1, \cdots, s_k$, the Coxeter complex is a $(k-1)$-dimensional simplicial complex which is a tiling of the sphere $\mathbb{S}^{k-1}$ by spherical simplices isometric to Δ^{k-1}. Coxeter complex is a *thin* building.

For a chamber system $\mathcal{C}$, there is a canonical associated complex $|\mathcal{C}|$, called the *geometric realization*. In particular, if $\mathcal{C}$ is a building, then $|\mathcal{C}|$ is a simplcial complex. A subcomplex of $|\mathcal{C}|$ isomorphic to a Coxeter complex (thin building) is called an *apartment*. We refer [4] for an equivalent definition of buildings in terms of axioms on apartments.

In the case that all proper residues are buildings, the so-called *universal cover* $\tilde{\mathcal{C}}$ can be viewed in the thin (length metric) topology on $\mathcal{C}$ as the usual topological universal cover. This cover clearly inherits the structure of a chamber system, and its fundamental group (deck transformations) acts freely as a group of automorphisms on it.

By invoking the following corollary of a profound theorem of Tits [36], Corollary 3 in Section 5.3 (cf. also [30], Theorem 4.9), we get

Theorem 4.3 (Tits). *The universal cover $\tilde{\mathcal{C}}$ of a connected chamber system $\mathcal{C}$ of type M over I is a building if and only if all residues of rank three are covered by buildings.*

From this we conclude that

Theorem 4.4 (Fang-Grove-Thorbergsson [13]). *Suppose M is a simply connected positively curved polar G manifold with orbit space M/G a simplex. If M is the Coxeter matrix of a spherical Coxeter group of rank at least 4, then the universal cover $\tilde{\mathcal{C}}(M; \mathsf{G})$ of the chamber system $\mathcal{C}(M; \mathsf{G})$ is a spherical building.*

We remark that the connectedness of the chamber system $\mathcal{C}(M; \mathsf{G})$ in the above theorem follows from a result of B. Wilking on dual foliations. Similarly, we have

Theorem 4.5. *Suppose M is a compact simply connected hyperpolar G manifold with orbit space M/G a simplex. If M is the Coxeter matrix of an affine Coxeter group of rank at least 3, then the universal cover $\tilde{\mathcal{C}}(M; \mathsf{G})$ of the chamber system $\mathcal{C}(M; \mathsf{G})$ is an affine building, provided $\mathcal{C}(M; \mathsf{G})$ is connected.*

Remark 4.6. The reflection group W in the above theorems may not be a Coxeter group, but its lifting $\hat{\mathsf{W}}$ is a Coxeter group of spherical type or affine type.

5. Polar actions on positively curved manifolds

In this section we present the rigidity theorem in [13] on polar actions on compact positively curved riemannian manifolds. The rigidity theorem completely classifies positively

curved polar manifolds of cohomogeneity at least 2, which in particular, broke the dream of constructing new examples of positively curved manifolds from polar actions in higher cohomogeneity.

Theorem 5.1 (Fang-Grove-Thorbergsson [13]). *A polar action on a simply connected, compact, positively curved manifold of cohomogeneity at least two is equivariantly diffeomorphic to a polar action on a compact rank one symmetric space.*

Remark 5.2. This is reminiscent of the situation for *isoparametric* submanifolds in euclidean spheres, where many isoparametric hypersurfaces are not homogeneous, whereas in higher codimensions by [34] they are the orbits of linear polar actions if they are irreducible or equivalently the orbits of isotropy representations of compact symmetric spaces by [9].

Remark 5.3. The above theorem is not true for cohomogeneity one actions. In fact infinite subfamilies of the Escheburg as well as of the Bazaikin spaces support cohomogeneity one actions as does a new example, non of which are even homogeneous spaces.

Remark 5.4. All polar actions on the simply connected, compact rank one symmetric spaces, i.e., the spheres and projective spaces, $\mathbb{S}^n, \mathbb{CP}^n, \mathbb{HP}^n$ and $\mathbb{OP}^2$ were classified in [9] and [26]. In all cases but the Cayley plane $\mathbb{OP}^2$ they are either linear polar actions on a sphere or they descend from such actions to a projective space.

For a polar G action on a positively curved manifold, notice that, since the sections are totally geodesic, hence positively curved as well. As a special case discussed in section 2 we have

Theorem 5.5 (Fang-Grove-Thorbergsson [13]). *The polar group* W *of a simply connected positively curved polar manifold of cohomogeneity at least two is a spherical Coxeter group or a* $\mathbb{Z}_2$ *quotient thereof. Moreover, the section with this action is equivariantly diffeomorphic to a sphere, respectively a real projective space with a linear action.*

• *Topological Tits Buildings of spherical type*. Recall, that a compact (spherical) building according to [BSp] is a Tits building $\tilde{\mathcal{C}}$ with a Hausdorff topology on the set $\mathrm{Vert}(\tilde{\mathcal{C}}) = V_1 \cup \cdots \cup V_{k+1}$ of all vertices such that the set $\tilde{\mathcal{C}}_{i_1,\cdots,i_{r+1}}$ of all simplices of type $(i_1, \cdots, i_{r+1})$ is closed in the product $V_{i_1} \times \cdots \times V_{i_{r+1}}$. With the induced topology on the k simplices $\tilde{\mathcal{C}}_{1,\cdots,k+1}$, $\tilde{\mathcal{C}}$ is called *compact, locally connected, infinite, metric* if $\tilde{\mathcal{C}}_{1,\cdots,k+1}$ has the appropriate property.

Example 5.6 (Classical Buildings). Let (U, K) be a symmetric pair, where U is a connected non-compact real semisimple Lie group without center. The isometric action of U on the symmetric space induces an action on the boundary sphere at infinity, $\mathbb{S}_\infty$, orbit equivalent to the subaction by K. The building at infinity of (U, K), or equivalently, the building associated to the polar K action on $\mathbb{S}_\infty$, is called the *classical building*.

The following main result of [6] is important for the proof of Theorem 5.1:

Theorem 5.7 (Burns-Spatzier). *An infinite, irreducible, locally connected, compact, metric, topologically Moufang building of rank at least* 2 *is classical.*

A key step in the proof of Theorem 5.1 is to endow a thick topology on the universal covering $\tilde{\mathcal{C}}(M, \mathsf{G})$ using the Hausdorff topology on compact subsets of M. We have

Theorem 5.8 (Fang-Grove-Thorbergsson [13]). *Whenever the universal cover $\tilde{\mathcal{C}}(M,\mathsf{G})$ of $\mathcal{C}(M,\mathsf{G})$ is a building, it admits the structure of a compact topological Moufang building.*

When the Coxeter diagram for M is connected of rank at least 4, by Theorem 4.4 and the above work of Burns and Spatzier [6], $\tilde{\mathcal{C}}(M;\mathsf{G})$ is a classical building. Our original polar G-action on the chamber system $\mathcal{C}(M,\mathsf{G})$ lifts to an action of $\tilde{\mathsf{G}}$ on the universal cover $\tilde{\mathcal{C}}(M;\mathsf{G})$, where $\tilde{\mathsf{G}}$ is a normal extension of the fundamental group π (of the chamber system $\mathcal{C}(M;\mathsf{G})$) by G. In the thick topology it turns out that both π and $\tilde{\mathsf{G}}$ are compact Lie subgroups of the topological automorphism group of the topological building $\tilde{\mathcal{C}}(M,\mathsf{G})$. Therefore, $\tilde{\mathcal{C}}(M,\mathsf{G})$ is homeomorphic to the sphere $\mathbb{S}_\infty$, and the action of $\tilde{\mathsf{G}}$ is orbit equivalent to a linear polar representation of the maximal compact subgroup of the topological automorphism group. In particular, π acts freely on the sphere in a continuous fashion, hence π is trivial, S^1 or S^3. From this Theorem 5.1 follows in the case M is irreducible of rank at least 4.

As a by-product we have the following simply connected C_3 geometry which is not building:

Theorem 5.9 (Fang-Grove-Thorbergsson [13]). *The universal cover $\tilde{\mathcal{C}}$ of the chamber system $\mathcal{C}(\mathbb{OP}^2,\mathsf{SU}(3)\,\mathsf{SU}(3))$ for the exceptional action on $\mathbb{OP}^2$ is a C_3 geometry which is not a building.*

The existence of such C_3 geometries is well known in the "real estate community" (see [23]), but this particular example which arises very naturally in our context was not known until now (see also [22]).

6. Bruhat-Tits buildings and hyperpolar actions on non-negatively curved manifolds

It is more flexible to construct hyperpolar actions on non-negatively curved manifolds using simple operations, e.g., the componentwise product of hyperpolar actions on M and N will be a hyperpolar action on the product $M\times N$; moreover, given a hyperpolar action by a compact Lie group K on M, and a compact Lie group $\mathsf{G}\supset\mathsf{K}$, the balanced product $\mathsf{G}\times_{\mathsf{K}}M$ with the action of G is again hyperpolar, with induced metric from the product metric on $\mathsf{G}\times M$. In order to concentrate to some fundamental building blocks, it is necessary to introduce

Definition 6.1. A hyperpolar G-action on M is *irreducible* if the lifted affine Coxeter group $\hat{\mathsf{W}}$ is irreducible.

The following is a special and more precisely stated version of a conjecture from [13]:

Conjecture. *An irreducible hyperpolar action of cohomogeneity at least 2 on a simply connected nonnegatively curved compact manifold is equivariantly diffeomorphic to a quotient of a polar action on a symmetric space.*

Remark 6.2. Recall that the conjugation action of a Lie group on itself, as well as the Hermann actions are all hyperpolar. An affirmative answer to the above conjecture provides essentially a riemannian geometric characterization of irreducible symmetric spaces. The conjecture however is not true for cohomogeneity 1 actions.

The following confirms the conjecture when the Coxeter diagram does not have double bonds of rank at least 3:

Theorem 6.3. *Assume (M, G) is an irreducible hyperpolar manifold of rank at least 3. If the chamber system $\mathcal{C}(M;\mathsf{G})$ is connected of types $\tilde{\mathsf{A}}_n$, $\tilde{\mathsf{D}}_n$, $\tilde{\mathsf{E}}_6$, $\tilde{\mathsf{E}}_7$ or $\tilde{\mathsf{E}}_8$. Then, there is a G-equivariant principal L-bundle $\hat{M} \to M$, such that $\hat{M}$ is $\mathsf{G}\times\mathsf{L}$-equivariant diffeomorphic to a compact Lie group with a hyperpolar isometric action.*

Remark 6.4. Results in the above theorem can be made more precisely, depending on the multiplicity m, as follows

(I) Assume $\mathcal{C}(M;\mathsf{G})$ is of type $\tilde{\mathsf{A}}_n$;

- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{SU}(n), \mathsf{SO}(n)\times\mathsf{SO}(n))$ if $m = 1$.
- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{SU}(n); \mathsf{SU}(n))$ or $(\mathsf{SU}(2n); \mathsf{Sp}(n)\times\mathsf{SO}(2n))$ if $m = 2$.
- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{SU}(2n); \mathsf{Sp}(n)\times\mathsf{Sp}(n))$ if $m = 4$.

(II) Assume $\mathcal{C}(M;\mathsf{G})$ is of type $\tilde{\mathsf{D}}_n$;

- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{SO}(2n), \mathsf{SO}(n)\times\mathsf{SO}(n)\times\mathsf{SO}(n)\times\mathsf{SO}(n))$ if $m = 1$.
- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{SO}(2n), \mathsf{SO}(2n))$ if $m = 2$.

(III) Assume $\mathcal{C}(M;\mathsf{G})$ is of type $\tilde{\mathsf{E}}_6$;

- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{E}_6, \mathsf{Sp}(4)/\mathbb{Z}_2\times\mathsf{Sp}(4)/\mathbb{Z}_2)$ if $m = 1$.
- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{E}_6, \mathsf{E}_6)$ if $m = 2$.

(IV) Assume $\mathcal{C}(M;\mathsf{G})$ is of type $\tilde{\mathsf{E}}_7$;

- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{E}_7, \mathsf{SU}(8)/\mathbb{Z}_2\times\mathsf{SU}(8)/\mathbb{Z}_2)$ if $m = 1$.
- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{E}_7, \mathsf{E}_7)$ if $m = 2$.

(V) Assume $\mathcal{C}(M;\mathsf{G})$ is of type $\tilde{\mathsf{E}}_8$;

- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{E}_8, \mathsf{SO}(16)/\mathbb{Z}_2\times\mathsf{SO}(16)/\mathbb{Z}_2)$ if $m = 1$.
- $(\hat{M}, \mathsf{G}\times\mathsf{L}) = (\mathsf{E}_8, \mathsf{E}_8)$ if $m = 2$.

Remark 6.5. For the remaining types $\tilde{\mathsf{B}}_n, \tilde{\mathsf{C}}_n, \tilde{\mathsf{F}}_4$ it is more difficult, but work is in progress.

Remark 6.6. If the chamber system $\mathcal{C}(M;\mathsf{G})$ is not connected, then $M = \mathsf{G}\times_{\mathsf{K}} N$, where K is a compact Lie subgroup, N is a polar K manifold whose chamber system $\mathcal{C}(N;\mathsf{K})$ is a gallery connected component of $\mathcal{C}(M;\mathsf{G})$. For this reason it suffices to consider connected chamber system.

According to Theorem 4.5, the universal cover $\tilde{\mathcal{C}}(M;\mathsf{G})$ of the chamber system $\mathcal{C}(M;\mathsf{G})$, associated to a hyperpolar G action on M, is an affine Bruhat-Tits building, if the rank is at least 3. However, we do not have an analouge of the Burns-Spatzier theory for "topological" Bruhat-Tits buildings.

• *Bruhat-Tits buildings of types $\tilde{\mathsf{A}}_n$, $\tilde{\mathsf{D}}_n$, $\tilde{\mathsf{E}}_6$, $\tilde{\mathsf{E}}_7$ or $\tilde{\mathsf{E}}_8$.*

For the affine building $\tilde{\mathcal{C}}(M;\mathsf{G})$, consider the maximal apartment system, the union of all Coxeter subbuildings in $\tilde{\mathcal{C}}(M;\mathsf{G})$. According to Bruhat-Tits ([30, 38]), there is a complete

discrete valuation field $\mathbb{K}$ associate to the building $\tilde{\mathcal{C}}(M;\mathsf{G})$, whose residue field k is determined by the spherical residues. It is obvious, in our cases, since the spherical residues are the buildings associated to the slice representations of compact Lie groups, that the residue field k must be $\mathbb{R}, \mathbb{C}$ or $\mathbb{H}$. Moreover, the non-commutative division algebra $\mathbb{H}$ occurs only when the type is $\tilde{\mathsf{A}}_n$ (since our diagram is simply laced, cf. [30, 35]). It follows from algebra that the complete discrete valuation field $\mathbb{K} = k((t))$, the field (division algebra) of formal Laurent series, and the valuation $v(f) \in \mathbb{Z}$ is the lowest degree of f with nonzero coefficient.

According to Bruhat-Tits, given a *special type* vertex o in the Coxeter diagram, the equivalence classes of sectors of type o defines a spherical building of type $\Pi - \{o\}$ where Π is the Coxeter diagram, called the spherical building at infinity. The spherical building at ∞ depends on the field $\mathbb{K}$. By Bruhat-Tits, if $\mathbb{K}$ is complete, then an affine building $\mathcal{C}$ of rank at least 4 is determined by the spherical building at ∞, denoted by $\mathcal{C}(\infty)$. The automorphism group $\mathcal{A}$ of the building $\mathcal{C}$ is the same as the automorphism group of the spherical building $\mathcal{C}(\infty)$, which is an algebraic group over $\mathbb{K}$.

It is an important feature of the affine buildings of types $\tilde{\mathsf{A}}_n$, $\tilde{\mathsf{D}}_n$, $\tilde{\mathsf{E}}_6$, $\tilde{\mathsf{E}}_7$, $\tilde{\mathsf{E}}_8$, that they are uniquely determined by the valuation field $(\mathbb{K}, v)$. Moreover, the residue field k is commutative unless the type is $\tilde{\mathsf{A}}_n$. The group $\mathcal{A}$ must be a split algebraic group over $\mathbb{K}$ of the same type. We point out that this is no longer true if the diagram has double bonds! In particular, for an $\tilde{\mathsf{A}}_n$ building, $\mathcal{A}$ contains $\mathsf{SL}_{n+1}(\mathbb{K})$ as a normal subgroup with quotient the automorphism group of the field $\mathbb{K}$ and diagram automorphism $\mathbb{Z}_2$.

• *Bruhat-Tits buildings and polar actions on Hilbert spaces*. As noticed above, spherical buildings over the classical fields are in 1-1 correspondence with polar representations. It is not completely clear yet, what should be the analogous correspondence for Bruhat-Tits buildings, but polar actions on Hilbert spaces first studied by Terng, seems to provide a model.

Let G be a compact connected semi-simple Lie group. The action of H^1-loops $H^1(\mathsf{S}^1, \mathsf{G})$ on the Hilbert space $V = H^1(\mathsf{S}^1, \mathfrak{g})$ given by the gauge transformation

$$g \cdot u = gug^{-1} - g'g^{-1}$$

is polar, and the constant loops with values in a Cartan subalgebra $\mathfrak{t}$ is a section. More generally, we have the following

Theorem 6.7 (Terng [33]). *Let G be a compact Lie group and $\mathsf{H} \subset \mathsf{G} \times \mathsf{G}$ be a closed Lie subgroup. Assume the biaction of H on G is polar with a flat torus section A. Let $\mathfrak{a}$ denote the Lie algebra of A. Let $P(\mathsf{G};\mathsf{H}) = \{g \in H^1([0,1];G) | (g(0);g(1)) \in \mathsf{H}\}$ and $V = H^0([0,1];\mathfrak{g})$. Then the gauge action of $P(\mathsf{G};\mathsf{H})$ on V is polar with section $\hat{\mathfrak{a}}$, the constant loops in $\mathfrak{a}$.*

We conjecture that every gallery connected component of the chamber system associated to the $P(\mathsf{G};\mathsf{H})$ action on V is an affine Bruhat-Tits building.

References

[1] M. Alexandrino, *On polar foliations and fundamental group,* Results Math. **60** (2011), no. 1-4, 213-223.

[2] E. M. Andreev, *Convex polyhedra in Lobačevskiĭ spaces* (Russian), Mat. Sb. (N.S.) **81 (123)** (1970), 445-478.

[3] M. Berger, *Riemannian geometry during the second half of the twentieth century*, University Lecture Series, 17. American Mathematical Society, Providence, RI, 2000.

[4] K.S. Brown, *Buildings,* Springer-Verlag, New York, 1989.

[5] N. Bourbaki, *Éléments de mathématique. Fasc. XXXIV. Groupes et algèbres de Lie. Chapitres 4, 5 et 6*, Hermann, Paris 1968.

[6] K. Burns and R. Spatzier, On topological Tits buildings and their classification. *Inst. Hautes etudes Sci. Publ. Math.* **65** (1987), 5-34.

[7] J. Cheeger and D. Gromoll, *On the structure of complete manifolds of nonnegative curvature*, Ann. of Math. **96** (1972), 413-443.

[8] H.S.M. Coxeter, *Discrete groups generated by reflections*, Annals of Math. **35** (1934), 588-621.

[9] J. Dadok, *Polar coordinates induced by actions of compact Lie groups*, Trans. Amer. Math. Soc. **288** (1985), 125–137.

[10] M.W. Davis, *The geometry and topology of Coxeter groups*, London Mathematical Society Monographs Series **32**. Princeton University Press, Princeton, NJ, 2008.

[11] W. von Dyck, *Gruppentheoretische Studien*, Math. Ann. **20** (1882), 1–44.

[12] F. Fang and K. Grove, *Reflection groups in non-negative curvature*, arXiv:1403.5019.

[13] F. Fang, K. Grove, and G. Thorbergsson, *Tits geometry and positive curvature*, arXiv: 1205.6222.

[14] F. Fang and X. Rong, *Positive pinching, volume and second Betti number*, Geom. Funct. Anal. **9** (1999), 641–674.

[15] ———, *The second twisted Betti number and the convergence of collapsing Riemannian manifolds*, Invent. Math. **150** (2002), 61–109.

[16] K. Grove, W. Ziller, *Curvature and symmetry of Milnor spheres*, Ann. of Math. **152** (2000), 331–367.

[17] L. Guijarro, *Improving the metric in an open manifold with nonnegative curvature*, Proc. Amer. Math. Soc., **126** (1998), 1541–1545

[18] V. Guillemin and S. Sternberg, *Convexity properties of the moment mapping, II*, Invent. Math. **77** (1984), 533–546.

[19] A.T. Huckleberry and T. Wurzbacher, *Multiplicity-free complex manifolds*, Math. Annalen **286** (1990), 261–280

[20] A. Khovanskii, *Combinatorics of sections of polytopes and Coxeter groups in Lobachevsky spaces*, The Coxeter legacy, 129–157, Amer. Math. Soc., Providence, RI, 2006.

[21] F. C. Kirwan, *Convexity properties of the momentum mapping, III,* Invent. Math. **77** (1984), 547–552.

[22] A. Lytchak, *Polar foliations on symmetric spaces*, arXiv:1204.2923v1 [math.DG].

[23] A. Neumaier, *Some sporadic geometries related to* $\mathrm{PGL}(3,2)$, Arch. Math. **42** (1984), 89–96.

[24] R. S. Palais and C.-L. Terng, *A general theory of canonical forms*, Trans. Amer. Math. Soc. **300** (1987), 771–789.

[25] A. Petrunin, W. Tuschmann, *Diffeomorphism finiteness, positive pinching and second homotopy,* Geom. Funct. Anal. **9** (1999), 736–774

[26] F. Podestà and G. Thorbergsson, *Polar actions on rank-one symmetric spaces*, J. Differential Geom. **53** (1999), 131–175.

[27] _____, *Polar and coisotropic actions on KÃd'hler manifolds*, Trans. Amer. Math. Soc. **354** (2002), 1759–1781

[28] H. Poincaré, *Théorie des groupes fuchsiens,* Acta Math. **1** (1882), 1–62.

[29] T. Püttmann, *Optimal pinching constants of odd-dimensional homogeneous spaces*, Invent. Math. **138**(1999), 631–684.

[30] M. Ronan, *Lectures on buildings,* Perspectives in Mathematics **7**. Academic Press, Inc., Boston, MA, 1989.

[31] J. Szenthe, *Orthogonally transversal submanifolds and the generalizations of the Weyl group*, Period. Math. Hungar. **15** (1984), 281–299.

[32] V. A. Sharafutdinov, *Convex sets in a manifold of nonnegative curvature*, (Russian) Mat. Zametki **26** (1979), 129–136.

[33] C.L. Terng, *Polar actions on Hilbert space*, J. Geom. Anal. **5** (1995), 129–150

[34] G. Thorbergsson, *Isoparametric foliations and their buildings*, Ann. of Math. **133** (1991), 429–446.

[35] J. Tits, *Buildings of spherical type and finite BN-pairs,* Lecture Notes in Mathematics **386**. Springer-Verlag, Berlin-New York, 1974.

[36] _____, A local approach to buildings, *The geometric vein*, The Coxeter Festschrift. Edited by C. Davis, B. Grünbaum, and F. A. Sherk, 519–547, Springer, New York-Berlin, 1981.

[37] E. B. Vinberg, *Discrete reflection groups in Lobachevsky spaces*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983), 593-601, Warsaw, 1984.

[38] R. Weiss, *The structure of affine buildings*, Annals of Mathematics Studies, 168. Princeton University Press, Princeton, NJ, 2009.

[39] A. Wörner, *A splitting theorem for nonnegatively curved Alexandrov spaces*, Geom. Topol. **16** (2012), 2391–2426.

Department of Mathematics, Capital Normal University, Beijing, 100048, China
E-mail: fuquan_fang@yahoo.com

Loop products, Poincaré duality, index growth and dynamics

Nancy Hingston

Abstract. A metric on a compact manifold M gives rise to a length function on the free loop space ΛM whose critical points are the closed geodesics on M in the given metric. Morse theory gives a link between Hamiltonian dynamics and the topology of loop spaces, between iteration of closed geodesics and the algebraic structure given by the Chas-Sullivan product on the homology of ΛM. Poincaré Duality reveals the existence of a related product on the cohomology of ΛM.

A number of known results on the existence of closed geodesics are naturally expressed in terms of nilpotence of products. We use products to prove a resonance result for the loop homology of spheres. There are interesting consequences for the length spectrum. We discuss briefly related results in Floer and contact theory.

Mark Goresky and Hans-Bert Rademacher are collaborators.

Mathematics Subject Classification (2010). Primary 58E10; Secondary 55P50.

Keywords. Closed geodesics, string topology, Morse theory.

1. Introduction

The search for closed geodesics on metric spheres has a long and very interesting history [10, 36, 51, 55, 57]. It features a number of beautiful and simple ideas whose details took decades to work out correctly. The history is also rich because of the variety of methods used to attack to the problem. The approach of dynamics is to look among the geodesics for those that close up, that is, to look for periodic orbits of the geodesic flow. The variational approach is to look among the closed curves for those that are geodesic. This latter method was developed extensively by Morse and is based on the fact that closed geodesics are critical points of the length or energy function.

Let M be a compact, simply connected manifold of dimension n with a Riemannian metric g. Let ΩM be the *based loop space* and ΛM the *free loop space* of M:

$$\begin{aligned} \Omega = &\quad \Omega M = H^{1,2}((S^1,*),(M,*)) \\ \subset &\quad \Lambda = \Lambda M = H^{1,2}(S^1,M). \end{aligned}$$

For technical reasons it turns out best to use not the length or the energy function, but the square-root of the energy:

$$F:\ \Lambda \to \mathbb{R}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

$$F(\gamma) = \sqrt{\int |\dot{\gamma}|^2 dt}.$$

Then $F(\gamma)$ has units of length, and $F(\gamma) \geq length(\gamma)$, with equality if and only if γ has constant speed. *The reader will not go wrong in thinking that F is the length function.* The critical points of F are precisely the closed gedesics on (M, g), and thus one can walk in the footsteps of Morse [55] and use the topology of Λ to find closed geodesics on M. The (approximate) correspondence between the topology of Λ and critical points/values of F works more specifically like this:

$$H_k(\Lambda) \approx \text{ critical points of index } k$$

(A *critical point* of F is a loop $\gamma \in \Lambda$ where all the first partial derivatives of F vanish. The *index* of γ is the maximal dimension of a subspace of the tangent space to Λ at γ where the second derivative is negative definite.) We get a map

$$X \in H_k(\Lambda) \to Cr(X) \in \mathbb{R}$$

where

$$Cr(X) =: \inf\{a \in \mathbb{R} : X \text{ has a representative in } \Lambda^{\leq a} =: F^{-1}[0, a]\}$$

is the *critical level* of the homology class X.

The "correspondence" follows from the fact that $Cr(X)$ is a critical value of F; and there is a critical point (or points) $\gamma \in \Lambda$ of length $Cr(X)$ with

$$index(\gamma) \leq \deg X \leq index(\gamma) + nullity(\gamma)$$

on which the homology class X is said to "lie hanging". Thus one can hope to get a rough "count" of closed geodesics by taking a "count" of homology classes. There is however a major difficulty: *iteration*. If $\gamma \in \Lambda$ is a closed geodesic, then so is each of its iterates $\gamma^m \in \Lambda$, where

$$\gamma^m(t) = \gamma(mt).$$

These iterates are geometrically indistinguishable, and should together contribute just 1 to our count of geodesics. But they are very different points in Λ; they have different lengths,

$$\ell(\gamma^m) = m\ell(\gamma)$$

and in general they have different indices as critical points. From the point of view of critical point theory, one closed geodesic looks like an army!

Bott [11] proved that the index of the iterates grows approximately linearly, and Gromoll and Meyer [34] that the contributions to the homology from the iterates of a single closed geodesic are bounded. Together these results led to the Theorem of Gromoll and Meyer: If the rank of $H_k(\Lambda; \mathbb{Q})$ is unbounded, then for any metric on M there must be infinitely many closed geodesics.

Sullivan and Vigué-Poirrier [66] proved that (for M compact and simply connected) the rank of $H_k(\Lambda; \mathbb{Q})$ is unbounded if and only if the cohomology ring $H^*(M; \mathbb{Q})$ is not a truncated polynomial ring in one generator; thus the Theorem of Gromoll and Meyer applies to "most" such manifolds. However for spheres and projective spaces the rank of $H_k(\Lambda; \mathbb{Q})$

is bounded (the groups $H_k(\Lambda;\mathbb{Q})$ are periodic in k for large k), and the Theorem does not apply. While for the standard metric on these spaces all geodesics are closed, it is not known whether there is a metric on a sphere or projective space of dimension > 2 with closed geodesics, or a metric with only one closed geodesic! It is known ([6, 8, 10, 27, 33, 39, 51]) that for any metric on S^2, and [38, 59] for a *generic* metric on a sphere or projective space, there are infinitely many closed geodesics. There are *Finsler* metrics on spheres and projective spaces with only finitely many closed geodesics [46, 68]. For our purposes the difference between Finsler metrics and Riemannian metrics is that for a Finsler metric the metric is not required to be reversible; thus reversing the direction will likely change the length of a path. Most of the methods from Riemannian geometry for proving the existence of closed geodesics also work for Finsler metrics, but "curve-shortening" [33] may introduce self-intersections.

If we hope to count closed geodesics on M via the correspondence between homology and critical points, we are led to the question: *Is there an algebraic operation on $H_k(\Lambda)$ that corresponds to iteration of closed geodesics?* In some critical cases the answer is yes.

2. Products

We begin with the Pontryagin product [58]

$$\bullet_\Omega : H_j(\Omega) \times H_k(\Omega) \to H_{j+k}(\Omega).$$

It is induced by the concatenation product on loops

$$\begin{aligned} \bullet = \ : \bullet_\Omega : \Omega \times \Omega &\to \Omega \\ (\alpha, \beta) &\to \alpha \bullet_\Omega \beta \end{aligned}$$

(First go around α, then β.) We (abusively) think of cycles X, Y in Ω as subsets: $X = \{\alpha\} \subset \Omega$, $Y = \{\beta\} \subset \Omega$; then in $H_*(\Omega)$ the product is given by

$$\begin{aligned} [X] \bullet_\Omega [Y] &= : [X \bullet_\Omega Y] \\ &= : [\{\alpha \bullet_\Omega \beta\}_{\alpha \in X; \beta \in Y}] \end{aligned}$$

Example 2.1. Let $M = S^n$ be the round sphere of radius 1 in $\mathbb{R}^{n+1}$ Then every geodesic on M is closed, and the critical values of F are the numbers $2m\pi$, $m \in \mathbb{Z}^{\geq 0}$ A *circle* on S^n is the intersection of S^n with a 2-plane in $\mathbb{R}^{n+1}$, parameterized with constant speed and with minimal period 1 (unless constant). Pick a vector $\overrightarrow{V}$ at the basepoint $*$ in S^n. Let A be the $(n-1)$-dimensional cycle on ΛS^n consisting of all circles on S^n beginning at $(*)$ with tangent vector $\lambda \overrightarrow{V}$ for some $\lambda \geq 0$, and let U be the 0-dimensional cycle on ΛS^n consisting of the constant loop at $(*)$. We invite the reader to check that

$$\begin{aligned} Cr[U] &= 0 \\ Cr[A] &= 2\pi \\ [U] \bullet_\Omega [A] &= [A] \bullet_\Omega [U] = [A], \end{aligned}$$

and that $[U]$ is the identity element in the Pontryagin ring.

Next what is

$$[A] \bullet_\Omega [A] =: [A]^2 \ ?$$

It should be clear that

$$Cr([A]^2) \leq 2Cr[A] = 4\pi;$$

indeed using the fact that $[A] \bullet_\Omega [A] = [A] \bullet_\Omega [A']$, where A' is the cycle on ΛS^n consisting of all circles on S^n beginning at $(*)$ with tangent vector $\lambda \overrightarrow{V'}$ with $\overrightarrow{V'} \neq \overrightarrow{V}$, and the fact that every loop in $A \bullet_\Omega A'$ of length $\geq 2\pi$ has a "corner" that can be cut to make it shorter, we see that $Cr[A \bullet_\Omega A] < 4\pi$, from which it follows that $Cr[A \bullet_\Omega A] \leq 2\pi$ (since $Cr[A \bullet_\Omega A]$ is a critical value of F). In fact $[A \bullet_\Omega A] = [B]$, where B is the $(2n-1)$-cycle consisting of all circles beginning at $(*)$, and the Pontryagin ring is the polynomial ring

$$H_*(\Omega S^n) = \mathbb{Z}[A].$$

Note that $[A]$ is *nonnilpotent*, that is for all m we have $[A]^m \neq [0]$, but $[A]$ is *level nilpotent*, by which we mean that, for some $m > 1$, $Cr\ [A]^m < mCr\ [A]$: One can show that

$$Cr\ [A]^{2m-1} = Cr\ [A]^{2m} = Cr\ [B]^m = 2m\pi.$$

So $[B]$ *is* level nonnilpotent.

Chas Sullivan product. This is a product on the homology of the free loop space $H_*(\Lambda)$, with degree $-n$:

$$\bullet_{CS} : H_j(\Lambda) \times H_k(\Lambda) \to H_{j+k-n}(\Lambda)$$

We give here the intuitive idea of the definition from the original paper [16] of Chas and Sullivan; for a more rigorous definition see [18, 19]. Let

$$e : \ \Lambda \to M$$
$$e(\gamma) = \gamma(0)$$

be the evaluation map. Let $X = \{\alpha\}$ and $Y = \{\beta\}$ be cycles in Λ. Assume that eX and eY intersect transversally in M. Then

$$[X] \bullet_{CS} [Y] =: [\{\alpha \bullet \beta : \alpha \in X,\ \beta \in Y, \text{ and } e\alpha = e\beta\}].$$

Example 2.2. Let $M = S^n$ be the round sphere of radius 1 in $\mathbb{R}^{n+1}$ Let $A, B, U \subset \Omega \subset \Lambda$ be as above. (But for degree reasons U is not the unit in the Chas-Sullivan ring, and $A^2 \neq B$.) Let C be the $(3n-2)-$ dimensional cycle in Λ consisting of all (parameterized) circles on S^n ("all circles great and small"), and E the $n-$ dimensional cycle in Λ consisting of all constant loops. We invite the reader to verify the following

(1) $[C] \bullet_{CS} [E] = [C]$

(2) $[E]$ is the identity element in the Chas-Sullivan ring.

(3) $[A] \bullet_{CS} [A] = 0$

(4) If X ,$Y \subset \Omega$, then $[X] \bullet_{CS} [Y] = 0$

(5) $[U] \bullet_{CS} [C] = [B]$

The Chas-Sullivan rings of spheres and projective spaces, which do not depend on the metric, were computed by [20], using somewhat more algebraic techniques. These rings are finitely generated (the generators listed above are almost enough) and the product is highly nontrivial; for example the element $[C]$ is nonnilpotent: $[C]^m \neq 0$, and with the standard metric on M the generator $[C]$ is level-nonnilpotent:

$$Cr\ [C]^m = mCr\,[C] = 2m\pi.$$

3. Poincaré duality

These cycles, and their Chas-Sullivan products, would have looked very familiar to Morse [55], who studied the topology of ΛS^n in the hope of finding closed geodesics. The cycles and products are also reminiscent of the work of of Bott-Samelson on manifolds all of whose geodesics are closed, which beautifully rounds out the Bott-Gromoll-Meyer-Vigué-Poirrier-Sullivan circle of ideas. When I first heard the definition of the Chas-Sullivan product (at the lunch table at the Institute for Advanced Study in 2005) I recognized it right away because I knew well these examples. I asked, "Where is the other product?"

There is a pervasive symmetry in the study of loop spaces and closed geodesics that can be thought of as Poincaré duality in the free loop space. If F is a Morse function on an oriented, compact manifold X of dimension N, then Poincaré duality can be thought of as "turning X upside down" like an hourglass: Following Milnor [54] we take the vertical axis to be the F-axis; thus turning the space X upside down is the same as replacing F with $-F$. The homology of X is computed by the cell complex of Morse-chains; for each critical point γ of index λ we have a disk of dimension λ that "hangs down" from γ and is attached below. The cohomology is computed by the cell complex of Morse cochains; for each critical point of coindex μ there is a a disk of dimension μ that "hangs up" from a critical point. The Kronecher product between homology and cohomology is given by the intersection pairing on Morse cycles and Morse cocycles. A homology class in dimension λ, when turned upside down, becomes the cohomology class of dimension $N - \lambda$ that is its Poincaré dual. As the free loop space is infinite dimensional, there is no duality map between homology and cohomology groups in complementary dimensions. However many things "work the same" if you "turn the free loop space upside down", that is if you do Morse theory on the free loop space with the function $-F$ instead of F. The guiding principle that every phenomenon in the free loop space has an dual counterpart has proven to be very powerful in a variety of contexts. We will present evidence of this duality principle and let you decide for yourself.

Mark Goresky and the author [32] looked for and found the product that is Poincaré dual to the Chas-Sullivan product, and investigated its properties. It was clear from the duality principle that the product should be of the form[1]

$$\circledast : H^j(\Lambda) \otimes H^k(\Lambda) \to H^{j+k+n-1}(\Lambda). \quad (*)$$

[1] In [32] the product is of the form

$$H^j(\Lambda, \Lambda^0) \otimes H^k(\Lambda, \Lambda^0) \to H^{j+k+n-1}(\Lambda, \Lambda^0)$$

but there is a way to extend the product to the form $(*)$.

The associated homology coproduct

$$\vee : H_i(\Lambda) \to \Sigma_{j+k=i+1-n} H_j(\Lambda) \otimes H_k(\Lambda)$$

had been previously introduced by Sullivan [62].

The "1" in the degree $(n-1)$ of the cohomology product is related to the index of the iterates of closed geodesics (see below), and to the observed fact that the duality principle works best not in the space Λ of *(parameterized)* $H^{2,1}$ loops, but in the space of *optimally parameterized* loops, the loops that are parameterized proportional to arclength. In this space F and $-F$ are on a more even footing, since we have thrown out all the nonoptimal parameterizations. These nonoptimal parameterizations are all directions in which F increases, for (arguably) no interesting reason if one is concerned with the geometry of loops. There is also a product on the based loops

$$\circledast_\Omega : H^j(\Omega) \otimes H^k(\Omega) \to H^{j+k+n-1}(\Omega).$$

whose relationship with the loop cohomology product $\circledast$ reflects the relationship [20] between the Chas-Sullivan and Pontryagin products. In [32] Goresky and the author showed that when M has a metric with all geodesics closed (for example if M is a sphere or projective space) and $n > 2$, the products $\circledast$ and $\circledast_\Omega$ are highly nontrivial; in particular the rings are finitely generated, as are the Pontryagin ring and the Chas-Sullivan rings computed by Cohen Jones and Yan [20]. Moreover the associated graded rings from the grading induced by the energy function (or F) in these cases can be computed by a general method based on Morse theory and geometry, and are also finitely generated and nontrivial. The complete ring structure is computed for spheres in [32]; they have the remarkable property that the rings are independent of the dimension n when n is odd, or when the coefficients are $\mathbb{Z}/2\mathbb{Z}$.

The *critical level of a cohomology class* is the reflected version of the critical level of a homology class:

$$Cr(x) = \sup\{a : x \text{ has support in } \Lambda^{\geq a}\}$$

where $\Lambda^{\geq a} = F^{-1}[a, \infty)$. The Chas-Sullivan product $\bullet$ and the loop cohomology product $\circledast$ satisfy the following dual basic inequalitites:

$$\begin{aligned} Cr(X \bullet Y) \leq & Cr(X) + Cr(Y) \quad \text{(homology)} \\ Cr(x \circledast y) \geq & Cr(x) + Cr(y) \quad \text{(cohomology)} \end{aligned}$$

On the based loop space, the Pontryagin product and the loop cohomology products satisfy the same inequalities. The cup-product does *not* satisfy such an inequality; see [32] 16.1. A cohomology class x is *level nilpotent* if for some m, $Cr(x^{\circledast m}) > mCr(x)$. The level nilpotence of a homology or cohomology product is related to the *index growth* of the critical point on which it lies hanging. Bott [11] proved, using a beautiful argument based on intersection theory in the symplectic group, that the index and nullity of the iterates γ^m of a closed geodesic γ satisfy the inequalities

$$\begin{aligned} m \cdot index(\gamma) - (m-1)(n-1) \leq & index(\gamma^m) \\ \leq & index + nullity(\gamma^m) \\ \leq & m \cdot index(\gamma) + (m-1)(n-1) \quad (**) \end{aligned}$$

If *equality*

$$m \cdot index(\gamma) - (m-1)(n-1) = index(\gamma^m)$$

holds for $m = m_0$, then if holds for $1 \leq m \leq m_0$, and we say the index growth is *minimal* up to $m = m_0$. If equality

$$index + nullity(\gamma^m) = m \cdot index(\gamma) + (m-1)(n-1)$$

holds for $m = m_0$, then if holds for $1 \leq m \leq m_0$, and we say the index growth is *maximal* up to $m = m_0$. *In the extreme cases the Chas-Sullivan powers, and the loop cohomology powers correspond to the iteration of closed geodesics*: When a homology class X lies hanging on a closed geodesic γ with minimal index growth, the Chas-Sullivan product X^{*m} lies hanging on the iterate γ^m; when a cohomology class x lies hanging on a closed geodesic γ with maximal index growth, the cohomology product x^{*m} lies hanging on the iterate γ^m.

As further evidence of Poincaré Duality here are some restatements of old theorems in terms of loop products. When these theorems were proved, the products had not been discovered. In retrospect it is clear that these theorems are naturally expressed in terms of products.

Bott [11]: *Let M be a compact Riemannian manifold. If all closed geodesics are non-degenerate (in the sense of Morse theory), then every homology class in $H_*(\Lambda)$ is level-nilpotent, and every cohomology class in $H^*(\Lambda)$ is level-nilpotent.*

Hingston [39]: *Let M be a compact Riemannian manifold. If there is a homology class in $H_*(\Lambda)$ that is not level nilpotent, then M has infinitely many closed geodesics. In particular, there is a closed geodesic of length L, $m_0 \in \mathbb{Z}^+$, and a sequence $\sigma_m \downarrow 0$ so that if $m \geq m_0$, M has a closed geodesic with length $\ell \in (mL, mL + \sigma_m)$.*

Hingston [40]:*Let M be a compact Riemannian manifold. If there is a cohomology class in $H^*(\Lambda)$ that is not level nilpotent, then M has infinitely many closed geodesics. In particular, there is a closed geodesic of length L, $m_0 \in \mathbb{Z}^+$, and a sequence $\sigma_m \downarrow 0$ so that if $m \geq m_0$, M has a closed geodesic with length $\ell \in (mL - \sigma_m, mL)$.*

The proofs in [39] and [40] are quite different, and the second is much more difficult. They are both variations on "Bangert's Lemma" [7]. There are many possibilities for the local geometry of a closed geodesic whose index for large m lies between the two bounds in $(**)$. But in the limiting cases the geometry is forced into a rigid mold which can be described exactly. The proof in [39] was discovered by accident; the guiding light of the Poincaré Duality principle then motivated the author to find the proof of [40].

Many of the early work on loop spaces was done not in the infinite dimensional setting, but in the finite dimensional approximation described by Morse [55] and refined by Bott [12] and Milnor [54]. The finite dimensional approximation space is a space of piecewise geodesic loops consisting of N minimal geodesic pieces, and parameterized by the vertices $\{(x_1, , , , , x_N)\} \subset M^N$. The finite dimensional approximation space is a finite dimensional manifold that for N large enough carries the topology and the Morse theory of $\Lambda^{\leq L}$ for L arbitrarily large. The Chas-Sulllivan product and the loop cohomology product are *almost* Poincaré dual in the finite dimensional approximation. The catch is that duality works best in the optimally parameterized loops, but the space of optimally parameterized piecewise geodesic loops (i.e. those for which the N pieces all have the same length) has singularities and (worse) spurious critical points. The loop cohomology product was originally defined by Goresky and the author using Poincaré duality in the finite dimansional approximation. When transformed to the infinite dimensional setting, the definition took it current form.

4. Applications to dynamics

In the search for periodic points it often turns out that the (necessarily degenerate) cases of minimal and maximal index growth for all m are precisely the cases not covered by other methods and thus the last cases to be proved. One example is the earlier mentioned Theorem [6, 8, 10, 27, 33, 39, 51]: Any Riemannian metric on S^2 has infinitely many closed geodesics. This was first proved by arguments of Lusternick-Schnirelmann/Ballmann/Grayson, Birkhoff, Bangert, and Franks. The last case to be proved was the case of extremal index growth. The author in [39] gave an alternative argument for the last step using the ideas outlined above. In his beautiful 2005 Annals paper [5] Angenent gives a new proof of the existence of infinitely many closed geodesics on S^2 except in the cases covered in [39]. Another example: The arguments in [39] and [40] are also valid for Finsler metrics. In 2010 Bangert and Long [9] proved the existence of at least two closed geodesics for any Finsler metric on S^2 by reducing the proof to the case of extremal growth.This result is sharp due to the wonderful examples of Katok [46, 69].

5. Related ideas in Floer / symplectic / contact theory

We mention briefly some related results and refer the reader to the references for detail [4, 22, 23, 26].

(1) Let M be a closed, symplectic manifold of dimension $2n$ with (for the sake of simplicity, though this is just the tip of the iceberg) $\pi_2 M = 0$. Let

$$H : S^1 \times M \to \mathbb{R}$$

be a periodic Hamiltonian. Following Arnold, Conley, Zehnder, Salamon, Floer, and many others, we look for periodic points of the Hamiltonian flow, which is given in local coordinates by

$$\dot{x} = J\nabla H(x,t).$$

These periodic points are critical points of the action function $\mathcal{A} : \Lambda M \to \mathbb{R}$. The critical points all have infinite index and coindex, but Floer theory was invented to circumvent this difficulty. When M is a torus, one can use a finite dimensional approximation. It was proved by Conley-Zehnder (in the case where M is a torus), and Salamon-Zehnder, that if all period 1 orbits are nondegenerate, then there are periodic orbits of arbitrarily high minimal integer period. The role of the nondegeneracy hypothesis is to ensure that the growth of the index (the Conley-Zehnder index in this case) is not extremal. An argument in the spirit of Bangert's Lemma, and in the spirit of [39, 40], applies in the case of extremal index growth, and was used by the author [41] (in the case where M is a torus) and by Ginzburg [28] to prove Conley's Conjecture: There are always periodic orbits of arbitrarily high minimal integer period. In [30] Ginzburg and Gürel state these results in terms of the pair-of-pants product on the Floer homology of the free loop space, and nonnilpotence of certain level homology classes occurs precisely when there is maximal or minimal index growth. One also has the exact analogs of the statements regarding the critical values: $\ell \in (mL, mL + \sigma_m)$ and $\ell \in (mL - \sigma_m, mL)$ in the cases of slow and fast growth.

(2) If M is a compact manifold, the cotangent bundle T^*M carries a natural symplectic structure. Let $H : S^1 \times T^*M \to \mathbb{R}$ be a periodic Hamiltonian of quadratic type. (The geodesic flow on T^*M is given by the Hamiltonian $H = |p|^2$.) There is an isomorphism between the Floer homology of the cotangent bundle of M and the homology of the free loopspace of M, in which the pair-of-pants product corresponds to the Chas-Sullivan product. [1, 2, 60, 64]. Abbondandolo and Schwarz [3] have described the product on the Floer side that corresponds to the loop coproduct.

(3) The contact setting. A contact form on M gives rise to an action function A on the free loop space ΛM, whose critical points are the closed Reeb orbits. Contact homology is the associated Morse homology, introduced using Floer theoretic tools by Eliashberg-Giventhal-Hofer in [25]. The iteration properties of local Floer homology were studied by Ginzburg-Gurel [30] and analogs of the theorem of Gromoll-Meyer were found in [45], McLean [1, 2, 17, 53, 60, 64, 65]. There are analogs of the theorem of Bangert-Long [9] mentioned above by reducing to the case of extremal growth: [31, 48]. See also [24, 29]. There are products in this context; see [17].

6. Critical levels of spheres

If M is compact and simply connected and carries a Riemannian metric, Gromov [35] (see also [56]) proved that the set of points

$$(C, d) = (Cr(X), \deg X) \in \mathbb{R}^2,$$

where $X \in H_d(\Lambda)$, $Cr(X)$ is the critical level of X, and $Cr(X) > 0$, lie between two lines: There are numbers $\mu, \nu > 0$ so that

$$\mu Cr(X) < \deg X < \nu Cr(X).$$

Note the critical levels have units of length. For simplicity all homology and cohomology will have rational coefficients. In [42] the author and Rademacher prove the

Theorem 6.1 (Resonance Theorem for Spheres). *If M is a sphere of dimension > 2, with a fixed Riemannian or Finsler metric g , then the points $(Cr(X), \deg X)$ lie at a finite distance from a line: there are constants$\overline{\alpha}$ and $\beta > 0$ so that*

$$\overline{\alpha} Cr(X) - \beta < \deg X < \overline{\alpha} Cr(X) + \beta.$$

The slope

$$\overline{\alpha} = \overline{\alpha}_g$$

is called the *global mean frequency* of of (M, g) and has units of conjugate points per unit length. While we have a proof of this theorem we still lack an explanation, and the theorem still seems quite surprising!

There is a spectral sequence bigraded by (length, degree), converging to $H_*(\Lambda)$, whose $\mathcal{E}^1$ page is a direct sum of the local level homology of the all the iterates of all the closed geodesics on M in the given metric. By a theorem of Bott [11] mentioned earlier $(**)$, the contributions of all the iterates of a single closed geodesic γ to the $\mathcal{E}^1$ page lie at a finite distance from the line

$$d = \overline{\alpha}_\gamma C$$

whose slope is the mean frequency

$$\overline{\alpha}_\gamma = \frac{\alpha_\gamma}{\ell(\gamma)},$$

where

$$\alpha_\gamma = \lim_{m\to\infty} \frac{index(\gamma^m)}{m}$$

is the *average index* of γ.

If the sectional (or, in the Finsler case, flag) curvature K of M is bounded between constants

$$K_1 < K < K_2$$

then the mean frequency is bounded as follows:

$$\frac{\sqrt{K_1}(n-1)}{\pi} < \overline{\alpha}_\gamma < \frac{\sqrt{K_2}(n-1)}{\pi}$$

For the round metric, and for the Katok metrics [46, 69] the curvature K is constant and thus the mean frequency of every closed geodesic satisfies

$$\overline{\alpha}_\gamma = \overline{\alpha}_g = \frac{\sqrt{K}(n-1)}{\pi},$$

though in general the closed geodesics in the Katok metrics have different lengths and different average indices. For these metrics all nonzero terms in the $\mathcal{E}^1$ page lie at a finite distance from the line

$$d = \frac{\sqrt{K}(n-1)}{\pi} C$$

and the theorem follows immediately. However an *open mapping theorem* [42] tells us that the average indices of geometrically distinct closed geodesics can be perturbed independently. Thus we expect that for a generic metric the $\mathcal{E}^1$ page has entries lying (approximately) along a union of lines, one line for each closed geodesic γ, with the spacing along each line determined by $\ell(\gamma)$. One proceeds from one page of the spectral sequence to the next by allowing certain generators to "cancel out" in pairs. Somehow in this process the generators that remain at the $\mathcal{E}^\infty$ page all lie with in a finite distance of the line $d = \overline{\alpha}C$. Unless a closed geodesic γ has mean frequency $\overline{\alpha}_\gamma$ *exactly* equal to the global mean frequency $\overline{\alpha}$, only a finite number of homology classes can lie hanging on the iterates of γ.

The Katok examples include nondegenerate Finsler metrics on spheres with only finitely many closed geodesics. If $M = S^n$ with $n > 2$ odd, it is proved in [42] :

Corollary 6.2. *For a metric g in a neighborhood of the Finsler metrics, and not too far from the round metric, at least one of the following holds:*

(1) *There are at least two closed geodesics γ with mean frequency $\overline{\alpha}_\gamma = \overline{\alpha}_g$.*

(2) *There is a sequence of closed geodesics $\{\gamma_j\}$ with mean frequencies $\overline{\alpha}_j \neq \overline{\alpha}_g$, satisfying*

$$\lim_{j\to\infty} \overline{\alpha}_j = \overline{\alpha}_g.$$

This is reminiscent of the conjecture of Hofer-Wysocki-Zehnder [44] to the effect that every Reeb flow for the tight 3-sphere has exactly two or infinitely many closed Reeb orbits. There is a conjecture that on S^2 a Finsler metric carries exactly two or infinitely many closed geodesics. This is proved in [37] near the standard metric. For $n > 2$ there is hope that the number "two" in the Corollary could be improved to the minimal number of geodesics for a Katok metric on S^n. A famous open problem is to prove the existence of infinitely many closed geodesics for any Riemannian metric on a sphere of dimension $n > 2$. As mentioned above, when $n = 2$ this statement has been proved for Riemannian metrics and is false for Finsler metrics by the Katok examples. An even more beautiful theorem would include the Finsler metrics and say that the only exceptions to the existence of infinitely many closed geodesics are very simple metrics in the Katok model.

Acknowledgements. The author is grateful to the Institute for Advanced Study, and especially to Helmut Hofer for the vibrant mathematical community they have fostered. Thanks to Umberto Hryniewicz for numerous discussions and for his help in putting together this paper.

References

[1] Abbondandolo, B., Schwarz,M., *Notes on Floer homology and loop space homology, Morse theoretic methods in nonlinear analysis and in symplectic topology*, Proceedings of the NATO Advanced Study Institute, Montreal, Canada, July 2004, Springer, Dordrecht, 2006.

[2] _____, *On the Floer homology of cotangent bundles*, Comm. Pure. Appl. Math. **59** (2005), 254–316.

[3] _____, *On product structures on Floer homology of cotangent bundles*, Preprint MPI MIS no.76, 2010.

[4] Amann, H., and E. Zehnder, *Nontrivial solutions for a class of non-resonance problems and applications to nonlinear differential equations*, Annali Sculoa sup. Pisa Cl. Sc. Serie IV, VII **4** (1980), 539–603.

[5] Angenent, S.B., *Curve shortening and the topology of closed geodesics on surfaces*, Annals of Mathematics **162** (2005), 1187–1241.

[6] Ballmann, W., *Der Satz von Lusternik und Schnirelmann*, Bonner Mathematische Schriften **102** (1978), Universität Bonn, Bonn, 1–25.

[7] Bangert, V., *Closed geodesics on complete surfaces*, Math. Ann. **251** (1980), no. 1, 83–96.

[8] _____, *On the existence of closed geodesics on two-spheres*, Int. J. Math. **4** (1993), 1–10.

[9] Bangert,V., Long,Y., *The existence of two closed geodesics on every Finsler 2-sphere*, Math. Ann. **346** (2010), 335–366

[10] Birkhoff, G.D., *Dynamical systems*. Amer. Math. Soc. Collqo. Pub., vol.9, NewYork, 1927.

[11] Bott, R., *On the iteration of closed geodesics and the Sturm intersection theory*, Comm. Pure Appl. Math. **9** (1956), 171–206.

[12] _____, *Morse theory and its application to homotopy theory*, Lectures delivered in Bonn, 1958, notes by A. van de Ven. Harvard lecture notes; Reprinted in Collected Papers, Vol. 1-4, Birkhäuser Boston, 1994, 1995.

[13] Bott, R., and H. Samelson, *The Pontryagin product in spaces of paths*, Comment. Math. Helv. **27** (1953), 320–337.

[14] Bredon, G., *Topology and Geometry*, Springer Verlag, New York, 1993.

[15] Chang, K-C., *Infinite Dimensional Morse Theory and Multiple Solution Problems*, PNLDE **6**, Birkhäuser, Boston, 1993.

[16] Chas, M., D. Sullivan, *String topology*, preprint, math.GT/9911159 (1999).

[17] Cieliebak, K., Latschev, J., *The role of string topology in symplectic field theory*, New perspectives and challenges in symplectic field theory, CRM Proc. Lecture Notes, **49** (2009), Amer. Math. Soc., Providence, RI, 113–146.

[18] Cohen,R., *Homotopy and geometric perspectives on string topology*, Lecture notes, Stanford University, 2005.

[19] Cohen, R., Jones, J., *A homotopy theoretic realization of string topology*, Math. Ann. **324** (2002), 773–798.

[20] Cohen, R., Jones, J.,Yan, J., *The loop homology algebra of spheres and projective spaces*, Progr. Math. **215**(2003), Birkhauser, Basel, 77–92.

[21] Cohen, R., Klein, J., Sullivan,D., *The homotopy invariance of the string topology loop product and string bracket*, Journal of Topology, **2** (2008), 391–408.

[22] Conley, C., Zehnder, E., *The Birkhoff-Lewis Fixed Point Theorem and a Conjecture of V. I. Arnold*, Invent. Math. **73** (1983), 33–49.

[23] _____, *A global fixed point theorem for symplectic maps and subharmonic solutions of Hamiltonian equations on tori*, Proc. Symp. Pure Math. **45** (1986), 283–299.

[24] Cristofaro-Gardiner, D., Hutchings, M., *From one Reeb orbit to two*, arXiv:1202.4839.

[25] Eliashberg, Y., Givental, A., Hofer, H., *Introduction to symplectic field theory*, Geom. Funct. Anal. 2000, Special Volume, Part II, 560–673.

[26] Floer, A., *Proof of the Arnold conjecture for surfaces and generalization to certain Kahler manifolds*, Duke Math. J. **53** (1986), 1–32.

[27] Franks, J., *Geodesics on S^2 and periodic points of annulus homeomorphisms*, Invent. Math. **108** (1992), 403–418.

[28] Ginzburg,V. L., *The Conley conjecture*, Ann. of Math. **172** (2010) 1127–1180.

[29] Ginzburg, V., Gören,Y., *Iterated Index and the Mean Euler Characteristic*, arXiv:1311.0547.

[30] Ginzburg, V., Gürel, B., *Local Floer homology and the action gap*, J. Symplectic Geom. **8** (2010), no. 3, 323–357.

[31] Ginzburg, V., Hein, D., Hryniewicz, U., Macarini, L., *Closed Reeb orbits on the sphere and symplectically degenerate maxima*, Acta Math. Vietnam. **38** (2013), no. 1, 55–78.

[32] Goresky, M., Hingston, N., *Loop products and closed geodesics*, Duke Math. J. **150** (2009), no. 1, 117–209.

[33] Grayson, M., *Shortening embedded curves*, Ann. Math. **129** (1989) 71–111.

[34] Gromoll, D., and W. Meyer, *Periodic geodesics on compact Riemannian manifolds*, J. Diff. Geom. **3** (1969), 493–510.

[35] Gromov, M., *Homotopical effects of dilatation*, J. Differential Geom. **13** (1978) no. 3, 303–310.

[36] Hadamard, J., *Les surfaces à courbures opposées et leur lignes géodesiques*, J. Math. Pures Appl (5), **4** (1898), 27–73.

[37] Harris, A., Paternain, G., *Dynamically convex Finsler metrics and J-holomorphic embedding of asymptotic cylinders*, Ann. Global Anal. Geom. **34** (2008), no. 2, 115–134.

[38] Hingston, N., *Equivariant Morse theory and closed geodesics*, J Diff. Geom. **19** (1984), 85–116.

[39] _____, *On the lengths of closed geodesics on a two-sphere*, Proc. Amer. Math. Soc. **125** (1997), 3099–3106.

[40] _____, *On the growth of the number of closed geodesics on the two-sphere*, Int. Math. Res. Not. **9** (1993), 253–262.

[41] _____, *Subharmonic solutions of Hamiltonian equations on tori*, Ann. of Math. (**2**) 170 (2009), no. 2, 529–560.

[42] Hingston, N., Rademacher, H-B., *Resonance for loop homology of spheres*, J. Differential Geom. **93** (2013), no. 1, 133–174.

[43] Hofer, H., Zehnder, E., *Symplectic Invariants and Hamiltonian Dynamics*, Birkhäuser Verlag, 1994.

[44] Hofer, H., Wysocki, K., Zehnder, E., *The dynamics on three-dimensional strictly convex energy surfaces*, Ann. of Math. (**2**) 148 (1998), no. 1, 197–289.

[45] Hryniewicz, U., Macarini, L., *Local contact homology and applications*, arXiv:1202.3122.

[46] Katok, A., *Ergodic perturbations of degenerate integrable Hamiltonian systems,Izv*, Akad. Nauk SSSR. **37** (1973), 539-576. Engl. transl., Math. USSR-Izv.**7** (1973), 535–571.

[47] Klingenberg, W., *Lectures on Closed Geodesics*, Grundlehren der mathematischen Wissenschaften **230**, Springer Verlag, Berlin, 1978.

[48] Liu, H., Long, Y., *On the existence of two closed characteristics on every compact star-shaped hypersurface in* $\mathbf{R^4}$, arXiv:1308.3904.

[49] Long, Y., *A Maslov-type index theory for symplectic paths*, Topological Methods in Nonlinear Analysis **10** (1997), 47–78.

[50] ———, *Index Theory for Symplectic Paths with Applications*, Progress in Mathematics **207**, Birkhäuser, Basel, 2002.

[51] Lusternik, L., Schnirelmann, L., Sur la problème de trois géodésiques fermées sur les surfaces de genus 0, *C.R.Acad. Sci. Paris* **189** (1929), 269–271.

[52] McDuff, D., Salamon, D., *Introduction to Symplectic Topology*, Oxford University Press, 1995.

[53] McLean, M., *Local Floer homology and infinitely many simple Reeb orbits*, Algebr. Geom. Topol. **12** (2012), no. 4, 1901–1923.

[54] Milnor, J., *Morse Theory*, Annals of Mathematics Studies **51**, Princeton University Press, Princeton N.J., 1963.

[55] Morse, M., *Calculus of Variations in the Large*, Amer. Math. Soc. Colloquium Publications, XVIII, Providence, R.I., 1934.

[56] Paternain, G. P., *Geodesic flows.* Progress in Mathematics, **180**, Birkhäuser Boston, Inc., Boston, MA, 1999.

[57] Poincaré, H., Sur les lignes géodesiques des surfaces convexes. *Trans. Amer. Math. Soc*. **6**, 237–274 (1905).

[58] Pontrjagin, L. , *Homologies in compact Lie groups*, Rec. Math. N. S. [Mat. Sbornik], **6**(48) (1939), 389–422.

[59] Rademacher, H.B., *On the average indices of closed geodesics*, J. Diff. Geom. **29** (1989), 65–83.

[60] Salamon, D. and J. Weber, *Floer homology and the heat flow*, Geometric and Functional Analysis (GAFA) **16** (2005), 1050–1138.

[61] Salamon, D. and Zehnder,E., *Morse theory for periodic solutions of Hamiltonian systems and the Maslov index*, Comm. Pure Appl. Math. **45** (1992), 1303–1360.

[62] Sullivan, D., *Open and closed string field theory interpreted in classical algebraic topology*, Topology, Geometry and Quantum Field Theory, Procdeedings of the 2002 Oxford Symposium, London Mathematical Society Lecture Note Series **308** (2004), Cambridge University Press, Cambridge, 344–357.

[63] Vigué-Poirrier, M., Sullivan, D., *The homology theory of the closed geodesic problem*, J. Differential Geometry **11** (1976), no. 4, 633–644.

[64] Viterbo, C., *Functors and computations in Floer homology with applications II*, preprint, 1996, revised 2003.

[65] ———, *Functors and computations in Floer homology with applications. I*, Geom. Funct. Anal. **9** (1999), 985–1033.

[66] Vigué-Poirrier, M., and D. Sullivan, *The homology theory of the closed geodesic problem*, J. Diff. Geom. **11** (1976), 633–644.

[67] Wang, W., *Closed geodesics on Finsler spheres*, Calculus of Variations and Partial Differential Equations Volume 45, Issue 1-2(2012), 253–272.

[68] Ziller, W., *Geometry of the Katok examples*, Ergod.Th. Dyn.Syst. **3** (1982), 135–157.

[69] ———, *The free loop space of globally symmetric spaces*, Invent. Math. **41** (1977), 1–22.

Department of Mathematics and Statistics,The College of New Jersey,Ewing, NJ 08628-0718, USA
E-mail: hingston@tcnj.edu

The surface subgroup and the Ehrenepreis conjectures

Jeremy Kahn and Vladimir Markovic

Abstract. We survey our recent results including the surface subgroup theorem and the Ehrenpreis conjecture. applications and future direction are discussed.

Mathematics Subject Classification (2010). 20H10, 57M50.

Keywords. Surface subgroup theorem, Ehrenepreis.

1. Introduction

1.1. The surface subgroup theorem. One of the corollaries of the Geometrization Theorem is that most 3-manifolds admit hyperbolic structure. Therefore when studying topology of a 3-manifold one can often assume that the manifold is endowed with a hyperbolic structure. This greatly expands the tool-kit that is available bringing hyperbolic geometry, analysis and dynamics into play.

An essential step in the eventual proof of the Virtual Haken and the Virtual Fibering Conjectures is the Surface Subgroup Theorem:

Theorem 1.1 (Kahn-Markovic). *Every closed hyperbolic 3-manifold contains a quasifuchsian surface subgroup.*

Recall that every hyperbolic manifold $\mathbf{M}^3$ can be represented as the quotient $\mathbf{M}^3 = \mathbb{H}^3/\mathcal{G}$, where $\mathbb{H}^3$ is the hyperbolic 3-ball and $\mathcal{G}$ a Kleinian group. Using geometry and relying on fine statistical properties of the frame flow on hyperbolic manifolds we proved in [14] the following result which implies the Surface Subgroup Theorem:

Theorem 1.2 (Kahn-Markovic). *Let $\mathbf{M}^3 = \mathbb{H}^3/\mathcal{G}$ denote a closed hyperbolic 3-manifold. Given any $\epsilon > 0$, there exists a $(1+\epsilon)$-quasifuchsian group $G < \mathcal{G}$.*

(Recall that a group is K-quasifuchsian if it is K-quasiconformal deformation of a Fuchsian group.)

The nearly geodesic surfaces we constructed in [14] have large genus (it can be shown that the genus of S grows polynomially with $\frac{1}{\epsilon}$). Moreover, each such surface $f(S) \subset \mathbf{M}^3$ represents the trivial homology class in $H_2(\mathbf{M}^3, \mathbb{Z})$.

A three holed sphere with a hyperbolic metric and geodesic boundary is called a pair of pants (after Thurston). Given $R > 0$, the pair of pants whose all 3 cuffs have the same length

Proceedings of the International Congress of Mathematicians, Seoul, 2014

$2R$ is called the R-perfect pair of pants. Let $\mathbf{S}(R)$ denote a genus 2 Riemann surface that is obtained by gluing two R-perfect pairs of pants along their cuffs with the twist of $+1$. The induced orbifold is denoted by $\mathrm{Orb}(R)$.

Theorem 1.2 is proved by the showing that for a given closed hyperbolic 3-manifold $\mathbf{M}^3$ and for every $\epsilon > 0$ and every large enough R, there exists a Riemann surface $S = S(\epsilon, R)$ and a continuous map $f : S \to \mathbf{M}^3$ such that the induced map between universal covers $\partial f : \partial\mathbb{H}^2 \to \partial\mathbb{H}^3$ is $(1+\epsilon)$-quasisymmetric. Moreover, the surface S admits a decomposition into R-perfect pants that are glued to each other with the twist of $+1$. In particular, such a Riemann surface S is a regular holomorphic cover of the Model Orbifold $\mathrm{Orb}(R)$.

For fixed $\epsilon, R > 0$, a good pair of pants (in a given hyperbolic 3-manifold) is a pair of pants whose cuffs have complex half-length ϵ close to R (see Section 3). In order to find the map $f : S \to \mathbf{M}^3$ one is guided by the following principles:

1. Do not start by trying to specify the surface S.
2. Instead, consider the good pants $\mathbf{\Pi}$ immersed in $\mathbf{M}^3$ as the building blocks and eventually construct the surface $f(S) \subset \mathbf{M}^3$ by appropriately assembling together all the good pants from $\mathbf{\Pi}$.

Typically there are many ways in which one can assemble the pants and get an immersed surface.

Consider any finite formal sum $W \in \mathbb{N}\mathbf{\Pi}$. Taking two copies of each pair of pants (with opposite orientations) one obtains the new formal sum $2W \in \mathbb{N}\mathbf{\Pi}$. Then along every geodesic in $\mathbf{M}^3$ that appears as a boundary curve of some of the pants from $2W$ one can pair off the pairs of pants that contain that geodesic as a boundary component (there may be many ways in which one can pair off these pants and we choose one way of doing it for each such geodesic). One can now assemble these pairs of pants according to the instructions to construct a closed surface in $\mathbf{M}^3$.

So, we have constructed a closed surface S and a map $f : S \to \mathbf{M}^3$, but the induced map between fundamental groups is not necessarily injective. For example, if W denotes a single pair of good pants then the surface S is a genus two surface obtained by gluing together two pairs of pants. However, the corresponding map $f : S \to \mathbf{M}^3$ collapses one pair of pants onto the other and therefore the induced map between fundamental groups is not injective.

Observe that every such surface $f(S) \subset \mathbf{M}^3$ represents the trivial homology class in $H_2(\mathbf{M}^3, \mathbb{Z})$. This is because each pair of pants from W is used twice and with different orientations.

It is clear from the previous discussion that if one wants to glue pairs of good pants in $\mathbf{M}^3$ to get a nearly geodesic surface (and thus a quasifuchsian surface) then any two pairs of pants that are glued along a geodesic should meet at an angle that is close to π. It turns out that in order to assemble good pants and construct a nearly geodesic surface in $\mathbf{M}^3$, what is needed is that the good pants are equidistributed in $\mathbf{M}^3$, which follows from the exponential mixing of the frame flow. We will explain this in more details in the next section, but here we state the mixing principle:

Lemma 1.3 (Exponential Mixing). *Let $\mathbf{M}^3$ denote a closed hyperbolic manifold (in particular a hyperbolic surface). There exists $\mathbf{q} > 0$ that depends only on $\mathbf{M}^3$ such that the following holds. Let $\psi, \phi : \mathcal{F}(\mathbf{M}^3) \to \mathbb{R}$ be two C^1 functions (here $\mathcal{F}(\mathbf{M}^3)$ denotes the frame bundle, if $\mathbf{M}^3$ is a Riemann surface this is just the tangent bundle). Then assuming*

that the volume of the frame bundle $\mathcal{F}(\mathbf{M}^3)$ is equal to 1, for every $r \in \mathbb{R}$ the inequality

$$\left| \int_{\mathcal{F}(\mathbf{M}^3)} (\mathbf{g}_r^* \psi)(x)\phi(x)\, d\Lambda(x) - \int_{\mathcal{F}(\mathbf{M}^3)} \psi(x)\, d\Lambda(x) \int_{\mathcal{F}(\mathbf{M}^3)} \phi(x)\, d\Lambda(x) \right| \leq Ce^{-\mathbf{q}|r|},$$

holds, where $C > 0$ only depends on the C^∞ norm of ψ and ϕ.

2. The Ehrenpreis conjecture

The Ehrenpreis conjecture was an old conjecture in the theory of Riemann surfaces. The idea is that although two Riemann surfaces S and T do not have a common finite cover (all covers in this proposal are regular and unbranched) one should still be able to interpolate between certain finite covers of S and T respectively (according to Gromov this statement goes back to Riemann). The precise formulation of the conjecture is as follows:

Conjecture 2.1 (Ehrenpreis Conjecture). *Let S and T denote two closed Riemann surfaces of genus at least 2 and let $\epsilon > 0$. Then there exists finite covers S_1 and T_1 of S and T respectively, such that S_1 and T_1 are $(1+\epsilon)$ quasiconformal to each other (that is there exists $(1+\epsilon)$-quasiconformal map $f : S_1 \to T_1$).*

In [16] J. Kahn and I have announced a proof of this conjecture.

Remark 2.2. The Enhrenpreis Conjecture is harder to prove because there may be more pants on one side of a closed geodesic than the other. So we need to add in a signed sum of pants so that there are an equal number on both sides of every good geodesic. Computing this correction requires the "good pants homology", which we develop in [16].

In fact, we prove this conjecture by proving the statement that every closed hyperbolic Riemann surface has a virtual decomposition into good pairs of pants that are glued by a twist that is nearly equal to $+1$. Recall that (ϵ, R)-good pair of pants is a pair of pants whose cuffs have the length ϵ-close to R. We prove the following virtual decomposition type theorem:

Theorem 2.3 (Kahn-Markovic). *Let S be a hyperbolic surface and let $\epsilon > 0$. Then for every large enough $R > 0$, the surface S has a finite cover S_1 that can be decomposed into (ϵ, R)-good pants such that every two adjacent pairs of good pants are glued with the twist that is $\frac{\epsilon}{R}$ close to $+1$.*

We then show that this surface S_1 is quasiconformally close to a finite cover of the model orbifold $\mathrm{Orb}(R)$ (see above for the definition of the Model orbifold $\mathrm{Orb}(R)$):

Theorem 2.4 (Kahn-Markovic). *Let S be a closed hyperbolic Riemann surface. Then for every $K > 1$, and every large enough $R > 0$ there are finite covers S_1 and O_1 of the surface S and the model orbifold $\mathrm{Orb}(R)$ respectively, and a K-quasiconformal map $f : S_1 \to O_1$.*

The Ehrenpreis conjecture is an immediate corollary of this theorem.

Theorem 2.5 (Kahn-Markovic). *Let S and M denote two closed Riemann surfaces. For any $K > 1$, one can find finite degree covers S_1 and M_1 of S and M respectively, such that there exists a K-quasiconformal map $f : S_1 \to M_1$.*

The proof of the above Virtual Decomposition theorem follows from the equidistribution of the good pants (in much the same way as in the proof of the Surface Subgroup Theorem) and from the Correction theory that we will outline below.

3. The setup and main ideas

3.1. The feet of a pair of pants. A pair of pants is a compact hyperbolic Riemann surface with geodesic boundary that is homeomorphic to the sphere minus three disjoint round open disks. Any such pair of pants is determined by the lengths of the three boundary components, which are called cuffs. For reasons which will become clear in the next section, we will prefer to work with the half-lengths, which of course are half the lengths of the three cuffs. In particular, an R-perfect pair of pants is a pair of pants whose three half-lengths are equal to R (for a given $R > 0$).

An orthogeodesic for a compact hyperbolic surface S with geodesic boundary is a proper geodesic arc which is orthogonal to the boundary of S at both endpoints. The long orthogeodesics for a pair of pants are the three embedded orthogeodesics which divide S into two components—these are the embedded orthogeodesics from a cuff to itself. The short orthogeodesics are the three other embedded orthogeodesics (from one cuff to another); the three short orthogeodesics together divide S into two right-angled hexagons. Because a right-angled hexagon is determined by the lengths of three alternating sides, these two right-angled hexagons must be isometric. It follows that the six endpoints of the three short orthogeodesics divide the three cuffs into six segments such that each cuff is divided into two equal segments. At each endpoint of an orthogeodesic η, there is a unique normal vector to the boundary that generates η (via the geodesic flow); we call this normal vector the foot of η at that endpoint. We say that the feet of a pair of pants at a given cuff are the feet of the two short orthogeodesics from that cuff to the two other cuffs. Thus there are two feet of a pair of pants in the normal bundle of each cuff of the pants.

3.2. Good and perfect panted surfaces. Now suppose that we are given a closed hyperbolic Riemann surface S of genus $g > 1$, and a maximal collection $\mathcal{C}$ of disjoint curves on S. (By a *curve* we mean an (smooth) isotopy class of smoothly embedded closed curves). Each of these curves can be uniquely realized as a closed geodesic on S; together they divide S into $2g - 2$ pairs of pants. For each closed geodesic γ, there are two of these pairs of pants with γ as boundary (or γ appears as two boundaries of the same pair of pants). We can then find two pairs of feet, and holding up the (universal cover of the) cuff vertically, we see that the two feet on the right are a certain distance above the two feet on the left—except that this distance is only defined up to the half-length of the cuff. Therefore, to each cuff C, we have two invariants: the positive real half-length of the geodesic representative γ of C, and the shear, which is defined up to the half-length of γ. There is a natural topology on panted surfaces (of a given genus), for which these $6g - 6$ invariants provide local coordinates.

An R-perfect panted surface is one for which all of the cuffs of the pants have half-length R, and all of the shears are equal to 1. An R, ϵ-good pair of pants is one for which all three cuffs have half-length within ϵ of R, and an R, ϵ-good panted surface is one made of out of good pants, for which all of the shears are within ϵ/R of 1. (Sometimes we will write perfect for R-perfect, and good for R, ϵ-good). For any good panted surface $S, \mathcal{C}$, there is a path through good panted surfaces to a perfect panted panted surface $S', \mathcal{C}'$ and a homeomorphism

$h\colon S \to S'$ determined (up to isotopy) by that path. (The path is determined up to homotopy rel endpoints, and hence the homeomorphism is determined up to isotopy). We say that $S', \mathcal{C}'$ is the *perfect version* of $S, \mathcal{C}$, and that h is the *perfecting homeomorphism*.

We prove the following theorem which provides a criterion for two large genus surfaces to be close to each other in the corresponding Moduli space with respect to the Teichmüller metric.

Theorem 3.1. *There exists R_0, K_0, and $\epsilon_0 > 0$ such that the following holds. Suppose that $S, \mathcal{C}$ is an R, ϵ good panted surface, and $R > R_0$, $\epsilon < \epsilon_0$. Let $S', \mathcal{C}'$ be the perfect version of $S, \mathcal{C}$. Then there is a $K_0\epsilon$-quasiconformal diffeomorphism $h\colon S \to S'$ that is homotopic to the perfecting homeomorphism.*

The proof of this theorem is very delicate and we omit it here (the reader can see Section 2 in [14]). It should be stressed that the requirement that pants are glued with the twist by $+1$ plays a vital and subtle role and the criterion would not hold without it.

Theorem 2.5 follows if we can prove the following:

Theorem 3.2. *For every closed hyperbolic Riemann surface S we can find a finite cover $\hat{S}$ and a maximal set $\mathcal{C}$ of disjoint curves on $\hat{S}$ such that $\hat{S}, \mathcal{C}$ is a good panted surface.*

or, more precisely, if we can prove the following:

Theorem 3.3. *For every closed hyperbolic Riemann surface S, $\epsilon > 0$, and $R > R_0(S, \epsilon)$, we can find a finite cover $\hat{S}$ and a maximal set $\mathcal{C}$ of disjoint curves on $\hat{S}$ such that $\hat{S}, \mathcal{C}$ is an R, ϵ good panted surface.*

Let us briefly explain this implication. We glue two R-perfect pairs of pants together with a shear of 1 at each cuff to obtain an R-perfect surface S_R with an orientation-preserving isometry group of size 12. The model orbifold O_R is the quotient of S_R by this group of isometries; the three cuffs of half-length R on S_R map to a single segment η_R of length $R/2$ on O_R connecting two of the the order 2 points on O_R. *Any R-perfect panted surface S is a finite cover of O_R* in such a way that the R-cuffs of S are the components of the pre-image of η_R by the cover. It follows that any two R-perfect panted surfaces have a common finite cover. Then given two surfaces S and T, and $\epsilon > 0$, we find $R, \epsilon/2K_0$ good panted covers $\hat{S}$ and $\hat{T}$ (for any R sufficiently large). By Theorem 3.1, these are each $\epsilon/2$ close to perfect surfaces, which then have a common cover. Therefore $\hat{S}$ and $\hat{T}$ have common covers within ϵ of each other in the Teichmüller metric, and we are finished.

Recall that the Teichmüller metric on the moduli space of compact Riemann surfaces of genus g is defined so that the distance between S and S' is $\log K$, where K is the infimum of K for which there exists a K-quasiconformal diffeomorphism $h\colon S \to S'$. We will often write $1 + \epsilon$-quasiconfomal when we should really be writing e^{ϵ}-quasiconformal, and so forth—the reader can make the necessary modifications.

3.3. Building a good cover. We can now begin to describe how we prove Theorem 3.3. Recall that S is our given closed hyperbolic Riemann surface. A good curve for S (really an R, ϵ good curve) will be a closed geodesic γ (or the associated free homotopy class) whose half-length $\mathbf{hl}(\gamma)$ is within ϵ of R. We will denote the set of R, ϵ good curves by $\Gamma_{\epsilon,R}$; it is a finite set, with size asymptotic to $4\epsilon e^{2R}/2R$ when R is large.

Now let Π be a topological pair of pants, and let $f\colon \Pi \to S$ be a π_1 injective immersion. Then there is a unique hyperbolic metric on Π (up to pullback by a diffeomorphism isotopic

to the identity) such that Π becomes a geometric pair of pants (with geodesic boundary) and f is isotopic to an isometric immersion. If Π is then a good pair of pants (for some R, ϵ), then we say that f represents an immersed good pair of pants in S. For any R and ϵ, there is a finite set $\Pi_{\epsilon,R} \equiv \Pi_{\epsilon,R}(S)$ of good pairs of pants in S. Using the exponential mixing of the geodesic flow on S, and the consequent estimates on the number of long orthogeodesic segments connecting a pair of geodesic segments on S, we prove that the feet of good pants are evenly distributed in the normal bundle of every good geodesic:

Theorem 3.4. *Suppose that $\gamma \in \Gamma_{\epsilon,R}$, and let I be an interval in the (square root of the) normal bundle for γ. The number $n(\gamma, I)$ of feet of pants in $\Pi_{\epsilon,R}$ that lie in I is estimated by*

$$n(\gamma, I) = \frac{n(\gamma)|I|}{2\mathbf{l}(\gamma)} + O(e^{(1-\alpha)R}),$$

where $n(\gamma)$ is the total number of pairs of feet on both sides on γ, and $\alpha \equiv \alpha(S)$. Moreover,

$$n(\gamma) \sim 2\epsilon^2 R e^R / \operatorname{Area}(S),$$

for R large given S and ϵ.

What is important in this statement is that the error term for $n(\gamma, I)$ is exponentially small (in R) compared to $n(\gamma)$. Up to this error term, the feet of the pants with γ as a boundary are evenly distributed on the normal bundle of γ. Let us suppose, by some miracle, that the distribution of feet is also *balanced*: that there are exactly as many feet on one side of γ as the other. (By the two sides of γ we mean the two components of the unit normal bundle for γ). Then it is a simple and elementary exercise to show that there is a bijection $\sigma\colon \Pi_\gamma^+ \to \Pi_\gamma^-$ (where Π_γ^+ and Π_γ^- are the pants with feet on the left and right sides of the unit normal bundle of the oriented geodesic γ) such that for any pair of pants $\pi \in \Pi_\gamma^+$, the feet of $\sigma(\pi)$ on γ are $1 + O(e^{-\alpha R})$ above the feet of π on γ. We then use this bijection to glue the pants in Π_γ^- to the pants of Π_γ^+ (along the cuffs that map to γ), and doing this with every $\gamma \in \Gamma_{\epsilon,R}$, we obtain a closed surface, made of the pants in $\Pi_{\epsilon,R}$, that is a finite cover of S. Because the shears are exponentially close to 1, and an exponentially small number is less than ϵ/R when R is large, we have obtained a good panted cover of S, and have thereby proven Theorem 3.3.

Of course, we have no reason to believe that there are exactly the same number of pants on the two sides of γ. We will describe in a few paragraphs how to correct this imbalance, but first we will describe the analogous construction in a closed hyperbolic three-manifold M, and we will see that in three dimensions, the work is a bit easier, because there is no imbalance to correct.

3.4. Working in three dimensions. Suppose that $f\colon \Pi \to M$ is a π_1-injective map from a topological pair of pants Π to M. We are interested in describing f up to homotopy. We can assume that f maps the boundaries of Π to closed geodesics $\gamma_0, \gamma_1, \gamma_2$ in M. We can also assume that f maps three disjoint arcs in Π (connecting the three boundary components) into three geodesic segments η_0, η_1, η_2 such that η_0 connects γ_1 and γ_2 and meets both geodesics orthogonally (and similarly for η_1 and η_2). These three arcs will divide Π into two (filled) hexagons, and f will map the boundary of each of these hexagons into skew right-angled hexagons.

Skew right-angled hexagons in $\mathbb{H}^3$ are very much like right-angled hexagons in $\mathbb{H}^2$, with $\mathbb{R}$ replaced by $\mathbb{C}$. That is, a skew right-angled hexagon is determined by the complex length

of three alternating sides. The real part of the complex length is the real length, and the imaginary part, which is defined up to multiples of $2\pi i$, is the amount of rotation from one adjacent side to the other adjacent side, after one adjacent side has been translated along the given side to meet the other adjacent side. Because the complex lengths of the η_i are the same in both skew right-angled hexagons, the two hexagons are isometric, and hence each γ_i is divided into two segments by the endpoints of $\eta_{i\pm1}$, and these two segments have equal complex length (with respect to the $\eta_{i\pm1}$). We call this complex length the complex half-length $\mathbf{hl}(\gamma)$ of γ. The *feet* or initial vectors of the orthogeodesics $\eta_{i\pm1}$ are elements of the unit normal bundle $N^1(\gamma)$, which is a torsor for $\mathbb{C}/(2\pi i\mathbb{Z} + 2\mathbf{hl}(\gamma)\mathbb{Z})$, and the difference of between the two feet is exactly $\mathbf{hl}(\gamma)$. Thus we can think of the unordered pair of feet as living in $N^1(\sqrt{\gamma})$, the set of unordered pairs that differ by $\mathbf{hl}(\gamma)$; it is a torsor for $\mathbb{C}/(2\pi\mathbb{Z} + \mathbf{hl}(\gamma)\mathbb{Z})$.

We let $\Gamma_{\epsilon,R}$ be the good closed geodesics in M (so $\gamma \in \Gamma_{\epsilon,R}$ if the complex length $\mathbf{l}(\gamma)$ satisfies $|\mathbf{l}(\gamma) - 2R| < 2\epsilon$), and we let $\mathbf{\Pi}_{\epsilon,R}$ be the good pants in M (so $f\colon \Pi \to M$ is in $\mathbf{\Pi}_{\epsilon,R}$ if for each $\gamma \in f(\partial\Pi)$ we have $|\mathbf{hl}(\gamma) - R| < \epsilon$).

We can then prove the analogue of Theorem 3.4 for distribution of the feet of good pants:

Theorem 3.5. *Suppose that $\gamma \in \Gamma_{\epsilon,R}$, and let $I \times J$ be a rectangle in the (square root of the) normal bundle for γ. The number $n(\gamma, I)$ of feet of pants in $\mathbf{\Pi}_{\epsilon,R}$ that lie in I is estimated by*

$$n(\gamma, I) = \frac{n(\gamma)|I \times J|}{2\pi\Re(\mathbf{hl}(\gamma))} + O(e^{(1-\alpha)R}),$$

where $n(\gamma)$ is the total number of pairs of feet in $N^1(\sqrt{\gamma})$, and $\alpha \equiv \alpha(M)$. Moreover,

$$n(\gamma) \sim 8\epsilon^4 R e^{2R}/\mathrm{Vol}(M)$$

for R large given M and ϵ.

It then follows that if $A_\gamma \subset N^1(\sqrt{\gamma})$ is the set of pairs of feet of good pants on γ, then we can find a permutation $\sigma\colon A_\gamma \to A_\gamma$ such that

$$|\sigma(x) - x - \pi i - 1| < \epsilon/R$$

for every $x \in A$.

Then we can assemble the pants of $\mathbf{\Pi}_{\epsilon,R}$ into a "good panted surface group representation" using the "doubling trick". We take two copies of every pair of pants in $\mathbf{\Pi}_{\epsilon,R}$, and give them the two possible orientations. Then for any $\gamma \in \Gamma_{\epsilon,R}$, we have two sets, Π^+_γ and Π^-_γ, of oriented pants with γ as boundary, where each pair of pants in Π^+_γ induces a "positive" orientation on γ (arbitrarily chosen), and the opposite holds for Π^-_γ. We then find $\hat{\sigma}\colon \Pi^+_\gamma \to \Pi^-_\gamma$ such that the pair of feet of $\hat{\sigma}(\Pi)$ on γ is σ applied to the pair of feet of Π on γ. In this way we pair off all of the boundary components of the two copies of the good pants.

We then obtain an immersed panted surface $f\colon S \to M$ (with a maximal set $\mathcal{C}$ of curves on S). It is an ϵ, R good panted surface group representation in the following sense: the restriction of f to every component of $S - \bigcup\mathcal{C}$ is a pair of good pants, and for every $C \in \mathcal{C}$, the complex shear coordinates—the difference (in $N^1(\sqrt{\gamma})$) between the feet of the pants on one side of C and the other—is within ϵ of $i\pi + 1$.

It follows that f is essential by the following theorem (closely analogous to Theorem 3.1), which gives a way to certify the injectivity of the induced homomorphism $\rho\colon \pi_1(S) \to \mathrm{Isom}(\mathbb{H}^3)$.

Theorem 3.6. *There exists R_0, K_0, and $\epsilon_0 > 0$ such that the following holds. Suppose that $\rho\colon \pi_1(S) \to \mathrm{Isom}(\mathbb{H}^3)$ is an R, ϵ good panted surface group representation, and $R > R_0$, $\epsilon < \epsilon_0$. Then we can find an R-perfect panted Fuchsian group (which we then think of as acting on $\mathbb{H}^3$), and an equivariant map $h\colon \mathbb{H}^3 \to \mathbb{H}^3$ that extends to be $K_0\epsilon$-quasiconformal on the boundary. In particular, ρ is a faithful, discrete, and quasifuchsian representation.*

3.5. The good pants homology and the Ehrenpreis conjecture. We now return the the problem of proving Theorem 3.3, which implies the Ehrenpreis conjecture. We will let $\Gamma_{\epsilon,R}$ denote the set of *oriented* geodesics, and we will let $\mathbb{Z}\Gamma_{\epsilon,R}$ denote the set of integral formal sums of elements of $\Gamma_{\epsilon,R}$, where we will think of opposite orientations of the same geodesic as summing to zero. Let $\partial\colon \Pi_{\epsilon,R} \to \mathbb{Z}\Gamma_{\epsilon,R}$ be the obvious boundary map. We prove that when R is large given S and ϵ, there is a map $q\colon \mathbb{Q}\Gamma_{\epsilon,R} \to \mathbb{Q}\Pi_{300\epsilon,R}$ such that, for any $\alpha \in \mathbb{Q}\Pi_{\epsilon,R}$,

1. $\partial q(\partial\alpha) = \partial\alpha$, and
2. $||q(\alpha)||_\infty \leq e^{-R}P(R)||\alpha||_\infty$ for any weighted sum α of good curves.

(Where $P(R)$ is a polynomial in R that depends only on S and ϵ). Letting $\alpha \equiv \Sigma\Pi_{\epsilon,R}$ be the formal sum of the good pants, we replace α with $\alpha' = \alpha - q(\partial\alpha)$ to obtain a "balanced" sum of pants ($\partial\alpha' = 0$) with the same equidistribution properties[1] as in Theorem 3.4 (because $q(\partial\alpha)$ is small compared to α). We can then pair these pants across every good geodesic to obtain an immersed (or covering) panted surface which, by Theorem 3.1, is $1+\epsilon$ quasiconformally equivalent to the corresponding perfect surface, thus proving the Ehrenpreis conjecture.

We will briefly outline the construction of the map q and the demonstration of the estimate (2). We define the "good pants homology" as $\mathbb{Q}\Gamma_{\epsilon,R}/\partial\mathbb{Q}\mathbf{\Pi}_{\epsilon,R}$; if two sums of good curves differ by an element of $\partial\mathbb{Q}\mathbf{\Pi}_{\epsilon,R}$, we will say that they are $\mathbf{\Pi}_{\epsilon,R}$ homologous. We prove that, if A_i, B_j, U, V $(i,j = 0,1)$ are elements of $\pi_1(S,*)$ such that the broken closed geodesic $[\cdot A_i \cdot U \cdot B_j \cdot V \cdot]$ has "bounded inefficiency" and the geodesic segments $\cdot U \cdot$ and $\cdot V \cdot$ are sufficiently long, then

$$\sum_{i,j=0,1} (-1)^{i+j}[A_iUB_jV] \equiv 0$$

in $\mathbf{\Pi}_{\epsilon,R}$ (really $\Pi_{300\epsilon,R}$) homology, provided, of course, that the $[A_iUB_jV]$ are the free homotopy classes (or, if you like, conjugacy classes in $\pi_1(S,*)$) of good curves.

This then permits us to define, for $A, T \in \pi_1(S,*)$,

$$A_T \equiv \frac{1}{2}\left([TAT^{-1}U] - [TA^{-1}T^{-1}U]\right),$$

where U is fairly arbitrary. Then A_T in good pants homology is independent of the choice of U. We can show through a series of lemmas (see [16]) that $(XY)_T \equiv X_T + Y_T$ in good pants homology; this then implies that any element of $\mathbb{Q}\mathbf{\Pi}_{\epsilon,R}$ that is zero in $H_1(S)$ is zero in $\mathbf{\Pi}_{\epsilon,R}$ homology.

We have not yet said anything about the function q. The idea is that whenever we prove that two formal sums of curves are equal in good pants homology, we produce a sum of good

[1]We should observe as well that α' is positive!

pants (the "witness" to the homology) whose boundary is equal to the difference of the two formal sums. When we make an arbitrary choice in determining the sum of good pants, we take the average of the results of our choices as our witness. When one identity in good pants homology is proved using another one, the witness for the latter is used to build the witness for the former. In this way, when we prove that

$$(XY)_T \equiv X_T + Y_T$$

in good pants homology, we can explicitly produce a function $g\colon \pi_1(S) \times \pi_1(S) \to \mathbb{Q}\mathbf{\Pi}_{\epsilon,R}$ such that $(XY)_T - X_T - Y_T = \partial g(X, Y)$.

We then let $g_1, \dots g_{2n}$ be a standard set of generators for $\pi_1(S, *)$; then $[g_1], \dots, [g_{2n}]$ also form a basis for $H_1(S)$, and so does $\mathfrak{g} = \{(g_1)_T, \dots, (g_{2n})_T\}$, because $X \equiv X_T$ in H_1. For any $\gamma \in \Gamma_{\epsilon,R}$, we can find a unique $\hat{\gamma} \in \mathbb{Z}\mathfrak{g}$ that is equal to γ in $H_1(S)$. Then in the course of proving that $\gamma \equiv \hat{\gamma}$ in $\mathbf{\Pi}_{\epsilon,R}$ homology, we produce $q\colon \mathbb{Q}\Gamma_{\epsilon,R} \to \mathbb{Q}\mathbf{\Pi}_{\epsilon,R}$ such that

$$\partial q(\gamma) = \gamma - \hat{\gamma}.$$

The identity (1) then follows for q, because $\widehat{\partial\alpha} = 0$ (because $\partial\alpha \equiv 0$ in H_1) for any $\alpha \in \mathbb{Q}\mathbf{\Pi}_{\epsilon,R}$ (where we have extended $\gamma \mapsto \hat{\gamma}$ to $\mathbb{Q}\mathbf{\Pi}_{\epsilon,R}$).

It remains only to show the estimate (2) for this q. Again, for each identity that we prove in [16], and each resulting implicit definition of a witness, we produce a corresponding estimate for the "witness function", using the previous estimates. Is this manner we produce the desired inequality.

4. Applications

4.1. Virtual classification of 3-Manifolds. A subsurface $S \subset \mathbf{M}^3$ (here S is a compact surface, possibly with boundary) is **essential** if the induced map between fundamental groups is an injection. The surface is **incompressible** in $\mathbf{M}^3$ if it is embedded in $\mathbf{M}^3$ and if every homotopically non-trivial simple loop on S is mapped onto a homotopically non-trivial closed curve in $\mathbf{M}^3$. Every essential embedded surface is incompressible, and the converse is a well-known theorem.

Machinery has been developed to study hyperbolic 3-manifolds that are Haken. A manifold is Haken if it contains an incompressible surface. If $\mathbf{M}^3$ is Haken, one can cut $\mathbf{M}^3$ along its incompressible surface to obtain a 3-manifold with boundary (which may be disconnected). Furthermore, hyperbolic 3-manifolds with boundary are known to be Haken so one can continue to cut until arriving at indecomposable pieces. This is known as the Haken hierarchy and it is a cornerstone of 3-dimensional topology. Although many 3-manifolds are not Haken it was conjectured by Thurston that every such manifold has a finite degree cover that is. This was known as the Virtual Haken Conjecture.

Thurston made an even stronger conjecture called the Virtual Fibering Conjecture. Let $\mathbb{S}_g$ denote a closed surface of genus $g \geq 2$. Given a homeomorphism $f : \mathbb{S}_g \to \mathbb{S}_g$, let $\mathbf{M}^3_f$ denote the corresponding mapping torus. Then $\mathbf{M}^3_f$ is a closed 3-manifold that fibers over the circle. Thurston proved that $\mathbf{M}^3_f$ is hyperbolic if and only if f is homotopic to a pseudo-Anosov homeomorphism of $\mathbb{S}_g$. The Virtual Fibering Conjecture Thurston stated that every hyperbolic 3-manifold has a finite degree cover that fibers over the circle.

These two conjectures were major driving forces behind the research in three dimensional topology in recent decades. Building on the Surface Subgroup Theorem of Kahn-Markovic and the work by Wise [23] and Haglund-Wise [12], Agol [2] completed the proofs of both conjectures. Below we state the main steps in the proof.

A group is cubulated if it is acting properly and co-compactly on a CAT(0) cube complex. It turns out that each cubulated hyperbolic group has a rich (hidden) underlying stucture. Wise developed this theory [23], although in his work he used an additional assumption that cubulated hyperbolic groups have a certain Haken hierarchy. Under this assumption he, and in collaboration with Haglund, Hruska, Hsu and others, showed that such groups can be embedded in Right Angled Artin groups which implies that such cubulated hyperbolic groups have many deep properties like being linear, LERF (that is, finitely generated subroups are separable), etc. In particular, if the fundamental group of a hyperbolic 3-manifold satisfies these assumption, it follows from Wise's theory that this 3-manifold is Virtually Haken. Moreover, using the Agol's criterion for virtual fibering [1], it also follows that such a 3-manifold virtually fibers over the circle.

In order for Wise's theory to be applied to hyperbolic 3-manifolds it has to be shown that 3-manifold groups are cubulated and that Wise's assumption on the Haken hierarchy can be dropped.

In the course of proving the Surface Subgroup Theorem we proved that given any two points on the 2-sphere there is a surface subgroup whose limit set separates these two points. Combining this fact and the Sageev construction [22], Bergeron-Wise showed:

Theorem 4.1 (Bergeron-Wise). *The fundamental group of a closed hyperbolic 3-manifold is cubulated.*

Finally, Agol [1] proved Wise's conjecture that cubulated hyperbolic groups are virtually special (which in particular means that Wise's assumption on the Haken hierarchy is not needed for his theory to work), and thus he was able to prove the Virtual Haken Theorem and the Virtual Fibering Theorem:

Theorem 4.2 (Agol). *Every closed hyperbolic 3-manifold has a finite cover that fibers over the circle. In particular, every hyperbolic 3-manifold has a finite cover that is Haken.*

The reader may want to consult the comprehensive survey article [3] by Aschenbrenner-Friedl-Wilton for a complete overview of these theories.

4.2. Counting Problems for Essential Surfaces and Moduli spaces. Counting closed geo-desics in negatively curved manifold is an old and profound subject. Standard results (that are essentially corollaries of the mixing properties of geodesic flows on negatively curved manifolds) state that the number of closed geodesics of length at most L grows exponentially with L. For hyperbolic manifolds (and in particular for Riemann surfaces) this asymptotic is precisely known (Margulis [19]) with excellent bounds on error terms.

Analogously, in a given hyperbolic 3-manifold $\mathbf{M}^3$ one can count the number of essential surfaces (up to homotopy) live inside $\mathbf{M}^3$. Let $s(\mathbf{M}^3, g)$ denote the number (up to homotopy) of genus g incompressible surfaces of $\mathbf{M}^3$. The following counting result was proved in [15]:

Theorem 4.3 (Kahn-Markovic). *Let $\mathbf{M}^3$ be a closed hyperbolic 3-manifold. There exist constants $0 < c_1 \leq c_2$, such that the inequality $(c_1 g)^{2g} \leq s(\mathbf{M}^3, g) \leq (c_2 g)^{2g}$, holds for every large g.*

- A difficult (and perhaps deep) conjecture is to prove that for some constant $c = c(\mathbf{M}^3) > 0$ the formula
$$\lim_{g\to\infty} \frac{\sqrt[2g]{s(\mathbf{M}^3, g)}}{g} = c$$
holds. A positive answer to this conjecture would represent a kind of the Prime Number Theorem for counting essential surfaces in 3-manifolds analogous to the Margulis' Prime Number Theorem for counting closed geodesics [19].
- Another important question is: What does a random essential surface of genus g (for some large g) inside $\mathbf{M}^3$ look like? Is this a quasifuchsian surface or is it a geometrically infinite surface (according to Thurston, Bonahon and Canary a geometrically infinite closed surface in $\mathbf{M}^3$ is a virtual fiber)?

4.3. Homology Of curves And surfaces in hyperbolic 3-Manifolds. In manifolds of negative curvature each homotopy class of closed curves can be realized by a unique geodesic. Given that closed curves have such nice geometric representatives, Thurston recently asked if one can represent each homology class in $H_2(\mathbf{M}^3, \mathbb{Z})$ by a nearly geodesic representative. The following theorem is proved using the methods from [14].

Theorem 4.4 (Liu-Markovic). *Every rational second homology class of a closed hyperbolic 3-manifold has a positive integral multiple represented by an oriented connected closed quasi-Fuchsian subsurface.*

It is well known that every homology class in $H_2(\mathbf{M}^3, \mathbb{Z})$ can be represented as a sum (with integer coefficients) of connected incompressible surfaces in $\mathbf{M}^3$. Such an incompressible surface may be quasifuchsian but it also can be a non-geometrically finite (and thus non quasifuchsian) incompressible surface. At any rate, this result shows that we can replace any such sum of incompressible surfaces with a connected quasifuchsian surface without changing the homology class.

Let γ_1 and γ_2 denote two oriented closed curves inside a closed hyperbolic 3-manifold $\mathbf{M}^3$. Moving into a general position one can show that if γ_1 and γ_2 are homologous in $\mathbf{M}^3$ then γ_1 and $-\gamma_2$ bound an immersed surface in $\mathbf{M}^3$. Topologically it is much more significant when two homologous closed curves γ_1 and $-\gamma_2$ bound an essential surface inside $\mathbf{M}^3$ (a surface, possibly with boundary, is essential in $\mathbf{M}^3$ if its fundamental group injects into the fundamental group of $\mathbf{M}^3$). The following claim asserts that this property is always true in the rational homology $H_1(\mathbf{M}^3, \mathbb{Q})$. In particular, every closed homologically trivial curve in a closed hyperbolic 3-manifold $\mathbf{M}^3$ rationally bounds an essential surface in $\mathbf{M}^3$. This answers a question of D. Calegari in the case of hyperbolic 3-manifolds (this problem is wide open for hyperbolic groups for example).

Theorem 4.5 (Liu-Markovic). *Every rationally null-homologous, π_1 injectively immersed oriented closed 1-submanifold in a closed hyperbolic 3-manifold has an equidegree finite cover which bounds an oriented connected compact immersed quasi-Fuchsian subsurface.*

The following two very recent results by Hongbin Sun are heavily dependent on the Virtual Haken Theorem and Theorem 4.5.

Theorem 4.6 (Sun). *Let A be a finite Abelian group. Then every closed hyperbolic 3-manifold $\mathbf{M}^3$ has a finite degree cover $\mathbf{M}^3_1$ such that A is a direct summand in $\mathrm{Tor}\big(H_1(\mathbf{M}^3, \mathbb{Z})\big)$.*

Theorem 4.7 (Sun). *For any closed oriented hyperbolic 3-manifold* $\mathbf{M}^3$*, and any closed oriented 3-manifold* $\mathbf{N}^3$*, there exists a finite cover* $\mathbf{M}^3_1$ *of* $\mathbf{M}^3$*, and a degree-2 map* $f : \mathbf{M}^3_1 \to \mathbf{N}^3$*, i.e.* $\mathbf{M}^3$ *virtually 2-dominates* $\mathbf{N}^3$.

Very recently, Ursula Hamenstead (see [13]) showed that most closed, rank-one locally symmetric spaces contain surface subgroups. In particular, she proves that every closed complex hyperbolic contains a surface subgroup.

Acknowledgements. JK is supported by NSF grant number DMS 0905812; VM is supported by NSF grant number DMS 1201463.

References

[1] I. Agol, *Criteria for virtual fibering*, J. Topol. **1** (2008), no. 2, 269–284.

[2] _____, *The virtual Haken conjecture*, with an appendix by I. Agol, D. Groves and J. Manning, arXiv:1204.2810.

[3] M. Aschenbrenner, S. Friedl, and H. Wilton, *3-Manifolds Groups*, arXiv:1205.0202.

[4] N. Bergeron and D. Wise, *A boundary criterion for cubulation*, Amer. J. Math. **134** (2012), 843–859.

[5] L. Bowen, *Weak Forms of the Ehrenpreis conjecture and the surface subgroup conjecture*, arXiv:math/0411662.

[6] M. Brin and M. Gromov, *On the ergodicity of frame flows*, Inventiones Math. **60** (1980), no. 1, 1–7.

[7] D. Calegari, *SCL*. MSJ Memoirs, 20. Mathematical Society of Japan, Tokyo, 2009.

[8] R. Canary, D. Epstein, and P. Green, *Notes on notes of Thurston*, With a new foreword by Canary. London Math. Soc. Lecture Note Ser., **328**, Fundamentals of hyperbolic geometry: selected expositions, 1–115, Cambridge Univ. Press, Cambridge, 2006.

[9] D. Cooper, D. Long, and A. Reid, *Essential closed surfaces in bounded* 3*-manifolds*, Journal American Mathematical Society **10** (1997), no. 3, 553–563.

[10] D. Epstein, A. Marden, and V. Markovic, *Quasiconformal homeomorphisms and the convex hull boundary*, Ann. of Math. (2) **159** (2004), no. 1, 305–336.

[11] A. Eskin and C. McMullen, *Mixing, counting, and equidistribution in Lie groups*, Duke Math. J. **71** (1993), no. 1, 181–209.

[12] F. Haglund and D. Wise, *A combination theorem for special cube complexes*, Ann. Math. (2) **176** (2012), No. 3, 1427–1482.

[13] U. Hamenstead, *Incompressible surfaces in rank one locally symmetric spaces*, arXiv: 1402.1704.

[14] J. Kahn and V. Markovic, *Immersing Nearly Geodesic Surfaces in a Closed Hyperbolic 3-manifold*, Annals of Mathematics, **175** (2012), 1127–1190.

[15] ———, *Counting essential surfaces in a closed hyperbolic 3-manifold*, Geometry and Topology **16** (2012), no. 1, 601–624.

[16] ———, *The good pants homology and a proof of the Ehrenpreis conjecture preprint*, arXiv:1101.1330.

[17] M. Lackenby, *Surface subgroups of Kleinian groups with torsion*, Invent. Math. **179** (2010), no. 1, 175–190

[18] Y, Liu and V. Markovic, *Homology of curves and surfaces in closed hyperbolic 3-manifolds*, arXiv:1309.7418.

[19] G. Margulis, *Certain applications of ergodic theory to the investigation of manifolds of negative curvature*, Funkcional. Anal. i Priloen. **3** (1969), no. 4, 89–90.

[20] C. Moore, *Exponential decay of correlation coefficients for geodesic flows*, Group representations, ergodic theory, operator algebras, and mathematical physics (Berkeley, Calif., 1984), 163–181, Math. Sci. Res. Inst. Publ., 6, Springer, New York, 1987.

[21] M. Pollicott, *Exponential mixing for the geodesic flow on hyperbolic three-manifolds*, J. Statist. Phys. **67** (1992), no. 3-4, 667–673.

[22] M. Sageev, *Ends of groups pairs and non-positively curved cube complexes*, Proc. London. Math. Soc. (3) **71** (1995), no. 3, 585–617.

[23] D. Wise, *The structure of groups with a quasi-convex hierarchy*, 189 pages, preprint, 2012.

CUNY Graduate Center, 365 5th Ave, New York, NY 10016, USA
E-mail: jkahn@gc.cuny.edu

Pure Mathematics, University of Cambridge, Wilberforce Road, CB3 0WB, Cambridge, UK
E-mail: v.markovic@dpmms.cam.ac.uk

The Geometry of Ricci Curvature

Aaron Naber

Abstract. This is an overview of recent developments in geometry and analysis of Riemannian manifolds with lower and bounded Ricci curvature.

Mathematics Subject Classification (2010). Primary 53-02.

Keywords. ricci, curvature, regularity, path space,

1. Introduction

This paper overviews various recent developments in analysis and geometry related to Ricci curvature. To begin with, if one were to consider a Riemannian manifold (M^n, g) then its curvature operator Rm may best be interpreted as the *hessian* of the metric. This is of course not literal, the actual hessian of the metric vanishes, but in many ways one can expect from a Riemannian manifold with bounded sectional curvature the same type of control one expects from a function with a C^2 bound. It is therefore not surprising a fairly complete understanding of such spaces exists.

The Ricci curvature Ric of the Riemannian manifold is the trace of the curvature. If we therefore interpret the curvature as the hessian of the metric, we should then interpret the Ricci curvature as the laplacian of the metric. Ricci curvature is a prevalent concept in geometric analysis, and by this interpretation one can understand why. Everywhere in analysis one deals with the laplacians of functions, it is therefore not surprising that one would want to deal with the laplacian of the metric. This intuition only goes so far, but it makes for a good starting point.

Though much of this paper will describe recent results in the development of Ricci curvature, we will see that many of the themes and ideas have applications to a broader class of equations, geometric and not. For good reviews on older developments in Ricci curvature we refer the reader to [6],[36].

Let us begin by breaking down the layout of this overview. Section 2 begins with a brief background and overview of Ricci curvature. This includes basic definitions and previous results in the literature which are of importance. Section 3 discusses recent results for spaces with bounded Ricci curvature. In particular, new analytic estimates are discussed which are equivalent to the corresponding bounds on Ricci curvature. Section 4 discusses recent advances in the structure of spaces with lower Ricci curvature bounds. This includes an outline of the constant dimension conjecture, and Hölder continuity of tangent cones. Section 5 discusses a recent technique in the proving of *a priori* estimates for solutions of nonlinear equations. The notion is that of quantitative stratification, and though the primary

Proceedings of International Congress of Mathematicians, 2014, Seoul

application discussed is to prove *a priori* curvature estimates for Einstein manifolds, there have since been applications in many unrelated areas, for instance minimal surfaces, mean curvature flow, harmonic maps between Riemannian manifolds, and critical sets of elliptic equations. Section 6 ends the paper with a list of some open problems and conjectures from the area of Ricci curvature.

2. Background and Overview of Ricci Curvature

There are many ways in which the Ricci tensor arises in both geometry and physics, but the primary interest of this article is in how Ric is tied in with the analysis and geometric structure of M. In this section we review some results, new and old, which help to give the appropriate intuition for the Ricci curvature.

2.1. Lower Ricci Curvature and The Heat Flow. In this section we review the results of Bakry-Emery-Ledoux on characterizing lower Ricci curvature bounds in terms of the behavior of the heat flow. These estimates, and related estimates, not only play a crucial role in the structure theory of sections 2.3 and 4, but are also the first motivation for related estimates for bounded Ricci curvature given in section 3.

Recall first that given a Riemannian manifold (M^n, g) we have associated to it the laplace operator Δ. From this we have the heat flow operator $H_t : L^2(M) \to L^2(M)$ induced by $\frac{1}{2}\Delta$, and the corresponding heat kernel $\rho_t(x, dy) = \rho_t(x, y)dv_g(y)$ defined by

$$H_t u(x) = \int_M u(y)\rho_t(x, dy)\,. \tag{2.1}$$

To understand the role of the Ricci curvature tensor on the analysis of M one begins with a simple computation with the laplacian and gradient to obtain the Bochner formula

$$\Delta|\nabla u|^2 = \langle \nabla u, \nabla \Delta u\rangle + 2|\nabla^2 u|^2 + 2\mathrm{Ric}(\nabla u, \nabla u)\,, \tag{2.2}$$

from which if we assume the lower Ricci bound $\mathrm{Ric} \geq -\kappa g$ we get the Bochner inequality

$$\Delta|\nabla u|^2 \geq \langle \nabla u, \nabla \Delta u\rangle - 2\kappa|\nabla u|^2\,. \tag{2.3}$$

These inequalities are equivalent to the lower bounds on the Ricci curvature, and are the basis for the definition of lower Ricci curvature given by Bakry-Emery [4]. It should be pointed out that their precise condition applies to a much broader situation.

From the Bochner formula many important estimates on the heat flow H_t can be proved, which themselves turn out to be equivalent to the lower Ricci bound. We summarize the results of [4],[5] in the following theorem.

Theorem 2.1 (Bakry-Emery-Ledoux). *Let (M^n, g) be a smooth Riemannian manifold, then the following are equivalent:*

(1) *Ric* $\geq -\kappa g$.
(2) $|\nabla H_t u| \leq e^{\frac{\kappa}{2}t} H_t|\nabla u|$.
(3) $|\nabla H_t u|^2 \leq e^{\kappa t} H_t|\nabla u|^2$.

(4) $\lambda_1(-\Delta_{x,t}) \geq \kappa(e^{\kappa t}-1)^{-1}$.

(5) $\int_M u^2 \ln u^2 \rho_t(x,dy) \leq 2\kappa^{-1}\,(e^{\kappa t}-1)\int_M |\nabla u|^2 \rho_t(x,dy)$ *if* $\int_M u^2\,\rho_t = 1$.

In condition (4) above we are using the heat kernel laplacian, defined as the laplace operator associated to the metric measure space $(M^n, g, \rho_t(x,dy))$. Explicitly, we have that

$$\Delta_{x,t} u = \Delta u - \langle \nabla_y \rho_t(x,y), \nabla u\rangle\,. \tag{2.4}$$

There are various consequences of Theorem 2.1, the most important of which is that a solution of the heat flow has a priori gradient estimates. There are dimensional versions of Theorem 2.1 which give rise to laplacian bounds as well.

2.2. Rigidity Theorems for Lower Ricci Curvature. In the previous subsection we discussed the relationship between lower Ricci curvature and the heat flow. This will eventually give the analytic control needed in the structure theory. However, as for many areas, structure theories for lower Ricci curvature also require a notion of rigidity. In this subsection we briefly review the two main such rigidities that play a role for lower Ricci curvature, namely volume monotonicity and the splitting theorem. Throughout the paper various generalizations of these rigidity theorems will be presented, and in essence we will see that each time one can improve on the rigidity it allows one to prove stronger structure theorems.

We begin with the Bishop-Gromov volume monotonicity and the associated rigidity statement. Throughout let us consider a manifold (M^n, g) with the lower Ricci bound $\mathrm{Ric} \geq (n-1)\kappa$. Then we denote by M^n_κ the space form of curvature κ. That is, M^n_κ is the unique simply connected n-manifold whose sectional curvatures are all κ. We denote by $\mathrm{Vol}_\kappa(B_r)$ the volume of a ball of radius $r>0$ in M^n_κ.

To explain the result in its full strength let us discuss the notion of a metric cone space. Namely, let Y be a compact Riemannian manifold, and let $C_\kappa(Y) \equiv Y\times[0,\infty)/\{Y\times\{0\}\}$ be the topological cone of Y. We can equipped $C_\kappa(Y)$ with a natural metric given by $g_\kappa \equiv dr\otimes dr + sn_\kappa(r)^2 g_Y$, where

$$sn_\kappa(r) \equiv \begin{cases} |\kappa|^{-1/2} sinh(\sqrt{|\kappa|}r), & \text{if } \kappa<0\\ r, & \text{if } \kappa = 0\\ |\kappa|^{-1/2} sin(\sqrt{|\kappa|}r), & \text{if } \kappa>0 \end{cases} \tag{2.5}$$

When $\kappa\equiv 0$ we will sometimes just write $C(Y) = C_0(Y)$ as the cone space. It is not so hard to generalize this construction when Y is an arbitrary metric space. Let us now state the Bishop-Gromov volume monotonicity, and the rigidity statement associated to it:

Theorem 2.2 (Bishop-Gromov). *Let (M^n,g) be complete with Ric $\geq (n-1)\kappa$. Then for each $x\in M$ if we consider the volume ratio $V_r(x) \equiv \frac{Vol(B_r(x))}{Vol_\kappa(B_r)}$, then we have that $\frac{d}{dr}V_r(x)\leq 0$. Further, if there exists $0<r_0<r_1$ such that $V_{r_0}(x) = V_{r_1}(x)$, then the annulus $A_{r_0,r_1}(x)$ is isometric to an annulus $A_{r_0,r_1}(0_\kappa)$ in M^n_κ.*

A quantitative version of the above is the basis for the stratification theory of the next section, as well as the starting point for the quantitative stratification and curvature estimates for Einstein manifolds of section 5.

To exploit the volume monotonicity and its associated rigidity of Theorem 2.2, one will need to be in the noncollapsed setting. That is, if (M^n,g,p) is a pointed Riemannian manifold with a lower Ricci bound, the estimates and structure of sections 2.3 and 5 depend on

a lower volume bound $\mathrm{Vol}(B_1(p)) > v > 0$ of the unit ball around p. On the other hand, there is another form of rigidity which will allow one to prove structure theorems even in the collapsed setting. The Cheeger-Gromoll splitting theorem is the starting point for this point of view:

Theorem 2.3 (Cheeger-Gromoll). *Let (M^n, g) be a complete Riemannian manifold with $\mathrm{Ric} \geq 0$. Assume there exists a line $\gamma : (-\infty, \infty) \to M$, that is, γ is a minimizing geodesic satisfying $d(\gamma(s), \gamma(t)) = |t - s|$ for all $s, t \in (-\infty, \infty)$. Then there exists a smooth manifold N with $\mathrm{Ric} \geq 0$ such that M is isometric to the product space $\mathbb{R} \times N$.*

The first structure theorems for spaces with lower Ricci curvature bounds follow in [7] by first proving an effective version of the splitting theorem, see section 2.3. The proofs of several conjectures remaining in [7] are proved by first building further refinements of the splitting theorem, see section 4 for more on this. Recently the splitting theorem has been proved in [22] for general metric-measure spaces which satisfy a nonnegative Ricci curvature assumption in the weak sense. In the paper [29], this is used to prove a structure theory for such spaces which extends the results of section 2.3 and 4.

2.3. Limit Spaces and Basic Structure Theory of Lower Ricci Curvature. In this section we review the basic structure of spaces with lower Ricci curvature bounds. If one wants to study the structure of such spaces, then it is equally important to study limits of sequences of such spaces. The right notion of limit here is given by Gromov-Hausdorff convergence. That is, let $(X, d_X, p), (Y, d_Y, q)$ be pointed metric spaces. We call a mapping $f : B_{\epsilon^{-1}}(p) \to B_{\epsilon^{-1}}(q)$ an ϵ-isometry if

(1) $f\big(B_{\epsilon^{-1}}(p)\big)$ is ϵ-dense in $B_{\epsilon^{-1}}(q)$

(2) $|d_X(x, y) - d_Y(f(x), f(y))| < \epsilon$ for all $x, y \in B_{\epsilon^{-1}}(p)$.

Then we say

$$d_{GH}(X, Y) \leq \epsilon\,, \tag{2.6}$$

if there exists ϵ-isometries $f : B_{\epsilon^{-1}}(p) \to B_{\epsilon^{-1}}(q)$ and $h : B_{\epsilon^{-1}}(q) \to B_{\epsilon^{-1}}(p)$. See [31] for a good introduction to the subject. The importance of the Gromov-Hausdorff convergence in the subject begins with the following classical compactness theorem by Gromov:

Theorem 2.4 (Gromov). *Let (M_j^n, g_j, p_j) be a sequence of complete, pointed Riemannian manifolds with $\mathrm{Ric}_j \geq -(n-1)$, then there exists a subsequence and a pointed metric space (X, d, p) such that*

$$(M_j^n, g_j, p_j) \to (X, d, p)\,, \tag{2.7}$$

where the convergence is in the (pointed) Gromov-Hausdorff sense.

In the above theorem X is a priori just an arbitrary metric space. The regularity and structure theory of spaces with lower Ricci curvature bounds comes down to the ability to say more about X. The beginning point is the following result by Cheeger and Colding [7], which gives quantitative refinements of Theorems 2.2 and 2.3, which in particular allow one to conclude the results for limit spaces.

Theorem 2.5 (Cheeger-Colding). *Let (M^n, g) be a complete manifold. Then for each $\epsilon > 0$ there exists $\delta(n, \epsilon) > 0$ such that if $\mathrm{Ric} \geq -\delta$ with $x \in M$, then the following hold:*

(1) *(Almost Volume Cone) If* $|V_{\delta^{-1}}(x) - V_{\delta}(x)| < \delta$, *then there exists a metric space* Y *with diam* $\leq \pi$ *such that* $d_{GH}\left(B_{\epsilon^{-1}}(x), B_{\epsilon^{-1}}(0_y)\right) < \epsilon$, *where* $0_y \in C(Y)$ *is the cone point.*

(2) *(Almost Splitting Theorem) If* $\gamma : (-\delta^{-1}, \delta^{-1}) \to M$ *is a minimizing geodesic with* $\gamma(0) = x$, *then there exists a metric space* Y *such that* $d_{GH}\left(B_{\epsilon^{-1}}(x), B_{\epsilon^{-1}}(y)\right) < \epsilon$, *where* $y \in \mathbb{R} \times Y$.

When combined with some geometric measure theory there are many applications to the above statement. To discuss them let us separate into cases, the collapsed and noncollapsed:

2.3.1. Noncollapsed Limit Spaces. In this subsection we consider a limit space $(M^n_j,g_j,p_j) \to (X, d, p)$ where $\mathrm{Ric}_j \geq -(n-1)$ and $\mathrm{Vol}(B_1(p_j)) > v > 0$. We call such a limit space noncollapsed. This is the scenario in which the strongest regularity results are available. The regularity theorems of [7] take two forms. The first is about the stratification of the singular set of X, and the second is about regular part of X. The results of sections 4 and 5 will generalize these, so we will discuss both.

Stratifying a limit space is based on the classical idea of blow up. Namely:

Definition 2.6. Given a metric space (X_x, d_x, x), we call X_x a tangent cone for X at $x \in X$ if there exists a sequence $r_j \to 0$ such that $(X, r_j^{-1}d, x) \to (X_x, d_x, x)$.

By Gromov's theorem we have that for every sequence $r_j \to 0$ there exists a convergent subsequence. In particular, tangent cones exist at every point. Unfortunately, they may be highly non-unique, which is to say the we can get different tangent cones at the same point by taking different blow up sequences. For instance, in [14] examples of limit spaces are constructed where the tangent cones at a fixed point are not only not isometric, they are not homeomorphic and the dimension of the singular set varies, see section 4.1. However, the following is an important structural theorem for noncollapsed spaces, and it follows immediately from theorem 2.5 and the volume monotonicity:

Theorem 2.7. *In a noncollapsed limit* X *every tangent cone* X_x *is a metric cone. That is, there exists a compact metric space* Y_x *such that* $X_x = C(Y_x)$.

This is the starting point for the notion of a stratification. The idea of a stratification is a pretty standard one in the study of geometric nonlinear equations. In spirit, the idea is to separate points of X based on how singular the points are. To make this more precise we define the notion of k-symmetric:

Definition 2.8. Consider the following definitions:

(1) We call a metric space Y k-symmetric for $k \geq 0$ if there exists a metric space Z such that $Y = C(Z) \times \mathbb{R}^k$.

(2) Given a metric space X, we define for each $k \in \mathbb{N}$ the closed strata

$$S^k(X) \equiv \{x \in X : \text{ no tangent cone at } x \text{ is k-symmetric}\}. \tag{2.8}$$

Note then that theorem 2.7 is the statement that every tangent cone is 0-symmetric. Standard blow up techniques from geometric measure theory can now be used to prove the following:

Theorem 2.9 ([7]). *If $(M^n_j, g_j, p_j) \to (X, d, p)$ where $\mathrm{Ric}_j \geq -(n-1)$ and $\mathrm{Vol}(B_1(p_j)) > v > 0$, then we have the following Hausdorff dimension estimate:*

$$\dim_H S^k(X) \leq k\,. \tag{2.9}$$

In section 5 an entirely new proof of theorem 2.9 will be described which will allow for a quantitative refinement. This quantitative refinement is crucial in the proof of L^p-curvature estimates for Einstein manifolds.

Let us end this section with the following regularity statement from [7]. By generalizing an argument of Reifenberg, combined with volume convergence and theorem 2.5 one obtains:

Theorem 2.10 ([7]). *If $(M^n_j, g_j, p_j) \to (X, d, p)$ where $\mathrm{Ric}_j \geq -(n-1)$ and $\mathrm{Vol}(B_1(p_j)) > v > 0$, then there exists an open dense subset $X_0 \subseteq X$ which is a topological manifold.*

2.3.2. Collapsed Limit Spaces. The case of collapsed limits is much worse. In the noncollapsed scenario one focused on theorem 2.5.1 and the volume monotonicity. For collapsed limits the only rigidity theorem one can exploit is the almost splitting of theorem 2.5.2. Now we are considering a sequence $(M^n_j, g_j, p_j) \to (X, d, p)$ satisfying $\mathrm{Ric}_j \geq -(n-1)$. Now in the case when $\mathrm{Vol}(B_1(p_j)) \to 0$, it is better to consider the normalized measure $\nu_j \equiv \mathrm{Vol}(B_1(p_j))^{-1} dv_{g_j}$ on M. In this case the notion of Gromov-Hausdorff convergence can be extended to measured Gromov-Hausdorff convergence to give us a limit (after possible passing to a subsequence)

$$(M^n_j, g_j, \nu_j, p_j) \to (X, d, \nu, p)\,, \tag{2.10}$$

where ν is a measure on X with the property that if $f_j : B_{\epsilon_j^{-1}}(p_j) \to B_{\epsilon_j^{-1}}(p)$ is the sequence of Gromov Hausdorff maps with $\epsilon_j \to 0$, then

$$f_{j,*}\nu_j \to \nu\,. \tag{2.11}$$

To prove structural theorems on X one begins by showing that almost every point of x lies in the interior of a minimizing geodesic. Note then that by theorem 2.5.2 this implies that every tangent cone at such points split an $\mathbb{R}$-factor. Then one starts to iterate this construction by applying it to the tangent cone itself to see that almost every double tangent cone splits $\mathbb{R}^2$. By repeating this process one can prove the following:

Theorem 2.11 ([7]). *If X is a potentially collapsed limit, then for ν-a.e. $x \in X$ we have that the tangent cone at x is unique and isometric to $\mathbb{R}^{k_x}$.*

It was conjectured that k_x could be taken independent of x in the above. In section 4 we will explain how this conjecture is proved.

3. Bounded Ricci Curvature and Analysis on Path Space

In this section we overview recent results of [30] concerned with a new class of estimates for spaces with bounded Ricci curvature. Recall from Section 2.1 the results of Bakry-Emery-Ledoux, which showed that lower bounds on the Ricci curvature are equivalent to

certain estimates on the heat flow on M. These estimates took several forms, but all were essentially concerned with the gradient behavior of functions on M. These estimates, and similar such estimates, play a hugely important role in the structure theory of spaces with lower Ricci curvature bounds. These estimates may also be used to make sense of lower Ricci curvature bounds on general metric measure spaces, see [3] in particular.

Using the results of Bakry-Emery-Ledoux as a moral starting point, the goal of this section is to explain how to generalize these estimates to the context of two sided bounds on the Ricci curvature. That is, we want estimates of an analytical nature, which are not only implied by bounds on the Ricci curvature, but are sufficiently strong that they are equivalent to bounds on the Ricci curvature. It will turn out that for each of the estimates of Theorem 2.1, there will be a corresponding stronger estimate which is equivalent to bounded Ricci curvature. In particular, we see how bounded Ricci curvature on M controls the analysis of path space $P(M)$ in a manner analogous to how lower Ricci curvature controls the analysis on M. Though we will not discuss it, we refer the reader to [30] to see how these ideas may be used to make sense of a notion of bounded Ricci curvature on metric-measure spaces.

In more detail, in this section we see that bounded Ricci curvature can be characterized in terms of the metric-measure geometry of path space $P(M)$. The correct notion of geometry on path space is one induced by what we call the parallel gradient, and the measures on path space of interest are the classical Wiener measures. Our first such characterization is in section 3.2 and shows that bounds on the Ricci curvature are equivalent to certain parallel gradient estimates on path space. These turn out to be infinite dimensional analogues of the Bakry-Emery gradient estimates, which are the finite dimensional gradient estimates on the heat flow well known to characterize lower Ricci bounds. Our second characterization is in section 3.3 and relates bounded Ricci curvature to the stochastic analysis of path space. In particular, we see that bounds on the Ricci curvature are equivalent to the appropriate $C^{\frac{1}{2}}$-time regularity of martingales on $P(M)$. Our final characterization of bounded Ricci curvature relates Ricci curvature to the analysis on path space. Specifically, in section 3.4 we study the Ornstein-Uhlenbeck operator L_x, a form of infinite dimensional laplacian on path space, and prove sharp spectral gap and log-sobolev estimates under the assumption of bounded Ricci curvature. Further we show these estimates on L_x are again equivalent to bounds on the Ricci curvature. For analogous results for d-dimensional bounded Ricci curvature see [30].

3.1. Basics of Path Space. Given a Riemannian manifold (M^n, g) we have associated to it the laplace operator Δ. From this we have the heat flow operator $H_t : L^2(M) \to L^2(M)$ induced by $\frac{1}{2}\Delta$, and the corresponding heat kernel $\rho_t(x, dy) = \rho_t(x, y)dv_g(y)$ defined by

$$H_t u(x) = \int_M u(y)\rho_t(x, dy)\,. \tag{3.1}$$

Now recall that path space

$$P(M) \equiv C^0([0, \infty), M)\,, \tag{3.2}$$

comes equipped with a canonical collection of mappings. Namely, for each partition $\mathbf{t} \equiv \{0 \le t_1 < t_2 < \cdots < t_k < \infty\}$ we have an induced mapping $e_{\mathbf{t}} : P(M) \to M^k$ given by

$$e_{\mathbf{t}}(\gamma) = (\gamma(t_1), \ldots, \gamma(t_k))\,. \tag{3.3}$$

These evaluation maps play two important roles. The first is to define a class of functions on $P(M)$ that one can work with to do analysis, the cylinder functions. That is, given $u : M^k \to \mathbb{R}$ and a partition $\mathbf{t}$ we define the cylinder function $U : P(M) \to \mathbb{R}$ by the formula $U(\gamma) \equiv e_{\mathbf{t}}^* u(\gamma) \equiv u(\gamma(t_1), \dots, \gamma(t_k))$.

The second role of the evaluation maps is in the construction of the Wiener measure. Namely, for each $x \in M$ it is possible to prove the existence of a measure Γ_x on path space $P(M)$ which is uniquely determined by the pushforwards

$$e_{\vec{t},*}\Gamma_x = \rho_{t_1}(x, dy_1)\rho_{t_2-t_1}(y_1, dy_2)\cdots\rho_{t_k-t_{k-1}}(y_{k-1}, dy_k)\,. \tag{3.4}$$

The complicated aspect of this theorem is that the measure Γ_x does in fact concentrate on the continuous curves, however it turns out this may be proved in vast generality, see [21].

Finally, let us recall that if we consider the L^2 functions on path space $L^2(P_x(M), \Gamma_x)$, then this Hilbert space comes canonically equipped with a weakly continuous, nested, 1-parameter family of closed subspaces $L^2_t(P_x(M)) \subseteq L^2(P_x(M))$ defined by $F \in L^2_t$ iff $F(\gamma) = F(\sigma)$ whenever $\gamma(s) = \sigma(s)$ for $s \leq t$. We may equivalently describe the subspaces $L^2_t(P_x(M))$ in the following manner. Note that path space comes equipped with a one parameter family of σ-algebras $\mathcal{F}^t$, where $\mathcal{F}^t$ is the weak σ-algebra generated by the valuation maps $e_{\mathbf{t}}$ with $\mathbf{t}$ a partition of $[0, t]$. Then $L^2_t(P_x(M))$ is the closed subspace of $L^2(P_x(M))$ of $\mathcal{F}^t$-measurable functions. With this in mind, we say a one parameter family of functions $F^t \in L^2(P(M))$ is a martingale iff for each $s < t$ we have that F_s is the projection of F^t to L^2_s. One can often view a martingale as the appropriate generalization of the heat flow on path space.

3.2. The Gradient Estimate.

3.2.1. The Parallel Gradient. In order to explain the estimates in [30] we have to define the right notion of geometry on path space. More specifically, we need to define the right notion of gradient. As in any situation, a gradient requires two pieces of information. First we need a notion of a directional derivative. However since $P(M)$ is a Banach manifold whose tangent space at γ may be identified with the continuous vector fields at γ, the notion of a directional derivative is well defined. Additionally, to define a gradient we need a restricted class of vectors, for instance those of norm 1 in some sense or another, in order to turn the directional derivative into a gradient.

The notion introduced in [30], and the one relevant to Ricci curvature, is that of the parallel gradient. In fact, there are a 1-parameter family of such gradients. In essence, each is a gradient which depends on only a finite dimensional amount of information. However, the whole family can be used to recover some more standard infinite dimensional notions of gradient. More specifically, given a cylinder function $F : P(M) \to \mathbb{R}$ we define the norm of its 0-parallel gradient $|\nabla_0 F| : P(M) \to \mathbb{R}^+$ by

$$|\nabla_0 F|(\gamma) \equiv \sup_V \{D_V F : \dot{V} = 0\,, |V|(0) = 1\}\,. \tag{3.5}$$

That is, we are looking at all parallel translation invariant vector fields along γ of norm 1, and taking the supremum of the directional derivatives. This is a n-dimensional space of directions. Though we will not discuss this point, it should be pointed out that there is a subtlety in the above definition. Namely, the curves γ are continuous, and thus it may not be clear what is meant by a parallel translation invariant vector field. This issue can be

resolved, though it requires the notion of the stochastic parallel translation map, which we will not discuss here.

Similarly, let us introduce the notion of the t-parallel gradient. We call a left continuous vector field V along γ a t-parallel vector field if $V(s) = 0$ if $s < t$, and if $\dot{V}(s) = 0$ for $s \geq t$. Note that for a cylinder function F the directional derivative $D_V F$ is still well defined. Then we define

$$|\nabla_t F|(\gamma) \equiv \sup_V \{D_V F : V \text{ is } t\text{-parallel}\,, |V|(t) = 1\}\,. \tag{3.6}$$

Note that though we have only defined the norm, it is easy to see that this supremem is obtained for some t-parallel translation invariant vector field, which we may define as being *the* t-parallel gradient $\nabla_t F$. It is not so hard to check that these all extend to closed linear operators on $L^2(P(M), \Gamma_x)$, see [30] for more on the parallel gradient.

3.2.2. The Estimate. Now we are in a position to discuss our first characterization of bounded Ricci curvature on M. Let us begin by recalling the classic gradient estimates of Bakry-Emery on the heat flow. Their estimates tell us that the lower Ricci curvature bound $\text{Ric} \geq -\kappa$ is equivalent to the gradient estimate on the heat flow given by

$$|\nabla H_t u| \leq e^{\frac{\kappa}{2}t} H_t |\nabla u|\,, \tag{3.7}$$

where H_t is the heat flow associated to the operator $\frac{1}{2}\Delta$ on M. The first characterization of bounded Ricci curvature will be a path space version of this estimate. In fact, one may recover (3.7) by applying the path space estimate to essentially the simplest type of function on path space.

To describe the characterization let $F \in C^0(P(M))$ be a continuous function on path space, for instance a smooth cylinder function. In section 3.1 we described for each $x \in M$ the construction of the Wiener measure Γ_x on path space. Let us observe that by letting the measures Γ_x act on F we can construct a continuous function on M by considering

$$\int_{P(M)} F\, d\Gamma_x\,, \tag{3.8}$$

as a function of x. This method takes continuous functions on $P(M)$ to continuous functions on M, and it is reasonable to ask what else we know about $\int F\, d\Gamma_x$ as a function on M in terms of F as a function on $P(M)$. In particular, when is it a lipschitz function on M, and can we control the gradient of $\int F\, d\Gamma_x$ as a function on M in terms of the gradient of F as a function on $P(M)$. In this case it of course matters a great deal what we mean by *gradient* of F on $P(M)$. It turns out that if we mean the parallel gradient as defined in section 3.2.1, then the estimate

$$|\nabla \int_{P(M)} F\, d\Gamma_x| \leq \int_{P(M)} |\nabla_0 F|\, d\Gamma_x\,, \tag{3.9}$$

is equivalent to the smooth metric measure space being Ricci flat. That is, we have that (3.9) holds if and only if $\text{Ric} \equiv 0$. More generally, we have that $|\text{Ric}| \leq \kappa$ if and only if

$$|\nabla \int_{P(M)} F\, d\Gamma_x| \leq \int_{P(M)} |\nabla_0 F| + \int_0^\infty \frac{\kappa}{2} e^{\frac{\kappa}{2}s} |\nabla_s F|\, ds \cdot d\Gamma_x\,. \tag{3.10}$$

As a first application, it is a worthwhile computation to see what happens when we apply the above estimate to the simplest function on path space, namely a cylinder function $F(\gamma) \equiv u(\gamma(t))$ where u is a smooth function on M and $t > 0$ is fixed. In this case we see that we exactly recover (3.7), see [30].

3.3. The $C^{1/2}$-Martingale Estimate. Our second characterization of bounded Ricci curvature relates the bounds on the Ricci curvature to the stochastic analysis on M. Recall from section 3.1 that a family of functions $F^t : P_x(M) \to \mathbb{R}$ on based path space $P_x(M) \equiv \{\gamma \in P(M) : \gamma(0) = x\}$ is a martingale if for each $s < t$ we have that F^s is the projection of F^t to the $\mathcal{F}^s$-measurable functions. Again, one should view this as an appropriate extension of the heat flow on M to path space.

A particularly natural construction of a martingale is to begin with an L^2 function $F : P_x(M) \to \mathbb{R}$ and let F^t be the projection of F to $L^2_t(P_x(M)$. The martingale F^t is then a measurement of how much of F, as a function on path space, depends only on first $[0, t]$ of a curve. As a family of functions F^t is highly nondifferentiable. To see this note that for any partition $\mathbf{t} = \{0 \leq t_1 < \cdots < t_{|\mathbf{t}|} < \infty\}$ we have the identity

$$||F||^2_{L^2} = \sum ||F^{t_{k+1}} - F^{t_k}||^2_{L^2} \,. \tag{3.11}$$

From this it is clear that not only is the family F^t not differentiable in the t-variable, but what we may hope to converge is the quadratic limit

$$\lim_{s\to 0} \frac{\left(F^{t+s} - F^t\right)^2}{s} = [dF^t] \,, \tag{3.12}$$

for at least $a.e.\ t > 0$ in L^1. One should be very careful that this is not a pointwise limit. One may rephrase this for a martingale as follows. If we consider the quadratic variation

$$[F^t] \equiv \lim_{\mathbf{t}} \sum \left(F^{t_{j+1}} - F^{t_j}\right)^2 , \tag{3.13}$$

where the limit is taken over partitions of $[0, t]$, then for a martingale $[F^t]$ is of bounded variation and is $\mathcal{F}^t$-measurable. In particular, its time derivative $[dF^t]$ exists for almost every time. This is highly nontrue for general stochastic processes.

The infinitesmal quadratic variation $[dF^t]$ is the correct replacement for the time derivative of F^t. A reasonable question then is what properties of F control the quadratic variation $[F^t]$ and its infinitesmal $[dF^t]$. The main estimate of this section is that the estimate

$$\int_{P(M)} [dF^t] \, d\Gamma_x \leq \int_{P(M)} |\nabla_t F|^2 \, d\Gamma_x \,, \tag{3.14}$$

is equivalent to M being Ricci flat $\mathrm{Rc} \equiv 0$. More generally, the estimate

$$\int_{P(M)} \sqrt{[dF^t]} \, d\Gamma_x \leq \int_{P(M)} |\nabla_t F| + \int_t^T \frac{\kappa}{2} e^{\frac{\kappa}{2}(s-t)} |\nabla_s F| \, d\Gamma_x \,, \tag{3.15}$$

$$\int_{P(M)} [dF^t] \, d\Gamma_x \leq e^{\frac{\kappa}{2}(T-t)} \int_{P(M)} |\nabla_t F|^2 + \int_t^T \frac{\kappa}{2} e^{\frac{\kappa}{2}(s-t)} |\nabla_s F|^2 \, d\Gamma_x \,, \tag{3.16}$$

is equivalent to the bound $-\kappa g \leq \text{Ric} \leq \kappa g$ on the Ricci curvature. There are several stronger versions of this estimate, but we refer the reader to [30] for more on this.

As a last remark let us compare these estimates to the lower Ricci curvature context, and in particular let us note that (3.16) implies a generalization of the Bakry-Emery gradient estimate when applied to the simplest functions on path space. Specifically, when we apply (3.16) to the functions of the form $F(\gamma) \equiv u(\gamma(t))$, where u is a smooth function on M and t is fixed, then we will get the estimate

$$H_t|\nabla H_{T-t}u|^2(x) \leq e^{\kappa(T-t)}H_T|\nabla u|^2(x)\,, \tag{3.17}$$

for every smooth u and all times $0 \leq t \leq T$. It is not hard to see that (3.17) is equivalent to (3.7), and in particular is itself equivalent to the Ricci curvature lower bound $Rc \geq -\kappa g$.

3.4. Spectral Gap and Log-Sobolev Estimates on the Ornstein-Uhlenbeck Operator. The third characterization of bounded Ricci curvature shows how to equate bounds on the Ricci curvature of a smooth manifold with the analysis on path space. Specifically, we will recall below how to define the Ornstein-Uhlenbeck operators $L_x : L^2(P_x(M), \Gamma_x) \to L^2(P_x(M), \Gamma_x)$, which are infinite dimensional laplacians on path space, and see how the spectral properties of L_x are equivalent to bounds on the Ricci tensor.

Spectral gap and log-Sobolev inequalities for the Ornstein-Uhlenbeck operator on path space have a long history. In the context of path space on $\mathbb{R}^n$ they were first proved by Gross [23]. In this case one can approximate in a very strong sense the Ornstein-Uhlenbeck operator by finite dimensional operators and thus prove the estimate rather directly by more classical arguments. In the case of path space on a smooth Riemannian manifold the Ornstein-Uhlenbeck operator was first defined in [17], and its spectral gap and log-sobolev properties were first studied in [18, 24] and [1]. In [1] it was proven that such estimates existed for an arbitrary compact Riemannian manifold. To prove the result the manifold was isometrically embedded in Euclidean space, and therefore the spectral gap itself depended on the embedding. In [18, 24] it was first understood that Ricci curvature could also be used to control the spectral gap and log-Sobolev inequalities. The proof in [18] was based on a clever manipulation of the martingale representation formula for manifolds, which itself was based on a combination of the classic Clark-Ocone-Haussmann formula and Driver's integration by parts formula for the Malliavin gradient. The proof in [24] is based on a more inductive procedure. We refer the reader to the useful book [25] for a more complete reference.

In this section we will discuss the estimates of [30], which give sharp spectral gap and log-sobolev estimates, and show these estimates are equivalent to Ricci curvature bounds. To explain all of this more carefully let us briefly discuss the H^1_x-gradient on path space, which was first introduced by Malliavin. We will be interested in what's to come in studying functions F which are defined on based path space $P_x(M)$. Normally it is easier to consider the constructions on smooth cylinder functions first, and then to extend more arbitrarily. Classically, one defines the H^1_x-gradient on based path space in a manner similar to the parallel gradient of section 3.2.1 by

$$|\nabla F|_{H^1_x}(\gamma) \equiv \sup\left\{D_V F : \int_\gamma |\dot{V}|^2 = 1, V(0) = 0\right\}. \tag{3.18}$$

Now on based path space $P_x(M)$ we have introduced both a natural geometry given by the H^1_x-gradient, and we have a canonical measure given by the Wiener measure Γ_x. This allows us to define a Dirichlet form, from which the Ornstein-Uhlenbeck operator will be

defined. Namely, we define the closed symmetric bilinear form on $L^2(P_x(M),\Gamma_x)$ by the formula

$$E[F] \equiv \frac{1}{2}\int_{P_x(M)} |\nabla F|^2_{H^1_x} d\Gamma_x \,. \tag{3.19}$$

In fact, we have that the energy functional $E[F]$ on a smooth manifold is a Dirichlet form, see [17]. In particular, by the standard theory of Dirichlet forms [28], there exists a unique, closed, nonnegative, self-adjoint operator

$$L_x : L^2(P_x(M),\Gamma_x) \to L^2(P_x(M),\Gamma_x)\,, \tag{3.20}$$

such that

$$E[F] = \int_{P_x(M)} \langle F, L_x F\rangle \, d\Gamma_x \,. \tag{3.21}$$

The operator L_x is the Ornstein-Uhlenbeck operator on $P_x(M)$. Let us remark that it is not hard to show that L_x preserves $\mathcal{F}^T$-measurable functions F on $P_x(M)$. One can rephrase this by saying that L_x restricts to a well defined operator on time restricted based path space

$$P^T_x(M) \equiv \{\gamma \in C^0([0,T],M) : \gamma(0) = x\}\,. \tag{3.22}$$

It is clear from the definition that $L_x 1 = 0$, and since L_x is a self-adjoint operator it has a well defined spectral theory. It is then reasonable to ask about the behavior of this spectrum, and in particular whether or not there exists a spectral gap for the operator. It is shown in [30] that M satisfies the Ricci curvature bound $|\mathrm{Ric}| \leq \kappa$ if and only if the Ornstein-Uhlenbeck operator L_x on time restricted path space $P^T_x(M)$ satisfies the spectral gap

$$\lambda_1(L_x) \geq 2\left(e^{\kappa T}+1\right)^{-1}. \tag{3.23}$$

In fact, it turns out to also be the case that a smooth metric measure space satisfies the Ricci curvature eigenvalue bound $|\mathrm{Ric}| \leq \kappa$ if and only if we have the seemingly stronger log-Sobolev estimate

$$\int_{P_x(M)} F^2 \ln F^2 \, d\Gamma_x \leq \left(e^{\kappa T}+1\right)\int_{P_x(M)} |\nabla F|^2_{H^1_x}\, d\Gamma_x \,, \tag{3.24}$$

where F is any $\mathcal{F}^T$-measurable function on path space such that $\int_{P_x(M)} F^2\, d\Gamma_x = 1$. It is fairly standard that a log-Sobolev estimate of the form (3.24) implies the spectral gap (3.23). A consequence of Theorem 3.1 is the converse statement.

We end the section by discussing the relationship between this estimate and the lower Ricci curvature version. As in the previous estimates, the goal is to apply the estimates (3.23) and (3.24) to a function $F(\gamma) \equiv u(\gamma(t))$, where u is a fixed function on M and t is fixed. In this case we recover the spectral gap

$$\lambda_1(-\Delta_{x,t}) \geq \kappa\left(e^{\kappa t}-1\right)^{-1}\,, \tag{3.25}$$

on the heat kernel laplacian, as well as the log-Sobolev estimate

$$\int_M u^2 \ln u^2 \rho_t(x, dy) \leq 2\kappa^{-1} \left(e^{\kappa t} - 1\right) \int_M |\nabla u|^2 \rho_t(x, dy)\,, \tag{3.26}$$

where u is any function such that $\int_M u^2(y)\rho_t(x, dy) = 1$. A consequence of [5] is that these estimates are themselves equivalent to the lower Ricci bound $Rc \geq -\kappa g$, and therefore we have again easily recovered the lower Ricci curvature from the path space estimate.

3.5. Summary of Main Results. In this section we simply summarize the results of the previous sections. Let us remark that the results of [30] are more general in several directions. To begin with, in [30] it is seen that the results hold for metric-measure spaces, not just Riemannian manifolds. Additionally, a dimensional version of the inequalities is proved. However, for simplicity sake we restrict ourselves to the context of a Riemannian manifold for the next statement:

Theorem 3.1. *Let (M^n, g) be a smooth complete metric measure space, then the following are equivalent:*

(R1) *The Ricci curvature satisfies the bound*

$$-\kappa g \leq Ric \leq \kappa g\,. \tag{3.27}$$

(R2) *For any function $F \in L^2(P(M), \Gamma_g)$ on the total path space $P(M)$ we have the estimate*

$$\left|\nabla \int_{P(M)} F\, d\Gamma_x\right| \leq \int_{P(M)} \left(|\nabla_0 F| + \int_0^\infty \frac{\kappa}{2} e^{\frac{\kappa}{2}s} |\nabla_s F|\, ds\right) d\Gamma_x\,, \tag{3.28}$$

(R3) *For any function $F \in L^2(P(M), \Gamma_g)$ on the total path space $P(M)$ which is $\mathcal{F}^T$-measurable we have the estimate*

$$\left|\nabla \int_{P(M)} F\, d\Gamma_x\right|^2 \leq e^{\frac{\kappa}{2}T} \int_{P(M)} |\nabla_0 F|^2 + \int_0^T \frac{\kappa}{2} e^{\frac{\kappa}{2}s} |\nabla_s F|^2\, ds \cdot d\Gamma_x\,. \tag{3.29}$$

(R4) *For any function $F \in L^2(P(M), \Gamma_x)$ on based path space $P_x(M)$ we have the estimate*

$$\int_{P(M)} \sqrt{[dF^t]}\, d\Gamma_x \leq \int_{P(M)} |\nabla_t F| + \int_t^T \frac{\kappa}{2} e^{\frac{\kappa}{2}(s-t)} |\nabla_s F|\, d\Gamma_x\,, \tag{3.30}$$

(R5) *For any function $F \in L^2(P(M), \Gamma_x)$ on based path space $P_x(M)$ which is $\mathcal{F}^T$-measurable we have the estimate*

$$\int_{P(M)} [dF^t]\, d\Gamma_x \leq e^{\frac{\kappa}{2}(T-t)} \int_{P(M)} |\nabla_t F|^2 + \int_t^T \frac{\kappa}{2} e^{\frac{\kappa}{2}(s-t)} |\nabla_s F|^2\, d\Gamma_x\,, \tag{3.31}$$

(R6) *The Ornstein-Uhlenbeck operator* $L_x : L^2(P_x^T(M),\Gamma_x) \to L^2(P_x^T(M),\Gamma_x)$ *on based path space* $P_x^T(M)$ *satisfies the spectral gap estimate*

$$\lambda_1(L_x) \geq 2\left(e^{\kappa T}+1\right)^{-1}. \tag{3.32}$$

(R7) *On based path space* $P_x(M)$*, equipped with the Wiener measure* Γ_x *and the* H_x^1*-gradient, if* F *is a* $\mathcal{F}^T$ *measurable function then we have the log-Sobolev inequality*

$$\begin{aligned}\int_{P_x(M)} F^2\ln F^2\,d\Gamma_x &\leq 2e^{\frac{\kappa}{2}T}\int_{P(M)}\left(\int_0^T \cosh(\frac{\kappa}{2}t)|\nabla_t F|^2\,dt\right)d\Gamma_x \\ &\leq \left(e^{\kappa T}+1\right)\int_{P_x(M)}|\nabla F|^2_{H_x^1}\,d\Gamma_x\,,\end{aligned} \tag{3.33}$$

whenever F *satisfies* $\int_{P_x(M)} F^2\,d\Gamma_x = 1$.

An obvious but interesting corollary of the above is the following characterization of Ricci flat manifolds.

Corollary 3.2. *Let* $(M^n,g,e^{-f}dv_g)$ *be a smooth complete metric measure space, then the following are equivalent:*

(1) *The space is Ricci flat, that is,* $Ric+\nabla^2 f = 0$

(2) *For any function* F *on the total path space* $P(M)$ *we have the estimate*

$$\left|\nabla\int_{P(M)} F\,d\Gamma_x\right| \leq \int_{P(M)}|\nabla_0 F|\,d\Gamma_x.$$

(3) *For any function* F *on based path space* $P_x(M)$ *we have the estimate*

$$\int_{P(M)}\sqrt{[dF^t]}\,d\Gamma_x \leq \int_{P(M)}|\nabla_t F|\,d\Gamma_x.$$

(4) *The Ornstein-Uhlenbeck operator* $L_x : L^2(P_x(M),\Gamma_x)\to L^2(P_x(M),\Gamma_x)$ *on based path space satisfies the spectral gap estimate* $\lambda_1(L_x)\geq 1$.

(5) *For any function* $F\in L^2(P_x(M),\Gamma_x)$ *with* $\int_{P_x(M)}F^2\,d\Gamma_x = 1$ *on based path space we have the estimate*

$$\int_{P_x(M)} F^2\ln F^2\,d\Gamma_x \leq 2\int_{P_x(M)}|\nabla F|^2_{H_x^1}\,d\Gamma_x.$$

Remark 3.3. By twisting the Wiener measure one can find versions of the above estimates which characterize general Einstein manifolds.

4. Structure of Lower Ricci Curvature

In this section we discuss some recent results of [15] which prove various conjectures from [7, 20] on the structure of limit spaces with lower Ricci curvature bounds. In particular, we discuss how such limits have a well defined dimension and have isometry groups which are lie groups. The proofs revolve around new estimates, which in short state that the geometry of metric balls can only change at a Holder rate along minimizing geodesics. In particular, tangent cones *at the same scale* change along a geodesic at most at a Hölder rate. This Hölder rate turns out to be sharp. The following is the main estimate of [15].

Theorem 4.1 ([15]). *Let (M^n, g) be a Riemannian manifold which satisfies Ric $\geq -(n-1)$ with $\gamma : [0, \ell] \to M$ a minimizing geodesic. There there exists $C(n), \alpha(n) > 0$ such that for any $s, t \in (\delta, (1-\delta)\ell)$ and $r \leq r(n)\ell$ we have that*

$$d_{GH}(B_r(\gamma(s)), B_r(\gamma(t))) < \frac{C}{\delta \ell}\, r\, |s-t|^{\alpha(n)}\,. \tag{4.1}$$

Remark 4.2. The most direct application of the above is to see that tangent cones at the same scale along minimizing geodesics in limit spaces X must change at a continuous rate, see [15].

In fact the Hölder exponent $\alpha(n)$ is *effectively* $\frac{1}{2}$, and examples are constructed to show that the result fails for $\alpha > \frac{1}{2}$, see Theorem 4.10. The proof of Theorem 4.1 involves many new estimates which we will not discuss here and we refer the reader to [15] for more on this.

Theorem 4.1 can be compared to an important theorem of Petrunin in [32], where it is shown that on Alexandrov spaces (i.e. spaces with lower sectional bounds), tangent cones on minimizing geodesics are unique. Here we see that we can't expect uniqueness under only a lower Ricci assumption, but that at least tangent cones do change at a continuous rate along a minimizing geodesics. Let us state the main applications of this estimate. The following was a conjecture from [7], which informally states that limit spaces have a well defined dimension.

Theorem 4.3 ([15]). *Let $M_i^n \to X$ be a limit of manifolds satisfying Ric $\geq -(n-1)$. Let $\mathcal{R}^k \subseteq X$ be defined by $x \in \mathcal{R}^k$ iff the tangent cone at x is unique and equal to $\mathbb{R}^k$. Then there exists a unique constant $k \leq n$ such that $X \setminus \mathcal{R}^k$ has measure zero.*

Remark 4.4. More precisely, the measure μ_X in the above theorem is the canonical limit measure on X gotten by limiting the normalized volume forms on M_i, see [CoNa1].

Proof. Instead of giving the full proof let us consider a simple example. The moral of the construction gives the idea for the general proof without the necessary technical complications needed for the general case. Thus let X be a trumpet space. That is, let $X \equiv (-\infty, 0] \cup_{(0,0_c)} C(S^1)$, where we have identified the end of the interval $(-\infty, 0]$ with the cone point 0_c of $C(S^1)$. So X is a connected space with a one dimensional part and a two dimensional part. Let us see why X cannot arise as a limit of spaces with lower Ricci bounds by using Theorem 4.1. Indeed, assume it is, and let $x_0 \in (-\infty, 0)$ be a point in the one dimensional part and $x_1 \in C(S^1)/0_c$ be a point in the two dimensional part. Let γ be a minimizing geodesic connecting x_0 to x_1. Now part of γ lives in $(-\infty, 0)$ and the other part in $C(S^1)$, hence there is t_0 such that $\gamma(t) \in (-\infty, 0)$ for $t < t_0$ and $\gamma(t) \in C(S^1)/\{0_c\}$

for $t > t_0$. In particular, for $t < t_0$ we have the tangent cone at $\gamma(t)$ is unique and isometric to $\mathbb{R}$, and for $t > t_0$ the tangent cone at $\gamma(t)$ is unique and isometric to $\mathbb{R}^2$. However, by Theorem 4.1, and the remark following it, the tangent cones of $\gamma(t)$ must be changing continuously along γ, which is a contradiction. □

Now that we have seen that limit spaces X have a well defined regular set $\mathcal{R}^k$, the structure of this regular set is of interest. It is known that if X is a limit space of manifolds with bounded sectional then the regular set $\mathcal{R}$ is totally convex. Under only a lower Ricci curvature assumption we prove that $\mathcal{R}$ is weakly convex (see [15] for a precise definition). Under the additional assumptions of bounded Ricci curvature and noncollapsing we can improve this to the statement that $\mathcal{R}$ is totally convex.

Theorem 4.5 ([15]). *Let $M_i^n \to X$ be a limit of manifolds satisfying $|Ric| \leq n-1$ and $Vol(B_1(p)) > v > 0$. Then the regular set $\mathcal{R}^n$ is totally geodesic. That is, if $p, q \in \mathcal{R}$ and γ is a minimizing geodesic connecting p and q, then $\gamma \subseteq \mathcal{R}$.*

Finally, we state one last application of Theorem 4.1. Namely, it is conjectured in [20] and [7] that the isometry group of a limit space X is a Lie group. This conjecture is based on the proof in [20] of this statement when X is Alexandrov, and the proof in [7] of this statement when X is a noncollapsed limit of manifolds with a lower Ricci bound. Using Theorem 4.1 we can prove this conjecture. Essentially, it is a consequence of the weak convexity of the regular set which was alluded to before:

Theorem 4.6 ([15]). *Let $M_i^n \to X$ be a limit of manifolds satisfying $Ric \geq -(n-1)$. Then the isometry group of X is a Lie group.*

4.1. Examples with Lower Ricci Curvature. Throughout this section we let $M_i^n \to Y$ be a limit of manifolds satisfying the lower Ricci bound $\mathrm{Ric} \geq -(n-1)$ and the noncollapsing assumption $\mathrm{Vol}(B_1(p)) > v > 0$.

In a different direction than the previous subsection we want to understand to what extent examples of limit spaces Y can be as degenerate as possible. Unlike the Alexandrov case, e.g. limits with lower sectional curvature, it is not the case that tangent cones even need to be unique anymore, though it is always an open question as to what extent this nonuniqueness can be pushed. In [14] we give a *characterization* for the families of tangent cones which may appear at a point in a noncollapsed limit. To make this precise note that it does hold that every tangent cone Y_p at $p \in Y$ in a noncollapsed limit is a metric cone, so $Y_p = C(X_p)$ where $X_p \in \mathcal{M}_{GH}$ is a compact metric space. Given this we let

$$\overline{\Omega}_{Y,p} \equiv \{X \in \mathcal{M}_{GH} : \text{ some tangent cone } Y_p \text{ at } p \text{ satisfies } Y_p = C(X)\}, \tag{4.2}$$

be the space of all such cross sections. It is easy to check that $\overline{\Omega}_{Y,p}$ is closed and path connected in $\mathcal{M}_{GH}$ with respect to the Gromov-Hausdorff topology. It also holds that if $X \in \overline{\Omega}_{Y,p}$ then

$$\mathrm{Ric}(X) \geq n-1, \tag{4.3}$$

and that

$$\mathrm{Vol}(X) = V(p), \tag{4.4}$$

is independent of $X \in \bar{\Omega}_{Y,p}$. The second statement is a consequence of the Bishop-Gromov monotonicity. In fact, we see in [14] that this almost characterizes subsets of $\mathcal{M}_{GH}$ which can appear as $\overline{\Omega}_{Y,p}$ for some limit space Y. All that is missing is what we call Ricci closability. We will not give a precise definition here, but essentially this is nothing more than a form of geometric cobordism statement. The precise characterization is the following.

Theorem 4.7 ([14]). *Let Ω be an open connected manifold, our parameter space. Let $\{(X^{n-1}, g_s)\}_{s\in\Omega} \subseteq \mathcal{M}_{GH}$, with $n \geq 3$, be a smooth family of closed manifolds such that (4.3) and (4.4) hold and such that for some s_0 we have that X_{s_0} is Ricci closable. Then there exists a sequence of complete manifolds $(M^n_\alpha, g_\alpha, p_\alpha) \overset{GH}{\to} (Y, d_Y, p)$ which satisfy $Ric_\alpha \geq 0$ and $Vol(B_1(p)) > v > 0$ for which $\overline{\{X_s\}} = \overline{\Omega}_{Y,p}$, where $\overline{\{X_s\}}$ is the closure of the set $\{X_s\}$ in the Gromov-Hausdorff topology.*

We have two primary applications of this. First, we construct limit spaces whose tangent cones at a point have singular sets of varying dimensions. In particular, this rules out the possibility of stratifying limit spaces based on tangent cones. Secondly, we construct examples where differing tangent cones at a point may not even be homeomorphic. Specifically we have

Theorem 4.8 ([14]). *For every $n \geq 3$, there exists a limit space $(M^n_\alpha, g_\alpha, p_\alpha) \overset{GH}{\to} (Y, d_Y, p)$ where each M_α satisfies $Ric \geq 0$ and $Vol(B_1(p_\alpha)) > v > 0$, and such that for each $0 \leq k \leq n-2$ there exists a tangent cone at p which is isometric to $\mathbb{R}^k \times C(X)$, where X is a smooth closed manifold not isometric to the standard sphere.*

This example tells us that stratifying limit spaces Y by the singular behavior of tangent cones is not possible. The next example gives even more degenerate behavior.

Theorem 4.9 ([14]). *There exists a limit space $(M^5_\alpha, g_\alpha, p_\alpha) \overset{GH}{\to} (Y^5, d_Y, p)$ of a sequence M_α satisfying $Ric \geq 0$ and $Vol(B_1(p_\alpha)) > v > 0$, and such that there exists distinct tangent cones $C(X_0)$, $C(X_1)$ at $p \in Y$ with X_0 homeomorphic to $\mathbb{C}P^2 \sharp \overline{\mathbb{C}P}^2$ and X_1 homeomorphic to S^4.*

The final example shows that Theorem 4.1 is sharp.

Theorem 4.10 ([15]). *For every $\epsilon > 0$ there exists a noncollapsed limit space Y_ϵ with a minimizing geodesic $\gamma \subseteq Y_\epsilon$ such that*

(1) *The tangent cones $Y_{\gamma(s)}$ at each of the points $\gamma(s)$ are unique.*

(2) *The path $B_1 Y_{\gamma(s)}$ in the set of compact metric spaces $\mathcal{M}_{GH}$ with the Gromov Hausdorff topology is $C^{1/2}$-Hölder continuous.*

(3) *The path $B_1 Y_{\gamma(s)}$ in the set of compact metric spaces $\mathcal{M}_{GH}$ with the Gromov Hausdorff topology is not $C^{1/2+\epsilon}$-Hölder continuous.*

5. Quantitative Stratification and Regularity of Nonlinear PDE's

In this section we will outline new methods for proving L^p-estimates on the solutions of nonlinear equations. The method originated in [12] with the proving of *a priori* curvature

estimates for Einstein manifolds, however the technique is very general and has since been applied to minimal surfaces [11], harmonic maps between Riemannian manifolds [11], mean curvature flow [9], harmonic map flow, and the study of critical sets of elliptic equations [10]. In spirit, the new ideas involve improvements on the idea of stratification, and combining these improvments with ϵ-regularity theorems that exploit the symmetry produced by the stratification.

In the study of nonlinear partial differential equations, and in particular the study of geometric pde's, a very common tool is that of a stratification. We saw an example of this in section 2.3, and the general philosophy remains the same in other cases. That is, typically solutions of such nonlinear equations are not smooth and there is some singular set Sing of the solution. However, as a general principle it is helpful to not only separate out the smooth points from the singular points, but also to separate points based on the *degree* of the singularity. Thus, one usually has a stratification $S^0 \subseteq S^1 \subseteq \cdots \subseteq S^{n-1} \equiv \text{Sing}$ of the singular set where S^k is defined as the set of points which are not of degree $k+1$. The precise meaning of this varies from situation to situation, see section 2.3 for an example of this in the context of lower Ricci curvature.

The typical result of a stratification theory of a nonlinear pde is to prove an estimate of the form

$$\dim S^k \leq k\,, \tag{5.1}$$

where almost always the meaning of dimension is in the Hausdorff sense. As powerful as such results are, there is an underlying weakness to them. Fundamentally, this stems from the simple fact that the rationals are a dense subset of Euclidean space which have Hausdorff dimension zero. In particular, (5.1) cannot prevent the singular set from being dense, or from being arbitrarily dense. Thus, despite the many applications of (5.1), any questions which involve control of a more effective nature cannot be answered with (5.1).

In particular, what one typically would like to prove for solutions of differential equations are Schauder type estimates. In the context of Ricci curvature this means one would like to conclude *a priori* L^p curvature bounds for Einstein manifolds, for minimal surfaces this means showing L^p bounds on the second fundamental form, and for harmonic maps between Riemannian manifolds this means proving *a priori* L^p estimates on the gradient or hessian of such a map. We will see how to prove such estimates in this section. In the spirit of this article we will focus on the proof for Einstein manifolds, however the basic outline is the same in each situation (even if the details vary).

The outline of this section is the follows. In section 5.1 we give the rigorous definition of the quantitative stratification, and a statement of the main results concerning it. In section 5.2 we outline the proof of these estimates. The ideas involved turn out to be different from the standard proof of (5.1) by dimension reduction, and interestingly enough can be used to give a new proof of (5.1). In section 5.3 we apply these results to Einstein manifolds to conclude *apriori* L^p estimates on the curvature. In fact, we will see that the results are significantly stronger than this, and actually give estimates for the regularity scale. Finally in section 5.4 we outline the proof of the estimates of section 5.3.

5.1. Quantitative Stratification and Statement of Main Results. The goal of the quantitative stratification is to see that under reasonable circumstances, most points have balls of some definite size around them which contain a lot of symmetries. Of course we will make this more precise. The standard notion of stratification, which is reviewed in section 2.3, is

based on the idea of grouping points of a limit space X by the number of symmetries of its tangent cones. This is an infinitesmal notion which depends on the blow up behavior at a point. In particular, on a smooth manifold the stratification is always trivial as every tangent cone is $\mathbb{R}^n$ and thus has maximal symmetry. On the other hand, the quantitative stratification introduced in this section is based on grouping points based on the number of *almost* symmetries of balls of definite size. In particular, this stratification is not based on infinitesmal behavior, and even for a smooth manifold the quantitative stratification is nontrivial. The applications of section 5.3 depend on this.

To define the quantitative stratification we proceed by making rigorous the notion of *almost* symmetries:

Definition 5.1. Let (X, d) be a metric space, we make the following definitions:

(1) Given $x \in X$ we say X is k-symmetric at x if there exists a compact metric space Y such that $X \equiv C(Y) \times \mathbb{R}^k$ is isometric to a cone space over Y cross $\mathbb{R}^k$, and if x is the cone tip under this isometry.

(2) Given $x \in X$ and $0 < r \leq 1$ and $\epsilon > 0$ we say that X is (k, ϵ, r)-symmetric at x if there exists a metric space Z which is k-symmetric at $z \in Z$ such that

$$d_{GH}\left(B_{\epsilon^{-1}r}(x), B_{\epsilon^{-1}r}(z)\right) < \epsilon r\,. \tag{5.2}$$

To state the second definition in words, we say that X is (k, ϵ, r)-symmetric if the ball $B_r(x)$ looks very close to having k-symmetries. The quantitative stratification is then defined as follows:

Definition 5.2. Given a metric space X with $\epsilon, r > 0$ fixed and $k \in \mathbb{N}$, we define the closed (k, ϵ, r)-strata $S^k_{\epsilon,r}(X)$ by

$$S^k_{\epsilon,r}(X) \equiv \{x \in X : \text{ for no } r \leq s \leq 1 \text{ is } X \ \ (k+1, \epsilon, s) - \text{symmetric at } x\}\,. \tag{5.3}$$

Thus, the strata $S^k_{\epsilon,r}(X)$ is the collection of points such that no ball of size at least r is almost $k+1$-symmetric. The first main result of [12] is to show that for manifolds which are noncollapsed and have lower Ricci curvature bounds, the set $S^k_{\epsilon,r}(X)$ is small in a very strong sense. To say this a little more carefully, if one imagines the k-strata as being a well behaved k-dimensional submanifold (for simplicity sake), then one would expect the volume of the r-tube around the set to behave like r^{n-k}. We show this almost holds:

Theorem 5.3 ([12]). *Let (M^n, g, p) be a complete manifold with $Rc \geq -(n-1)$ and $Vol(B_1(p)) > v > 0$. Then for every $\epsilon > 0$ there exists $C(n, v, \epsilon)$ such that*

$$Vol(B_r(S^k_{\epsilon,r}(M)) \cap B_1(p)) \leq C r^{n-k-\epsilon}\,. \tag{5.4}$$

Remark 5.4. Though one can take $\epsilon > 0$ to be arbitrarily small, there is an ϵ loss in the power $r^{n-k-\epsilon}$ beyond what one might hope for. In fact, some loss is necessary as if one only assumes a lower bound on the Ricci curvature then the result fails for r^{n-k}. Under a full bound on the Ricci curvature it is conceivable the result may be improved however.

5.2. Outline of Proof of Theorem 5.3. We now give an outline of the proof of Theorem 5.3. The classical dimension estimates on the singular set (5.1) are given by a dimension reduction argument. Such an argument cannot be quantified. Morally, this is because the basis for the argument is to take multiple tangent cones. The quantification of such an argument leads to the comparison of scales which are not comparible, and a loss of effective control. Instead, we give an entirely new proof of (5.1) which is more adaptable to being made quantitative. The next subsections will outline the proof and the main ideas involved.

5.2.1. Cone Splitting. The essence of any stratification is symmetry. A stratification is a measurement of how many points (and scales) do not have a lot of symmetry. To provide estimates on such things one needs a way of producing symmetry. The basic idea in [12] is that of cone splitting.

To explain this, recall that a metric space X is 0-symmetric if $X = C(Y)$ is a cone space over another metric space. More generally X is 0-symmetric with respect to $x \in X$ if $X = C(Y)$ such that x is the cone point of $C(Y)$. Imagine now that X is 0-symmetric with respect to two distinct points $x_0, x_1 \in X$. In fact, it is not a hard exercise that this is only possible if $X \equiv \mathbb{R} \times C(Y')$ is 1-symmetric with x_0 and x_1 cone points lying on the same $\mathbb{R}$-factor. Said more simply, two 0-symmetries imply a 1-symmetry. This may be generalized simply into a process we call cone splitting:

Cone Splitting: Let $X = \mathbb{R}^k \times C(Y)$ be k-symmetric, and assume there exists $x_0 \notin \mathbb{R}^k \times \{0_y\}$, where 0_y is the cone point of $C(Y)$, such that X is 0-symmetric with respect to x_0. Then X is $k+1$-symmetric.

Thus the key step in producing symmetry of a metric space is simply to produce lots of 0-symmetry. More simply, k independent 0-symmetries imply a k-symmetry. We will see in the next subsection how to do this. Let us end this subsection with the effective version of the above cone splitting for spaces with lower Ricci curvature bounds:

Theorem 5.5 ([12]). *Let (M^n, g) be a Riemannian manifold with Ric $\geq -(n-1)$. For every $\epsilon, \tau > 0$ there exists $\delta(n, \epsilon, \tau) > 0$ such that the following holds. If M is $(10r, k, \delta)$-symmetric at $x \in M$ with Gromov-Hausdorff map $\phi : B_{10r}(x) \to \mathbb{R}^k \times C(Y)$, and if there exists $x_0 \in B_r(x)$ such that $\phi(x_0) \notin B_\tau(\mathbb{R}^k \times \{0_y\})$ with M being $(10r, 0, \delta)$-symmetric at x_0, then M is $(5r, k+1, \epsilon)$-symmetric at x.*

5.2.2. Energy Decomposition. In this last subsection we saw the key to forcing k-symmetries was to construct k indepedent 0-symmetries. To begin this process we want to consider the case of just finding which scales are 0-symmetric. The key ingredient here is the existence of a monotone quantity. Namely, if M satisfies Ric $\geq -(n-1)$ then for each $x \in M$ we have the monotone quantity

$$V_r(x) = -\ln \frac{\text{Vol}(B_r(x))}{\text{Vol}_{-1}(B_r)}\,, \tag{5.5}$$

where B_r is a ball of radius r in hyperbolic space. One can rephrase theorem 2.5.2 in the following manner:

A consequence of the Cheeger-Colding theory is the following quantitative form of the rigidity of theorem 2.2:

Theorem 5.6. *Let (M^n, g) be a Riemannian manifold with Ric $\geq -(n-1)$, then for each $\epsilon > 0$ there exists $\delta(n, \epsilon) > 0$ such that if for $r < \delta$ we have that $|V_r(x) - V_{\delta r}(x)| < \delta$, then M is $(r, 0, \epsilon)$-symmetric at x.*

Now let us denote by $r_\alpha = 2^{-\alpha}$ the sequence of scales which drop by a factor of $\frac{1}{2}$. For a given point $x \in M$ we will call the scale r_α a *good* scale if x is $(r_\alpha, 0, \epsilon)$-symmetric at scale r_α, and a *bad* scale otherwise. Now let us check the following easy, but highly important, consequence of the monotone quantity:

Theorem 5.7 ([12]). *Let (M^n, g, p) be a Riemannian manifold with Ric $\geq -(n-1)$ and Vol$(B_1(p)) > v > 0$. Then for each $\epsilon > 0$ there exists $N(n, \epsilon, v) > 0$ such that for every $x \in B_1(p)$ we have that there are at most N bad scales at x.*

Proof. Note first that if Vol$(B_1(p)) > v > 0$, then the doubling property of the volume tells us that Vol$(B_1(x)) > C(n)^{-1}v > 0$ for each $x \in B_1(p)$. Now it follows from theorem 5.6 that for r_α to be a good scale it suffices that $V_{r_\alpha}(x) - V_{r_{\alpha+\beta}}(x) < \delta$, where $\beta(n, \epsilon)$ and $\delta(n, \epsilon)$ are fixed constants. But now let us observe that

$$\sum |V_{r_\alpha}(x) - V_{r_{\alpha+1}}(x)| = \sum V_{r_\alpha}(x) - V_{r_{\alpha+1}}(x) = V_1(x) - V_0(x) \leq -\log(C^{-1}v)\,. \tag{5.6}$$

On the other hand imagine there are at least N scales for which $V_{r_\alpha}(x) - V_{r_{\alpha+\beta}}(x) > \delta$. Then we get that

$$\sum |V_{r_\alpha}(x) - V_{r_{\alpha+1}}(x)| \geq \frac{N\delta}{\beta}\,. \tag{5.7}$$

Combining the last two inequalities gives the required upper bound on N. □

The last Theorem tells us that at every point most scales are almost 0-symmetric. What we want to see now is that at most points and scales there are other points nearby which are also *good* scales. By the cone-splitting this would give us our desired symmetries.

We begin with the following, for each $x \in B_1(p)$ let us associate to x the infinite $\mathbb{Z}_2$-tuple given by

$$T_\alpha(x) = \begin{cases} 0, & \text{if } x \text{ is } (r_\alpha, 0, \epsilon) - \text{symmetric} \\ 1, & \text{otherwise} \end{cases} \tag{5.8}$$

The content of the last Theorem is that

$$\sum T_\alpha(x) \leq N(n, v, \epsilon)\,. \tag{5.9}$$

Now the energy decomposition of M is as follows. Let us fix $\beta \in \mathbb{N}$, and for each β-tuple of $\mathbb{Z}_2$, $T \in \mathbb{Z}_2^\beta$, let us define

$$M_T \equiv \{x \in B_1(p) : T_\alpha(x) = T_\alpha \text{ for all } 0 \leq \alpha \leq \beta\}\,. \tag{5.10}$$

That is, we are grouping points of M by which scales they are good and bad on. Now the key result is the following:

Theorem 5.8 ([12]). *Let (M^n, g, p) be a Riemannian manifold with Ric $\geq -(n-1)$ and Vol$(B_1(p)) > v > 0$. Then for each $\epsilon, \epsilon' > 0$ and $T \in \mathbb{Z}_2^\beta$ there exists $C(n, \epsilon) > 0$ such that*

$$\text{Vol}(B_r(S^k_{\epsilon,r} \cap M_T)) < Cr^{n-k-\epsilon'}, \tag{5.11}$$

for $r \geq r_\beta$.

The proof is involved, and based on an induction on β. The key idea which brings the energy decomposition into play is that by only comparing points which are good on the same scales, either we can locally cover M_T by a sufficiently small number of balls for the estimate, or if we cannot then this forces splitting by the cone-splitting principle.

5.2.3. Proof of Theorem 5.3. The proof of Theorem 5.3 now follows almost immediately from Theorem 5.8. There is no harm in assuming $r = r_\beta$ for some β, otherwise there exists a unique r_β such that $r_\beta \leq r < r_{\beta-1}$ and by volume doubling we can still use the r_β covering.

Now for each β-tuple $T \in \mathbb{Z}_2^\beta$ we have the desired estimate

$$\text{Vol}(B_{r_\beta}(S^k_{\epsilon,r_\beta} \cap M_T)) < Cr_\beta^{n-k-\epsilon'}, \tag{5.12}$$

from Theorem 5.8. In general, to estimate Vol$(B_r(S^k_{\epsilon,r} \cap B_1(p)))$ we would then need to multiply this estimate by the number of β-tuples T, which is $2^\beta = r_\beta^{-1}$. Of course this is too large a loss. However, by Theorem 5.7 we only need to consider those β-tuples T for which $\sum T_\alpha \leq N$, which is on the order of $\approx \beta^N \approx C \log r_\beta$. Thus if we let $\epsilon' = \epsilon/2$ we have that

$$\text{Vol}(B_{r_\beta}(S^k_{\epsilon,r_\beta} \cap B_1(p))) \leq \sum_T \text{Vol}(B_{r_\beta}(S^k_{\epsilon,r_\beta} \cap M_T)) < Cr_\beta^{n-k-\epsilon/2} \log r_\beta \leq Cr_\beta^{n-k-\epsilon}, \tag{5.13}$$

which proves Theorem 5.3.

5.3. Application to Einstein Manifolds. Now let us outline how Theorem 5.3 may be used to prove *a priori* L^p bounds on the curvature, and in fact for the regularity scale, for Einstein manifolds. As was discussed, the same moral works in many other areas to provide the first Schauder estimates for various nonlinear pde's.

To begin with, so that we may state the result in its full strength, let us define the notion of regularity scale:

Definition 5.9. Given a smooth Riemannian manifold (M^n, g), for each point $x \in M$ let us define the regularity scale $r_{|Rm|}(x)$ as

$$r_x = r_{|\text{Rm}|}(x) \equiv \max\left\{0 < r \leq 1 : \sup_{B_r(x)} |\text{Rm}| \leq r^{-2}\right\}. \tag{5.14}$$

Let us begin with some remarks on the regularity scale. For starters, note the easy estimate $|\text{Rm}|(x) \leq r_x^{-2}$, and thus lower bounds on the regularity scale correspond to upper bounds on the curvature. However, control over the curvature at a single point gives no a priori control of the geometry in a neighborhood of that point. On the other hand, the regularity scale bounds the curvature in a definite size neighborhood, which tells us everything about

the possible geometries. In particular, if M is Einstein we can immediately conclude that on a ball of half the size we have control over all derivatives of the curvature, that is:

$$\sup_{B_{r_x/2}(x)} |\nabla^{(k)}\mathrm{Rm}| \leq C_k r_x^{-k}\,. \tag{5.15}$$

Additionally it is worth remarking that the definition of the regularity scale is chosen to be scale invariant. That is, if we rescale the geometry by r_x^{-1}, so that $B_{r_x}(x) \to B_1(x)$, then we have that the curvature $|\mathrm{Rm}| \leq 1$ is bounded by one on the ball $B_1(x)$ of radius one.

Now the main result of this section and the main application of the quantitative stratification is the following:

Theorem 5.10 ([12]). *Let (M^n, g, p) be an Einstein manifold with $|Ric| \leq n-1$ and $Vol(B_1(p)) > v > 0$. Then the following hold:*

(1) *For each $p < 1$ we have that*

$$\int_{B_1(p)} |Rm|^p \leq \int_{B_1(p)} r^{-2p} < C(n, v, p)\,, \tag{5.16}$$

and in particular we have the Minkowski estimate

$$Vol(B_r(\{x : r_x \leq r\})) < C(n, v, p) r^{2p}\,. \tag{5.17}$$

(2) *If M is Kähler, then for each $p < 2$ we have that*

$$\int_{B_1(p)} |Rm|^p \leq \int_{B_1(p)} r^{-2p} < C(n, v, p)\,, \tag{5.18}$$

and in particular we have the Minkowski estimate

$$Vol(B_r(\{x : r_x \leq r\})) < C(n, v, p) r^{2p}\,. \tag{5.19}$$

(3) *Finally if we assume $\int |Rm|^p \equiv \Lambda < \infty$, then we can conclude the much stronger estimate*

$$Vol(B_r(\{x : r_x \leq r\})) < C(n, \Lambda, p) r^{2p}\,. \tag{5.20}$$

Theorem 5.10.3 is particularly useful for Kähler Einstein manifolds, where one has the L^2 bound based on topological assumptions.

5.4. Outline of Proof of the Quantitative Estimates. Now we end this section by outlining the proof of Theorem 5.10. In particular we focus on Theorem 5.10.2, the proofs of the other statements are similar up to technical details. The idea is very straightforward given the quantitative stratification of Theorem 5.3 and the ϵ-regularity of Cheeger-Tian. Namely, let us state their theorem in the following manner: If M^n is Kähler Einstein with $|\mathrm{Rc}| \leq n-1$ and $\mathrm{Vol}(B_1(p)) > v > 0$, then there exists $\epsilon(n, v) > 0$ such that if p is $(n-3, \epsilon, 2r)$-symmetric, then $r_x \geq r$.

Now let us fix $r > 0$ and $\epsilon > 0$ from the above statement and $\delta > 0$ aribitrarily. Then Theorem 5.3 tells us in particular that

$$\mathrm{Vol}(B_r(S^{n-4}_{\epsilon,2r}(M) \cap B_1(p))) < C(n, \epsilon, \delta) r^{4-\delta}\,. \tag{5.21}$$

However, if $x \notin S^{n-4}_{\epsilon,2r}$ then this is exactly the statement that there exists $s \geq 2r$ such that x is $(n-3,\epsilon,s)$-symmetric. In particular, by the ϵ-regularity theorem this immediately gives $r_x \geq r$. Combining this with the volume estimate (5.21) immediately gives the estimate (5.19), as claimed.

6. Open Problems and Conjectures

We end this write up with a small list of open problems and conjectures involving spaces with lower Ricci curvature bounds. Some of these are old and some are new. Let us begin with the basic issue of regularity of spaces with lower Ricci curvature bounds:

Open Question. *Let X be a Gromov-Hasudorff limit of manifolds M^n_j satisfying Ric $\geq -(n-1)$. More generally let X be a metric-measure space satisfying a lower Ricci curvature bound in the generalized sense, for instance X could be a $RCD(n,K)$-space as in [3]. Then is X a topological manifold on an open dense subset?*

In a recent paper [29] it has been shown that such a $RCD(n,K)$ space is rectifiable, which is roughly the statement that X may be obtained from gluing together subsets of Euclidean space. The statement of manifold roughly requires these subsets to be *open*. This is still unknown, and one of the more interesting problems in the area.

In another direction it has been discussed that even for noncollapsed limit spaces the tangent cones need not be unique, and in fact by [15] pretty much any subset which could be the collection of tangent cones at a point in some limit is. However, it would be of interest to understand how big a subset of a limit space does have unique tangent cones. One might hope for the following:

Conjecture 6.1. *Let (X,d,p) be a Gromov-Hasudorff limit of manifolds (M^n_j,g_j,p_j) satisfying Ric$_j \geq -(n-1)$ and Vol$(B_1(p_j)) > v > 0$. Then the set of points of X which have nonunique tangent cones has dimension $\leq n-3$.*

It is a corollary of [7] that away from a set of dimension $\leq n-2$ the tangent cones are unique and $\mathbb{R}^n$. It is also known that away from a set of dimension $\leq n-3$ that there exists *some* tangent cone which is $\mathbb{R}^{n-2}\times C(S^1(r))$, where $S^1(r)$ is the circle of radius $r \leq 1$. An equivalent version of the above conjecture is therefore the following

Conjecture 6.2. *Let (X,d,p) be a Gromov-Hasudroff limit of manifolds (M^n_j,g_j,p_j) satisfying Ric$_j \geq -(n-1)$ and Vol$(B_1(p_j)) > v > 0$. Then away from a set of Hausdorff dimension $\leq n-3$ the tangent cone is unique and isometric to $\mathbb{R}^{n-2}\times C(S^1(r))$ for some $r \leq 1$.*

A corollary of the above would be the $n-2$-rectifiability of the singular set.

Instead of considering just spaces with a lower Ricci curvature bound we could consider Einstein manifolds, and in particular noncollapsed Einstein manifolds. There are many structural conjectures out there at the moment, and the following is a strengthened version which would imply many of them:

Conjecture 6.3. *Let (M^n,g,p) be an Einstein manifold satisfying $|$Ric$| \leq n-1$ and*

Vol$(B_1(p)) > v > 0$. *Then there exists* $C(n, v) > 0$ *such that*

$$\int_{B_1(p)} |Rm|^2 < C\,. \tag{6.1}$$

Note that if M^n is Kähler then the conjecture is known if we further assume topological restrictions on M. Without any such topological restrictions it has been proved in Theorem 5.10 that $\int_{B_1(p)} |\mathrm{Rm}|^{2-\epsilon} < C_\epsilon$ for any $\epsilon > 0$. On the other hand, for four dimensional Einstein manifolds the conjecture is again known to hold under topological restrictions (in this case the L^2 norm of the curvature is equivalent to the Euler characteristic). Finally, if M^4 is Kähler and four dimensional, then the above conjecture can be shown to hold in full.

We end with the following structural conjecture for noncollapsed limits of Einstein manifolds:

Conjecture 6.4. *Let* (X, d, p) *be a Gromov-Hasudorff limit of Einstein manifolds* (M^n_j, g_j, p_j) *satisfying* $|Ric_j| \leq n - 1$ *and Vol*$(B_1(p_j)) > v > 0$. *Then* X *is homeomorphic to a real-analytic variety.*

A great deal of progress has been made recently in the above conjecture, in particular it has been proven for compact Kähler manifolds with positive first chern class, see [35],[16].

References

[1] S. Aida and K. D. Elworthy, *Differential calculus on path and loop spaces I. Logarithmic Sobolev inequalities on path spaces*, C. R. Acad. Sci. Paris **321**, (1995) 97–102.

[2] M. T. Anderson, *Convergence and rigidity of metrics under Ricci curvature bounds*, Invent. Math. **102** (1990), 429–445.

[3] Luigi Ambrosio and Nicola Gigli, and Giuseppe Savare, *Metric measure spaces with Riemannian Ricci curvature bounded from below*, preprint, 2012.

[4] D. Bakry and M. Emery, *Diffusions hypercontractives*, Seminaire de Probabilites XIX, Lecture Notes in Math., Springer-Verlag, New York. 1123 (1985), 177–206.

[5] D. Bakry and M. Ledoux, *A logarithmic Sobolev form of the Li-Yau parabolic inequality*, Rev. Mat. Iberoamericana **22** (2006), no. 2, 683–702.

[6] J. Cheeger, *Degeneration of Riemannian Metrics under Ricci Curvature Bounds*, Springer: Publications of the Scuola Normale Superiore.

[7] J. Cheeger and T. Colding, *On the Structure of Spaces with Ricci Curvature Bounded Below I*, J. Differential Geometry **45** (1997), 406–480.

[8] J. Cheeger, T.H. Colding, and G. Tian, *On the singularities of spaces with bounded Ricci curvature*, Geom. Functional Analysis **12** No. 5 (2002), 873–914.

[9] J. Cheeger, R. Haslhofer, and A. Naber, *Quantitative Stratification and the Regularity of Mean Curvature Flows*, preprint (2012).

[10] J. Cheeger, A. Naber, and D. Valtorta, *Quantitative Stratification and Critical Sets of Elliptic Equations*, preprint (2012).

[11] J. Cheeger and A. Naber, *Quantitative Stratification and the Regularity of Harmonic Maps and Minimal Currents*, accepted to Communications on Pure and Applied Mathematics, (2011).

[12] ———, *Lower Bounds on Ricci Curvature and Quantitative Behavior of Singular Sets*, Inventiones Math. **191** (2013), 321–339.

[13] T. Colding, A. Naber, and Lower Ricci, *Curvature, Branching, and Bi-Lipschitz Structure of Uniform Reifenberg Spaces*, accepted to Advances in Mathematics (2012).

[14] ———, *Characterization of Tangent Cones of Noncollapsed Limits with Lower Ricci Bounds and Applications*, Geometric and Functional Analysis **23**, Issue 1 (2013), 134–148.

[15] ———, *Sharp Hölder continuity of tangent cones for spaces with a lower Ricci curvature bound and applications*, Annals of Mathematics **176**, Issue 2 (2012), 1173–1229.

[16] X. Chen, S. Donaldson, and S. Sun, *Kahler-Einstein metrics and stability*, preprint, http://arxiv.org/abs/1210.7494.

[17] B. Driver and M. Röckner, *Construction of Diffusions on path and loop spaces on compact Riemannian manifolds*, C. R. Acad. Sci. (Paris) **315**, Serié I, 603–608 (1992)

[18] S. Fang, *Inégalité du type de Poincaré sur l'espace des chemins riemanniens*, C. R. Acad. Sci. (Paris),**318**, Serié I, 257–260 (1994)

[19] K. Fukaya, *The fundamental groups of almost nonnegatively curved manifolds*, Ann. of Math. **136**(2) (1992), 253–333.

[20] K. Fukaya and T. Yamaguchi, *Isometry groups of singular spaces*, Math. Z. **216** (1994), no. 1, 31–44.

[21] M. Fukushima, Y. Oshima, and M. Takeda, *Dirichlet Forms and Symmetric Markov Processes,* de Gruyter Studies in Mathematics **19**.

[22] N. Gigli, *The splitting theorem in non-smooth context*, preprint.

[23] L. Gross, *Logarithmic Sobolev inequalities*, Am. J. of Math. **97** (1975), 1061–1083.

[24] E. Hsu, *Logarithmic Sobolev Inequalities on Path Spaces Over Riemannian Manifolds*, Comm. Math. Phys. **189** (1997), 9–16.

[25] ———, *Stochastic Analysis on Manifolds*, Graduate Studies in Mathematics, AMS, 2000.

[26] V. Kapovitch, A. Petrunin, and W. Tuschmann, *Nilpotency, almost nonnegative curvature and the gradient push*, Annals of Mathematics, **171**(1) (2010), 343–373.

[27] V. Kapovitch and B. Wilking, *Structure of Fundamental Groups of Manifolds with Ricci Curvature Bounded Below*, preprint.

[28] Z. Ma and M. Röckner, *Dirichlet Forms*, Springer-Verlag, 1991.

[29] A. Mondino and A. Naber, *Structure Theory for Metric-Measure Spaces with Lower Ricci Curvature Bounds I*, preprint (2014).

[30] A. Naber, *Characterizations of Bounded Ricci Curvature on Smooth and NonSmooth Spaces*, preprint, 2013.

[31] P. Petersen, *Riemannian Geometry*, Springer, Graduate Texts in Mathematics, Vol 171.

[32] A. Petrunin, *Parallel transportation for Alexandrov space with curvature bounded below*, Geom. Funct. Anal. **8** (1998), no. 1, 123–148.

[33] K.-T. Sturm, *On the geometry of metric measure spaces*, Acta Math. **196** (2006), 65–131.

[34] G. Tian, *On Kähler-Einstein metrics on certain Kähler manifolds with* $c_1(M) > 0$, Invent. Math. **89** (1987), no. 2, 225–246.

[35] ———, *K-stability and Kähler-Einstein metrics*, preprint, arXiv:1211.4669.

[36] G. Wei, *Manifolds with A Lower Ricci Curvature Bound*, Survey, http://www.math.ucsb.edu/~wei/paper/06survey.pdf.

220 Lunt Hall, 2033 Sheridan Rd, Evanston Il, 60208

E-mail: anaber@math.northwestern.edu

New applications of Min-max Theory

André Neves

Abstract. I will talk about my recent work with Fernando Marques where we used Almgren–Pitts Min-max Theory to settle some open questions in Geometry: The Willmore conjecture, the Freedman–He–Wang conjecture for links (jointly with Ian Agol), and the existence of infinitely many minimal hypersurfaces in manifolds of positive Ricci curvature.

Mathematics Subject Classification (2010). Primary 53C42; Secondary 49Q05.

Keywords. Minimal surfaces, Willmore energy, conformal geometry, Min-max Theory.

1. Introduction

I will start by introducing three problems in Geometry which, while quite distinct, have in common the fact that their solution comes from understanding unstable critical points in the space of all embedded hypersurfaces on a given manifold. A panoramic overview of variational methods in Geometry can be found in the contribution of Fernando C. Marques [36].

1.1. Willmore Conjecture. A central question in Mathematics has been the search for the "optimal" representative within a certain class of objects. Partially motivated by this principle, Thomas Willmore started in the 60's the quest for the "optimal" immersion of a surface in space.

With that in mind, he associated to every compact surface $\Sigma \subset \mathbb{R}^3$ the quantity (now known as the *Willmore energy*),

$$\mathcal{W}(\Sigma) = \int_\Sigma \left(\frac{k_1 + k_2}{2}\right)^2 d\mu,$$

where k_1, k_2 are the principal curvatures of Σ.

The Willmore energy is invariant under rigid motions, scaling, and is large when the surface contains long thin tubes or long thin holes thus detecting how "bended" the surface Σ is in space. Less obvious, is the fact that the Willmore energy is also invariant under the inversion $x \mapsto x/|x|^2$ and thus invariant under conformal transformations. Willmore himself only found this some years later but, as we explain soon, this was know already since the twenties.

It is worthwhile to remark that in applied sciences the Willmore energy had already made its appearance a long time ago, under the name of *bending energy*, in order to study vibrating

Proceedings of the International Congress of Mathematicians, Seoul, 2014

properties of thin plates. In 1810's, Marie-Sophie Germain proposed, as the bending energy of a thin plate, the integral with respect to the surface area of an even, symmetric function of the principal curvatures, in which case the Willmore energy is the simplest possible example (excluding the area). Similar quantities were also considered by Poisson around the same time.

Moreover, the Willmore energy had also appeared in Mathematics through the work of Thomsen [52] and Blaschke [7] in the 1920's but their findings were forgotten and only brought to light after the interest in the Willmore energy increased. In particular, Thomsen and Blaschke were already aware of the conformal invariance of the Willmore energy.

Back to Willmore's quest for the "best" possible immersion, he showed that round spheres have the least possible Willmore energy among all compact surfaces in space. More precisely, ever compact surface $\Sigma \subset \mathbb{R}^3$ has

$$\mathcal{W}(\Sigma) \geq 4\pi$$

with equality only for round spheres.

Having found the compact surface with least possible energy, he tried to find the torus in space with smaller energy than any other tori. It is interesting to note that, just by looking at the shape of tori in space, no obvious candidate stands out. Hence, Willmore fixed a circle on a plane and considered tubes Σ_r of a constant radius r around that circle. When r is very small, Σ_r is a thin tube around the planar circle and thus its energy $\mathcal{W}(\Sigma_r)$ will be very large. If we keep increasing the value of r, the "hole" centered at the axis of revolution of the torus decreases and eventually disappears for some r_0. Thus $\mathcal{W}(\Sigma_r)$ will be arbitrarily large for r close to r_0. Therefore $\mathcal{W}(\Sigma_r)$ must have an absolute minimum as r ranges from 0 to r_0, which Willmore computed to be $2\pi^2$.

Up to scaling, the "optimal" torus that Willmore found has generating circle with radius 1 and center at distance $\sqrt{2}$ from the axis of revolution:

$$(u, v) \mapsto \big((\sqrt{2} + \cos\, u) \cos\, v, (\sqrt{2} + \cos\, u) \sin\, v, \sin\, u\big) \in \mathbb{R}^3.$$

In light of his findings, Willmore conjectured [55]:

Willmore Conjecture (1965). *Every compact surface Σ of genus one has*

$$\int_\Sigma \left(\frac{k_1 + k_2}{2}\right)^2 d\mu \geq 2\pi^2.$$

It seems at first rather daring to make such conjecture after having tested it only on a very particular one parameter family of tori. On the other hand, the torus Willmore found is special and had already appeared in Geometry: Inside the unit 3-sphere S^3 in $\mathbb{R}^4$ there is a highly symmetric torus, the Clifford torus, which is given by $S^1(\frac{1}{\sqrt{2}}) \times S^1(\frac{1}{\sqrt{2}})$. There is a stereographic projection from the 3-sphere onto space that sends the Clifford torus to the "optimal" torus found by Willmore.

The richness of the Willmore conjecture derives partially from the fact that the Willmore energy is invariant under conformal maps. One immediate consequence is that the conjecture can be restated for surfaces in the unit 3-sphere S^3. Indeed, if Σ is a compact surface in S^3 and $\tilde{\Sigma}$ its image in $\mathbb{R}^3$ under stereographic projection, then one has

$$\mathcal{W}(\tilde{\Sigma}) = \int_\Sigma 1 + \left(\frac{k_1 + k_2}{2}\right)^2 d\mu,$$

where k_1, k_2 are the principal curvatures of Σ *with respect* to the standard metric on S^3. For this reason, one calls the left-hand side of the above equation the *Willmore energy* $\mathcal{W}(\Sigma)$ of $\Sigma \subset S^3$.

The conjecture had been verified in many special cases borrowing inspiration from several distinct areas such as integral geometry, algebraic geometry, minimal surfaces, analysis or conformal geometry. We refer the reader to [38] for the history of partial results and mention simply the ones relevant to our work.

In 1982 it was proven by Li and Yau [33] that the Willmore energy of any non-embedded surface must be at least 8π (which is strictly bigger than $2\pi^2$. In particular, it suffices to check the Willmore conjecture for embedded tori.

Ros [46] proved in 1999 the Willmore conjecture for tori in S^3 that are preserved by the antipodal map and his method motivated our approach.

Curiously, two biologists, Bensimon and Mutz [40], verified the Willmore conjecture with the aide of a microscope while studying the physics of membranes! They produced toroidal vesicles in a laboratory and observed that they assumed the shape, which according to the Helfrich model [23] should be the minimizer for the Willmore energy, of the Clifford torus or its conformal images (called Dupin cyclides).

Jointly with Fernando Marques [38] we showed that

Theorem 1.1. *Every embedded compact surface Σ of S^3 with positive genus has*

$$\mathcal{W}(\Sigma) \geq 2\pi^2.$$

Equality only holds, up to rigid motions, for stereographic projections of the Clifford torus.

The rigidity statement characterizing the equality case in Theorem 1.1 is optimal because, as we have mentioned, the Willmore energy is conformal invariant.

Using the Li–Yau result previously mentioned we obtain

Corollary. *The Willmore conjecture holds.*

1.2. Energy of links. The second application comes from the theory of links in $\mathbb{R}^3$. Let $\gamma_i : S^1 \to \mathbb{R}^3$, $i = 1, 2$, be a 2-component link, i.e., a pair of closed curves in Euclidean three-space with $\gamma_1(S^1) \cap \gamma_2(S^1) = \emptyset$.

A 2-component link is said to be *nontrivial* if it cannot be deformed without intersecting itself into two curves contained in disjoint balls. To every link (γ_1, γ_2) one associates an integer invariant, called the linking number $\mathrm{lk}(\gamma_1, \gamma_2)$, that intuitively measures how many times each curve winds around the other.

To every 2-component link (γ_1, γ_2), O'Hara [44] associated an energy, called the *Möbius cross energy*. Its definition is reminiscent of the electrostatic potential energy and is given by ([15, 44]):

$$E(\gamma_1, \gamma_2) = \int_{S^1 \times S^1} \frac{|\gamma_1'(s)||\gamma_2'(t)|}{|\gamma_1(s) - \gamma_2(t)|^2}\, ds\, dt.$$

Freedman, He, and Wang studied this energy in detail and found that it has the remarkable property of being invariant under conformal transformations of $\mathbb{R}^3$ [15], just like the Willmore energy.

Using Gauss formula for the linking number, one can see that

$$E(\gamma_1, \gamma_2) \geq 4\pi |\mathrm{lk}(\gamma_1, \gamma_2)|$$

and so it is then natural to search for optimal configurations, i.e., minimizers of the Möbius energy. This question can be given the following nice physical interpretation (see [44]). Assuming that each curve in the link is non-conductive, charged uniformly and subject to a Coulomb's repulsive force, the equilibrium configuration the link will assume should minimize the Möbius energy.

Freedman, He and Wang [15] considered this question and after looking at the particular the case where one of the link components is a planar circle, they made the following conjecture.

Freedman–He–Wang Conjecture (1994). *The Möbius energy is minimized, among the class of all nontrivial links in $\mathbb{R}^3$, by the stereographic projection of the standard Hopf link in S^3.*

The standard Hopf link $(\hat{\gamma}_1, \hat{\gamma}_2)$ in S^3 is described by

$$\hat{\gamma}_1(s) = (\cos s, \sin s, 0, 0) \in S^3 \quad \text{and} \quad \hat{\gamma}_2(t) = (0, 0, \cos t, \sin t) \in S^3,$$

and it is simple to check that $E(\hat{\gamma}_1, \hat{\gamma}_2) = 2\pi^2$.

In a joint work with Ian Agol and Fernando Marques [1] we showed that:

Theorem 1.2. *Let (γ_1, γ_2) be a 2-component link in $\mathbb{R}^3$ with $|\mathrm{lk}(\gamma_1, \gamma_2)| = 1$. Then $E(\gamma_1, \gamma_2) \geq 2\pi^2$.*

Equality only holds, up to rigid motions and orientation, for stereographic projections of the Hopf link.

It follows from a result of He [22] that it suffices to prove the conjecture for links (γ_1, γ_2) with linking number $\mathrm{lk}(\gamma_1, \gamma_2) = \pm 1$. Thus, we obtained the following corollary

Corollary. *The conjecture made by Freedman, He, and Wang holds.*

1.3. Existence of embedded minimal hypersurfaces. A question lying at the core of Differential Geometry, asked Poincaré [41] in 1905, is whether every closed Riemann surface always admits a closed geodesic.

If the surface is not simply connected then we can minimize length in a nontrivial homotopy class and produce a closed geodesic. Therefore the question becomes considerably more interesting on a two-sphere, and the first breakthrough was in 1917, due to Birkhoff [6], who found a closed geodesic for any metric on a two-sphere.

Later, in a remarkable work, Lusternik and Schnirelmann [35] showed that every metric on a 2-sphere admits three simple (embedded) closed geodesics (see also [4, 17, 27, 30, 34, 51]). This result is optimal because there are ellipsoids which admit no more than three simple closed geodesics.

This suggests the question of whether we can find an infinite number of geometrically distinct closed geodesics in any closed surface. It is not hard to find infinitely many closed geodesics when the genus of the surface is positive.

The case of the sphere was finally settled in 1992 by Franks [14] and Bangert [5]. Their works combined imply that every metric on a two-sphere admits an infinite number of closed geodesics. Later, Hingston [24] estimated the number of closed geodesics of length at most L when L is very large.

Likewise, one can ask whether every closed Riemannian manifold admits a closed minimal hypersurface. When the ambient manifold has topology one can find minimal hypersurfaces by minimization and so, like in the surface case, the question is more challenging

when every hypersurface is homologically trivial. Using min-max methods, and building on earlier work of Almgren, Pitts [42] in 1981 proved that every compact Riemannian $(n+1)$-manifold with $n \leq 5$ contains a smooth, closed, embedded minimal hypersurface. One year later, Schoen and Simon [48] extended this result to any dimension, proving the existence of a closed, embedded minimal hypersurface with a singular set of Hausdorff codimension at least 7.

When M is diffeomorphic to a 3-sphere, Simon–Smith [49] showed the existence of a minimal embedded sphere using min-max methods (see also [8]).

Motivated by these results, Yau made the following conjecture [58] (first problem in the Minimal Surfaces section):

Yau's Conjecture (1982). *Every compact 3-manifold (M, g) admits an infinite number of smooth, closed, immersed minimal surfaces.*

Lawson [32] showed in 1970 that the round 3-sphere admits embedded minimal surfaces of every possible genus.

When M is a compact hyperbolic 3-manifold, Khan and Markovic [28] found an infinite number of incompressible surfaces in M of arbitrarily high genus. One can then minimize energy in their homotopy class and obtain an infinite number of smooth, closed, immersed minimal surfaces.

Jointly with Fernando Marques [39] we showed

Theorem 1.3. *Let (M^{n+1}, g) be a compact Riemannian manifold with $2 \leq n \leq 6$ and a metric of positive Ricci curvature.*

Then M contains an infinite number of distinct, smooth, embedded, minimal hypersurfaces.

Until Theorem 1.3 was proven, it was not even known whether metrics on the 3-sphere arbitrarily close to the round metric also admit an infinite number of minimal surfaces.

I find a fascinating problem to shed some light into the asymptotic behaviour of the minimal surfaces given by Theorem 1.3.

2. Almgren–Pitts Min-max Theory

As mentioned in the Introduction, Theorem 1.1, Theorem 1.2, and Theorem 1.3, follow from understanding the topology of the space of all embedded hypersurfaces.

In very general terms, the guiding principle of Morse Theory is that given a space and a function defined on that space, the topology of the space forces the function to have certain critical points. For instance, if the space has a k-dimensional nontrivial cycle, then the function should have a critical point of index at most k.

The space we are interested is $\mathcal{Z}_n(M)$, the space of all orientable compact hypersurfaces with possible multiplicities in a compact Riemannian $(n+1)$-manifold (M, g) with $n \geq 2$. If we allow for non-orientable hypersurfaces as well, the space is denoted by $\mathcal{Z}_n(M; \mathbb{Z}_2)$.

These spaces are studied in the context of Geometric Measure Theory and come with a well understood topology (flat topology) and equipped with a natural functional which associates to every element in $\mathcal{Z}_n(M)$ (or $\mathcal{Z}_n(M; \mathbb{Z}_2)$) its n-dimensional volume.

I will try to keep the discussion with as little technical jargon as possible in order to convey the main ideas and thus ignore almost all technical issues.

Critical points of the volume functional are called *minimal hypersurfaces* and their index is the number of independent deformations that decrease the area. For instance, on the 3-torus with the flat metric, there is a natural flat 2-torus which minimizes area in its homology class and thus it is a minimal surface of index zero. Likewise, on the 3-sphere with the round metric, the equator (which has area 4π) is a minimal sphere and has the property that if we "push" it up into the northern hemisphere then its area is decreased. Hence, its index is at least one. On the other hand, one can check that a deformation of the equator that preserves the enclosed volume is never area decreasing. Thus the equator has index one.

Almgren [3] started in the 60's the study of Morse Theory for the volume functional on $\mathcal{Z}_n(M)$ (or $\mathcal{Z}_n(M;\mathbb{Z}_2)$) and that continued through the 70's jointly with Pitts, his Phd student. I present now the basic principles of Almgren–Pitts Min-max Theory.

Suppose X is a topological space and $\Phi : X \to \mathcal{Z}_n(M)$ a continuous function. Consider

$$[\Phi] = \{\Psi : X \to \mathcal{Z}_n(M) : \ \Psi \text{ homotopic to } \Phi \text{ relative to } \partial X\}.$$

Note that if $\Psi \in [\Phi]$ then $\Phi = \Psi$ on ∂X. To the homotopy class $[\Phi]$ we associate the number, called the *width*,

$$\mathbf{L}([\Phi]) = \inf_{\Psi\in[\Phi]} \sup_{x\in X} \mathrm{vol}(\Psi(x)).$$

The Almgren-Pitts Min-max Theorem [42] can be stated as

Theorem 2.1 (Min-max Theorem). *Assume that* $\mathbf{L}([\Phi]) > \sup_{x\in\partial X} \mathrm{vol}(\Phi(x))$.

There is a compact embedded minimal hypersurface Σ *(with possible multiplicities) such that*

$$\mathbf{L}([\Phi]) = \mathrm{vol}(\Sigma).$$

The theorem also holds for $\mathcal{Z}_n(M;\mathbb{Z}_2)$ with no modifications.

The support of Σ is smooth outside a set of codimension 7 and thus smooth if $n \geq 6$ (for $n \geq 5$ the regularity theory was done by Schoen and Simon [48]). The Min-max Theorem allows for Σ to be a union of disjoint hypersurfaces, each with some multiplicity. More precisely

$$\Sigma = n_1\Sigma_1 + \ldots + n_k\Sigma_k,$$

where $n_i \in \mathbb{N}$, $i = 1, \ldots, k$, and $\{\Sigma_1, \ldots, \Sigma_k\}$ are embedded minimal surfaces with disjoint supports.

Naturally, if the space of parameters X is a k-dimensional, we expect the index of Σ to be at most k but this fact has not been proven.

The condition $\mathbf{L}([\Phi]) > \sup_{x\in\partial X} \mathrm{vol}(\Phi(x))$ means that $[\Phi]$ is capturing some nontrivial topology of $\mathcal{Z}_n(M)$. The guiding philosophy behind Min-max Theory consists in finding examples of homotopy classes satisfying this condition and then use the Min-max Theorem to deduce geometric consequences.

The next example illustrates well this methodology. Given $f : M \to [0,1]$ a Morse function, consider the continuous map

$$\Phi : [0,1] \to \mathcal{Z}_n(M), \quad \Phi(t) = \partial\{x \in M : f(x) < t\}.$$

We have $\Phi(0) = \Phi(1) = 0$ because all elements in $\mathcal{Z}_n(M)$ with zero volume are identified to be the same and Almgren showed in [2] that $L([\Phi]) > 0$, i.e., $[\Phi]$ is a nontrivial element of $\pi_1(\mathcal{Z}_n(M), \{0\})$. Using Min-max Theorem one obtains the existence of a minimal

hypersurface in (M, g) whose volume realizes $L([\Phi])$ and thus (M, g) admits a minimal embedded hypersurface which is smooth outside a set of codimension 7. This application was one of the motivations for Almgren and Pitts to develop their Min-max Theory.

3. The $2\pi^2$ Theorem

Let I^k denote a closed k-dimensional cube and $B_r(p)$ denote the geodesic ball in S^3 of radius r centered at p.

We present a criteria due to Marques and myself [38] to ensures that a map $\Phi : I^5 \to \mathcal{Z}_2(S^3)$ determines a nontrivial 5-dimensional homotopy class in $\mathcal{Z}_\in(\mathcal{S}^\ni)$.

Let $\mathcal{G}_o$ be the set of all oriented geodesic spheres in $\mathcal{Z}_2(S^3)$. Each nonzero element in $\mathcal{G}_o$ is determined by its center and radius. Thus this space is homeomorphic to $S^3 \times [-\pi, \pi]$ with an equivalence relation that identifies $S^3 \times \{-\pi\}$ and $S^3 \times \{\pi\}$ all with the zero in $\mathcal{Z}_2(S^3)$.

The maps Φ we consider have the property that

$$\Phi(I^4 \times \{1\}) = \Phi(I^4 \times \{0\}) = \{0\} \quad \text{and} \quad \Phi(I^4 \times I) \subset \mathcal{G}_o.$$

Hence $\Phi(I^5)$ can be thought of a 5-cycle in $\mathcal{Z}_2(S^3)$ whose boundary lies in $\mathcal{G}_o$ and thus $[\Phi]$ can be seen as an element of $\pi_5(\mathcal{Z}_2(S^3), \mathcal{G}_o)$. The next theorem gives a condition under which $[\Phi] \neq 0$ in $\pi_5(\mathcal{Z}_2(S^3), \mathcal{G}_o)$, i.e., the image of Φ cannot be homotoped into the set of all geodesic spheres.

Theorem 3.1. *Let $\Phi : I^5 \to \mathcal{Z}_2(S^3)$ be a continuous map such that:*

(1) $\Phi(x, 0) = \Phi(x, 1) = 0$ *for any* $x \in I^4$*;*

(2) *for any $x \in \partial I^4$ fixed we can find $Q(x) \in S^3$ such that*

$$\Phi(x, t) = \partial B_{\pi t}(Q(x)), \quad 0 \leq t \leq 1.$$

In particular, $\Phi(I^5) \subset \mathcal{G}_o$.

(3) *the center map $Q : \partial I^4 \to S^3$ has* $\deg(Q) \neq 0$.

Then

$$\mathbf{L}([\Phi]) > 4\pi = \sup_{x \in \partial I^5} \mathbf{M}(\Phi(x)).$$

Condition (3) is crucial, as the next example shows. Consider

$$\Phi : I^5 \to \mathcal{Z}_2(S^3), \quad \Phi(x, t) = \partial B_{\pi t}(p),$$

where p is a fixed point in S^3. Conditions (1) and (2) of Theorem 3.1 are satisfied but $\mathbf{L}([\Phi]) = 4\pi$ because $\Phi(I^5) \subset \mathcal{G}_o$.

Sketch of proof. The idea for the proof is, in very general terms, the following. Let $\mathcal{R} \subset \mathcal{G}_o$ denote the space of all oriented great spheres (which is homeomorphic to S^3). With this notation, note that from condition (2) we have that

$$\Phi(\partial I^5 \times \{1/2\}) \subset \mathcal{R}.$$

Moreover, the map $\Phi : \partial I^5 \times \{1/2\} \to \mathcal{R} \approx S^3$ has degree equal to $\deg(Q)$ and thus nonzero by condition (3). For simplicity, suppose we can find $\Psi \in [\Phi]$ so that

$$\mathbf{L}([\Phi]) = 4\pi = \sup_{x \in I^5} \text{area}(\Psi(x)).$$

Then $S = \Psi^{-1}(\mathcal{R})$ should be a 4-dimensional cycle in I^5 with $\partial S = \partial I^4 \times \{1/2\}$. Thus

$$\Psi_*[\partial S] = \partial[\Psi(S)] = \partial[\mathcal{R}] = 0 \text{ in } H_3(\mathcal{R}, \mathbb{Z}).$$

On the other hand, $\Phi = \Psi$ on $\partial S = \partial I^4 \times \{1/2\}$ and so

$$\Psi_*[\partial S] = \Phi_*[\partial I^4 \times \{1/2\}] = \deg(Q)[\mathcal{R}] \neq 0$$

and this is a contradiction. □

Suppose now that Φ is a map satisfying the hypothesis of Theorem 3.1. From the Min-max Theorem we obtain the existence of Σ, an embedded minimal surface, such that $\mathbf{L}([\Phi]) = \text{area}(\Sigma) > 4\pi$. Moreover, it is natural to expect that Σ has index at most 5 because we are dealing with a 5-parameter family of surfaces.

Urbano [54] in 1990 classified minimal surface of S^3 with low index and he gave a rather elegant and short proof of

Theorem 3.2 (Urbano's Theorem). *Assume S is a closed embedded minimal surface in S^3 having* $\text{index}(S) \leq 5$.

Then, up to ambient isometries, S is either a great sphere (index one) or the Clifford torus (index five) up to ambient isometries.

Remark 3.3. We already argued that the great sphere has index one. The Clifford torus has index five because unit speed normal deformations decreases area, the four parameter space of conformal dilations (to be seen later) also decrease area, and these five deformations are linearly independent.

Going back to our discussion, we see that Σ cannot be a great sphere because its area is $\mathbf{L}([\Phi]) > 4\pi$ and so it has to be a Clifford torus with area $2\pi^2$. This heuristic discussion motivates the

Theorem 3.4 ($2\pi^2$ Theorem). *Assume that Φ satisfied the hypothesis of Theorem 3.1. Then*

$$\sup_{x \in I^5} \text{area}(\Phi(x)) \geq 2\pi^2.$$

A proof can be found in [38]. Because the Almgren–Pitts theory does not provide us with the fact that the index of Σ is at most 5, we had to use a new set of arguments to prove the index estimate in the case we were interested.

4. Strategy to prove Theorem 1.1

We sketch the proof of the inequality in Theorem 1.1. The complete argument can be found in [38].

The conformal maps of S^3 (modulo isometries) can be parametrized by the open unit 4-ball B^4, where to each nonzero $v \in B^4$ we consider the conformal dilation F_v centered at $\frac{v}{|v|}$ and $-\frac{v}{|v|}$. Composing with the stereographic projection $\pi : S^3 \setminus \{-\frac{v}{|v|}\} \to \mathbb{R}^3$, the map

$$\pi \circ F_v \circ \pi^{-1} : \mathbb{R}^3 \to \mathbb{R}^3$$

corresponds to a dilation in space centered at the origin, where the dilation factor tends to infinity as $|v|$ tends to one.

Given a compact embedded surface $S \subset S^3$ and $-\pi \leq t \leq \pi$, we denote by S_t the surface at distance $|t|$ from S, where S_t lies in the exterior (interior) of S if $t \geq 0$ ($t \leq 0$). Naturally, S_t might not be a smooth embedded surface due to the existence of possible focal points but it will always be well defined in the context of Geometric Measure Theory.

We can now define the following 5-parameter family $\{\Sigma_{(v,t)}\}_{(v,t)\in B^4\times[\pi,\pi]}$ of surfaces in S^3 given by

$$\Sigma_{(v,t)} = (F_v(\Sigma))_t \in \mathcal{Z}_2(S^3).$$

One crucial property of this 5-parameter family is the following.

Theorem 4.1 (Heintze–Karcher Inequality). *For every $(v,t) \in B^4 \times [\pi,\pi]$ we have*

$$\operatorname{area}(\Sigma_{(v,t)}) \leq \mathcal{W}(\Sigma).$$

A related result was proven by Ros in [46].

In order to apply the $2\pi^2$ Theorem it is important that we understand the behaviour $\Sigma_{(v,t)}$ as (v,t) approaches the boundary of $B^4 \times [\pi,\pi]$. The fact that the diameter of S^3 is π implies that $\Sigma_{(v,\pm\pi)} = 0$ for all $v \in B^4$ and so we are left to analyze what happens when v approaches S^3.

Assume v in the 4-ball tends to $p \in S^3$. If p does not belong to Σ, then $F_v(\Sigma)$ is "pushed" into $\{-p\}$ as v tends to p and so $\operatorname{area}(F_v(\Sigma))$ tends to zero. When p lies in Σ the situation is considerably more subtle. Indeed, if v approaches p radially, i.e., $v = sp$ with $0 < s < 1$, then $F_{sp}(\Sigma)$ converges, as s tends to 1, to the unique great sphere tangent to Σ at p. Thus the continuous function in S^3 given by $p \mapsto \operatorname{area}(\Sigma_{sp})$ tends, as $s \to 1$, to a discontinuous function that is zero outside Σ and 4π along Σ. Hence, for any $0 < \alpha < 4\pi$, there must exist a sequence $\{v_i\}_{i\in\mathbb{N}}$ in B^4 tending to Σ so that $\operatorname{area}(\Sigma_{v_i})$ tends to α and so it is natural to expect that the convergence of $F_v(\Sigma)$ depends on how v approaches $p \in \Sigma$. A careful analysis revealed that, depending on the angle at which v tends to p, $F_v(\Sigma)$ tends to a geodesic sphere tangent to Σ at p, with radius and center depending on the angle of convergence.

Initially this behaviour was a source of perplexity but then we realized that, even if the parametrization was becoming discontinuous near the boundary of the parameter space, the closure of the family $\{\Sigma_{(v,t)}\}_{(v,t)\in B^4\times[\pi,\pi]}$ in $\mathcal{Z}_2(S^3)$ was a "nice" continuous 5-cycle in $\mathcal{Z}_2(S^3)$. Hence, we were able to reparametrize this family and obtain a continuous map $\Phi : I^5 \to \mathcal{Z}_2(S^3)$ with $\Phi(I^5)$ equal to the closure of $\{\Sigma_{(v,t)}\}_{(v,t)\in B^4\times[\pi,\pi]}$ in $\mathcal{Z}_2(S^3)$ and satisfying conditions (1) and (2) of Theorem 3.1.

Finally, and most important of all, we showed that the degree of the center map Q in condition (2) of Theorem 3.1 is exactly the genus of Σ. This point is absolutely crucial because it showed us that the map Φ "remembers" the genus of the surface Σ. Thus, when the genus is positive, condition (3) of Theorem 3.1 is also satisfied and we obtain from the $2\pi^2$ Theorem that

$$2\pi^2 \leq \sup_{x\in I^5} \operatorname{area}((\Phi(x)).$$

On the other hand, because $\Phi(I^5)$ is equal to the closure of $\{\Sigma_{(v,t)}\}_{(v,t)\in B^4\times[\pi,\pi]}$ in $\mathcal{Z}_2(S^3)$, we obtain from the Heintze–Karcher Inequality that

$$\sup_{x\in I^5} \operatorname{area}(\Phi(x)) \leq \mathcal{W}(\Sigma).$$

This means that $\mathcal{W}(\Sigma) \geq 2\pi^2$, which is the statement we wanted to prove.

5. Strategy to prove Theorem 1.2

The approach to prove Theorem 1.2 is similar to the one used in Theorem 1.1. The conformal invariance of the energy implies that it suffices to consider links (γ_1, γ_2) in S^3. For each link (γ_1, γ_2) in S^3 we construct a suitable family $\Phi : I^5 \to \mathcal{Z}_2(S^3)$ that satisfies conditions (1) and (2) of Theorem 3.1. Moreover, we will also show that if $|\mathrm{lk}(\gamma_1, \gamma_2)| = 1$ then condition (3) of Theorem 3.1 is will also be satisfied. Hence we can apply the $2\pi^2$ Theorem and conclude that

$$\sup_{x\in I^5} \operatorname{area}(\Phi(x)) \geq 2\pi^2.$$

On the other hand, the map Φ is constructed so that $\operatorname{area}(\Phi(x)) \leq E(\gamma_1, \gamma_2)$ for each $x \in I^5$ and this implies the inequality in Theorem 1.2.

We give a brief indication of how the map Φ is constructed.

To every pair of curves in $\mathbb{R}^4$ there is a natural way to construct a "torus" in S^3. More precisely, given two curves (γ_1, γ_2) in $\mathbb{R}^4$, the *Gauss map* is denoted by

$$G(\gamma_1, \gamma_2) : S^1 \times S^1 \to S^3, \quad (s,t) \mapsto \frac{\gamma_1(s) - \gamma_2(t)}{|\gamma_1(s) - \gamma_2(t)|}$$

and we consider $G(\gamma_1, \gamma_2)_\#(S^1 \times S^1)$ in $\mathcal{Z}_2(S^3)$. Furthermore, one can check that

$$\operatorname{area}(G(\gamma_1, \gamma_2)_\#(S^1 \times S^1)) \leq E(\gamma_1, \gamma_2).$$

For instance, if (γ_1, γ_2) is the Hopf link then $G(\gamma_1, \gamma_2)_\#(S^1 \times S^1)$ is the Clifford torus and the inequality above becomes an equality.

Given $v \in B^4$, we consider the conformal map F_v of $\mathbb{R}^4$ given by an inversion centered at v. The conformal map F_v sends the unit 4-ball B^4 to some other ball centered at $c(v) = \frac{v}{1-|v|^2}$. We consider

$$g : B^4 \times (0, +\infty) \to \mathcal{Z}_2(S^3)$$

given by

$$g(v,z) = G\left(F_v \circ \gamma_1, \lambda(F_v \circ \gamma_2 - c(v)) + c(v)\right)_\# (S^1 \times S^1).$$

Intuitively, $g(v,z)$ is the image of the Gauss map of the link obtained by applying the conformal transformation F_v to (γ_1, γ_2) and then dilating the curve $F_v \circ \gamma_2$ with respect to the center $c(v)$ by a factor of λ. Note that both curves $F_v \circ \gamma_1$ and $\lambda(F_v \circ \gamma_2 - c(v)) + c(v)$ are contained in spheres centered at $c(v)$.

The 5-parameter family we just described also enjoys a Heintze–Karcher type-inequality, meaning that for all $(v, \lambda) \in B^4 \times (0, +\infty)$ we have

$$\operatorname{area}(g(v,z)) \leq E(\gamma_1, \gamma_2).$$

The map Φ is constructed via a reparametrization of g.

6. Gromov–Guth families

In order to apply the Min-max Theorem on a general manifold M, it is important that one understands the homotopy groups of the space $\mathcal{Z}_n(M;\mathbb{Z}_2)$. This was done by Almgren [2] in 1962 and he showed that

$$\pi_1(\mathcal{Z}_n(M;\mathbb{Z}_2)) = \mathbb{Z}_2 \quad \text{and} \quad \pi_i(\mathcal{Z}_n(M;\mathbb{Z}_2)) = 0 \quad \text{if } i > 1.$$

Thus $\mathcal{Z}_n(M;\mathbb{Z}_2)$ is weakly homotopic equivalent to $\mathbb{RP}^\infty$ and so we should expect that $\mathcal{Z}_n(M,\mathbb{Z}_2)$ contains, for every $p \in \mathbb{N}$, an homotopically nontrivial p-dimensional projective space.

From the weak homotopy equivalence with $\mathbb{RP}^\infty$, we have that

$$H^k(\mathcal{Z}_n(M;\mathbb{Z}_2)) = \mathbb{Z}_2 \quad \text{for all } k \in \mathbb{N} \text{ with generator } \bar{\lambda}^k.$$

We are interested in studying maps $\Phi : X \to \mathcal{Z}_n(M;\mathbb{Z}_2)$ whose image detects $\bar{\lambda}^p$ for some $p \in \mathbb{N}$.

Given a simplicial complex X, a continuous map $\Phi : X \to \mathcal{Z}_n(M;\mathbb{Z}_2)$ is called a *p-sweepout* if $\Phi^*(\bar{\lambda}^p) \neq 0$ in $H^p(X;\mathbb{Z}_2)$. Heuristically, a continuous map $\Phi : X \to \mathcal{Z}_n(M;\mathbb{Z}_2)$ is called a *p-sweepout* if for every set $\{x_1,\dots,x_p\} \subset M$, there is $\theta \in X$ so that $\{x_1,\dots,x_p\} \subset \Phi(\theta)$.

Gromov [18–20] and Guth [21] studied p-sweepouts of M.

We now check that p-sweepouts exists for all $p \in M$. Let $f \in C^\infty(M)$ be a Morse function and consider the map

$$\Phi : \mathbb{RP}^p \to \mathcal{Z}_n(M;\mathbb{Z}_2),$$

given by

$$\Phi([a_0,\dots,a_p]) = \partial\left\{x \in M : a_0 + a_1 f(x) + \ldots + a_p f^p(x) < 0\right\}.$$

Note that the map Φ is well defined because opposite orientations on the same hypersurface determine the same element in $\mathcal{Z}_n(M;\mathbb{Z}_2)$. A typical element of $\Phi([a_0,\dots,a_p])$ will be $f^{-1}(r_1) \cup \ldots \cup f^{-1}(r_j)$ where $r_1,\dots,r_j$ are the real roots of the polynomial $p(t) = a_0 + a_1 t + \ldots + a_p t^p$. It easy to see that Φ satisfies the heuristic definition of a p-sweepout given above and in [39] we check that the map is indeed a p-sweepout.

Denoting the set of all p-sweepouts of M by $\mathcal{P}_p$, the *p-width* of M is defined as

$$\omega_p(M) = \inf_{\Phi \in \mathcal{P}_p} \sup\{\mathrm{vol}(\Phi(x)) : x \in \mathrm{dmn}(\Phi)\},$$

where $\mathrm{dmn}(\Phi)$ is the domain of Φ.

It is interesting to compare the p-width with the min-max definition of the p^{th}-eigenvalue of (M,g). Set $V = W^{1,2}(M) \setminus \{0\}$ and recall that

$$\lambda_p = \inf_{(p+1)-\text{plane } P \subset V} \max\left\{\frac{\int_M |\nabla f|^2 dV_g}{\int_M f^2 dV_g} : f \in P\right\}.$$

Hence one can see $\{\omega_p(M)\}_{p\in\mathbb{N}}$ as a nonlinear analogue of the Laplace spectrum of M, as proposed by Gromov [18].

The asymptotic behaviour of $w_p(M)$ is governed by the following result proven by Gromov [18] and Guth [21].

Theorem 6.1 (Gromov and Guth's Theorem). *There exists a positive constant $C = C(M, g)$ so that, for every $p \in \mathbb{N}$,*

$$C^{-1} p^{\frac{1}{n+1}} \leq \omega_p(M) \leq C p^{\frac{1}{n+1}}.$$

The idea to prove the lower bound is, roughly speaking, the following. Choose p disjoint geodesic balls $B_1, \dots, B_p$ with radius proportional to $p^{-\frac{1}{n+1}}$. For every p-sweepout Φ one can find $\theta \in \mathrm{dmn}(\Phi)$ so that $\Phi(\theta)$ divides each geodesic ball into two pieces with almost identical volumes. Hence, when p is sufficiently large, the isoperimetric inequality implies that $\Phi(\theta) \cap B_i$ has volume no smaller than $c(n)p^{-\frac{n}{n+1}}$ for all $i = 1, \dots, p$, where $c(n)$ is a universal constant. As a result $\Phi(\theta)$ has volume greater or equal than $c(n)p^{\frac{1}{n+1}}$.

The upper bound can be proven using a very nice bend-and-cancel argument introduced by Guth [21]. In the case when M is a $n+1$-dimensional sphere S^{n+1}, the upper bound has the following simple explanation. If we consider the set of all homogenous harmonic polynomials in S^{n+1} with degree less or equal than $d \in \mathbb{N}$, we obtain a vector space of dimension $p(d) + 1$, where $p(d)$ grows like d^{n+1}. Considering the zero set of all these polynomials we obtain a map Φ from a $p(d)$-dimensional projective plane into $\mathcal{Z}_n(S^{n+1}; \mathbb{Z}_2)$. Crofton formula implies that the zero-set of each of these polynomials has volume at most $\omega_n d$, where ω_n is the volume of an n-sphere. Thus, for every $\theta \in \mathbb{RP}^{p(d)}$, $\mathrm{vol}(\Phi(\theta))$ is at most a fixed multiple of $p(d)^{\frac{1}{n+1}}$.

7. Strategy to prove Theorem 1.3

The idea to find an infinite number of minimal surfaces consists in applying the Min-max Theorem to the family of p-sweepouts $\mathcal{P}_p$ for all $p \in \mathbb{N}$.

The first thing we show is that if $\omega_p(M) = \omega_{p+1}(M)$ for some $p \in \mathbb{N}$, then M admits an infinite number of minimal embedded hypersurfaces. We achieve this using Lusternick-Schnirelman and, roughly speaking, the idea is as follows (for details see [39]):

Suppose for simplicity that $\omega_p(M) = \omega_{p+1}(M) = \sup_{x \in X} \mathrm{vol}(\Phi(x))$ for some $p+1$-sweepout Φ. We argue by contradiction and assume that Ω, the set of all embedded minimal hypersurfaces (with possible multiplicities) and volume at most $\omega_{p+1}(M)$, is finite.

Let $K = \Psi^{-1}(\Omega)$ and $\lambda = \Phi^*(\bar{\lambda})$, where $\bar{\lambda}$ generates $H^1(\mathcal{Z}_n(M; \mathbb{Z}_2); \mathbb{Z}_2)$.

We must have λ vanishing on K because otherwise there would exist a curve γ in K so that $\lambda(\gamma) = \bar{\lambda}(\Phi \circ \gamma) \neq 0$, i.e. $\Phi \circ \gamma$ would be a nontrivial element of $\pi_1(\mathcal{Z}_n(M; \mathbb{Z}_2))$. But $\Phi \circ \gamma$ has image contained in the finite set Ω, which implies it is constant, and thus contractible.

Therefore λ^p cannot vanish on $X \setminus K$ because otherwise $\lambda^{p+1} = \lambda^p \smile \lambda$ would be zero on $(X \setminus K) \cup K = X$ and this is impossible because Φ is a $p+1$-sweepout. As a result $\Phi_{|X \setminus K}$ is a p-sweepout whose image contains no minimal hypersurfaces (with possible multiplicities), and so we pull-tight the family to obtain another p-sweepout Ψ so that

$$\omega_p(M) \leq \sup_{x \in X \setminus K} \mathrm{vol}(\Psi(x)) < \sup_{x \in X \setminus K} \mathrm{vol}(\Phi(x)) \leq \sup_{x \in X} \mathrm{vol}(\Phi(x)) = \omega_p(M).$$

This gives us the desired contradiction.

Hence we can assume that the sequence $\{\omega_p(M)\}_{p \in \mathbb{N}}$ is strictly increasing.

We then argue again by contradiction and assume that there exist only finitely many smooth, closed, embedded minimal hypersurfaces with multiplicity one, and we call this set

Λ. Using the Min-max Theorem we have that

$$\omega_p(M) = n_{p,1}\mathrm{vol}(\Sigma_{p,1}) + \ldots + n_{p,k}\mathrm{vol}(\Sigma_{p,k}), \quad n_{p,1}, \ldots, n_{p,k} \in \mathbb{N},$$

where $\{\Sigma_{p,1}, \ldots, \Sigma_{p,k}\}$ are multiplicity one minimal hypersurfaces with disjoint support. Because M has positive Ricci curvature, Frankel's Theorem [13] says that any two minimal embedded hypersurfaces intersect and so $\omega_p(M) = n_p\mathrm{vol}(\Sigma_p)$ for some $\Sigma_p \in \Lambda$ and $n_p \in \mathbb{N}$. The fact that $\omega_p(M)$ is strictly increasing and a counting argument shows that Λ being finite implies that $\omega_p(M)$ must grow linearly in p. This is in contradiction with the sublinear growth of $\omega_p(M)$ in p given by the Gromov and Guth's Theorem.

8. Open problems

Min-max Theory is an exciting technique which I think can be used not only to solve other open questions in Geometry but also to provide some new directions. Some of these questions are well-known and others arose from extensive discussions with Fernando Marques.

Min-max questions. The Almgren-Pitts Min-max Theory does not provide index estimates for the min-max minimal hypersurface. The importance of this issue was already clear to Almgren [3] who wrote

> *"The chief utility of the homology approach would lie in the attempt to assign a topological index to stationary integral varifolds in some analytically useful way."*

A folklore conjecture states that the if the homotopy class in the Min-max Theorem is defined with k-parameters, then the minimal hypersurface given by the Min-max Theorem has index at most k. It is implicit in the conjecture that one finds a meaningful way of assigning an index to a minimal embedded hypersurface with multiplicities. When $k = 1$, the conjecture was confirmed by Marques and myself if the ambient manifold is three dimensional and the metric has positive Ricci curvature [39]. Later this was extended to the case where the ambient manifold has dimension between three and seven and the metric has positive Ricci curvature [60].

Naively, one should also expect that for bumpy metrics the index is bounded from below by the number of parameters.

Many of the subtle issues in Min-max Theory are related with the fact that the min-max minimal hypersurface can have multiplicities. That said, one does not know an example where the width of some homotopy class is realized by an unstable minimal hypersurface with multiplicity. For instance, does the equator with multiplicity two (and so area 8π) realizes the width of some homotopy class in the round 3-sphere? It is highly conceivable that unstable minimal surfaces with higher multiplicity can be approximated (in the varifold norm) by a sequence of embedded minimal surfaces with smaller area and this is one of the reasons the question is interesting.

Old questions. We now mention four well known open problems which could be answered using Min-max Theory.

The first two are natural generalizations of the Willmore conjecture.

The Willmore conjecture in S^4 states that among all tori in S^4, the Clifford torus minimizes the Willmore energy. It is interesting that in this case there are minimal embedded projective planes (Veronese surface) which have area (6π) smaller than the Clifford torus and bigger than the equator.

For higher genus surfaces, Kusner [31] conjectured that the Lawson minimal surface $\xi_{1,g}$ minimizes the Willmore energy among all surfaces of genus g (numerical evidence was provided in [25]). It would be extremely interesting to find the index of $\xi_{1,2}$. Wishful thinking would suggest 9 but there is no real evidence.

The third problem consists of finding, among all non-totally geodesic minimal hypersurfaces in the unit n-sphere S^n, the one with least possible volume. The conjecture, due to Solomon, is that these minimal hypersurface are given by

$$S^{m-1}\left(\sqrt{\frac{m-1}{2m-1}}\right)\times S^m\left(\sqrt{\frac{m}{2m-1}}\right)\subset S^n$$

if $n=2m$, and by

$$S^{m-1}\left(\frac{1}{\sqrt{2}}\right)\times S^{m-1}\left(\frac{1}{\sqrt{2}}\right)\subset S^n$$

if $n=2m-1$. In S^3 this conjecture was confirmed by Marques and myself in [38] and in the general case there was some progress due to White and Ilmanen [26].

It would be desirable to have a sharp index characterization similar Urbano's Theorem for each of these three problems. In the third problem, Perdomo [45] achieved that assuming the hypersurfaces are preserved by the antipodal map.

The fourth and final problem is a beautiful conjecture of White [56] which says that any metric on a 3-sphere has five distinct minimal embedded tori and he proved this for small perturbations of the round metric [57]. An easier conjecture would be to say that any metric on a 3-sphere has nine distinct minimal surfaces of genus either zero or one.

Some new questions. For the purpose of applications in Geometry and Topology, it is important to estimate the topology of the minimal hypersurface given by the Min-max Theorem.

For ambient 3-manifolds, due to the combined work of Simon–Smith [49], De Lellis–Pellandini [10], and Ketover [29], it is now known that, roughly speaking, if the Min-max technique is applied to continuous one-parameter family of embedded surfaces of genus g, then the min-max minimal surface has at most genus g.

In higher dimensions it is not so clear how to control the topology of the min-max minimal hypersurface by the same methods.

An alternative approach would be to try to characterize the topology of the min-max minimal hypersurface via its index. Note that if the minimal hypersurfaces are produced via Min-max methods then one should expect some control on the index.

For ambient 3-manifolds, Ejiri–Micallef [12] showed that the index of a minimal orientable surface is bounded from above by a multiple of area plus genus and if the metric has positive Ricci curvature then it is known [59] that index one orientable minimal surfaces have genus 3 at most (a conjecture that I heard from Rick Schoen states that the genus should be two at most).

For higher dimensions, Rick Schoen conjectured that index one embedded orientable compact minimal hypersurfaces in ambient manifolds with positive Ricci curvature have

bounded first Betti number. We conjecture that if the ambient manifold has positive Ricci curvature then an index k embedded orientable compact minimal hypersurface has first Betti number bounded by fixed multiple of k. Savo [47] showed this holds on round spheres of any dimension.

Another direction of research would be to understand the p-widths $\omega_p(M)$ of (M^{n+1}, g) for $n \geq 2$. The sequence $\{\omega_p(M)\}_{p\in\mathbb{N}}$ can be thought of as a nonlinear spectrum for the manifold and we would expect it to be asymptotically related with the spectrum of the Laplacian. Taking this perspective, many interesting questions arise.

For instance, does the nonlinear spectrum satisfy a Weyl Law? More precisely, can we find a universal constant $a(n)$ so that

$$\lim_{p\to\infty} \omega_p(M) p^{-\frac{1}{n+1}} = a(n)(\mathrm{vol}(M,g))^{\frac{n}{n+1}}?$$

This question has been suggested by Gromov in [19, Section 8] and in [20, Section 5.2]. A classical result of Uhlenbeck [53] states that generic metrics have simple eigenvalues. Likewise, we expect that for generic metrics at least, $\omega_p(M)$ is achieved by a multiplicity one minimal hypersurface with index p. Can we say anything about how they look like? For instance, are they becoming equidistributed in space? The proof of Gromov-Guth's Theorem suggests that. Do they behave like nodal sets of eigenfunctions? Is their first betti number proportional to p?

Nodal sets of eigenfunctions provide a natural upper bound for $\omega_p(M)$. Making this more precise, denote by $\phi_0, \ldots, \phi_p$ the first $(p+1)$-eigenfunctions for the Laplace operator, where ϕ_0. Consider

$$\Phi_p : \mathbb{RP}^p \to \mathcal{Z}_n(M; \mathbb{Z}_2),$$
$$\Phi_p([a_0, \ldots, a_p]) = \partial\{x \in M : a_0\phi_0(x) + \ldots + a_p\phi_p(x) < 0\}.$$

and set

$$\bar{\omega}_p(M) = \sup_{\theta\in\mathbb{RP}^p} \mathrm{vol}(\Phi_p(\theta)) \geq \omega_p(M).$$

It seems a challenging question to determine how close to one $\frac{\bar{\omega}_p(M)}{\omega_p(M)}$ is getting as p tends to infinity. If the quotient is bounded, that would imply a conjecture of Yau regarding the asymptotic growth of the volume of nodal sets (which was proven in the analytic case by Donnelly and Fefferman [11] and for recent progress see [9, 50]). Can we determine that quotient on an n-sphere?

Acknowledgements. The author was partly supported by Marie Curie IRG Grant and ERC Start Grant.

References

[1] I. Agol, F. C. Marques, and A. Neves, *Min-max theory and the energy of links*, arXiv:1205.0825 [math.GT] (2012), 1–19.

[2] F. Almgren, *The homotopy groups of the integral cycle groups,* Topology (1962), 257–299.

[3] ———, *The theory of varifolds*, Mimeographed notes, Princeton, 1965.

[4] W. Ballmann, *Der Satz von Lusternik und Schnirelmann,* (German) Beiträge zur Differentialgeometrie, Heft 1, pp. 1–25, Bonner Math. Schriften, 102, Univ. Bonn, Bonn, 1978.

[5] V. Bangert, *On the existence of closed geodesics on two-spheres,* Internat. J. Math. **4** (1993), 1–10.

[6] G. D. Birkhoff, *Dynamical systems with two degrees of freedom,* Trans. Amer. Math. Soc. **18** (1917), no. 2, 199–300.

[7] W. Blaschke, *Vorlesungen Über Differentialgeometrie III,* Berlin: Springer, 1929.

[8] T. Colding and C. De Lellis, *The min-max construction of minimal surfaces,* Surveys in Differential Geometry VIII , International Press, (2003), 75–107.

[9] T. H. Colding and W. P. Minicozzi II, *Lower bounds for nodal sets of eigenfunctions,* Comm. Math. Phys. **306** (2011), no. 3, 777–784.

[10] C. De Lellis and F. Pellandini, *Genus bounds for minimal surfaces arising from min-max constructions,* J. Reine Angew. Math. **644** (2010), 47–99.

[11] H. Donnelly and C. Fefferman, *Nodal sets of eigenfunctions on Riemannian manifolds,* Invent. Math. **93** (1988), 161–183.

[12] N. Ejiri and M. Micallef, *Comparison between second variation of area and second variation of energy of a minimal surface*, Adv. Calc. Var. 1 (2008), 223–239.

[13] T. Frankel, *On the fundamental group of a compact minimal submanifold,* Ann. of Math. **83** (1966), 68–73.

[14] J. Franks, *Geodesics on S^2 and periodic points of annulus homeomorphisms,* Invent. Math. **108** (1992), 403–418.

[15] M. Freedman, Z-X. He, and Z. Wang, *Möbius energy of knots and unknots,* Ann. of Math. (2) **139** (1994), no. 1, 1–50.

[16] S. Germain, *Recherches sur la théorie des surfaces élastiques,* Paris, 1921.

[17] M. Grayson, *Shortening embedded curves,* Ann. Math. **120** (1989) 71–112.

[18] M. Gromov, *Dimension, nonlinear spectra and width,* Geometric aspects of functional analysis, (1986/87), 132–184, Lecture Notes in Math., 1317, Springer, Berlin, 1988.

[19] ———, *Isoperimetry of waists and concentration of maps,* Geom. Funct. Anal. **13** (2003), 178–215.

[20] ———, *Singularities, expanders and topology of maps. I. Homology versus volume in the spaces of cycles*, Geom. Funct. Anal. **19** (2009), 743–841.

[21] L. Guth, *Minimax problems related to cup powers and Steenrod squares,* Geom. Funct. Anal. **18** (2009), 1917–1987.

[22] Zheng-Xu He, *On the minimizers of the Möbius cross energy of links*, Experiment. Math. **11** (2002), no. 2, 244–248.

[23] W. Helfrich, *Elastic properties of lipid bilayers: Theory and possible experiments*, Z. Naturforsch. **28** (1973), 693–703.

[24] N. Hingston, *On the growth of the number of closed geodesics on the two-sphere,* Internat. Math. Res. Notices (1993), 253–262.

[25] L. Hsu, R. Kusner, and J. Sullivan, *Minimizing the squared mean curvature integral for surfaces in space forms,* Experiment. Math. **1** (1992), 191–207.

[26] T. lmanen and B. White, *Sharp Lower Bounds on Density of Area-Minimizing Cones*, preprint.

[27] J. Jost, *A nonparametric proof of the theorem of Lusternik and Schnirelman,* Arch. Math. (Basel) **53** (1989), 497–509.

[28] J. Kahn and V. Marković, *Counting essential surfaces in a closed hyperbolic three-manifold,* Geom. Topol. 16 (2012), 601–624.

[29] D. Ketover, *Degeneration of min-max sequences in 3-manifolds,* arXiv:1312.2666 [math.DG] (2013).

[30] W. Klingenberg, *Lectures on closed geodesics,* Grundlehren der Mathematischen Wissenschaften, Vol. 230. Springer-Verlag, Berlin-New York, 1978.

[31] R. Kusner, *Estimates for the biharmonic energy on unbounded planar domains, and the existence of surfaces of every genus that minimize the squared-mean-curvature integral,* Elliptic and parabolic methods in geometry, A K Peters, 67–72, (1996).

[32] B. Lawson, *Complete minimal surfaces in S^3,* Ann. of Math. (2) **92** (1970), 335–374.

[33] P. Li and S-T. Yau, *A new conformal invariant and its applications to the Willmore conjecture and the first eigenvalue of compact surfaces,* Invent. Math. **69** (1982), 269–291.

[34] L. Lusternik, *Topology of functional spaces and calculus of variations in the large,* Trav. Inst. Math. Stekloff **19** (1947).

[35] L. Lusternik and L. Schnirelmann, *Topological methods in variational problems and their application to the differential geometry of surfaces*, Uspehi Matem. Nauk (N.S.) **2**, (1947), 166–217.

[36] F. C. Marques, *Minimal surfaces - variational theory and applications*, Proceedings of the ICM, Seoul, Korea (2014).

[37] F. C. Marques and A. Neves, *Rigidity of min-max minimal spheres in three-manifolds*, Duke Math. J. **161** (2012), no. 14, 2725–2752.

[38] _____, *Min-max theory and the Willmore conjecture*, Ann. of Math. (2) **179** (2014), 683–782.

[39] F. C. Marques and A. Neves, *Existence of infinitely many minimal hypersurfaces in positive Ricci curvature*, preprint.

[40] M. Mutz and D. Bensimon, *Observation of toroidal vesicles,* Phys. Rev. A, **43** (1991), 4525–4527.

[41] H. Poincaré, *Sur les lignes godsiques des surfaces convexes,* Trans. Amer. Math. Soc. **6** (1905), no. 3, 237–274.

[42] J. Pitts, *Existence and regularity of minimal surfaces on Riemannian manifolds,* Mathematical Notes 27, Princeton University Press, Princeton, 1981.

[43] S. D. Poisson, *Mémoire sur les surfaces élastiques,* Mem. Cl. Sci. Math. Phys., Inst. de France (1812), 167–225.

[44] J. O'Hara, *Energy of a knot,* Topology **30** (1991), 241–247.

[45] O. Perdomo, *Low index minimal hypersurfaces of spheres,* Asian J. Math. **5** (2001), 741–749.

[46] A. Ros, *The Willmore conjecture in the real projective space,* Math. Res. Lett. **6** (1999), 487–493.

[47] A. Savo, *Index bounds for minimal hypersurfaces of the sphere,* Indiana Univ. Math. J. **59** (2010), 823–837.

[48] R. Schoen and L. Simon, *Regularity of stable minimal hypersurfaces,* Comm. Pure Appl. Math. **34** (1981), 741–797.

[49] F. Smith, *On the existence of embedded minimal 2-spheres in the 3-sphere, endowed with an arbitrary Riemannian metric*, PhD thesis supervised by L. Simon, University of Melbourne (1982).

[50] C. D. Sogge and S. Zelditch, *Lower bounds on the Hausdorff measure of nodal sets,* Math. Res. Lett. **18** (2011), no. 1, 25–37.

[51] I. Taimanov, *On the existence of three nonintersecting closed geodesics on manifolds that are homeomorphic to the two-dimensional sphere,* (Russian) Izv. Ross. Akad. Nauk Ser. Mat. **56** (1992), 605–635.

[52] G. Thomsen, *Über Konforme Geometrie, I: Grundlagen der Konformen Flächentheorie,* Abh. Math. Sem. Hamburg (1923), 31–56.

[53] K. Uhlenbeck, *Generic properties of eigenfunctions,* Amer. J. Math. **98** (1976), 1059–1078.

[54] F. Urbano, *Minimal surfaces with low index in the three-dimensional sphere,* Proc. Amer. Math. Soc. **108** (1990), 989–992.

[55] T. J. Willmore, *Note on embedded surfaces,* An. Sti. Univ. "Al. I. Cuza" Iasi Sect. I a Mat. (N.S.) **11B** (1965) 493–496.

[56] B. White, *Every three-sphere of positive Ricci curvature contains a minimal embedded torus,* Bull. Amer. Math. Soc. (N.S.) **21** (1989), 71–75.

[57] ———, *The space of minimal submanifolds for varying Riemannian metrics,* Indiana Univ. Math. J. 40 (1991), 161–200.

[58] S.-T. Yau, *Problem section,* Seminar on Differential Geometry, pp. 669–706, Ann. of Math. Stud., 102, Princeton Univ. Press, Princeton, N.J., 1982.

[59] ———, *Nonlinear analysis in geometry,* Enseign. Math. **33** (1987), 109–158.

[60] X. Zhou, *Min-max minimal hypersurface in* (M^{n+1}, g) *with* $Ric_g > 0$ *and* $2 \leq n \leq 6$, arXiv:1210.2112v2 [math.DG] (2012).

Imperial College, Huxley Building, 180 Queen's Gate, London SW7 2RH, United Kingdom
E-mail: a.neves@imperial.ac.uk

When symplectic topology meets Banach space geometry

Yaron Ostrover

Abstract. In this paper we survey some recent works that take the first steps toward establishing bilateral connections between symplectic geometry and several other fields, namely, asymptotic geometric analysis, classical convex geometry, and the theory of normed spaces.

Mathematics Subject Classification (2010). 53D35, 52A23, 52A40, 37D50, 57S05.

Keywords. Symplectic capacities, Viterbo's volume-capacity conjecture, Mahler's conjecture, Hamiltonian diffeomorphisms, Hofer's metric.

1. Introduction

In the last three decades, symplectic topology has had an astonishing amount of fruitful interactions with other fields of mathematics, including complex and algebraic geometry, dynamical systems, Hamiltonian PDEs, transformation groups, and low-dimensional topology; as well as with physics, where, for example, symplectic topology plays a key role in the rigorous formulation of mirror symmetry.

In this survey paper, we present some recent works that take first steps toward establishing novel interrelations between symplectic geometry and several fields of mathematics, namely, asymptotic geometric analysis, classical convex geometry, and the theory of normed spaces. In the first part of this paper (Sections 2 and 3) we concentrate on the theory of symplectic measurements, which arose from the foundational work of Gromov [34] on pseudoholomorphic curves; followed by the seminal works of Ekeland and Hofer [24] and Hofer and Zehnder [42] on variational theory in Hamiltonian systems, and Viterbo on generating functions [89]. This theory – also known as the theory of "symplectic capacities" – lies nowadays at the core of symplectic geometry and topology.

In Section 2, we focus on an open symplectic isoperimetric-type conjecture proposed by Viterbo in [88]. It states that among all convex domains with a given volume in the classical phase space $\mathbb{R}^{2n}$, the Euclidean ball has the maximal "symplectic size" (see Section 2 below for the precise statement). In a collaboration with S. Artstein-Avidan and V. D. Milman [6], we were able to prove an asymptotic version of Viterbo's conjecture, that is, we proved the conjecture up to a universal (dimension-independent) constant. This has been achieved by adapting techniques from asymptotic geometric analysis and adjusting them to a symplectic context, while working exclusively in the linear symplectic category.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

The fact that one can get within a constant factor to the full conjecture using only linear embeddings is somewhat surprising from the symplectic-geometric point of view, as in symplectic geometry one typically needs highly nonlinear tools to estimate capacities. However, this fits perfectly into the philosophy of asymptotic geometric analysis. Finding dimension independent estimates is a frequent goal in this field, where surprising phenomena such as concentration of measure (see e.g. [67]) imply the existence of order and structures in high dimensions, despite the huge complexity it involves. It would be interesting to explore whether similar phenomena also exist in the framework of symplectic geometry. A natural important source for the study of the asymptotic behavior (in the dimension) of symplectic invariants is the field of statistical mechanics, where one considers systems with a large number of particles, and the dimension of the phase space is twice the number of degrees of freedom. It seems that symplectic measurements were overlooked in this context so far.

In Section 3 we go in the opposite direction: we show how symplectic geometry could potentially be used to tackle a 70-years-old fascinating open question in convex geometry, known as the Mahler conjecture. Roughly speaking, Mahler's conjecture states that the minimum of the product of the volume of a centrally symmetric convex body and the volume of its polar body is attained (not uniquely) for the hypercube. In a collaboration with S. Artstein–Avidan and R. Karasev [8], we combined tools from symplectic geometry, classical convex analysis, and the theory of mathematical billiards, and established a close relation between Mahler's conjecture and the above mentioned symplectic isoperimetric conjecture by Viterbo. More preciesly, we showed that Mahler's conjecture is equivalent to a special case of Viterbo's conjecture (see Section 3 for details).

In the second part of the paper (Section 4), we explain how methods from functional analysis can be used to address questions regarding the geometry of the group $\mathrm{Ham}(M,\omega)$ of Hamiltonian diffeomorphisms associated with a symplectic manifold (M,ω). One of the most striking facts regarding this group, discovered by Hofer in [40], is that it carries an intrinsic geometry given by a Finsler bi-invariant metric, nowadays known as Hofer's metric. This metric measures the time-averaged minimal oscillation of a Hamiltonian function that is needed to generate a Hamiltonian diffeomorphism starting from the identity. Hofer's metric has been intensively studied in the past twenty years, leading to many discoveries covering a wide range of subjects from Hamiltonian dynamics to symplectic topology (see e.g., [43, 59, 75] and the references therein). A long-standing question raised by Eliashberg and Polterovich in [26] is whether Hofer's metric is the only bi-invariant Finsler metric on the group $\mathrm{Ham}(M,\omega)$. Together with L. Buhovsky [17], and based on previous results by Ostrover and Wagner [72], we used methods from functional analysis and the theory of normed function spaces to affirmatively answer this question. We proved that any non-degenerate bi-invariant Finsler metric on $\mathrm{Ham}(M,\omega)$, which is generated by a norm that is continuous in the C^∞-topology, gives rise to the same topology on $\mathrm{Ham}(M,\omega)$ as the one induced by Hofer's metric.

As mentioned before, the outlined interdisciplinary connections described above are just the first few steps in what seems to be a promising new direction. We hope that further exploration of these connections will strengthen the dialogue between these fields and symplectic geometry, and expand the range of methodologies alongside research questions that can be tackled through these means.

We end this paper with several open questions and speculations regarding some of the mentioned topics (see Section 5).

2. A symplectic isoperimetric inequality

A classical result in symplectic geometry (Darboux's theorem) states that symplectic manifolds - in a sharp contrast to Riemannian manifolds - have no local invariants (except, of course, the dimension). The first examples of global symplectic invariants were introduced by Gromov in his seminal paper [34], where he developed and used pseudoholomorphic curve techniques to prove a striking symplectic rigidity result. Nowadays known as Gromov's "non-squeezing theorem", this result states that one cannot map a ball inside a thinner cylinder by a symplectic embedding. This theorem paved the way to the introduction of global symplectic invariants, called symplectic capacities which, roughly speaking, measure the symplectic size of a set.

We will focus here on the case of the classical phase space $\mathbb{R}^{2n} \simeq \mathbb{C}^n$ equipped with the standard symplectic structure $\omega = dq \wedge dp$. We denote by $B^{2n}(r)$ the Euclidean ball of radius r, and by $Z^{2n}(r)$ the cylinder $B^2(r) \times \mathbb{C}^{n-1}$. Gromov's non-squeezing theorem asserts that if $r < 1$ there is no symplectomorphism ψ of $\mathbb{R}^{2n}$ such that $\psi(B^{2n}(1)) \subset Z^{2n}(r)$. The following definition, which crystallizes the notion of "symplectic size", was given by Ekeland and Hofer in their influential paper [24].

Definition 2.1. A symplectic capacity on $(\mathbb{R}^{2n}, \omega)$ associates to each subset $U \subset \mathbb{R}^{2n}$ a number $c(U) \in [0, \infty]$ such that the following three properties hold:

(P1) $c(U) \leq c(V)$ for $U \subseteq V$ (monotonicity);

(P2) $c\big(\psi(U)\big) = |\alpha|\, c(U)$ for $\psi \in \mathrm{Diff}(\mathbb{R}^{2n})$ such that $\psi^*\omega = \alpha\omega$ (conformality);

(P3) $c\big(B^{2n}(r)\big) = c\big(Z^{2n}(r)\big) = \pi r^2$ (nontriviality and normalization).

Note that (P3) disqualifies any volume-related invariant, while (P1) and (P2) imply that for $U, V \subset \mathbb{R}^{2n}$, a necessary condition for the existence of a symplectomorphism ψ with $\psi(U) = V$, is $c(U) = c(V)$ for any symplectic capacity c.

It is a priori unclear that symplectic capacities exist. The above mentioned non-squeezing result naturally leads to the definition of two symplectic capacities: the Gromov radius, defined by $\underline{c}(U) = \sup\{\pi r^2 \,|\, B^{2n}(r) \overset{\mathrm{s}}{\hookrightarrow} U\}$; and the cylindrical capacity, defined by $\overline{c}(U) = \inf\{\pi r^2 \,|\, U \overset{\mathrm{s}}{\hookrightarrow} Z^{2n}(r)\}$, where $\overset{\mathrm{s}}{\hookrightarrow}$ stands for symplectic embedding. It is easy to verify that these two capacities are the smallest and largest possible symplectic capacities, respectively. Moreover, it is also known that the existence of a single capacity readily implies Gromov's non-squeezing theorem, as well as the Eliashberg-Gromov C^0-rigidity theorem, which states that for any closed symplectic manifold (M, ω), the symplectomorphism group $\mathrm{Symp}(M, \omega)$ is C^0-closed in the group of all diffeomorphisms of M (see e.g., Chapter 2 of [43]).

Shortly after Gromov's work, other symplectic capacities were constructed, such as the Hofer-Zehnder [43] and the Ekeland-Hofer [24] capacities, the displacement energy [40], the Floer-Hofer capacity [27, 28], spectral capacities [29, 70, 89], and, more recently, Hutchings's embedded contact homology capacities [44]. Nowadays, symplectic capacities are among the most fundamental objects in symplectic geometry, and are the subject of intensive research efforts (see e.g., [45, 47, 52, 55–57, 60, 63, 82], and [20] for a recent detailed survey and more references). However, in spite of the rapidly accumulating knowledge regarding symplectic capacities, they are notoriously difficult to compute, and there are no general methods even to effectively estimate them.

In [88], Viterbo investigated the relation between the symplectic way of measuring the size of sets using symplectic capacities, and the classical approach using volume. Among many other inspiring results, in that work he conjectured that in the class of convex bodies in $\mathbb{R}^{2n}$ with fixed volume, the Euclidean ball B^{2n} maximizes any given symplectic capacity. More precisely,

Conjecture 2.2 (Viterbo's volume-capacity inequality conjecture). *For any convex body K in $\mathbb{R}^{2n}$ and any symplectic capacity c,*

$$\frac{c(K)}{c(B)} \leq \left(\frac{\mathrm{Vol}(K)}{\mathrm{Vol}(B)}\right)^{1/n}, \text{ where } B = B^{2n}(1).$$

Here and henceforth a convex body of $\mathbb{R}^{2n}$ is a compact convex set with non-empty interior. The isoperimetric inequality above was proved in [88] up to a constant that depends linearly on the dimension using the classical John ellipsoid theorem. In a joint work with S. Artstein-Avidan and V. D. Milman (see [6]), we made further progress towards the proof of the conjecture. By customizing methods and techniques from asymptotic geometric analysis and adjusting them to the symplectic context, we were able to prove Viterbo's conjecture up to a universal (i.e., dimension-independent) constant. More precisely, we proved that

Theorem 2.3. *There is a universal constant A such that for any convex domain K in $\mathbb{R}^{2n}$, and any symplectic capacity c, one has*

$$\frac{c(K)}{c(B)} \leq A \left(\frac{\mathrm{Vol}(K)}{\mathrm{Vol}(B)}\right)^{1/n}, \text{ where } B = B^{2n}(1).$$

We emphasize that in the proof of Theorem 2.3 we work exclusively in the category of linear symplectic geometry. It turns out that even in this limited category of linear symplectic transformations, there are tools which are powerful enough to obtain a dimension-independent estimate as above. While this fits with the philosophy of asymptotic geometric analysis, it is less expected from a symplectic geometry point of view, where one expects that highly nonlinear methods, such as folding and wrapping techniques (see e.g., the book [82]), would be required to effectively estimate symplectic capacities.

The proof of Theorem 2.3 above is based on two ingredients. The first is the following simple geometric observation (see Lemma 3.3 in [6], cf. [1]).

Lemma 2.4. *If a convex body $K \subset \mathbb{C}^n$ satisfies $K = iK$, then $\overline{c}(K) \leq \frac{4}{\pi}\, \underline{c}(K)$.*

Sketch of Proof. Let rB^{2n} be the largest multiple of the unit ball contained in K, and let $x \in \partial K \cap rS^{2n-1}$ be a contact point between the boundary of K and the boundary of rB^{2n}. It follows from the convexity assumption that the body K lies between the hyperplanes $x + x^{\perp}$ and $-x + x^{\perp}$. Moreover, since $K = iK$, it lies also between $-ix + ix^{\perp}$ and $ix + ix^{\perp}$. Thus, the projection of K onto the plane spanned by x and ix is contained in a square of edge length $2r$. This square can be turned into a disc with area $4r^2$, after applying a non-linear symplectomorphism which is essentially two-dimensional. Therefore, K is contained in a symplectic image of the cylinder $Z^{2n}(\sqrt{4/\pi}\, r)$, and the lemma follows. □

Since by monotonicity, Conjecture 2.2 trivially holds for the Gromov radius $\underline{c}$, it follows from Lemma 2.4 that

Corollary 2.5. *Theorem 2.3 holds for convex bodies $K \subset \mathbb{C}^n$ such that $K = iK$.*

The second ingredient in the proof is a profound result in asymptotic geometric analysis discovered by V.D. Milman in the mid 1980's called the "reverse Brunn-Minkowski inequality" (see [65, 66]). Recall that the classical Brunn-Minkowski inequality states that if A and B are non-empty Borel subsets of $\mathbb{R}^n$, then

$$\mathrm{Vol}(A+B)^{1/n} \geq \mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^{1/n},$$

where $A + B = \{x + y \,|\, x \in A,\, y \in B\}$ is the Minkowski sum. Although at first glance it seems that one cannot expect any inequality in the reverse direction (consider, e.g., two very long and thin ellipsoids pointing in orthogonal directions in $\mathbb{R}^2$), it turns out that for convex bodies, if one allows for an extra choice of "position", i.e., a volume-preserving linear image of the bodies, then one can reverse the Brunn-Minkowski inequality up to a universal constant factor.

Theorem 2.6 (Milman's reverse Brunn-Minkowski inequality). *For any two convex bodies K_1, K_2 in $\mathbb{R}^n$, there exist linear volume preserving transformations T_{K_i} $(i = 1, 2)$, such that for $\widetilde{K}_i = T_{K_i}(K_i)$ one has*

$$\mathrm{Vol}(\widetilde{K}_1 + \widetilde{K}_2)^{1/n} \leq C\left(\mathrm{Vol}(\widetilde{K}_1)^{1/n} + \mathrm{Vol}(\widetilde{K}_2)^{1/n}\right),$$

for some absolute constant C.

We emphasize that the transformation $T_{K_i}(i = 1, 2)$ in Theorem 2.6 depends solely on the body K_i, and not on the joint configuration of the bodies K_1 and K_2. For more details on the reverse Brunn-Minkowski inequality see [66, 75].

We can now sketch the proof of Theorem 2.3 (for more details see [6]). Since every symplectic capacity is bounded above by the cylindrical capacity $\overline{c}$, it is enough to prove the theorem for $\overline{c}$. For the sake of simplicity, we assume in what follows that K is centrally symmetric, i.e., $K = -K$. This assumption is not too restrictive, since by a classical result of Rogers and Shephard [79] one has that $\mathrm{Vol}(K + (-K)) \leq 4^n \mathrm{Vol}(K)$. After adjusting Theorem 2.6 to the symplectic context, one has that for any convex body $K \subset \mathbb{R}^{2n}$, there exists a linear symplectomorphism $S \in \mathrm{Sp}(2n)$ such that SK and iSK satisfy the reverse Brunn-Minkowski inequality, that is, the volume $\mathrm{Vol}(SK + iSK)^{1/n}$ is less than some constant times $\mathrm{Vol}(K)^{1/n}$. Combining this with the properties of symplectic capacities and Corollary 2.5, we conclude that

$$\frac{\overline{c}(K)}{\overline{c}(B)} \leq \frac{\overline{c}(SK+iSK)}{\overline{c}(B)} \leq A\left(\frac{\mathrm{Vol}(SK+iSK)}{\mathrm{Vol}(B)}\right)^{\frac{1}{n}} \leq A'\left(\frac{\mathrm{Vol}(K)}{\mathrm{Vol}(B)}\right)^{\frac{1}{n}},$$

for some universal constant A', and thus Theorem 2.3 follows.

In the next section we will show a surprising connection between Viterbo's volume-capacity conjecture and a seemingly remote open conjecture from the field of convex geometric analysis: the Mahler conjecture on the volume product of centrally symmetric convex bodies.

3. A symplectic view on Mahler's conjecture

Let $(X, \|\cdot\|)$ be an n-dimensional normed space and let $(X^*, \|\cdot\|^*)$ be its dual space. Note that the product space $X \times X^*$ carries a canonical symplectic structure, given by the skew-symmetric bilinear form $\omega\big((x,\xi),(x',\xi')\big) = \xi(x') - \xi'(x)$, and a canonical volume form, the *Liouville* volume, given by $\omega^n/n!$. A fundamental question in the field of convex geometry, raised by Mahler in [58], is to find upper and lower bounds for the Liouville volume of $B \times B^\circ \subset X \times X^*$, where B and B° are the unit balls of X and X^*, respectively. In what follows we shall denote this volume by $\nu(X)$. The quantity $\nu(X)$ is an affine invariant of X, i.e. it is invariant under invertible linear transformations. We remark that in the context of convex geometry $\nu(X)$ is also known as the *Mahler volume* or the *volume product* of X.

The Blaschke-Santaló inequality asserts that the maximum of $\nu(X)$ is attained if and only if X is a Euclidean space. This was proved by Blaschke [14] for dimensions two and three, and generalized by Santaló [81] to higher dimensions. The following sharp lower bound for $\nu(X)$ was conjectured by Mahler [58] in 1939:

Conjecture 3.1 (Mahler's volume product conjecture). *For any n-dimensional normed space X one has $\nu(X) \geq 4^n/n!$.*

The conjecture has been verified by Mahler [58] in the two-dimensional case. In higher dimensions it is proved only in a few special cases (see e.g., [33, 49, 64, 69, 76–78, 80, 86]). A major breakthrough towards answering Mahler's conjecture is a result due to Bourgain and Milman [16], who used sophisticated tools from functional analysis to show that the conjecture holds asymptotically, i.e., up to a factor γ^n, where γ is a universal constant. This result has been re-proved later on, with entirely different methods, by Kuperberg [51], using differential geometry, and independently by Nazarov [68], using the theory of functions of several complex variables. A new proof using simpler asymptotic geometric analysis tools has been recently discovered by Giannopoulos, Paouris, and Vritsiou [32]. The best known constant today, $\gamma = \pi/4$, is due to Kuperberg [51].

Despite great efforts to deal with the general case, a proof of Mahler's conjecture has been insistently elusive so far, and is currently the subject of intensive research. A possible reason for this, as pointed out for example by Tao in [87], is that, in contrast with the above mentioned Blaschke-Santaló inequality, the equality case in Mahler's conjecture, which is obtained for example for the space l_∞^n of bounded sequences with the standard maximum norm, is not unique, and there are in fact many distinct extremizers for the (conjecturally) lower bound of $\nu(X)$ (see, e.g., the discussion in [87]). This practically renders impossible any proof based on currently known optimisation techniques, and a radically different approach seems to be needed.

We refer the reader to Section 5 below for further discussion on the characterization of the equality case of Mahler's conjecture, and its possible connection with symplectic geometry.

In a recent work with S. Artstein-Avidan and R. Karasev [8], we combined tools from symplectic geometry, convex analysis, and the theory of mathematical billiards, and established a close relationship between Mahler's conjecture and Viterbo's volume-capacity conjecture. More precisely, we proved in [8] that

Theorem 3.2. *Viterbo's volume-capacity conjecture implies Mahler's conjecture.*

In fact, it follows from our proof that Mahler's conjecture is equivalent to a special case of

Viterbo's conjecture, where the latter is restricted to the Ekeland-Hofer-Zehnder symplectic capacity, and to domains in the classical phase space of the form $\Sigma \times \Sigma^\circ \subset \mathbb{R}^{2n} = \mathbb{R}^n_q \times \mathbb{R}^n_p$ (for more details see [8], and in particular Remark 1.9 ibid.). Here, $\Sigma \subset \mathbb{R}^n_q$ is a centrally symmetric convex body, the space $\mathbb{R}^n_p$ is identified with the dual space $(\mathbb{R}^n_q)^*$, and

$$\Sigma^\circ = \{p \in \mathbb{R}^n_p \,|\, p(q) \leq 1 \text{ for every } q \in \Sigma\}$$

Theorem 3.2 is a direct consequence of the following result proven in [8].

Theorem 3.3. *There exists a symplectic capacity c such that $c(\Sigma \times \Sigma^\circ) = 4$ for every centrally symmetric convex body $\Sigma \subset \mathbb{R}^n_q$.*

With Theorem 3.3 at our disposal, it is not difficult to derive Theorem 3.2.

Proof of Theorem 3.2. Assume that Viterbo's volume-capacity conjecture holds. From Theorem 3.3 it follows that there exists a symplectic capacity c such that for every centrally symmetric convex body $\Sigma \subset \mathbb{R}^n_q$ one has

$$\frac{4^n}{\pi^n} = \frac{c^n(\Sigma \times \Sigma^\circ)}{\pi^n} \leq \frac{\mathrm{Vol}(\Sigma \times \Sigma^\circ)}{\mathrm{Vol}(B^{2n})} = \frac{n!\,\mathrm{Vol}(\Sigma \times \Sigma^\circ)}{\pi^n},$$

which is exactly the bound for $\mathrm{Vol}(\Sigma \times \Sigma^\circ)$ required by Mahler's conjecture. □

In the rest of this section we sketch the proof of Theorem 3.3 (see [8] for a detailed exposition). We remark that an alternative proof, based on an approach to billiard dynamics developed in [11], was recently given in [3]. We start with recalling the definition of the Ekeland-Hofer-Zehnder capacity, which is the symplectic capacity that appears in Theorem 3.3.

The restriction of the standard symplectic form $\omega = dq \wedge dp$ to a smooth closed connected hypersurface $\mathcal{S} \subset \mathbb{R}^{2n}$ defines a 1-dimensional subbundle $\ker(\omega|\mathcal{S})$, whose integral curves comprise the characteristic foliation of $\mathcal{S}$. In other words, a *closed characteristic* of $\mathcal{S}$ is an embedded circle in $\mathcal{S}$ tangent to the canonical line bundle

$$\mathfrak{S}_\mathcal{S} = \{(x, \xi) \in T\mathcal{S} \,|\, \omega(\xi, \eta) = 0 \text{ for all } \eta \in T_x\mathcal{S}\}.$$

Recall that the symplectic action of a closed curve γ is defined by $A(\gamma) = \int_\gamma \lambda$, where $\lambda = pdq$ is the Liouville 1-form. The action spectrum of $\mathcal{S}$ is

$$\mathcal{L}(\mathcal{S}) = \{\,|\,A(\gamma)\,|\,;\ \gamma \text{ closed characteristic on } \mathcal{S}\}\,.$$

The following theorem, which is a combination of results from [24] and [43], states that on the class of convex domains in $\mathbb{R}^{2n}$, the Ekeland-Hofer capacity c_{EH} and Hofer-Zehnder capacity c_{HZ} coincide, and are given by the minimal action over all closed characteristics on the boundary of the corresponding convex body.

Theorem 3.4. *Let $K \subseteq \mathbb{R}^{2n}$ be a convex bounded domain with smooth boundary. Then there exists at least one closed characteristic $\widetilde{\gamma} \subset \partial K$ satisfying*

$$c_{\mathrm{EH}}(K) = c_{\mathrm{HZ}}(K) = A(\widetilde{\gamma}) = \min \mathcal{L}(\partial K).$$

We remark that although the above definition of closed characteristics, as well as Theorem 3.4, were given only for the class of convex bodies with smooth boundary, they can naturally be generalized to the class of convex sets in $\mathbb{R}^{2n}$ with non-empty interior (see [7]). In what follows, we refer to the coinciding Ekeland-Hofer and Hofer-Zehnder capacities on this class as the Ekeland-Hofer-Zehnder capacity.

We turn now to show that for every centrally symmetric convex body $\Sigma \subset \mathbb{R}^n_q$, the Ekeland–Hofer–Zehnder capacity satisfies $c_{\rm EHZ}(\Sigma \times \Sigma^\circ) = 4$. For this purpose, we now switch gears and turn to mathematical billiards in Minkowski geometry.

It is folklore to people in the field that billiard flow can be treated, roughly speaking, as the limiting case of geodesic flow on a boundaryless manifold. Indeed, let Ω be a smooth plane billiard table, and consider its "thickening", i.e. an infinitely thin three dimensional body whose boundary Γ is obtained by pasting two copies of Ω along their boundaries and smoothing the edge. Thus, a billiard trajectory in Ω can be viewed as a geodesic line on the boundary of Γ, that goes from one copy of Ω to another each time the billiard ball bounces off the boundary. The main technical difficulties with this strategy is the rigorous treatment of the limiting process, and the analysis involved with the dynamics near the boundary. One approach to billiard dynamics and the existence question of periodic trajectories is an approximation scheme which uses a certain "penalization method" developed by Benci and Giannoni in [10] (cf. [5, 48]). In what follows we present an alternative approach, and use characteristic foliation on singular convex hypersurfaces in $\mathbb{R}^{2n}$ (see e.g., [21, 23, 50]) to describe Finsler type billiards for convex domains in the configuration space $\mathbb{R}^n_q$. The main advantage of this approach is that it allows one to use the natural one-to-one correspondence between the geodesic flow on a manifold and the characteristic foliation on its unit cotangent bundle, and thus provides a natural "symplectic setup" in which one can use tools such as Theorem 3.4 above in the context of billiard dynamics. In particular, we show that the Ekeland-Hofer-Zehnder capacity of certain Lagrangian product configurations $\mathcal{K} \times \mathcal{T}$ in the classical phase space $\mathbb{R}^{2n}$ is the length of the shortest periodic *$\mathcal{T}$-billiard trajectory in $\mathcal{K}$* (see e.g., [7, 88]), which we turn now to describe.

The general study of billiard dynamics in Finsler and Minkowski geometries was initiated by Gutkin and Tabachnikov in [36]. From the point of view of geometric optics, Minkowski billiard trajectories describe the propagation of light in a homogeneous anisotropic medium that contains perfectly reflecting mirrors. Below, we focus on the special case of Minkowski billiards in a smooth convex body $\mathcal{K} \subset \mathbb{R}^n_q$. We equip $\mathcal{K}$ with a metric given by a certain norm $\|\cdot\|$, and consider billiards in $\mathcal{K}$ with respect to the geometry induced by $\|\cdot\|$. More precisely, let $\mathcal{K} \subset \mathbb{R}^n_q$, and $\mathcal{T} \subset \mathbb{R}^n_p$ be two convex bodies with smooth boundary, and consider the unit cotangent bundle

$$U^*_{\mathcal{T}}\mathcal{K} := \mathcal{K} \times \mathcal{T} = \{(q,p) \,|\, q \in \mathcal{K}, \text{ and } g_{\mathcal{T}}(p) \leq 1\} \subset T^*\mathbb{R}^n_q = \mathbb{R}^n_q \times \mathbb{R}^n_p.$$

Here $g_{\mathcal{T}}$ is the gauge function $g_{\mathcal{T}}(x) = \inf\{r \,|\, x \in r\mathcal{T}\}$. When $\mathcal{T} = -\mathcal{T}$ is centrally symmetric one has $g_{\mathcal{T}}(x) = \|x\|_{\mathcal{T}}$. For $p \in \partial\mathcal{T}$, the gradient vector $\nabla g_{\mathcal{T}}(p)$ is the outer normal to $\partial\mathcal{T}$ at the point p, and is naturally considered to be in $\mathbb{R}^n_q = (\mathbb{R}^n_p)^*$.

Motivated by the classical correspondence between geodesics in a Riemannian manifold and characteristics of its unit cotangent bundle, we define $(\mathcal{K}, \mathcal{T})$-billiard trajectories to be characteristics in $U^*_{\mathcal{T}}\mathcal{K}$ such that their projections to $\mathbb{R}^n_q$ are closed billiard trajectories in $\mathcal{K}$ with a bouncing rule that is determined by the geometry induced from the body $\mathcal{T}$; and vice versa, the projections to $\mathbb{R}^n_p$ are closed billiard trajectories in $\mathcal{T}$ with a bouncing rule that is determined by $\mathcal{K}$. More precisely, when we follow the vector fields of the dynamics, we

move in $\mathcal{K} \times \partial\mathcal{T}$ from (q_0, p_0) to $(q_1, p_0) \in \partial\mathcal{K} \times \partial\mathcal{T}$ following the inner normal to $\partial\mathcal{T}$ at p_0. When we hit the boundary $\partial\mathcal{K}$ at the point q_1, the vector field changes, and we start to move in $\partial\mathcal{K} \times \mathcal{T}$ from (q_1, p_0) to $(q_1, p_1) \in \partial\mathcal{K} \times \partial\mathcal{T}$ following the outer normal to $\partial\mathcal{K}$ at the point q_1. Next, we move from (q_1, p_1) to (q_2, p_1) following the opposite of the normal to $\partial\mathcal{T}$ at p_1, and so on and so forth (see Figure 1). It is not hard to check that when one of the bodies, say $\mathcal{T}$, is a Euclidean ball, then when considering the projection to $\mathbb{R}^n_q$, the bouncing rule described above is the classical one (i.e., equal impact and reflection angles). Hence, the above reflection law is a natural variation of the classical one when the Euclidean structure on $\mathbb{R}^n_q$ is replaced by the metric induced by the norm $\|\cdot\|_{\mathcal{T}}$. We continue with a more precise definition.

Definition 3.5. Given two smooth convex bodies $\mathcal{K} \subset \mathbb{R}^n_q$ and $\mathcal{T} \subset \mathbb{R}^n_p$. A closed $(\mathcal{K}, \mathcal{T})$-billiard trajectory is the image of a piecewise smooth map $\gamma\colon S^1 \to \partial(\mathcal{K} \times \mathcal{T})$ such that for every $t \notin \mathcal{B}_\gamma := \{t \in S^1 \mid \gamma(t) \in \partial\mathcal{K} \times \partial\mathcal{T}\}$ one has

$$\dot{\gamma}(t) = d\,\mathfrak{X}(\gamma(t)),$$

for some positive constant d and the vector field $\mathfrak{X}$ given by

$$\mathfrak{X}(q,p) = \begin{cases} (-\nabla g_{\mathcal{T}}(p), 0), & (q,p) \in int(\mathcal{K}) \times \partial\mathcal{T}, \\ (0, \nabla g_{\mathcal{K}}(q)), & (q,p) \in \partial\mathcal{K} \times int(\mathcal{T}). \end{cases}$$

Moreover, for any $t \in \mathcal{B}_\gamma$, the left and right derivatives of $\gamma(t)$ exist, and

$$\dot{\gamma}^{\pm}(t) \in \{\alpha(-\nabla g_{\mathcal{T}}(p), 0) + \beta(0, \nabla g_{\mathcal{K}}(q)) \mid \alpha, \beta \geq 0,\ (\alpha, \beta) \neq (0,0)\}.$$

Although in Definition 3.5 there is a natural symmetry between the bodies $\mathcal{K}$ and $\mathcal{T}$, in what follows we shall assume that $\mathcal{K}$ plays the role of the billiard table, while $\mathcal{T}$ induces the geometry that governs the billiard dynamics in $\mathcal{K}$. We will use the following terminology: for a $(\mathcal{K}, \mathcal{T})$-billiard trajectory γ, the curve $\pi_q(\gamma)$, where $\pi_q\colon \mathbb{R}^{2n} \to \mathbb{R}^n_q$ is the projection of γ to the configuration space, shall be called a *$\mathcal{T}$-billiard trajectory in $\mathcal{K}$*. Moreover, similarly to the Euclidean case, one can check that $\mathcal{T}$-billiard trajectories in $\mathcal{K}$ correspond to critical points of a length functional defined on the j-fold cross product of the boundary $\partial\mathcal{K}$, where the distances between two consecutive points are measured with respect to the support function $h_{\mathcal{T}}$, where $h_{\mathcal{T}}(u) = \sup\{\langle x, u\rangle\,;\, x \in \mathcal{T}\}$.

Definition 3.6. A closed $(\mathcal{K}, \mathcal{T})$-billiard trajectory γ is said to be *proper* if the set $\mathcal{B}_\gamma$ is finite, i.e., γ is a broken bicharacteristic that enters and instantly exits the boundary $\partial\mathcal{K} \times \partial\mathcal{T}$ at the reflection points. In the case where $\mathcal{B}_\gamma = S^1$, i.e., γ is travelling solely along the boundary $\partial\mathcal{K} \times \partial\mathcal{T}$, we say that γ is a *gliding trajectory*.

The following theorem was proved in [7].

Theorem 3.7. *Let $\mathcal{K} \subset \mathbb{R}^n_q$, $\mathcal{T} \subset \mathbb{R}^n_p$ be two smooth convex bodies. Then, every $(\mathcal{K}, \mathcal{T})$-billiard trajectory is either a proper trajectory, or a gliding one. Moreover the Ekeland-Hofer-Zehnder capacity $c_{\mathrm{EHZ}}(\mathcal{K} \times \mathcal{T})$ of the Lagrangian product $\mathcal{K} \times \mathcal{T}$ is the length of the shortest periodic $\mathcal{T}$-billiard trajectory in $\mathcal{K}$ measured with respect to the support function $h_{\mathcal{T}}$.*

This theorem provides an effective way to estimate (and sometimes compute) the Ekeland-Hofer-Zehnder capacity of Lagrangian product configurations in the phase space. For example, in [8] (see Remark 4.2 therein) we used elementary tools from convex geometry to show

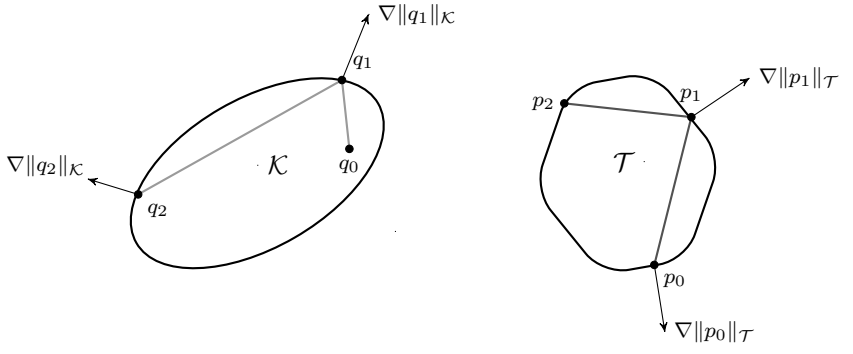


Figure 3.1. A proper $(\mathcal{K},\mathcal{T})$-Billiard trajectory.

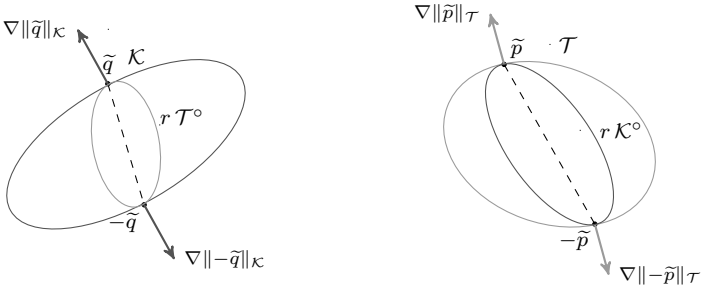


Figure 3.2. *$\mathcal{T}$-billiard trajectory in $\mathcal{K}$ of length* $4\,\mathrm{inrad}_{\mathcal{T}^\circ}(\mathcal{K})$.

that for centrally symmetric convex bodies, the shortest $\mathcal{T}$-billiard trajectory in $\mathcal{K}$ is a 2-periodic trajectory connecting a tangency point q_0 of $\mathcal{K}$ and a homotetic copy of $\mathcal{T}^\circ$ to $-q_0$ (see Figure 2). This result extends a previous result by Ghomi [31] for Euclidean billiards. In both cases, the main difficulty in the proof is to show that the above mentioned 2-periodic trajectory is indeed the shortest one. With this geometric observation at our disposal, we proved in [8] the following result: denote by $\mathrm{inrad}_{\mathcal{T}}(\mathcal{K}) = \max\{r \,|\, r\mathcal{T} \subset \mathcal{K}\}$.

Theorem 3.8. *If $\mathcal{K} \subset \mathbb{R}^n_q$, $\mathcal{T} \subset \mathbb{R}^n_p$ are centrally symmetric convex bodies, then*

$$c_{\mathrm{EHZ}}(\mathcal{K}\times\mathcal{T}) = \overline{c}(\mathcal{K}\times\mathcal{T}) = 4\,\mathrm{inrad}_{\mathcal{T}^\circ}(\mathcal{K})$$

Note that Theorem 3.8 immediately implies Theorem 3.3 above, which in turn implies Theorem 3.2. Thus, we have shown that Mahler's conjecture follows from a special case of Viterbo's conjecture. In fact, it follows immediately from the proof of Theorem 3.2 that Mahler's conjecture is equivalent to Viterbo's conjecture when the latter is restricted to the Ekeland-Hofer-Zehnder capacity, and to convex domains of the form $\Sigma\times\Sigma^\circ$, where $\Sigma \subset \mathbb{R}^n_q$ is a centrally symmetric convex body. We hope that further pursuing this line of research will lead to a breakthrough in understanding both conjectures.

3.1. Bounds on the length of the shortest billiard trajectory. Going somehow in the opposite direction, one can also use the theory of symplectic capacities to provide several bounds and inequalities for the length of the shortest periodic billiard trajectory in a smooth convex body in $\mathbb{R}^n$. In [7] we prove the following theorem, which for the sake of simplicity we state only for the case of Euclidean billiards (for several other related results see [3, 5, 11, 31, 47, 48, 88]).

Theorem 3.9. *Let $K \subset \mathbb{R}^n$ be a smooth convex body, and let $\xi(K)$ denote the length of the shortest periodic billiard trajectory in K. Then,*

(i) $\xi(K_1) \leq \xi(K_2)$*, for any convex domains $K_1 \subseteq K_2 \subseteq \mathbb{R}^n$ (monotonicity);*

(ii) $\xi(K) \leq C\sqrt{n}\,\mathrm{Vol}(K)^{\frac{1}{n}}$*, for some universal constant $C > 0$;*

(iii) $4\mathrm{inrad}(K) \leq \xi(K) \leq 2(n+1)\mathrm{inrad}(K)$*;*

(iv) $\xi(K_1 + K_2) \geq \xi(K_1) + \xi(K_2)$ *(Brunn-Minkowski type inequality).*

We remark that the inequality $4\mathrm{inrad}(K) \leq \xi(K)$ in (iii) above was proved already in [31], the monotonicity property was well known to experts in the field (although it has not been addressed in the literature to the best of our knowledge), and all the results in Theorem 3.9 were later recovered and generalized by different methods (see [3, 47, 48]). Moreover, in light of the "classical versus quantum" relation between the length spectrum in Riemannian geometry and the Laplace spectrum, via trace formulae and Poisson relations, Theorem 3.9 can be viewed as a classical counterpart of some well-known results for the first Laplace eigenvalue on convex domains. It is interesting to note that, to the best of the author's knowledge, the exact value of the constant C in part (ii) of Theorem 3.9 is unknown already in the two-dimensional case.

4. The Uniqueness of Hofer's Metric

One of the most striking facts regarding the group of Hamiltonian diffeomorphisms associated with a symplectic manifold is that it can be equipped with an intrinsic geometry given by a bi-invariant Finsler metric known as Hofer's metric [40]. In contrast to the case of finite-dimensional Lie groups, the existence of such a metric on an infinite-dimensional group of transformations is highly unusual due to the lack of local compactness. Hofer's metric is exceptionally important for at least two reasons: first, Hofer showed in [40] that this metric gives rise to an important symplectic capacity known as "displacement energy", which turns out to have many different applications in symplectic topology and Hamiltonian dynamics (see e.g., [18, 40, 43, 52, 53, 74, 75]). Second, it provides a certain geometric intuition for the understanding of the long-time behaviour of Hamiltonian dynamical systems.

In [26], Eliashberg and Polterovich initiated a discussion on the uniqueness of Hofer's metric (cf. [25, 75]). They asked whether for a closed symplectic manifold (M, ω), Hofer's metric is the only bi-invariant Finsler metric on the group of Hamiltonian diffeomorphisms. In this section we explain (following [17] and [72]) how tools from classical functional analysis and the theory of normed function spaces can be used to positively answer this question, and show that up to equivalence of metrics, Hofer's metric is unique. For this purpose, we now turn to more precise formulations.

Let (M,ω) be a closed $2n$-dimensional symplectic manifold, and denote by $C_0^\infty(M)$ the space of smooth functions that are zero-mean normalized with respect to the canonical volume form ω^n. With every smooth time-dependent Hamiltonian function $H : M\times[0,1] \to \mathbb{R}$, one associates a vector field X_{H_t} via the equation $i_{X_{H_t}}\omega = -dH_t$, where $H_t(x) = H(t,x)$. The flow of X_{H_t} is denoted by ϕ_H^t and is defined for all $t \in [0,1]$. The group of Hamiltonian diffeomorphisms consists of all the time-one maps of such Hamiltonian flows, i.e.,

$$\mathrm{Ham}(M,\omega) = \{\phi_H^1 \mid \phi_H^t \text{ is a Hamiltonian flow}\}.$$

When equipped with the standard C^∞-topology, the group $\mathrm{Ham}(M,\omega)$ is an infinite-dimensional Fréchet Lie group. Its Lie algebra, denoted here by $\mathcal{A}$, can be naturally identified with the space of normalized smooth functions $C_0^\infty(M)$. Moreover, the adjoint action of $\mathrm{Ham}(M,\omega)$ on $\mathcal{A}$ is the standard action of diffeomorphisms on functions, i.e., $\mathrm{Ad}_\phi f = f\circ\phi^{-1}$, for every $f \in \mathcal{A}$ and $\phi \in \mathrm{Ham}(M,\omega)$. For more details on the group of Hamiltonian diffeomorphisms see e.g., [43, 62, 75].

Next, we define a Finsler pseudo-distance on $\mathrm{Ham}(M,\omega)$. Given any pseudo-norm $\|\cdot\|$ on $\mathcal{A}$, we define the length of a path $\alpha : [0,1] \to \mathrm{Ham}(M,\omega)$ as

$$\mathrm{length}\{\alpha\} = \int_0^1 \|\dot\alpha\| dt = \int_0^1 \|H_t\| dt,$$

where $H_t(x) = H(t,x)$ is the unique normalized Hamiltonian function generating the path α. Here H is said to be normalized if $\int_M H_t\omega^n = 0$ for every $t \in [0,1]$. The distance between two Hamiltonian diffeomorphisms is given by

$$d(\psi,\varphi) := \inf \mathrm{length}\{\alpha\},$$

where the infimum is taken over all Hamiltonian paths α connecting ψ and φ. It is not hard to check that d is non-negative, symmetric, and satisfies the triangle inequality. Moreover, any pseudo-norm on the Lie algebra $\mathcal{A}$ that is invariant under the adjoint action yields a bi-invariant pseudo-distance function on $\mathrm{Ham}(M,\omega)$, i.e., $d(\psi,\phi) = d(\theta\,\psi,\theta\,\phi) = d(\psi\,\theta,\phi\,\theta)$, for every $\psi,\phi,\theta \in \mathrm{Ham}(M,\omega)$.

From here forth we deal solely with such pseudo-norms and we refer to d as the pseudo-distance generated by the pseudo-norm $\|\cdot\|$.

We remark in passing that a fruitful study of right-invariant Finsler metrics on $\mathrm{Ham}(M,\omega)$, motivated in part by applications to hydrodynamics, was initiated by Arnold [4]. In addition, non-Finslerian bi-invariant metrics on $\mathrm{Ham}(M,\omega)$ have been intensively studied in the realm of symplectic geometry, starting with the works of Viterbo [89], Schwarz [84], and Oh [70], and followed by many others.

Remark 4.1. When one studies geometric properties of the group of Hamiltonian diffeomorphisms, it is convenient to consider smooth paths $[0,1] \to \mathrm{Ham}(M,\omega)$, among which those that start at the identity correspond to smooth Hamiltonian flows. Moreover, for a given Finsler pseudo-metric on $\mathrm{Ham}(M,\omega)$, a natural geometric assumption is that every smooth path $[0,1] \to \mathrm{Ham}(M,\omega)$ has finite length. As it turns out, the latter assumption is equivalent to the continuity of the pseudo-norm on $\mathcal{A}$ corresponding to the pseudo-Finsler metric in the C^∞-topology (see [17]). Thus, in what follows we shall mainly consider such pseudo-norms.

It is highly non-trivial to check whether a distance function on the group of Hamiltonian diffeomorphisms generated by a pseudo-norm is non-degenerate, that is, $d(\mathrm{Id}, \phi) > 0$ for $\phi \neq \mathrm{Id}$. In fact, for closed symplectic manifolds, a bi-invariant pseudo-metric d on $\mathrm{Ham}(M, \omega)$ is either a genuine metric or identically zero. This is an immediate corollary of a well-known theorem by Banyaga [9], which states that $\mathrm{Ham}(M, \omega)$ is a simple group, combined with the fact that the null-set

$$\mathrm{null}(d) = \{\phi \in \mathrm{Ham}(M, \omega) \mid d(\mathrm{Id}, \phi) = 0\}$$

is a normal subgroup of $\mathrm{Ham}(M, \omega)$. A renowned result by Hofer [40] states that the L_∞-norm on $\mathcal{A}$ gives rise to a genuine distance function on $\mathrm{Ham}(M, \omega)$ known now as Hofer's metric. This was proved by Hofer for the case of $\mathbb{R}^{2n}$, then generalized by Polterovich [74], and finally proven in full generality by Lalonde and McDuff [53]. In a sharp contrast to the above, Eliashberg and Polterovich showed in [26] that for a closed symplectic manifold (M, ω) ons has

Theorem 4.2 (Eliashberg and Polterovich). *For $1 \leq p < \infty$, the pseudo-distances on $\mathrm{Ham}(M, \omega)$ corresponding to the L_p-norms on $\mathcal{A}$ vanish identically.*

The following question was asked in [26] (cf. [25, 75]):

Question 4.3. *What are the $\mathrm{Ham}(M, \omega)$-invariant norms on $\mathcal{A}$, and which of them give rise to genuine bi-invariant metrics on $\mathrm{Ham}(M, \omega)$?*

It was observed in [17] that any pseudo-norm $\|\cdot\|$ on the space $\mathcal{A}$ can be turned into a $\mathrm{Ham}(M, \omega)$-invariant pseudo-norm via a certain invariantization procedure $\|f\| \mapsto \|f\|_{\mathrm{inv}}$. The idea behind this procedure is based on the notion of infimal convolution (or epi-sum), from convex analysis. Recall that the infimal convolution of two functions f and g on $\mathbb{R}^n$ is defined by $(f \square g)(z) = \inf\{f(x) + g(y) \mid z = x + y\}$. This operator has a simple geometric interpretation: the epigraph (i.e., the set of points lying on or above the graph) of the infimal convolution of two functions is the Minkowski sum of the epigraphs of those functions. The invariantization $\|\cdot\|_{\mathrm{inv}}$ of $\|\cdot\|$ is obtained by taking the orbit of $\|\cdot\|$ under the group action, and consider the infimal convolution of the associated family of norms. More precisely, define

$$\|f\|_{\mathrm{inv}} = \inf\Big\{\sum \|\phi_i^* f_i\| \ ; \ f = \sum f_i, \text{ and } \phi_i \in \mathrm{Ham}(M, \omega)\Big\}.$$

We remark that in the above definition of $\|f\|_{\mathrm{inv}}$ the sum $\sum f_i$ is assumed to be finite. Note that $\|\cdot\|_{\mathrm{inv}} \leq \|\cdot\|$. Thus, if $\|\cdot\|$ is continuous in the C^∞-topology, then so is $\|\cdot\|_{\mathrm{inv}}$. Moreover, if $\|\cdot\|'$ is a $\mathrm{Ham}(M, \omega)$-invariant pseudo-norm, then:

$$\|\cdot\|' \leq \|\cdot\| \Longrightarrow \|\cdot\|' \leq \|\cdot\|_{\mathrm{inv}}.$$

In particular, the above invariantization procedure provides a plethora of $\mathrm{Ham}(M, \omega)$-invariant genuine norms on $\mathcal{A}$, e.g., by applying it to the $\|\cdot\|_{C^k}$-norms.

In [72] we made a first step toward answering Question 4.3 using tools from the theory of normed spaces and functional analysis. More precisely, regarding the first part of Question 4.3, we proved

Theorem 4.4 (Ostrover and Wagner). *Let $\|\cdot\|$ be a $\mathrm{Ham}(M, \omega)$-invariant norm on $\mathcal{A}$ such that $\|\cdot\| \leq C\|\cdot\|_\infty$ for some constant C. Then $\|\cdot\|$ is invariant under all measure preserving diffeomorphisms of M.*

In other words, any $\mathrm{Ham}(M,\omega)$-invariant norm on $\mathcal{A}$ that is bounded above by the L_∞-norm, must also be invariant under the much larger group of measure preserving diffeomorphisms. Theorem 4.4 plays an important role in the proof of the following result, which gives a partial answer to the second part of Question 4.3.

Theorem 4.5 (Ostrover and Wagner). *Let $\|\cdot\|$ be a $\mathrm{Ham}(M,\omega)$-invariant norm on $\mathcal{A}$ such that $\|\cdot\| \leq C\|\cdot\|_\infty$ for some constant C, but the two norms are not equivalent.*[1] *Then the associated pseudo-distance d on $\mathrm{Ham}(M,\omega)$ vanishes identically.*

Although Theorem 4.5 gives a partial answer to the second part of Question 4.3, prima facie, there might be $\mathrm{Ham}(M,\omega)$-invariant norms on $\mathcal{A}$ which are either strictly bigger than the L_∞-norm, or incomparable to it. In a joint work with L. Buhovsky [17] we showed that under the natural continuity assumption mentioned in Remark 4.1 above, this cannot happen. Hence, up to equivalence of metrics, Hofer's metric is unique. More precisely,

Theorem 4.6 (Buhovsky and Ostrover). *Let (M,ω) be a closed symplectic manifold. Any C^∞-continuous $\mathrm{Ham}(M,\omega)$-invariant pseudo-norm $\|\cdot\|$ on $\mathcal{A}$ is dominated from above by the L_∞-norm i.e., $\|\cdot\| \leq C\|\cdot\|_\infty$ for some constant C.*

Combining Theorem 4.6 and Theorem 4.5 above, we obtain:

Corollary 4.7. *For a closed symplectic manifold (M,ω), any bi-invariant Finsler pseudo-metric on $\mathrm{Ham}(M,\omega)$, obtained by a pseudo-norm $\|\cdot\|$ on $\mathcal{A}$ that is continuous in the C^∞-topology, is either identically zero, or equivalent to Hofer's metric. In particular, any non-degenerate bi-invariant Finsler metric on $\mathrm{Ham}(M,\omega)$ which is generated by a norm that is continuous in the C^∞-topology gives rise to the same topology on $\mathrm{Ham}(M,\omega)$ as the one induced by Hofer's metric.*

In the rest of this section we briefly describe the strategy of the proof of Theorem 4.6 in the two-dimensional case. For the proof of the general case see [17]. We start with two straightforward reduction steps. First, for technical reasons, we shall consider pseudo-norms on the space $C^\infty(M)$, instead of the space $\mathcal{A}$. The original claim will follow, since any $\mathrm{Ham}(M,\omega)$ invariant pseudo-norm $\|\cdot\|$ on $\mathcal{A}$ can be naturally extended to an invariant pseudo-norm $\|\cdot\|'$ on $C^\infty(M)$ by

$$\|f\|' := \|f - M_f\|, \text{ where } M_f = \tfrac{1}{\mathrm{Vol(M)}}\int_M f\omega^n.$$

Note that if $\|\cdot\|$ is continuous in the C^∞-topology, then so is $\|\cdot\|'$, and that the two norms coincide on the space $\mathcal{A}$. Second, by using a standard partition of unity argument, we can reduce the proof of Theorem 4.6 to a "local result", i.e., it is sufficient to prove the theorem for $\mathrm{Ham}_c(W,\omega)$-invariant pseudo-norms on the space of compactly supported smooth functions $C_c^\infty(W)$, where $W = (-L,L)^2$ is an open square in $\mathbb{R}^2$ (see [17] for the details).

The next step, which is one of the key ideas of the proof, is to define the "largest possible" $\mathrm{Ham}_c(W,\omega)$-invariant norm on the space of compactly supported smooth functions $C_c^\infty(W)$. To this end, we fix a (non-empty) finite collection of functions $\mathcal{F} \subset C_c^\infty(W)$, and define:

$$\mathcal{L}_{\mathcal{F}} := \Big\{\sum_{i,k} c_{i,k}\Phi_{i,k}^* f_i \,|\, c_{i,k} \in \mathbb{R}, \Phi_{i,k} \in \mathrm{Ham}_c(W,\omega), f_i \in \mathcal{F}, \text{ and } \#\{(i,k) \,|\, c_{i,k} \neq 0\} < \infty\Big\}.$$

[1] Two norms are said to be equivalent if $\frac{1}{C}\|\cdot\|_1 \leqslant \|\cdot\|_2 \leqslant C\|\cdot\|_1$ for some constant $C > 0$.

We equip the space $\mathcal{L}_{\mathcal{F}}$ with the norm

$$\|f\|_{\mathcal{L}_{\mathcal{F}}} = \inf \sum |c_{i,k}|,$$

where the infimum is taken over all the representations $f = \sum c_{i,k}\, \Phi_{i,k}^* f_i$ as above.

Definition 4.8. For any compactly supported function $f \in C_c^\infty(W)$, let

$$\|f\|_{\mathcal{F},\max} = \inf \big\{ \liminf_{i\to\infty} \|f_i\|_{\mathcal{L}_{\mathcal{F}}} \big\},$$

where the infimum is taken over all subsequences $\{f_i\}$ in $\mathcal{L}_{\mathcal{F}}$ which converge to f in the C^∞-topology. As usual, the infimum of the empty set is set to be $+\infty$.

The main feature of the norm $\|\cdot\|_{\mathcal{F},\max}$ is that it dominates from above any other $\mathrm{Ham}_c(W,\omega)$-invariant pseudo-norm that is continuous in the C^∞-topology.

Lemma 4.9. *Let $\mathcal{F} \subset C_c^\infty(W)$ be a non-empty finite collection of smooth compactly supported functions in W. Then any $\mathrm{Ham}_c(W,\omega)$-invariant pseudo-norm $\|\cdot\|$ on $C_c^\infty(W)$ that is continuous in the C^∞-topology satisfies*

$$\|\cdot\| \leqslant C\|\cdot\|_{\mathcal{F},\max},$$

for some absolute constant C.

Proof of Lemma 4.9. Since the collection $\mathcal{F}$ is finite, set $C = \max\{\|g\|;\, g \in \mathcal{F}\}$. For any $f = \sum c_{i,k}\, \Phi_{i,k}^* f_i \in \mathcal{L}_{\mathcal{F}}$, one has

$$\|f\| \leq \sum |c_{i,k}| \|\Phi_{i,k}^* f_i\| \leq C \sum |c_{i,k}|. \tag{4.1}$$

By the definition of $\|\cdot\|_{\mathcal{L}_{\mathcal{F}}}$, this immediately implies that $\|f\| \leq C\|f\|_{\mathcal{L}_{\mathcal{F}}}$. The lemma now follows by combining (4.1), the definition of $\|\cdot\|_{\mathcal{F},\max}$, and the fact that the pseudo-norm $\|\cdot\|$ is assumed to be continuous in the C^∞-topology. □

The next step, which is the main part of the proof, is to show that for a suitable collection of functions $\mathcal{F} \subset C_c^\infty(W)$, the norm $\|\cdot\|_{\mathcal{F},\max}$ is in turn bounded from above by the L_∞-norm. In light of the above, this would complete the proof of Theorem 4.6 in the two-dimensional case.

There are two independent components in the proof of this claim. First, we show that one can decompose any $f \in C_c^\infty(W^2)$ with $\|f\|_\infty \leqslant 1$ into a finite combination $f = \sum_{i=1}^{N_0} \epsilon_j \Psi_j^* g_j$. Here, $\epsilon_j \in \{-1, 1\}$, $\Psi_j \in \mathrm{Ham}_c(W^2,\omega)$, and g_j are smooth rotation-invariant functions whose L_∞-norm is bounded by an absolute constant, and which satisfy certain other technical conditions (see Proposition 3.5 in [17] for the precise statement). In what follows we call such functions "simple functions". We emphasize that N_0 is a constant independent of f. Thus, we can restrict ourselves to the case where f is a "simple function". In the second part of the proof, we construct an explicit collection $\mathcal{F} = \{\mathfrak{f}_0, \mathfrak{f}_1, \mathfrak{f}_2\}$, where $\mathfrak{f}_i \in C_c^\infty(W^2)$, and $i = 0, 1, 2$. Using an averaging procedure (see the proof of Theorem 3.4 in [17]), one can show that every "simple function" $f \in C_c^\infty(W^2)$ can be approximated arbitrarily well in the C^∞-topology by a sum of the form

$$\sum_{i,k} \alpha_{i,k} \widetilde{\Psi}_{i,k}^* \mathfrak{f}_k, \text{ where } \widetilde{\Psi}_{i,k} \in \mathrm{Ham}_c(W^2,\omega),\ k \in \{0,1,2\},$$

and such that $\sum |\alpha_{i,k}| \leq C\|f\|_\infty$ for some absolute constant C. Combining this with the above definition of $\|\cdot\|_{\mathcal{F},\max}$, we conclude that $\|f\|_{\mathcal{F},\max} \leq C\|f\|_\infty$ for every $f \in C_c^\infty(W^2)$. Together with Lemma 4.9, this completes the proof of Theorem 4.6 in the 2-dimensional case.

5. Some open questions and speculations

Do symplectic capacities coincide on the class of convex domains? As mentioned above, since the time of Gromov's original work, a variety of symplectic capacities have been constructed and the relations between them often lead to the discovery of surprising connections between symplectic geometry and Hamiltonian dynamics. In the two-dimensional case, Siburg [85] showed that any symplectic capacity of a compact connected domain with smooth boundary $\Omega \subset \mathbb{R}^2$ equals its Lebesgue measure. In higher dimensions symplectic capacities do not coincide in general. A theorem by Hermann [37] states that for any $n \geq 2$ there is a bounded star-shaped domain $S \subset \mathbb{R}^{2n}$ with cylindrical capacity $\overline{c}(S) \geq 1$, and arbitrarily small Gromov radius $\underline{c}(S)$. Still, for large classes of sets in $\mathbb{R}^{2n}$, including ellipsoids, polydiscs and convex Reinhardt domains, all symplectic capacities coincide [37]. In [88] Viterbo showed that for any bounded convex subset Σ of $\mathbb{R}^{2n}$ one has $\overline{c}(\Sigma) \leq 4n^2\underline{c}(\Sigma)$. Moreover, one has (see [37, 41, 88]) the following:

Conjecture 5.1. *For any convex domain Σ in $\mathbb{R}^{2n}$ one has $\underline{c}(\Sigma) = \overline{c}(\Sigma)$.*

This conjecture is particularly challenging due to the scarcity of examples of convex domains in which capacities have been computed. Moreover, note that Conjecture 5.1 is stronger than Viterbo's conjecture (Conjecture 2.2 above), as the latter holds trivially for the Gromov radius.

A somewhat more modest question in this direction is whether Conjecture 5.1 holds asymptotically, i.e., whether there is an absolute constant A such that for any convex domain $K \subset \mathbb{R}^{2n}$ one has $\overline{c}(K) \leq A\,\underline{c}(K)$. It would be interesting to explore whether methods from asymptotic geometric analysis can be used to answer this question.

Are Hanner polytopes in fact symplectic balls in disguise? Recall that Mahler's conjecture states that the minimum possible Mahler volume is attained by a hypercube. It is interesting to note that the corresponding product configuration, when looked at through symplectic glasses, is in fact a Euclidean ball in disguise. More precisely, it was proved in §7.9 of [82] (cf. Corollary 4.2 in [56]) that the interior of the product of a hypercube $Q \subset \mathbb{R}^n_q$ and its dual body, the cross-polytope $Q^\circ \subset \mathbb{R}^n_p$, is symplectomorphic to the interior of a Euclidean ball $B^{2n}(r) \subset \mathbb{R}^n_q \times \mathbb{R}^n_p$ with the same volume. On the other hand, as mentioned in Section 3 above, if Mahler's conjecture holds, then there are other minimizers for the Mahler volume aside of the hypercube. For example, consider the class of Hanner polytopes. A d-dimensional centrally symmetric polytope P is a Hanner polytope if either P is one-dimensional (i.e., a symmetric interval), or P is the free sum or direct product of two (lower dimensional) Hanner polytopes P_1 and P_2. Recall that the free sum of two polytopes, $P_1 \subset \mathbb{R}^n$, $P_2 \subset \mathbb{R}^m$ is a $n+m$ polytope defined by $P_1 \oplus P_2 = \mathrm{Conv}(\{P_1 \times \{0\}\} \cup \{\{0\} \times P_2\}) \subset \mathbb{R}^{n+m}$. It is not hard to check (see e.g. [80]) that the volume product of the cube is the same as that of Hanner polytopes. Thus every Hanner polytope is also a candidate for a minimizer of the volume product among symmet-

ric convex bodies. In light of the above mentioned result from [82], a natural question is the following:

Question 5.2. *Is every Hanner polytope a symplectic image of a Euclidean ball?*

More precisely, is the interior of every Hanner polytope symplectomorphic to the interior of a Euclidean ball with the same volume? A negative answer to this question would give a counterexample to Conjecture 5.1 above, since it would show that the Gromov radius must be different from the Ekeland-Hofer-Zehnder capacity.

Symplectic embeddings of Lagrangian products. Since Gromov's work [34], questions about symplectic embeddings have lain at the heart of symplectic geometry (see e.g., [12, 13, 35, 45, 56, 60, 61, 63, 82, 83]). These questions are usually notoriously difficult, and up to date most results concern only the embeddings of balls, ellipsoids and polydiscs. Note that even for this simple class of examples, only recently has it become possible to specify exactly when a four-dimensional ellipsoid is embeddable in a ball (McDuff and Schlenk [63]), or in another four-dimensional ellipsoid (McDuff [60]). For some other related works we refer the reader to [15, 19, 22, 30, 38, 39, 71].

Since symplectic capacities can naturally be used to detect symplectic embedding obstructions, and in light of the results mentioned in Section 3 (in particular, Theorem 3.8), it is only natural to try to extend the above list of currently-known examples, and study symplectic embeddings of convex "Lagrangian products" in the classical phase space. The main advantage of this class of bodies is that the action spectrum can be computed via billiard dynamics. This property would presumably make it easier to compute or estimate the Ekeland-Hofer capacities [24], or Hutchings' embedded contact homology capacities [44, 45], in this setting. A natural first step in this direction would be to consider the embedding of the Lagrangian product of two balls into a Euclidean ball. More precisely, to study the function $\sigma : \mathbb{N} \to \mathbb{R}$ defined by

$$\sigma(n) = \inf\big\{a \,|B_q^n(1) \times B_p^n(1) \overset{\text{symp}}{\hookrightarrow} B^{2n}(a)\big\}.$$

To the best of the author's knowledge, the value of $\sigma(n)$ is unknown already for the case $n = 2$.

Acknowledgements. I am deeply indebted to Leonid Polterovich for generously sharing his insights and perspective on topics related to this paper, as well as for many inspiring conversations throughout the years. I have also benefited significantly from an ongoing collaboration with Shiri Artstein-Avidan, I am grateful to her for many stimulating and enjoyable hours working together. I would also like to thank Felix Schlenk and Leonid Polterovich for their valuable comments on an earlier draft of this paper.

References

[1] Álvarez-Paiva, J.C. and Balacheff, F., *Optimalité systolique infinitésimale de l'oscillateur harmonique,* Séminaire de Théorie Spectrale et Géométrie **27** (2009), 11–16.

[2] ———, *Contact geometry and isosystolic inequalities,* to appear in Geom. Funct. Anal. Preprint. arXiv:1109.4253.

[3] Akopyan, A.V., Balitskiy, A. M., Karasev, R. N., and Sharipova, A., *Elementary results in non-reflexive Finsler billiards,* Preprint arXiv:1401.0442.

[4] Arnold, V. I., *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, (French) Ann. Inst. Fourier (Grenoble) **16** 1966 fasc. 1, 319–361.

[5] Albers, P. and Mazzucchelli, M., *Periodic bounce orbits of prescribed energy,* Int. Math. Res. Not. IMRN 2011, no. 14, 3289–3314.

[6] Artstein-Avidan, S., Milman, V., and Ostrover Y., *The M-ellipsoid, Symplectic Capacities and Volume*, Comment. Math. Helv. **83** (2008), no. 2, 359–369.

[7] Artstein-Avidan, S. and Ostrover Y., *Bounds for Minkowski billiard trajectories in convex bodies,* Intern. Math. Res. Not. (IMRN) (2012) doi:10.1093/imrn/rns216.

[8] Artstein-Avidan, S., Karasev, R., and Ostrover, Y., *From symplectic measurements to the Mahler conjecture,* to appear in Duke Math J., Preprint. arXiv:1303.4197.

[9] Banyaga, A., *Sur la structure du groupe des difféomorphisms qui préservent une forme symplectique*, Comment. Math. Helv. **53** (1978), no.2, 174–227.

[10] Benci, V. and Giannoni, F., *Periodic bounce trajectories with a low number of bounce points,* Ann. Inst. H. Poincaré Anal. Non Linéaire 6 (1989), no. 1, 73–93.

[11] Bezdek, D. and Bezdek, K., *Shortest billiard trajectories,* Geom. Dedicata 141 (2009), 197–206.

[12] Biran, P., *Symplectic packing in dimension 4,* Geom. Funct. Anal. **7** (1997), no. 3, 420–437.

[13] ———, *Lagrangian barriers and symplectic embeddings,* Geom. Funct. Anal. **11** (2001), no. 3, 407–464.

[14] Blaschke, W., *Über affine Geometrie VII: Neue Extremeigenschaften von Ellipse und Ellipsoid*, Ber. Verh. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Kl **69** (1917) 306–318, Ges. Werke **3** 246–258.

[15] Buse, O. and Hind, R., *Ellipsoid embeddings and symplectic packing stability*, Compos. Math. **149** (2013), no. 5, 889–902.

[16] Bourgain, J. and Milman, V. D., *New volume ratio properties for convex symmetric bodies in* $\mathbb{R}^n$*,* Invent. Math. **88** (1987), no. 2, 319–340.

[17] Buhovsky L. and Ostrover, Y., *Bi-invariant Finsler metrics on the group of Hamiltonian diffeomorphisms*, Geom. Funct. Anal. **21** (2011), no. 6, 1296–1330.

[18] Chekanov, Yu. V., *Lagrangian intersections, symplectic energy, and areas of holomorphic curves,* Duke Math. J. **95** (1998), no. 1, 213–226.

[19] Choi, K., Cristofaro-Gardiner, D., Frenkel, D., Hutchings, M., and Ramos, V.G.B., *Symplectic embeddings into four-dimensional concave toric domains,* arXiv:1310.6647.

[20] Cieliebak, T., Hofer, H., Latschev, and J., Schlenk F., *Quantitative symplectic geometry,* Dynamics, ergodic theory, and geometry, 1-44, Math. Sci. Res. Inst. Publ., 54, Cambridge Univ. Press, Cambridge 2007.

[21] Clarke, F. H., *Periodic solutions to Hamiltonian inclusions,* J. Differential Equations **40** (1981), no. 1, 1–6.

[22] Cristofaro-Gardiner, D. and Kleinman, A., *Ehrhart polynomials and symplectic embeddings of ellipsoids*, arXiv:1307.5493.

[23] Ekeland, I., *Convexity Methods in Hamiltonian Systems,* Ergeb. Math. Grenzgeb. **19**, Springer, Berlin, 1990.

[24] Ekeland, I. and Hofer, H., *Symplectic topology and Hamiltonian dynamics,* Mathematische Zeitschrift, **200** (1989), no. 3, 355–378.

[25] Eliashberg, Y., *Symplectic topology in the nineties,* Symplectic geometry. Differential Geom. Appl. **9** (1998), no. 1-2, 59–88.

[26] Eliashberg, Y. and Polterovich, L., *Bi-invariant metrics on the group of Hamiltonian diffeomorphisms.* Internat. J. Math. **4** (1993), 727–738.

[27] Floer, A. and Hofer, H., *Symplectic homology. I. Open sets in* $\mathbb{C}^n$*,* Math. Z. **215** (1994), no. 1, 37–88.

[28] Floer, A., Hofer, H., and Wysocki, K., *Applications of symplectic homology. I,* Math. Z. **217** (1994), no. 4, 577–606

[29] Frauenfelder, U., Ginzburg, V., and Schlenk, F., *Energy capacity inequalities via an action selector*, Geometry, spectral theory, groups, and dynamics, 129–152, Contemp. Math., 387, Amer. Math. Soc., Providence, RI, 2005.

[30] Frenkel, D. and Müller, D., *Symplectic embeddings of 4-dimensional ellipsoids into cubes*, arXiv:1210.2266.

[31] Ghomi, M., *Shortest periodic billiard trajectories in convex bodies,* Geom. Funct. Anal. **14** (2004), no. 2, 295–302.

[32] Giannopoulos, A., Paouris, G., and Vritsiou, B., *The isotropic position and the reverse Santaló inequality,* to appear in Israel J. Math. Preprint. arXiv:1112.3073.

[33] Gordon, Y., Meyer, M., and Reisner, S., *Zonoids with minimal volume product—a new proof,* Proc. Amer. Math. Soc. **104** (1988), no. 1, 273–276.

[34] Gromov, M., *Pseudoholomorphic curves in symplectic manifolds*, Invent. Math. **82** (1985), no. 2, 307–347.

[35] Guth, L., *Symplectic embeddings of polydisks,* Inven. Math. **172** (2008), 477–489.

[36] Gutkin, E. and Tabachnikov, S., *Billiards in Finsler and Minkowski geometries,* J. Geom. Phys. **40** (2002), no. 3-4, 277–301.

[37] Hermann, D., *Non-equivalence of symplectic capacities for open sets with restricted contact type boundary*, Prépublication d'Orsay numéro 32 (29/4/1998).

[38] Hind, R. and Kerman, E., *New obstructions to symplectic embeddings,* preprint. arXiv: 0906.4296.

[39] Hind, R. and Lisi, S., *Symplectic embeddings of polydisks,* To appear in Selecta Mathematica. Preprint. arXiv:1304.3065.

[40] Hofer, H., *On the topological properties of symplectic maps,* Proc. Roy. Soc. Edinburgh Sect. A 115 (1990), 25-38.

[41] ———, *Symplectic capacities*, Geometry of low-dimensional manifolds, 2 (Durham, 1989), 15-34, London Math. Soc. Lect. Note Ser., 151, Cambridge Univ. Press, 1990.

[42] Hofer, H. and Zehnder, E., *A new capacity for symplectic manifolds,* Analysis, et cetera, 405–427, Academic Press, Boston, MA, 1990.

[43] ———, *Symplectic invariants and Hamiltonian dynamics*, Birkhauser Advanced Texts, Birkhauser Verlag, 1994.

[44] Hutchings, M., *Quantitative embedded contact homology*, J. Differential Geom. **88** (2011), no. 2, 231–266.

[45] ———, *Recent progress on symplectic embedding problems in four dimensions,* Proc. Natl. Acad. Sci. USA 108 (2011), no. 20, 8093–8099.

[46] ———, Some open problems on symplectic embeddings and the Weinstein conjecture, http://floerhomology.wordpress.com/2011/09/14/open-problems/.

[47] Irie, K., *Symplectic capacity and short periodic billiard trajectory,* Math. Z. **272** (2012), no. 3-4, 1291–1320.

[48] ———, *Periodic billiard trajectories and Morse theory on loop spaces,* Preprint. arXiv:1403.1953.

[49] Kim, J., *Minimal volume product near Hanner polytopes,* J. Funct. Anal. **266** (2014), no. 4, 2360–2402.

[50] Künzle, A. F., *Singular Hamiltonian systems and symplectic capacities*, Singularities and differential equations (Warsaw, 1993), 171–187, Banach Center Publ., 33, Polish Acad. Sci., Warsaw, 1996.

[51] Kuperberg, G., *From the Mahler conjecture to Gauss linking integrals,* Geom. Funct. Anal., **18**, no. 3, (2008), 870–892.

[52] Lalonde, F., *Energy and capacities in symplectic topology,* in: Geometric topology (Athens, GA, 1993), 328-374, AMS/IP Stud. Adv. Math., **2.1**, Amer. Math. Soc., Providence, RI, 1997.

[53] Lalonde, F. and McDuff, D., *The geometry of symplectic energy,* Ann. of Math. (2) **141** (1995), no. 2, 349–371.

[54] ———, *Hofer's L^∞-geometry: Energy and stability of Hamiltonian flows*, parts I, II, Invent. Math. **122** (1995), 1–33, 35–69.

[55] Landry, M., McMillan, M., and Tsukerman, E., *On symplectic capacities of toric domains,* Preprint. arXiv:1309.5072.

[56] Latschev, J., McDuff, D., and Schlenk, F., *The Gromov width of 4-dimensional tori*, arXiv:1111.6566.

[57] Lu, G., *Gromov-Witten invariants and pseudo symplectic capacities,* Israel. J. Math. **156** (2006), 1–63.

[58] Mahler, K., *Ein Übertragungsprinzip für konvexe Korper,* Casopis Pyest. Mat. Fys. **68** (1939), 93–102.

[59] McDuff, D., *Geometric variants of the Hofer norm,* J. Symplectic Geom. 1 (2002), no. 2, 197–252.

[60] ———, *The Hofer conjecture on embedding symplectic ellipsoids,* J. Diff. Geom. **88** (2011), no. 3, 519–532.

[61] McDuff, D. and Polterovich, L., *Symplectic packings and algebraic geometry. With an appendix by Yael Karshon,* Invent. Math. **115** (1994), no. 3, 405–434.

[62] McDuff, D. and Salamon, D., *Introduction to Symplectic Topology,* Second edition. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 1998.

[63] McDuff, D. and Schlenk, F., *The embedding capacity of 4-dimensional symplectic ellipsoids,* Ann. of Math. (2) **175** (2012), no. 3, 1191–1282.

[64] Meyer, M., *Une caractérisation volumique de certains espaces normés de dimension finie,* Israel J. Math. **55** (1986), no. 3, 317–326.

[65] Milman, V.D., *An inverse form of the Brunn-Minkowski inequality with applications to the local theory of normed spaces,* C. R. Acad. Sci. Paris Sér. I Math. **302** (1986), no. 1, 25–28.

[66] ———, *Isomorphic symmetrizations and geometric inequalities,* in: Geometric aspects of functional analysis (1986/87), 107–131, Lecture Notes in Math., 1317, Springer, Berlin, 1988.

[67] Milman, V.D. and Schechtman, G., *Asymptotic Theory of Finite Dimensional Normed Spaces,* Lectures Notes in Math. 1200, Springer, Berlin (1986).

[68] Nazarov, F., *The Hörmander proof of the Bourgain-Milman theorem*, in: Geometric aspects of functional analysis, 335–343, Lecture Notes in Math., 2050, Springer, Heidelberg, 2012.

[69] Nazarov, F., Petrov, F., Ryabogin, D., and Zvavitch, A., *A remark on the Mahler conjecture: local minimality of the unit cube,* Duke Math. J. **154** (2010), no. 3, 419–430.

[70] Oh, Y-G., *Chain level Floer theory and Hofer's geometry of the Hamiltonian diffeomorphism group,* Asian J. Math. **6** (2002), no. 4, 579–624.

[71] Opshtein, E., *Symplectic packings in dimension 4 and singular curves,* arXiv:1110.2385.

[72] Ostrover, Y. and Wagner, R., *On the extremality of Hofer's metric on the group of Hamiltonian diffeomorphisms*, Int. Math. Res. Not. **35** (2005), 2123–2141.

[73] Pisier, G., *The Volume of Convex Bodies and Banach Space Geometry,* Cambridge University Press, Cambridge (1989).

[74] Polterovich, L., *Symplectic displacement energy for Lagrangian submanifolds,* Ergodic Theory Dynam. Systems **13** (1993), no. 2, 357–367.

[75] ———, *The Geometry of the Group of Symplectic Diffeomorphisms*, Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2001.

[76] Reisner, S., *Zonoids with minimal volume product,* Math. Z. **192** (1986), no. 3, 339–346.

[77] ———, *Minimal volume-product in Banach spaces with a 1-unconditional basis,* J. London Math. Soc. **36** (1987), no.1, 126–136.

[78] Reisner, S., Schütt, C., and Werner, E., *Mahler's conjecture and curvature,* Int. Math. Res. Not. **2012**, no. 1, 1–16.

[79] Rogers, C.A. and Shephard, C., *The difference body of a convex body,* Arch. Math. **8** (1957), 220–233.

[80] Saint Raymond, J., *Sur le volume des corps convexes symétriques,* in: Initiation Seminar on Analysis: G. Choquet–M. Rogalski–J. Saint-Raymond, 20th Year: 1980/1981, Exp. No. 11, Publ. Math. Univ. Pierre et Marie Curie, 46, Univ. Paris VI, Paris, 1981.

[81] Santaló, L.A., *Un invariante afin para los cuerpos convexos de espacio de n dimensiones,* Portugal. Math **8** (1949) 155–161.

[82] Schlenk. F., *Embedding Problems in Symplectic Geometry,* de Gruyter Expositions in Mathematics, **40**, Berlin, 2005.

[83] ———, *Symplectic embeddings of ellipsoids,* Israel J. Math. **138** (2003), 215–252.

[84] Schwarz, M., *On the action spectrum for closed symplectically aspherical manifolds*, Pacific J. Math. **193** (2000), 1046–1095.

[85] Siburg, K.F., *Symplectic capacities in two dimensions*, Manuscripta Math. **78** (1993), no. 2, 149–163.

[86] Stancu, A., *Two volume product inequalities and their applications,* Canad. Math. Bull. **52** (2009), no. 3, 464–472.

[87] Tao. T., *Structure and Randomness. Pages from Year One of a Mathematical Blog,* American Mathematical Society, Providence, RI, 2008.

[88] Viterbo, C., *Metric and isoperimetric problems in symplectic geometry,* J. Amer. Math. Soc. **13** (2000), no. 2, 411–431.

[89] ———, *Symplectic topology as the geometry of generating functions,* Math. Ann. **292** (1992), 685–710.

School of Mathematical Sciences, Tel Aviv University, Ramat Aviv 69978 Israel
E-mail: ostrover@post.tau.ac.il

On the future stability of cosmological solutions to Einstein's equations with accelerated expansion

Hans Ringström

Abstract. The solutions of Einstein's equations used by physicists to model the universe have a high degree of symmetry. In order to verify that they are reasonable models, it is therefore necessary to demonstrate that they are future stable under small perturbations of the corresponding initial data. The purpose of this contribution is to describe mathematical results that have been obtained on this topic. A question which turns out to be related concerns the topology of the universe: what limitations do the observations impose? Using methods similar to ones arising in the proof of future stability, it is possible to construct solutions with arbitrary closed spatial topology. The existence of these solutions indicate that the observations might not impose any limitations at all.

Mathematics Subject Classification (2010). Primary 83C05; Secondary 35Q76.

Keywords. General relativity, stability of solutions, Einstein-Vlasov system.

1. Introduction

In 1915, the interpretation of gravitational forces was fundamentally altered by the introduction of Einstein's general theory of relativity. The underlying mathematical structures of the theory were not well understood at the time, and as a consequence, some of the fundamental questions have only recently been phrased in the form of mathematical problems. Since Einstein's equations are not as commonly studied in mathematics as many other equations that appear in physics, we here wish to give a brief description of their origin and of how different perspectives on them have developed since the inception of general relativity. However, the main purpose of this contribution is more specific. Recent observational data indicate that the universe is expanding at an accelerated rate. As a consequence, physicists nowadays use solutions to Einstein's equations with accelerated expansion to model the universe. Since the model solutions are highly symmetric (they are spatially homogeneous and isotropic), a natural question to ask is: are they stable? In order to phrase this question in a more precise way, it is necessary to formulate Einstein's equations (coupled to various matter equations) as an initial value problem. It turns out that there is a natural and geometric notion of initial data, and that, given initial data, there is a uniquely associated maximal Cauchy development. A more precise formulation of the question of stability is then: given initial data corresponding to one of the standard models, do small perturbations thereof yield maximal Cauchy developments that are globally similar? The currently preferred models have a big bang type singularity and an expanding direction. Proving stability in the direction of the

Proceedings of the International Congress of Mathematicians, Seoul, 2014

singularity is quite difficult (there are some results in the case of special matter fields), but there are several results on stability in the expanding direction. For that reason we shall focus on the expanding direction here, and we shall think of it as corresponding to the future.

The outline of this contribution is as follows. We begin, in Section 2, by giving a brief description of the origin of the general theory of relativity. Moreover, we explain how the present contribution fits into the general context of mathematical studies of Einstein's equations. In Sections 3 and 4, we then discuss the formulation of the initial value problem, which is needed in order for us to be able to state a stability result. In Section 5, we then discuss the topic of stability in general. We give a rough description of some of the results that have been obtained in the past, as well as of some of the methods. However, we shall only formulate a theorem in the case of the Einstein-Vlasov system. In order to be able to do so, we devote Section 6 to a discussion of this system. In Sections 7 and 8, we then describe the background solutions we are interested in proving stability of, and state the relevant results. Finally, in Section 9, we discuss a construction which indicates that the observations do not impose any restrictions on the topology of the universe.

2. General relativity

In order to discuss the general theory of relativity, it is natural to begin with Einstein's paper on special relativity [6]. The starting point of the paper is the contemporary interpretation of electrodynamics. Noting that this interpretation involves asymmetries, and postulating that the speed of light is independent of inertial observer, Einstein was led to the Poincaré group of transformations, relating the observations of inertial observers. Due to the added insight of Poincaré and Minkowski, it was realized that this group is the group of isometries of Minkowski space; recall that Minkowski space is $\mathbb{R}^4$ with the inner product $\langle x, y\rangle = x^t \eta y$, where $\eta = \mathrm{diag}(-1, 1, 1, 1)$. This interpretation indicates the importance of geometry. As a next step, it is clear that Newtonian gravity has to be modified. Two important principles that guided Einstein in his search for a modified theory were the *equivalence principle* (the equality between inert and gravitational mass; this is roughly speaking the idea that it is not possible to distinguish between a coordinate system at rest in a uniform gravitational field and a uniformly accelerated coordinate system far away from all matter, for example) and the *principle of general covariance*, the idea that the equations should be independent of the choice of coordinate system. By a simple thought experiment involving rotating coordinate systems, it can be argued (heuristically) that acceleration distorts the geometry; cf. [7, pp. 58–59]. Combining this observation with the equivalence principle indicates that gravitation should affect the geometry. In fact, it is not unnatural to equate gravitation with distorsion of the geometry. Since the geometry at a point should be described by the Minkowski metric (with respect to a suitable choice of coordinates), the natural underlying object in general relativity is a Lorentz manifold; in other words, a manifold M on which a smooth symmetric covariant two-tensor field g is defined, where g is such that it, at each point of M, equals the Minkowski metric with respect to suitable coordinates. The standard notions and constructions in Riemannian geometry (Levi-Civita connection, curvature tensor, Ricci tensor, scalar curvature, geodesics etc.) can be defined in the same way in Lorentz geometry, and we shall use them below without further comment. The one question that remains is: what equation should (M, g) satisfy? In some way, the geometry should be related to the matter sources. On the level of special relativity, it was already clear that the matter should

be combined into the so-called stress-energy tensor; a symmetric covariant two-tensor field, the exact form of which depends on the specific matter model. Let us denote this object by T. It should be the source term in Einstein's equations (it can be thought of as a generalization of the matter density in Poisson's equation in Newtonian gravity). As a consequence, what remains is to determine what the left hand side of the equation should be. To begin with, it should clearly be symmetric. Due to the equations for the matter, the stress-energy tensor should be divergence free. As a consequence, the left hand side should be as well. Moreover, it should be such that the resulting equations are independent of the choice of coordinates. Finally, it is natural (for the sake of simplicity, and in analogy with the Poisson equation) to require that the left hand side should contain at most second order derivatives of the gravitational field (i.e., the metric). However, the only equations fulfilling these requirements are the ones of the form

$$G + \Lambda g = \alpha T, \tag{2.1}$$

where Λ and α are constants and

$$G = \mathrm{Ric} - \frac{1}{2} S g$$

is the *Einstein tensor*, defined in terms of the Ricci tensor, Ric, and the scalar curvature, S, of the metric g (the reader interested in a justification of this statement is referred to the corollary of [15, Theorem 1, p. 500]). In (2.1), we shall, for simplicity, assume $\alpha = 1$. Moreover, we shall refer to Λ as the *cosmological constant*. The resulting equations are

$$G + \Lambda g = T, \tag{2.2}$$

and we shall refer to them as *Einstein's equations*.

2.1. Historical development. It is of interest to say a few words concerning how different perspectives on these equations have developed over time. In the intial phase, physicists tried to find explicit solutions to the equations. In order to do so, they imposed symmetry assumptions adapted to the physical situation of interest. When considering physical objects such as a star, a galaxy, a globular cluster etc. (i.e., an *isolated system*), a natural first symmetry assumption to make is that of spherical symmetry. This assumption led to the class of Schwarzschild spacetimes, which can be used to model the gravitational field outside a non-rotating star or black hole. Much later, the Kerr family of solutions was found, describing the rotating case. When modelling the universe as a whole, another type of symmetry assumption is required. Guided by the Copernican principle, a natural starting point in this case is the assumption of spatial homogeneity and isotropy; this is the assumption that at 'one moment in time', it is not possible to distinguish between two points in space, nor is it possible to distinguish between two directions. Symmetry assumptions of this type (corresponding to the so-called *cosmological setting*) led to the Friedman-Lemaître-Robertson-Walker metrics, which are still used to this very day when modelling the universe (though the preferred matter models have changed over time). Even though mathematicians nowadays consider significantly less symmetric solutions, the problems considered can still be divided into ones concerning isolated systems and ones concerning the cosmological setting.

In the initial stages of the development of general relativity, when the emphasis was on finding explicit solutions, the geometry remained somewhat obscure. As a consequence, some of the features of, e.g., the Schwarzschild solutions were misunderstood for several

decades. In the 50's and 60', the geometry received more attention, and the so-called *singularity theorems* were proven. In order to give an idea of the statements of these results, it is necessary to introduce the notion of causal geodesics. To begin with, a vector v in Minkowski space is said to be *timelike* if $\langle v, v\rangle < 0$; *lightlike* or *null* if $\langle v, v\rangle = 0$ and $v \neq 0$; and *spacelike* if $\langle v, v\rangle > 0$ or $v = 0$. A vector which is either timelike or null is said to be *causal*. These notions can be generalized to Lorentz manifolds. Moreover, it makes sense to speak of timelike curves etc. as well as spacelike hypersurfaces. In particular, the character of a geodesic (timelike, null, spacelike) is preserved, so that it is meaningful to speak of timelike geodesics etc. In the interpretation of general relativity, a causal curve corresponds to an observer that travels at a speed less than or equal to that of light. Moreover, a timelike geodesic corresponds to a freely falling test particle, and a null geodesics corresponds to a light ray. Thus causal geodesics are of particular importance in general relativity. A notion which is also of importance is that of a time orientation. At a given spacetime point, the set of causal vectors based at that point has two components. A continuous choice of component corresponds to a *time orientation* (and we shall, from now on, assume all Lorentz manifolds to be time oriented). Vectors belonging to the chosen component will be referred to as *future oriented*.

The singularity theorems of Hawking and Penrose give general conditions that ensure the existence of incomplete causal geodesics. Since the existence of such a geodesic means that there is a freely falling test particle (or a light ray) which exits the spacetime in finite parameter time, Hawking and Penrose equated causal geodesic incompleteness with the existence of a singularity (examples illustrate that this is not always reasonable). Due to the results, it is to be expected that singularities, in the sense of causal geodesic incompleteness, occur generically in solutions to Einstein's equations. These results changed the perspective concerning the occurrences of singularities. Moreover, due to the methods used to prove them, the importance of the subject of Lorentz geometry became apparent.

In the early 50's, Yvonne Choquet-Bruhat formulated Einstein's equations as an initial value problem [8]. It took a significant amount of time before this perspective became a natural starting point in the subject. Since the initial data cannot be specified freely (they have to satisfy an underdetermined, non-linear system of elliptic PDE's, referred to as the *constraint equations*), and since, given initial data, the evolution problem typically involves proving global existence of solutions to a non-linear system of hyperbolic PDE's, this is perhaps not so surprising. In particular, the relevant PDE tools were not so well developed in the early 50's. Nevertheless, this perspective has become more and more important in the subject. This is, in particular, due to the fact that central questions such as that of stability are most naturally formulated using it.

With the above description in mind, the present contribution can be said to be concerned with the initial value formulation of Einstein's equations in the cosmological setting. Moreover, the precise notion of stability we shall use is highly dependent on a Lorentz geometric interpretation of the outcome of the PDE analysis.

3. On the character of Einstein's equations

In order to justify that it is meaningful to formulate Einstein's equations as an initial value problem, let us begin by focusing on the vacuum equations with a vanishing cosmological constant. Since these equations can be written $\mathrm{Ric} = 0$, it is of interest to know if Ric,

considered as a differential operator acting on the components of the metric, has a particular character (elliptic, hyperbolic etc.). Due to the diffeomorphism invariance of the equations, this is not to be expected. On the other hand, it is possible to break the diffeomorphism invariance by making a special choice of coordinates. In fact, choosing coordinates such that the contracted Christoffel symbols vanish, the Ricci tensor (schematically) takes the form

$$\mathrm{Ric}_{\alpha\beta} = -\frac{1}{2} g^{\mu\nu} \partial_\mu \partial_\nu g_{\alpha\beta} + F_{\alpha\beta}(g, \partial g), \tag{3.1}$$

where F is a quadratic expression in the first derivatives of the metric components. In this equation, we assume Greek indices to range from 0 to n, where $n+1$ is the dimension of the Lorentz manifold, and we tacitly assume that repeated indices are summed over (the Einstein summation convention). With respect to these coordinates, Einstein's vacuum equations can thus be thought of as a system of non-linear wave equations for the metric components. As a consequence, it seems natural to formulate a corresponding inital value problem.

It is of interest to note that the above issues arise not only in general relativity, but also in Riemannian geometry and in Ricci flow. In Riemannian geometry, it is sometimes convenient to think of the Ricci tensor as an elliptic differential operator acting on the components of the metric; this yields good control of the metric components, given information concerning the Ricci tensor. It is therefore of interest to consider the so-called *harmonic coordinates*, defined by the condition that the contracted Christoffel symbols vanish. The reason for referring to these coordinates as harmonic is that their defining requirement is equivalent to

$$\Delta_g x^\mu = 0,$$

where Δ_g is the scalar covariant Laplacian associated with the metric g and x^μ are the components of the coordinate system. The analogous coordinates in the Lorentzian setting are sometimes, by analogy, referred to as harmonic coordinates (and sometimes as wave coordinates). In Ricci flow, the relevant equation is $\partial_t g = -2\mathrm{Ric}[g]$, and when dealing with this equation analytically, it would be convenient if Ric were an elliptic operator. Hamilton's original idea concerning how to prove local existence was to appeal to the Nash-Moser inverse function theorem. However, later proofs instead relied on breaking the diffeomorphism invariance in order to obtain a strictly parabolic equation.

4. The initial value problem

With the above observations in mind, it seems natural to formulate an initial value problem. However, it is not so clear what the initial data should be, nor where they should be specified. It turns out that there are several ways of proceeding, but we shall here focus on the perspective that arises in analogy with the standard Cauchy problem for the ordinary wave equation. In that setting, the initial data are specified on a $t = \mathrm{const}$ hypersurface in Minkowski space. These hypersurfaces are special in several ways. First of all, they are *spacelike*, meaning that the induced metric is Riemannian (in this particular case, they are in fact the ordinary Euclidean metric). Moreover, they are intersected exactly once by every inextendible causal curve; cf. the above terminology. Hypersurfaces in Lorentz manifolds which are intersected exactly once by every inextendible casual curve are referred to as *Cauchy hypersurfaces*. They are natural surfaces on which to specify initial data, since given initial data on a Cauchy hypersurface (for the linear wave equation on the Lorentz

manifold), there is a unique corresponding solution. A Lorentz manifold which admits a Cauchy hypersurface is called *globally hyperbolic*.

Turning to the choice of the intial data, it would seem natural to specify the metric components and their normal derivative at the initial hypersurface (keeping (3.1) in mind). However, since Einstein's equations are geometric in nature, the initial data should be geometric as well. On the other hand, the induced metric and second fundamental form are geometric in nature and correspond to a part of the desired information; with respect to local coordinates, they yield some of the metric components and the normal derivative of some metric components. The induced metric and second fundamental form would thus seem to constitute minimal information needed in order to construct a solution. On the other hand, it unfortunately turns out that these initial data cannot be specified freely. In order to be more specific, let Σ be a spacelike hypersurface in a Lorentz manifold on which Einstein's equations (2.2) are satisfied. Contracting the equations twice with respect to the future directed unit normal, say N, yields

$$\frac{1}{2}[\bar{S} - \bar{k}_{ij}\bar{k}^{ij} + (\mathrm{tr}_{\bar{g}}\bar{k})^2] = \rho + \Lambda, \tag{4.1}$$

where $\bar{g}$ and $\bar{k}$ are the induced metric and second fundamental form on the hypersuface Σ respectively; cf. [18, Proposition 13.3, p. 149]. Moreover, $\bar{S}$ is the scalar curvature of the metric $\bar{g}$, indices are raised and lowered with $\bar{g}$ and $\rho = T(N, N)$. In particular, all the ingredients in (4.1) are intrinsic to the hypersurface. Contracting (2.2) once with respect to the future directed unit normal and once with respect to a tangential vector yields the equation

$$\overline{\nabla}^j \bar{k}_{ji} - \overline{\nabla}_i \mathrm{tr}_{\bar{g}}\bar{k} = -J_i, \tag{4.2}$$

where $\overline{\nabla}$ is the Levi-Civita connection associated with the metric $\bar{g}$ and J is the one-form field defined by $J = -T(N, \cdot)$; cf. [18, Proposition 13.3, p. 149]. Again, the ingredients of (4.2) are intrinsic to the hypersurface Σ. Clearly, the initial data have to satisfy (4.1) and (4.2), which are referred to as the *Hamiltonian* and *momentum constraints* respectively; collectively, we shall refer to them as the *constraint equations*. It is natural to ask whether the constraint equations are sufficient in order to guarantee the existence of a corresponding development. In the vacuum setting, this question was settled in the seminal result of Yvonne Choquet-Bruhat [8], which we now formulate.

Theorem 4.1. *Let $(\Sigma, \bar{g}, \bar{k})$ be initial data for Einstein's vacuum equations; i.e., Σ is an n-dimensional manifold, $\bar{g}$ is a Riemannian metric and $\bar{k}$ is a symmetric covariant 2-tensor field satisfying the vacuum constraint equations; i.e., (4.1) and (4.2) with $\Lambda = 0$, $\rho = 0$ and $J = 0$. Then there is a globally hyperbolic development of the initial data. In other words, a Lorentz manifold (M, g) satisfying Einstein's vacuum equations and an embedding $i : \Sigma \to M$ such that $i^* g = \bar{g}$ and $i^* \kappa = \bar{k}$, where κ is the second fundamental form of $i(\Sigma)$ in (M, g). Moreover, $i(\Sigma)$ is a Cauchy hypersurface in (M, g).*

This result has been generalized to include many different types of matter models. We shall not list them, but for all the matter models discussed in this contribution, there is a result analogous to Theorem 4.1.

Even though Theorem 4.1 is important, it does have one deficiency; there is no uniqueness statement. Given initial data, there are infinitely many inequivalent globally hyperbolic developments associated with it. In order to obtain uniqueness, it is necessary to require

some sort of maximality. In fact, the fundamental result, due to Yvonne Choquet-Bruhat and Robert Geroch [4], is the following.

Theorem 4.2. *Let $(\Sigma, \bar{g}, \bar{k})$ be initial data for Einstein's vacuum equations. Then there is a unique maximal globally hyperbolic development.*

Due to this theorem, it is clear that the notion of initial data introduced in the statement of Theorem 4.1 is meaningful. Unfortunately, there are examples of maximal globally hyperbolic developments that are extendible in the class of all (not necessarily globally hyperbolic) developments. In fact, there might even be inequivalent maximal developments, indicating that the general theory of relativity is not deterministic. Since the examples are very special, one is led to the strong cosmic censorship conjecture. However, that is not the main topic of this contribution. In fact, we shall here be content with the maximal globally hyperbolic development as *the* development of the intial data.

5. Stability

Since Einstein's equations can be formulated as an initial value problem, it is possible to phrase the stability question: Given initial data corresponding to a specific solution, do small perturbations thereof yield maximal globally hyperbolic developments which are globally similar? The question is still somewhat vague, since we have not specified what is meant by globally similar, nor what is meant by small perturbations. However, the precise meaning in practice depends on the particular solution under consideration, and even for a given solution it is sometimes possible to take different perspectives.

Turning to the stability results that have been obtained in the past, the first one is due to Helmut Friedrich; cf. [9], which contains a proof of stability of de Sitter space. In the same paper, he also proved future stability of Minkowski space, starting with hyperboloidal initial data. Later on, Demetrios Christodoulou and Sergiu Klainerman proved stability of Minkowski space [5]. That stability holds when using harmonic coordinates was only demonstrated much later by Hans Lindblad and Igor Rodnianski [13]. Another perspective on the stability of Minkowski space is given by the work of Lydia Bieri; cf. [3]. Even though all of the references [3, 5, 9, 13] pertain to the problem of stability of Minkowski space, they are very different in nature; the assumptions and conclusions are different in all of these references, and the results correspond to different notions of 'smallness' and 'global similarity'.

Minkowski space is a natural solution to start with when one is interested in isolated systems. However, the topic of the present contribution is cosmology. Of the references mentioned above, the one which is of interest in that setting is [9], in which Friedrich proves stability of de Sitter space. For an appropriate value of the cosmological constant, Λ, the metric of de Sitter space is given by

$$g_{\mathrm{dS}} = -dt \otimes dt + \cosh^2(t)\bar{g}_{\mathbb{S}^3}$$

on $\mathbb{R} \times \mathbb{S}^3$, where $\bar{g}_{\mathbb{S}^3}$ is the standard metric on $\mathbb{S}^3$. The de Sitter space is a solution to Einstein's vacuum equations with a positive cosmological constant; i.e.,

$$G + \Lambda g = 0.$$

The result of Friedrich is peculiar to $3+1$-dimensions, but Michael Anderson later generalized it to $n+1$-dimensions, with n odd; cf. [1]. Friedrich also generalized [9] to include matter of Maxwell and Yang-Mills type; cf. [10]. All of these references yield stability of cosmological solutions with accelerated expansion. They thus belong to the class of results we wish to discuss here. It is of interest to note that the proofs given in [1, 9, 10] are based on conformal reformulations of the equations. The idea is to first rescale the background spacetime by a conformal factor, so that what corresponds to past and future infinity in the physical spacetime is at a finite distance away with respect to the rescaled metric. The second step is then to derive a suitable system of equations for the rescaled metric and conformal factor (in reality, the variables might be quite different). This step can be expected to be very difficult, and the only cases in which it is known to be possible is when the matter sources have suitable conformal invariance properties. However, when it is possible, the problem of global existence and stability becomes an issue of continuous dependence on initial data. Assuming the conformally rescaled equations admit a well posed initial value problem (with respect to an appropriately chosen gauge; i.e., an appropriate choice of how to break the diffeomorphism invariance), this is, however, immediate, so that the desired result follows. It is of interest to note that, even though [1] yields stability in the case of higher dimensions, it can be used to prove stability in $3+1$-dimensions for spacetimes with a special type of matter source; cf. [12]. The results mentioned above are appealing due the the geometric nature of the arguments involved. However, the methods used seem to suffer from a lack of robustness. This leads us to the different perspective developed in [17].

In [17], we considered the case of Einstein's equations coupled to a non-linear scalar field. The relevant stress-energy tensor in that case is

$$T = d\phi \otimes d\phi - \left[\frac{1}{2} g(\mathrm{grad}\phi, \mathrm{grad}\phi) + V(\phi)\right] g, \tag{5.1}$$

where ϕ is a scalar valued function on the manifold (the so-called *scalar field*) and V is a smooth function on $\mathbb{R}$ referred to as the *potential*. The relevant matter field equations are

$$\Box_g \phi - V' \circ \phi = 0, \tag{5.2}$$

where $\Box_g$ is the scalar wave operator associated with g (defined in the same way as the scalar Laplacian in the case of Riemannian geometry). Note that (5.1) is divergence free if (5.2) holds. In [17], it is assumed that $V(0) > 0$, $V'(0) = 0$ and $V''(0) > 0$; in other words, that 0 is a positive non-degenerate local minimum of the potential. Moreover, the scalar field is assumed to be small initially.

The motivation for studying non-linear scalar fields is partly due to their interest in physics. Once it became clear that the observational data indicate that the universe is expanding at an accelerated rate, it was natural to try to find matter models that induce accelerated expansion. One possibility is to include a positive cosmological constant. Another is to add matter of non-linear scalar field type. The types of potentials considered above specialize to the case of a positive cosmological constant when demanding that $\phi = 0$ (a case which can already be handled using conformal methods). However, they are more general, and only in the case of special relations between $V(0)$ and $V''(0)$ do the conformal methods seem to work; cf. [11].

Following the appearance of [17], there were several results obtained using similar methods; cf. [19] (treating the case of an exponential potential and generalizing the results of

[12]), [16, 23] (in which an electromagnetic field was added), [21, 22] (in which the Euler-Einstein system was considered). However, the situation we focus on in what follows is the Einstein-Vlasov setting, discussed in [20].

6. The Einstein-Vlasov system

The physical situation matter of Vlasov type is supposed to represent is that of a gas. The fundamental assumption is that collisions are sufficiently rare that they can be negelected (including binary collisions would, e.g., lead to the Boltzmann equation, which we do not consider here). In the case of general relativity, each particle in the gas can thus be expected to behave as a freely falling test particle. A test particle with a non-zero rest mass (also referred to as a *massive* particle) can thus be expected to travel along a timelike geodesic. In the case of a zero rest mass (i.e., a *massless*) particle, the relevant curves are the null geodesics. On the other hand, the particles collectively generate a gravitational field which, in its turn, affects the geometry (and thereby the geodesics). In order to describe the gas, it is convenient to use a distribution function. The natural space on which this function is defined is the space of states of particles. Assuming all the particles to have rest mass 1, the space of states is given by the set of future directed unit timelike vectors. We shall denote this set by P, and we shall refer to it as the *mass shell*. The *distribution function*, say f, is then a function from P to the non-negative real numbers.

In order to couple Vlasov matter to Einstein's equations, it is necessary to explain how to construct a stress-energy tensor, given a distribution function. Moreover, it is necessary to formulate an evolution equation for the distribution function. The relevant stress-energy tensor is defined by

$$T|_{T_\xi M \times T_\xi M} = \int_{P_\xi} f(p) p^\flat \otimes p^\flat \mu_{P_\xi}(p). \tag{6.1}$$

In this equation, ξ is a spacetime point (i.e., an element of the spacetime manifold M); P_ξ is the mass shell above ξ (i.e., the elements of P based at ξ); p is an element of P_ξ; $p^\flat$ is the one-form metrically associated with p (i.e., $p^\flat(X) = g(p, X)$ for $X \in T_\xi M$); and μ_{P_ξ} is a volume form defined on P_ξ in the following way: the metric g induces a Lorentz metric g_ξ on $T_\xi M$, the Lorentz metric g_ξ induces a Riemannian metric on P_ξ, and this Riemannian metric induces a volume form on P_ξ (which we denote by μ_{P_ξ}). It is important to note that it is necessary to impose fall-off conditions on the distribution function in order for (6.1) to make sense. Often the requirement of compact support in the momentum directions is imposed, but we here prefer to demand that the distribution function belong to Sobolev spaces with appropriate weights in the momentum directions.

Turning to the equation for the distribution function, it is given by

$$\mathcal{L} f = 0, \tag{6.2}$$

and it is referred to as the *Vlasov equation*. Here $\mathcal{L}$ is a vector field on P defined as follows. Given an element of P, say v, there is a unique geodesic γ such that $\dot\gamma(0) = v$. Moreover, $\dot\gamma(s)$ is a curve in P, and its tangent vector at 0 (considered as a curve in P) is $\mathcal{L}_v$, the vector field $\mathcal{L}$ at the point v. Note that the Vlasov equation is equivalent to the requirement that $f(\dot\gamma)$ be constant for each geodesic γ with initial values on P. Moreover, this requirement corresponds to the assumption that collisions can be neglected, so that the particles travel

along timelike geodesics. It is of interest to note that (6.2) implies that the stress energy tensor defined by (6.1) is divergence free (regardless of whether Einstein's equations are satisfied or not).

Summing up the above discussion, the *Einstein-Vlasov* system with a positive cosmological constant is given by the equations

$$\begin{aligned} G + \Lambda g =& T, \\ \mathcal{L} f =& 0, \end{aligned}$$

where T is defined by (6.1). It is also possible to couple this system to a non-linear scalar field, but we shall focus on the above equations in what follows. There are results corresponding to Theorems 4.1 and 4.2 in this setting. We shall not write them down in detail, but it is of some interest to clarify what the initial data are.

Initial data for the Einstein-Vlasov system. For the geometry, the relevant initial data are the induced metric and second fundamental form, just as before. Since the Vlasov equation is a first order equation, we only need one initial datum for the distribution function. In order to explain how it is related to the spacetime picture, let us assume that we have a solution (M, g, f) and a spacelike hypersurface Σ in (M, g). Then there is a diffeomorphism from P_Σ (the mass shell above Σ) to $T\Sigma$ obtained by projecting orthogonally to the normal of Σ. Let us denote it by proj_Σ. The initial datum for the distribution function is given by $\bar{f} = f \circ \mathrm{proj}_\Sigma^{-1}$, and it is defined on $T\Sigma$. In the case of the Einstein-Vlasov system, the relevant initial data are $(\Sigma, \bar{g}, \bar{k}, \bar{f})$, where Σ is an n-dimensional manifold, $\bar{g}$ is a Riemannian metric on Σ, $\bar{k}$ is a symmetric covariant 2-tensor field on Σ and $\bar{f}$ is a smooth, non-negative function on $T\Sigma$. Moreover, these data should satisfy the constraint equations (4.1) and (4.2) (where the matter quantities should be expressed in terms of $\bar{g}$ and $\bar{f}$, something which can be done; cf. [20, (7.20) and (7.21), p. 92]). In order to phrase a stability result, we also need a notion of distance between initial data sets.

Distance between initial data sets. Let us assume Σ to be a closed manifold. Then we can use ordinary Sobolev norms on manifolds to measure the distance between two metrics and between two symmetric covariant 2-tensor fields. Since the tangent space of Σ is non-compact, we do, however, need a different norm to measure the difference between initial data for the distribution function. We shall use

$$\|\bar{f}\|_{H^l_{\mathrm{V}1,\mu}} = \left(\sum_{i=1}^{j} \sum_{|\alpha|+|\beta|\leq l} \int_{\bar{\mathsf{x}}_i(U_i)\times\mathbb{R}^n} \langle \bar{\varrho} \rangle^{2\mu+2|\beta|} \bar{\chi}_i(\bar{\xi}) (\partial_{\bar{\xi}}^{\alpha} \partial_{\bar{\varrho}}^{\beta} \bar{\mathsf{f}}_{\bar{\mathsf{x}}_i})^2 (\bar{\xi}, \bar{\varrho}) d\bar{\xi} d\bar{\varrho} \right)^{1/2}. \quad (6.3)$$

In this expression, $(U_i, \bar{\mathsf{x}}_i)$, $i = 1, \ldots, j$, is a covering of Σ by coordinate neighbourhoods, and $\{\bar{\chi}_i\}$ is a partition of unity subordinate to the covering $\{U_i\}$. The expression $\bar{\mathsf{f}}_{\bar{\mathsf{x}}_i}$ is the distribution function expressed with respect to the local coordinates on $T\Sigma$ induced by $(U_i, \bar{\mathsf{x}}_i)$; in particular, it is a function on $\bar{\mathsf{x}}_i(U_i) \times \mathbb{R}^n$, where the $\mathbb{R}^n$-factor corresponds to the tangential directions. Finally, we use the notation

$$\langle \bar{\varrho} \rangle = (1 + |\bar{\varrho}|^2)^{1/2}.$$

Considering the norm (6.3), there are two contributions to the power of the weight $\langle \bar{\varrho} \rangle$; 2μ and $2|\beta|$. The reason for including μ is that it yields an overall decay (assuming it to be positive). In fact, for $\mu > n/2 + 1$, the relevant matter quantities are well defined, assuming the

right hand side of (6.3) to be bounded (for a high enough l). The reason for including $2|\beta|$ is that it ensures that the notion of smallness obtained using (6.3) is geometrically meaningful; the exact value of the right hand side of (6.3) depends on the coordinates and the partition of unity, but different choices lead to equivalent norms, assuming we include $2|\beta|$ in the power of the weight. We shall refer to the space of functions $\bar{f}$ such that (6.3) is bounded for all l by $\bar{\mathfrak{D}}^{\infty}_{\mu}(T\Sigma)$ (this space can also be defined in case Σ is not compact; we then only require the integrals appearing in the definition of the norm to be bounded on compact subsets of Σ). The reader interested in a more detailed discussion of norms such as (6.3) is referred to [20]. In this reference, there is also a description of the relevant function spaces for the corresponding distribution functions on the maximal globally hyperbolic development associated with the initial data. The final ingredient we need before phrasing a stability result is a description of the relevant background solutions. We turn to this topic next.

7. Background solutions

Let us begin by describing the class of solutions to Einstein's equations which is currently preferred by physicists when modelling the universe. The geometry is taken to be spatially homogeneous and isotropic, as well as spatially flat. In other words, the relevant metrics take the form

$$g_{\mathrm{model}} = -dt \otimes dt + a^2(t)\bar{g}_{\mathrm{E}}$$

on $I \times \mathbb{R}^3$ (or $I \times \mathbb{T}^3$), where I is an open interval, $\bar{g}_{\mathrm{E}}$ is the standard Euclidean metric on $\mathbb{R}^3$, and a is a positive smooth function on I. Concerning the matter sources, they are usually taken to be a combination of so-called *perfect fluids*. In the case of a perfect fluid (and the above type of symmetry conditions), the stress energy tensor is of the form

$$T = (\rho + p)dt \otimes dt + p g_{\mathrm{model}}.$$

Here the functions ρ and p are referred to as the *energy density* and the *pressure* respectively. In order to obtain evolution equations for p and ρ, it is common to introduce an *equation of state*, giving p in terms of ρ. The condition that T be divergence free then yields an evolution equation for ρ. Two equations of state that are often used by physicists are *dust* (in which case $p = 0$) and *radiation* (in which case $p = \rho/3$). In fact, the early universe is expected to have been radiation dominated, and at late times, the matter is expected to behave as dust. Physicists often study one of these situations at at time, and then they include only dust or only radiation. However, it is possible to include both at the same time, and we shall take the matter content of the standard model to consist of a radiation fluid and dust. The corresponding stress energy tensors are required to be divergence free individually, and this yields evolution equations for the corresponding energy densities. Finally, a mechanism is required in order to produce the observed accelerated expansion. One possibility is to include a non-linear scalar field, but we shall here simply add a positive cosmological constant Λ to the above description. The relevant equations are then

$$\begin{aligned} G + \Lambda g_{\mathrm{model}} =& T_{\mathrm{rad}} + T_{\mathrm{dust}}, \\ T_{\mathrm{rad}} =& (\rho_{\mathrm{rad}} + p_{\mathrm{rad}})dt \otimes dt + p_{\mathrm{rad}} g_{\mathrm{model}}, \\ T_{\mathrm{dust}} =& \rho_{\mathrm{dust}} dt \otimes dt, \end{aligned}$$

$$\dot{\rho}_{\mathrm{rad}} = -4\frac{\dot{a}}{a}\rho_{\mathrm{rad}},$$
$$\dot{\rho}_{\mathrm{dust}} = -3\frac{\dot{a}}{a}\rho_{\mathrm{dust}},$$

where $p_{\mathrm{rad}} = \rho_{\mathrm{rad}}/3$ and G is the Einstein tensor of g_{model}. It should be pointed out that solutions of the above type are only relevant models after decoupling (i.e., the time at which matter and radiation decoupled). In particular, inflationary phases etc. are not included. The above matter models are not of Vlasov type. However, it turns out to be possible to approximate the above solutions arbitrarily well with solutions to the Einstein-Vlasov system with a positive cosmological constant and the above type of symmetry; cf. [20, Chapter 28]. Moreover, Vlasov matter is such that it naturally behaves as a radiation fluid close to the singularity and as dust in the expanding direction. In other words, it is not necessary to put in a dust and a radiation fluid by hand; Vlasov matter is such that this emerges naturally. Finally, Vlasov matter is conceptually natural in the later part of the evolution of the universe. As a consequnce, we shall prefer it here.

Spatial homogeneity. It is of interest to put the above example into a slightly bigger context, namely that of spatially homogeneous solutions. In [24], Robert Wald presented general ideas for how to analyze the future asymptotics of spatially homogeneous solutions to Einstein's equations with a positive cosmological constant (assuming the matter sources satisfy certain energy conditions). He did not address the issue of future global existence; this was taken for granted. However, he did obtain quite general results. The most fundamental ingredient of the argument is the Hamiltonian constraint (4.1). This equation can be written

$$(\mathrm{tr}_{\bar{g}}\bar{k})^2 = -\frac{3}{2}\bar{S} + \frac{3}{2}\bar{\sigma}_{ij}\bar{\sigma}^{ij} + 3\rho + 3\Lambda, \tag{7.1}$$

where $\bar{\sigma}_{ij}$ are the components of the trace free part of the second fundamental form. Assuming the matter to satisfy the *dominant energy condition* (i.e., the requirement that $T(u,v) \geq 0$ for future directed timelike vectors), the energy density ρ is non-negative. Considering (7.1), it is thus clear that the only term on the right hand side which might be negative is the first one. However, the sign of the scalar curvature of the metric induced on the hypersurfaces of spatial homogeneity is intimately connected with the symmetry type. Before describing this connection in detail, let us give a formal definition of a spatially homogeneous spacetime: it is the maximal globally hyperbolic development of homogeneous initial data (we assume the relevant matter model to be such that the initial value problem is well posed). Initial data, given by a manifold Σ, a metric $\bar{g}$, a symmetric covariant 2-tensor field, as well as matter fields, are said to be homogeneous if there is a smooth transitive Lie group action on Σ which leaves the initial data invariant. In the 3-dimensional case, there are two possibilities. Focusing on the simply connected setting for simplicity, Σ is either a Lie group or $\mathbb{S}^2 \times \mathbb{R}$. The latter case is referred to as Kantowski-Sachs in the physics community, and we shall ignore it in what follows, since the corresponding metrics have positive scalar curvature; cf. (7.1). That is not to say that it is not possible to obtain results in the Kantowski-Sachs setting, but rather that the statements of the corresponding results would be more involved. Turning to the Lie group setting, $\mathrm{SU}(2)$ constitutes a particular case; it is the only simply connected 3-dimensional Lie group which admits a left invariant metric with positive scalar curvature. Again, there are results in the $\mathrm{SU}(2)$ setting, but the statements of the results are more involved. Ignoring Kantowski-Sachs and $\mathrm{SU}(2)$ for the moment, the remaining symmetry

types are such that the corresponding invariant metrics have non-positive scalar curvature. Returning to (7.1), we conclude that the right hand side has a positive lower bound. On the other hand, since the left hand side is zero when the volume is at a local maximum (or minimum), this indicates that there is no local maximum or minimum. Naively, one would then expect that there is a big bang in one time direction and infinite expansion in the other. In fact, the corresponding solutions all have an expanding direction (which we shall refer to as the future). Moreover, it is possible to say something concerning the future asymptotics: the solution isotropizes and the matter content becomes irrelevant.

Spatially homogeneous solutions to the Einstein-Vlasov system. As mentioned above, the results in [24] were based on the assumption that the solution exists globally to the future. When studying a particular case, it thus has to be verified that this holds. In the case of the Einstein-Vlasov equations with a positive cosmological constant, this was done in [14] (the result was later extended to the case of non-compact support in the momentum directions in [20]). Moreover, asymoptotic information concerning the solution was obtained. In what follows, we wish to describe a stability result for these solutions.

8. Stability in the Einstein-Vlasov setting

Before stating the main stability result, let us define the relevant background initial data; the definition below is a specialization of [20, Definition 7.21, p. 107] to the case of the Einstein-Vlasov system with a positive cosmological constant.

Definition 8.1. Let G be a 3–dimensional Lie group and $5/2 < \mu \in \mathbb{R}$. Let $\bar{g}$ and $\bar{k}$ be a left invariant Riemannian metric and a left invariant symmetric covariant 2–tensor field on G respectively. Furthermore, let $\bar{f} \in \bar{\mathfrak{D}}^{\infty}_{\mu}(TG)$ be left invariant; in other words, if $h \in G$, then $\bar{f} \circ L_{h*} = \bar{f}$. Then $(G, \bar{g}, \bar{k}, \bar{f})$ are referred to as *Bianchi initial data* for the Einstein–Vlasov system with a positive cosmological constant, assuming they constitute initial data in the ordinary sense.

As discussed in the previous section, the corresponding solutions have an expanding direction (if the universal covering group of the Lie group is not isomorphic to $\mathrm{SU}(2)$), and it is of interest to prove global non-linear stability in that direction. It is also important to keep in mind that by letting $G = \mathbb{R}^3$ (or $G = \mathbb{T}^3$); taking $\bar{g}$ and $\bar{k}$ to be suitable multiples of the standard Euclidean metric; and by making an appropriate choice of $\bar{f}$, one obtains initial data corresponding to a solution which is consistent with observations. Future stability of solutions consistent with observations is thus a corollary of the result below. The following theorem is a specialization of [20, Theorem 7.22, p. 108] to the case of the Einstein-Vlasov system with a positive cosmological constant.

Theorem 8.2. *Let $5/2 < \mu \in \mathbb{R}$ and $(G, \bar{g}_{\mathrm{bg}}, \bar{k}_{\mathrm{bg}}, \bar{f}_{\mathrm{bg}})$ be Bianchi initial data for the Einstein–Vlasov system with a positive cosmological constant, where*

- *the universal covering group of G is not isomorphic to $\mathrm{SU}(2)$,*
- $\mathrm{tr}\bar{k}_{\mathrm{bg}} = \bar{g}^{ij}_{\mathrm{bg}} \bar{k}_{\mathrm{bg},ij} > 0$.

Assume that there is a cocompact subgroup Γ of the isometry group of the initial data. Let Σ be the compact quotient. Then the initial data induce initial data on Σ which, by abuse

of notation, will be denoted by the same symbols. Make a choice of Sobolev norms $\| \cdot \|_{H^l}$ *on tensor fields on* Σ *and a choice of norms* $\| \cdot \|_{H^l_{\mathrm{Vl},\mu}}$. *Then there is an* $\epsilon > 0$ *such that if* $(\Sigma, \bar{g}, \bar{k}, \bar{f})$ *are initial data for the Einstein–Vlasov system with a positive cosmological constant with the property that*

$$\|\bar{g} - \bar{g}_{\mathrm{bg}}\|_{H^5} + \|\bar{k} - \bar{k}_{\mathrm{bg}}\|_{H^4} + \|\bar{f} - \bar{f}_{\mathrm{bg}}\|_{H^4_{\mathrm{Vl},\mu}} \leq \epsilon,$$

then the maximal globally hyperbolic development of $(\Sigma, \bar{g}, \bar{k}, \bar{f})$ *is future causally geodesically complete.*

It is perhaps worth commenting on the requirement that there be a cocompact subgroup of the isometry group of the initial data. We expect this requirement to be unnecessary (though we have not proven this statment). However, it would then be necessary to introduce a more complicated notion of distance between initial data sets in the formulation of stability. The reason for focusing on future causal geodesic completeness in the conclusions is the physical interpretation that freely falling test particles (light) follow timelike (null) geodesics. Future causal geodesic completeness thus implies that freely falling test particles do not exit the spacetime in finite proper time to the future. In this geometric sense, the solution is thus future global. It is of course also of interest to write down estimates characterizing the asymptotic behaviour. This has been done in [20, Theorem 7.16, p. 104–106]; cf. [20, Theorem 7.22, p. 108]. We shall not repeat the technical details here.

In the presence of a positive cosmological constant, solutions are expected to homogenize and isotropize at late times. In fact, they are expected to appear de Sitter like, and this rough expectation goes under the name of the *cosmic no-hair conjecture*. A more precise formulation of this expectation is given in [2, Definition 8, p. 7]. We shall not write down the formal definition here, as it requires a somewhat technical discussion of the causal structure of solutions (the main point is to focus on the parts of the spacetime that can actually be seen by observers). However, the solutions that arise as a result of Theorem 8.2 become de Sitter like asymptotically to the future, in the sense of [2, Definition 8, p. 7].

Even though we have excluded Lie groups whose universal covers are isomorphic to $\mathrm{SU}(2)$, there are results in that setting. However, it is then necessary to impose additional conditions. An example of a result which holds when perturbing isotropic solutions is given in [20, Theorem 7.28, p. 109].

The $\mathbb{T}^3$-Gowdy symmetric setting. Beyond the above stability results concerning spatially homogeneous solutions, there are results in the $\mathbb{T}^3$-Gowdy symmetric setting. The main assumption that characterizes this symmetry class is the requirement that the initial data be invariant under a 2-torus action. In practice, the effective number of spacetime dimensions is thus 2. On the other hand, the symmetry class admits both inhomogeneities and anisotropies. Nevertheless, it turns out that solutions to the Einstein-Vlasov system in the $\mathbb{T}^3$-Gowdy symmetric setting homogeneize and isotropize. In fact, they are future asymptotically de Sitter like. Moreover, perturbing the initial data corresponding to a $\mathbb{T}^3$-Gowdy symmetric solution in the class of all solutions yields maximal globally hyperbolic developments with the same properties. The reader interested in a more detailed description is referred to [2].

9. On the topology of the universe

In Section 7, we described the solutions that physicists normally use to model the universe. Note that the justification for using them is based not only on observations, but also on the philosophical idea that all observers should see something which is roughly similar (an assumption which cannot be tested). In practice, the assumption that leads to the standard models is that every observer sees exactly the same spatially homogeneous and isotropic solution. Clearly, this is asking too much, since what we see is not exactly spatially homogeneous and isotropic. An assumption which would be slightly more reasonable would be to fix a standard model and to say that every observer should see something which is very close to that standard model. It is of interest to ask what limitations on the topology such an assumption imposes; note that the standard perspective, which implies a locally homogeneous and isotropic spatial geometry, is only consistent with a topology which is the 3-sphere, hyperbolic space or Euclidean space, or a quotient thereof. However, using methods similar to ones on which the future global non-linear stability result is based, it turns out to be possible to prove that, given

- a closed 3-manifold, say Σ,
- a standard solution (with flat spatial geometry and $\mathbb{R}^3$ spatial topology),
- a time t_0 in the existence interval of the standard solution (note that the matter models discussed here are only valid after decoupling, and we shall think of t_0 as representing decoupling),
- a choice of norm (say C^k-norm) and $\epsilon > 0$,

there is a solution to the Einstein-Vlasov system with a positive cosmological constant, such that

- it is the maximal globally hyperbolic development of initial data,
- it is future causally geodesically complete,
- it has spatial topology Σ (globally hyperbolic Lorentz manifolds have topology $\mathbb{R} \times \Sigma$, where Σ is a Cauchy hypersurface; Σ is referred to as the *spatial topology*),
- every observer considers the solution to be at distance ϵ away from the chosen standard solution to the future of t_0 and with respect to the chosen C^k-norm,
- the solution is stable with all these properties (in other words, if we perturb the corresponding initial data, we obtain a maximal globally hyperbolic development with the same properties).

Under the given assumptions, it is thus not possible to draw any conclusions concerning the topology of the universe.

The statement is still somewhat imprecise; it is not so clear how to measure the distance (as percieved by an observer) between the solution and the background solution. This is a somewhat technical issue, and we refer the interested reader to [20, Section 7.9] for a discussion.

The above description is somewhat brief, and we refer the reader interested in more details to [20, Section 7.9] for a mathematical statement of the result, and to [20, Chapter 34] for a proof.

Acknowledgements. The author would like to acknowledge the support of the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine, and the Swedish Research Council.

References

[1] Anderson, M. T., *Existence and Stability of even-dimensional asymptotically de Sitter spaces*, Ann. Henri Poincaré **6** (2005), 801–820.

[2] Andréasson, H. and Ringström, H., *Proof of the cosmic no-hair conjecture in the $\mathbb{T}^3$-Gowdy symmetric Einstein-Vlasov setting*, arXiv, (2013), http://de.arxiv.org/abs/1306.6223.

[3] Bieri, L. and Zipser, N., *Extensions of the stability theorem of the Minkowski space in general relativity,* AMS/IP Studies in Advanced Mathematics, **45**, American Mathematical Society, Providence, RI; International Press, Cambridge, MA, 2009.

[4] Choquet-Bruhat, Y. and Geroch, R., *Global aspects of the Cauchy problem in general relativity*, Commun. Math. Phys. **14** (1969), 329–335.

[5] Christodoulou, D. and Klainerman, S., *The global non-linear stability of the Minkowski space*, Princeton University Press, Princeton, N.J., 1993.

[6] Einstein, A., *Zur Elektrodynamik bewegter Körper*, Annalen der Physik **322** (10) (1905), 891–921.

[7] ———, *The Meaning of Relativity, sixth edition*, Methuen & Co. Ltd., London, 1956.

[8] Fourès-Bruhat, Y., *Théorème d'existence pour certains systèmes d'équations aux derivées partielles non linéaires*, Acta Math. **88** (1952), 141–225.

[9] Friedrich, H., *On the existence of n-geodesically complete or future complete solutions of Einstein's field equations with smooth asymtptotic structure*, Commun. Math. Phys. **107** (1986), 587–609.

[10] ———, *On the global existence and the asymptotic behavior of solutions to the Einstein–Maxwell–Yang–Mills equations*, J. Differential Geom. **34**, no. 2, (1991), 275–345.

[11] ———, *Non-zero rest-mass fields in cyclic cosmologies*, arXiv (2013), http://de.arxiv.org/abs/1311.0700.

[12] Heinzle, J. M. and Rendall, A. D., *Power-law Inflation in Spacetimes without Symmetry*, Commun. Math.Phys. **269** (2007), 1–15.

[13] Lindblad, H. and Rodnianski, I., *The global stability of Minkowski space-time in harmonic gauge*, Ann. of Math. (2) **171** (2010), no. 3, 1401–1477.

[14] Lee, H., *Asymptotic behaviour of the Einstein–Vlasov system with a positive cosmological constant*, Math. Proc. Camb. Phil. Soc. **137** (2004), 495–509.

[15] Lovelock, D., *The Einstein Tensor and Its Generalizations*, J. Mathematical Phys. **12** (1971), 498–501.

[16] Luo, X. and Isenberg, J., *Power Law Inflation with Electromagnetism*, arXiv (2012), http://de.arxiv.org/abs/1210.7566

[17] Ringström, H., *Future stability of the Einstein non-linear scalar field system*, Invent. math. **173** (2008), 123–208.

[18] ———, *The Cauchy Problem in General Relativity*, European Mathematical Society, Zürich, 2009.

[19] ———, *Power law inflation*, Commun. Math. Phys. **290** (2009), 155–218.

[20] ———, *On the Topology and Future Stability of the Universe*, Oxford University Press, Oxford, 2013.

[21] Rodnianski, I. and Speck, J., *The nonlinear future stability of the FLRW family of solutions to the irrotational Euler-Einstein system with a positive cosmological constant*, J. Eur. Math. Soc. (JEMS) **15**, no. 6, (2013), 2369–2462.

[22] Speck, J., *The nonlinear future stability of the FLRW family of solutions to the Euler-Einstein system with a positive cosmological constant*, Selecta Math. (N.S.) **18**, no. 3, (2012), 633–715.

[23] Svedberg, C., *Future Stability of the Einstein–Maxwell–Scalar Field System*, Ann. Henri Poincaré **12**, No. 5, (2011), 849–917.

[24] Wald, R., *Asymptotic behaviour of homogeneous cosmological models in the presence of a positive cosmological constant*, Phys. Rev. D **28** (1983), 2118–2120.

Department of Mathematics, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden
E-mail: hansr@kth.se

Solitons in geometric evolution equations

Natasa Sesum

Abstract. We will discuss geometric properties and classification of special solutions to geometric evolution equations called solitons. Our focus will be on the Ricci flow and the Yamabe flow solitons. These are very special solutions to considered geometric evolution equations that move by diffeomorphisms and homotheties. Solitons are very important solutions to our equations because very often they arise as singularity models. Therefore classifying the solitons helps us understand and classify encountered singularities in geometric flows.

Mathematics Subject Classification (2010). 53C44.

Keywords. Ricci flow, Yamabe flow, solitons.

1. Introduction

We will focus on two geometric evolution equations. The first one is the Ricci flow equation, that is,

$$\begin{aligned} \frac{\partial}{\partial t} g_{ij} &= -2R_{ij}, \\ g(\cdot, 0) &= g_0(\cdot), \end{aligned} \tag{1.1}$$

where (M, g_0) is an arbitrary Riemannian manifold. The second one is the Yamabe flow equation, that is,

$$\begin{aligned} \frac{\partial}{\partial t} g_{ij} &= -R g_{ij}, \\ g_0(\cdot, 0) &= g_0(\cdot). \end{aligned} \tag{1.2}$$

In both geometric equations a singularity is most likely to occur in finite time. The singularity in both flows is characterized by the norm of the curvature operator blowing up at a singular time. In order to understand any of those singularities one defines a sequence of parabolic dilations in a space-time neighborhood of a considered singularity. The limit of that sequence is called a *singularity model* and turns out to be either an *ancient* solution, which means a solution that lives from $(-\infty, T)$ where $T < \infty$, or an *eternal* solution, which is a solution that lives for all times, from $(-\infty, \infty)$. Hence, understanding the ancient and eternal solutions help us understand the singularities in geometric flows, which is crucial if we want to use the flow in order to better understand topological and geometric properties of our manifold M.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Special examples of ancient and eternal solutions are given by *solitons*. These are the solutions to the considered flow equations that move just by diffeomorphisms and homotheties. In this article we focus on solitons, their classification if possible and also their geometric and topological properties.

2. Ricci solitons

There has been a lot of interest in Ricci solitons recently. They are natural generalizations of Einstein metrics and play important role in the singularity analysis of the Ricci flow. The concept of Ricci solitons was introduced by Hamilton in mid 80's.

Definition 2.1. A complete Riemannian metric g on a smooth Riemannian manifold M is called a *Ricci soliton* if there exists a smooth vector field V such that

$$R_{ij} + (\mathcal{L}_V g)_{ij} = \lambda g_{ij},$$

for some constant λ and $(\mathcal{L}_V g)_{ij} = \frac{1}{2}(\nabla_i V_j + \nabla_j V_i)$. If $V = \nabla f$ for some smooth function f then

$$R_{ij} + \nabla_i \nabla_j f = \lambda g_{ij}, \tag{2.1}$$

and we say we have a *gradient Ricci soliton*. We call it *expanding*, *steady* or *shrinking*, depending on whether $\lambda < 0$, $\lambda = 0$ or $\lambda > 0$, respectively.

By a suitable rescaling of metric g, we can normalize $\lambda = -\frac{1}{2}, 0$ or $\frac{1}{2}$. Below we focus on gradient Ricci solitons and we will be denoting them as a triple (M, g, f) and refer to a function f as to the *potential* function.

2.1. Examples of Ricci solitons. There are not so many examples of Ricci solitons so far. Below we describe a few well known examples.

1. **(Gaussian solitons)** (R^n, g) with the flat Euclidean metric can be equipped with both, the shrinking and the expanding Ricci soliton. More precisely $(\mathrm{R}^n, g, \frac{|x|^2}{4})$ is a gradient Ricci shrinker and $(\mathrm{R}^n, g, -\frac{|x|^2}{4})$ is a gradient Ricci expander.
2. The generalized round cylinders $S^k \times \mathrm{R}^{n-k}$ are gradient Ricci shrinkers.
3. **(The cigar soliton)** In dimension two, Hamilton [32] discovered the first example of complete, noncompact steady soliton on R^2, called the *cigar soliton*, where

$$ds^2 = \frac{dx^2 + dy^2}{1 + x^2 + y^2}, \qquad \text{with} \quad f = -\log(1 + x^2 + y^2).$$

4. **(The Bryant soliton)** Higher dimensional ($n \geq 3$) noncompact gradient steady solitons on R^n were discovered by Bryant. They are rotationally symmetric, have positive sectional curvature and the geodesic sphere S^{n-1} of radius s has the diameter of order $\sqrt{s}$. Recently Brendle ([5]) showed they are the only steady κ-noncollapsed solitons with positive sectional curvature (noncollapsing here is in the sense of Perelman [44]).

2.2. Ricci solitons as singularity models. In order to study the Ricci flow and its singularity models Perelman ([44]) introduced the $\mathcal{W}$-functional

$$\mathcal{W}(g,f,\tau)=\int_M [\tau(R+|\nabla f|^2)+f-n](4\pi\tau)^{-\frac{n}{2}}e^{-f}\,dV_g,$$

where g is a Riemannian metric, f is a smooth function on M^n and τ a positive scale parameter. It is easy to see that the functional $\mathcal{W}$ is invariant under scaling and diffeomorphisms, that is, $\mathcal{W}(\alpha\phi^*g,\phi^*f,\alpha\tau)=\mathcal{W}(g,f,\tau)$, for any positive number α and any diffeomorphism ϕ. Perelman showed the Ricci flow is up to diffeomorphisms the gradient flow of functional $\mathcal{W}$, which is also monotone along the flow. More precisely, he finds that

$$\frac{d}{dt}\mathcal{W}=2\tau\int_M |R_{ij}+\nabla_i\nabla_j f-\frac{g_{ij}}{2\tau}|^2(4\pi\tau)^{-\frac{n}{2}}e^{-f}\,dV,$$

under the flow

$$\frac{\partial}{\partial t}g_{ij}=-2R_{ij},\qquad \frac{\partial f}{\partial t}=-\Delta f+|\nabla f|^2-R+\frac{n}{2\tau},\qquad \dot\tau=-1.$$

It is now even more apparent that Ricci solitons whose equation appears in the evolution of the monotone quantity $\mathcal{W}$ play an important role in the study of the Ricci flow. The solitons are related to the Li-Yau-Hamilton inequality, also called differential Harnack estimate. More precisely, the Li-Yau-Hamilton quantity vanishes on expanding solitons. More importantly, Ricci solitons often appear as blow-up limits around the singularities in the Ricci flow.

A complete solution $g(t)$ to (1.1) is called *ancient* if it is defined for $-\infty<t<T$. It turns out that Type I and Type II singularity models are ancient (those singularity models are defined to be the limits of dilations of solutions to the Ricci flow around their Type I and Type II space-time singularities, respectively, which are to be defined below). Steady and shrinking Ricci solitons are special examples of ancient solutions. Using Perelman's monotonicity formula for $\mathcal{W}$ in [48] we showed that if a singularity model is compact, it has to be a gradient shrinking Ricci soliton. In [42], Naber removed our assumption that a singularity model is compact but did not show the singularity model is always non-flat. Nonflatness of the singularity model has been obtained in [29] and [13] independently.

Consider a solution $(M,g(\cdot,t))$ to (1.1) for $t\in[0,T)$ and $T\le\infty$, where either M is compact or at each time the metric is complete and has bounded curvature. Hamilton [34] showed that if T is a *maximal* singularity time then either $T=\infty$ or the curvature tensor $|Rm|$ is unbounded as $t\to T$. More precisely we have showed that

Theorem 2.2 ([47]). *Let $(M,g(\cdot,t))$ be a compact solution to* (1.1) *for $t\in[0,T)$ and $T\le\infty$. Then if T is a maximal singularity time we have either $T=\infty$ or the Ricci curvature tensor $|Ric|$ is unbounded as $t\to T$.*

Intuitively it is clear that if we take a dumbbell metric in S^3 with a neck like $S^2\times B^1$, we expect the neck will shrink because the positive curvature in the S^2 direction will dominate the slightly negative curvature in the B^1 direction. These intuitive picture of a neck pinching off in some finite time and having the sequence of dilations around the singularity converging to a round infinite cylinder $S^2\times\mathrm{R}$ was justified in [1] (they actually showed the same phenomenon occurs in higher dimensions as well). They considered an open set of initial

metrics of *rotationally symmetric* metrics evolving on S^{n+1}. These are **Type I** singularities, meaning that

$$\limsup_{t\to T}\sup_{M}(T-t)\,|Rm(\cdot,t)|<\infty.$$

If the above quantity is equal to infinity we say we have a **Type II** singularity. In [37] we have recently considered the generalized Berger warped product metrics on $S^1\times S^3$ and we have constructed open sets of initial metrics (not necessarily rotationally symmetric) of the form

$$g=ds^2+f(s)^2w^1\otimes w^1+g(s)^2\,(w^2\otimes w^2+w^3\otimes w^3),$$

where s is the arc length from an arbitrary, but a fixed point s_0 on S^1 and where w^1,w^2,w^3 is a coframe, algebraically dual to a fixed Milnor frame on $S^3=SU(2)$. The behavior of the Ricci flow around the singularities is the same as the one in the intuitive picture of a Ricci flow behavior on a dumbbell. More precisely we have the following result.

Theorem 2.3 ([37]). *The eccentricity of every warped Berger solution of Ricci flow is uniformly bounded: there exists C_0 depending only on the initial data such that the estimate*

$$|f-g|\le C_0\min\{f,g\} \tag{2.2}$$

holds pointwise for as long as the solution exists, without additional assumptions.

(i) *There exist open sets of warped Berger metrics satisfying certain assumptions such that all solutions originating in these sets develop local neckpinch singularities at some $T<\infty$. Each such solution has the properties that*

(a) *the ordering $f\le g$ is preserved;*

(b) *the singularity is Type-I, with $|Rc|\le C\big(\min f(\cdot,t)\big)^{-2}$, and*

$$\frac{1}{C}\sqrt{T-t}\le\min f(\cdot,t)\le C\sqrt{T-t};$$

(c) *the diameter is bounded as $t\nearrow T$.*

(ii) *There exist open sets of warped Berger metrics satisfying certain assumptions such that as solutions originating in these sets become singular, they become asymptotically round at rates that break scale invariance. Specifically, in addition to the properties above, they satisfy the following C^0, C^1, and C^2 bounds at the neck:*

$$(T-t)^{-1/2}|f-g|\le C\sqrt{T-t}, \tag{2.3}$$

$$(T-t)|\kappa_{12}-\kappa_{23}|\le C\sqrt{T-t}, \tag{2.4}$$

$$(T-t)|\kappa_{01}-\kappa_{02}|\le C\sqrt{T-t}. \tag{2.5}$$

In a neighborhood of each smallest neck, where $\kappa_{01}<0$, there is the further bound

$$(T-t)\,\big(|\kappa_{01}|+|\kappa_{02}|\big)\le\frac{C}{|\log(T-t)|}, \tag{2.6}$$

where $\kappa_{12},\kappa_{31},\kappa_{23}$ are the curvatures of the corresponding vertical planes in the total space and $\kappa_{01},\kappa_{02},\kappa_{03}$ are the curvatures of mixed vertical-horizontal planes in the total space. As a corollary of the cylindrical estimate (2.6) *we have that Type-I blowups $\tilde G=(T-t)^{-1}G$ of the solution converge near each neck to the shrinking cylinder soliton.*

We say the examples constructed in [1] and [37] develop *non-degenerate neckpinch* singularities and the round cylinder, which is an example of a Ricci shrinking soliton, is their *singularity model*. In [34] Hamilton also described the intuitive picture of a degenerate neckpinch singularity. Imagine the dumbbell is not symmetric and that one side is much bigger than the other. If we choose the sizes of the spheres on both sides appropriately, we expect a *degenerate* singularity, which means pinching off the little sphere and there is nothing left on one side. In [31] it was verified that such a degenerate neck-pinching Type II singularity can be formed on S^n with suitable rotationally symmetric metric for all $n \geq 3$. The precise singularity formation of those Type II singularities was discussed in [2]. One of the conclusions is that the sequence of dilations around this Type II singularity converges to the Bryant soliton.

2.3. Classification results for Ricci solitons. As we have discussed above, gradient shrinking Ricci solitons arise as singularity models of Type I singularities of the Ricci flow and that is why understanding of those is important in studying the singularities of the Ricci flow. Classification of gradient shrinking Ricci solitons has been a subject of interest for many people.

Hamilton [32] showed that the only closed gradient shrinking Ricci solitons in two dimensions are Einstein. In the case of three dimensions, Ivey proved that all compact, gradient shrinking Ricci solitons must have constant positive curvature. The recent work of Böhm and Wilking [4] implies the compact gradient shrinking Ricci solitons with positive curvature operator in any dimension have to be of constant curvature, generalizing Ivey's result. Koiso ([38]), H.-D.Cao [8], Feldman, Ilmanen and Knopf [30] constructed examples of both, compact and complete non-compact gradient shrinking Ricci solitons that are not Einstein. Some properties of compact shrinking Ricci solitons have been proved in [28].

The classification of complete, noncompact gradient shrinking Ricci solitons has been recently studied by many people. The Hamilton-Ivey estimate shows that those three dimensional solitons have nonnegative sectional curvatures. Combining this with the results of Perelman yields that the three dimensional gradient shrinking solitons with bounded sectional curvatures are S^3, $\mathbb{R}^3$, $S^2 \times \mathbb{R}$ and the quotients of those. Recently Ni and Wallach ([43]) studied the classification of complete gradient shrinking Ricci solitons with vanishing Weyl curvature tensor, in any dimension, under the assumptions of nonnegative Ricci curvature and at most exponential growth of the norm of curvature operator. They showed under their assumptions we can only have S^n, $\mathbb{R}^n$, $S^{n-1} \times \mathbb{R}$ and the quotient of those. In [14] the assumption on nonnegative Ricci curvature has been relaxed to having the Ricci curvature bounded from below. In [45] Petersen and Wylie obtained the same classification under a different assumption. Besides the vanishing of the Weyl tensor they assumed a certain integral bound involving the potential function f,

$$\int_M |Ric|^2 e^{-f}\, dV < \infty. \tag{2.7}$$

The question whether certain integral curvature estimates including (2.7) for complete gradient shrinking Ricci solitons are true has been raised in [45] and [14]. The motivation, as pointed out also in [9], is to prove a classification result for complete gradient shrinkers which are locally conformally flat, thus extending a theorem of Perelman. In [50] Zhang proved that gradient shrinking Ricci solitons with vanishing Weyl tensor must have nonnegative curvature operator, which proved the classification of such solitons as finite quotients

of $\mathbb{R}^n$, $\mathbb{S}^{n-1}\times\mathbb{R}$ or $\mathbb{S}^n$.

Besides being interesting on their own, proving that certain curvature integral quantities (including (2.7)) are finite would have as a consequence an alternate, simpler proof for this classification. In [39] we prove the following.

Theorem 2.4. *Let M^n be a complete gradient shrinking Ricci soliton normalized such that*

$$Ric + Hess_f = \frac{1}{2}g$$

Then we have

$$\int_M |Ric|^2 e^{-f} < \infty.$$

Moreover, for any $\lambda > 0$ we have

$$\int_M |Ric|^2 e^{-\lambda f} < \infty.$$

As a consequence of Theorem 2.4 we have the following classification result for complete gradient shrinking Ricci solitons.

Theorem 2.5. *Any $n-$dimensional complete shrinking gradient Ricci soliton with harmonic Weyl tensor (meaning $div\, W = 0$) is a finite quotient of $\mathbb{R}^n$, $\mathbb{S}^{n-1}\times\mathbb{R}$ or $\mathbb{S}^n$.*

In [39] we show that the volume of a gradient steady soliton is infinite. More precisely, we show there are uniform constants $c, r_0 > 0$ so that for any $r > r_0$,

$$Vol(B_p(r)) \geq cr.$$

In [13] it was proved that a complete non-compact gradient shrinking Ricci soliton has at most Euclidean volume growth. In [40] was obtained that the volume of a non-compact shrinking Ricci soliton must be of at least linear growth and this is the best we can expect due to the examples of shrinking cylinders on which the volume of geodesic balls grow exactly linearly. This linear lower bound on volume was obtained as an application of the spectral analysis of weighted laplacian to the study of the geometry and topology of gradient Ricci solitons.

Under a restrictive assumption on scalar curvature we also show that a gradient shrinking Kähler-Ricci soliton is connected at infinity, that is, it has one end. In a recent work [41], Munteanu and Wang show that any shrinking Kähler-Ricci soliton must be connected at infinity.

Recall that a gradient shrinking Ricci soliton (M, g, f) satisfies the equation (after rescaling)

$$R_{ij} + \nabla_i\nabla_j f = \frac{1}{2}g_{ij}.$$

An important ingredient we used in [39] and that has been used in some of above mentioned results is the asymptotic behavior of the potential function f. It has been showed in [13] that the potential function f satisfies the estimate

$$\frac{1}{4}(r(x) - c_1)^2 \leq f(x) \leq \frac{1}{4}(r(x) + c_2)^2,$$

where $r(x) = dist(x_0, x)$ is the distance from some fixed point $x_0 \in M$, c_1 and c_2 are positive constants depending only on n and the geometry of g_{ij} on a unit ball $B_{x_0}(1)$.

2.4. Eternal solutions in two dimensions. In [22] we have considered *eternal* solutions (the ones that live for all $t \in \mathrm{R}$) of the logarithmic fast diffusion equation

$$\frac{\partial u}{\partial t} = \Delta \log u \qquad \text{on } \mathrm{R}^2 \times \mathrm{R}. \tag{2.8}$$

This equation represents the evolution of the conformally flat metric $g_{ij} = u\, I_{ij}$ under the *Ricci Flow*

$$\frac{\partial g_{ij}}{\partial t} = -2\, R_{ij}.$$

The equivalence follows from the observation that the metric $g_{ij} = u\, I_{ij}$ has scalar curvature $R = -(\Delta \log u)/u$ and in two dimensions $R_{ij} = \frac{1}{2}\, R\, g_{ij}$.

In [21] we introduced the width w of the metric $g = u\, I_{ij}$. Let $F : \mathrm{R}^2 \to [0, \infty)$ denote a proper function F, such that $F^{-1}(a)$ is compact for every $a \in [0, \infty)$. The width of F is defined to be the supremum of the lengths of the level curves of F, namely $w(F) = \sup_c L\{F = c\}$. The width w of the metric g is defined to be the infimum

$$w(g) = \inf_F w(F).$$

Theorem 2.6. *Assume that u is a positive smooth eternal solution of equation* (2.8) *which defines a complete metric and satisfies conditions*

$$w(g(t)) < \infty, \qquad \forall t \in \mathrm{R}, \tag{2.9}$$

$$\|R(\cdot, t)\|_{L^\infty(\mathrm{R}^2)} < \infty, \qquad \forall t \in \mathrm{R}. \tag{2.10}$$

Then, u is a gradient soliton of the form

$$U(x,t) = \frac{2}{\beta\,(|x - x_0|^2 + \delta\, e^{2\beta t})} \tag{2.11}$$

for some $x_0 \in \mathrm{R}^2$ and some constants $\beta > 0$ and $\delta > 0$. It is known as the cigar soliton on R^2.

It is shown in [21] that maximal solutions u of the initial value problem $u_t = \Delta \log u$ on $\mathrm{R}^2 \times [0, T)$, $u(x, 0) = f(x)$ which vanish at time $T < \infty$ satisfy the width bound $c\,(T - t) \le w(g(t)) \le C\,(T - t)$ and the maximum curvature bound $c\,(T - t)^{-2} \le R_{\max}(t) \le C\,(T - t)^{-2}$ for some constants $c > 0$ and $C < \infty$, independent of t. Hence, one may rescale u near $t \to T$ and pass to the limit to obtain an eternal solution of equation (2.8) which satisfies the bounds (2.9) and (2.10). Theorem 2.6 provides then a classification of the limiting solutions.

The bounded width assumption (2.9) is necessary for the conclusion of the Theorem. If this condition is not satisfied, then (2.8) admits for example the flat (constant) solutions. This has been discussed in [18]. He shows that all eternal solutions with bounded curvature at each time slice are either the plane or the cigar soliton.

In [25] we consider ancient compact solutions to the Ricci flow in dimension two. We show the following result.

Theorem 2.7. *Let $g(\cdot, t)$ be an ancient compact two dimensional solution to the Ricci flow. Then, it is either one of the contracting spheres or one of the King-Rosenau solutions.*

3. Yamabe solitons

We start with the definition of a *Yamabe soliton*.

Definition 3.1. A Riemannian manifold (M^n, g_{ij}) is called a Yamabe gradient soliton if there exists a smooth scalar (potential) function $f : M^n \to \mathrm{R}$ and a constant $\rho \in \mathbb{R}$ such that

$$(R - \rho)\, g_{ij} = \nabla_i \nabla_j f. \tag{3.1}$$

If $\rho > 0$, $\rho < 0$ or $\rho = 0$, then g is called a Yamabe shrinker, Yamabe expander or Yamabe steady soliton respectively. By scaling the metric, we may assume with no loss of generality that $\rho = 1, -1, 0$ respectively.

When f is a constant function in (3.1) we say that the corresponding Yamabe soliton is a trivial Yamabe soliton. It has been known (see [17], [27], [36]) that every compact Yamabe soliton is of constant scalar curvature, hence trivial, since in this case the potential function f turns out to be constant.

Yamabe solitons are special solutions to the Yamabe flow

$$\frac{\partial}{\partial t} g_{ij} = -R\, g_{ij}. \tag{3.2}$$

This flow was introduced by R. Hamilton [35] as an approach to solve the *Yamabe problem* on manifolds of positive conformal Yamabe invariant. It is the negative L^2-gradient flow of the total scalar curvature, restricted to a given conformal class. Hamilton [35] showed the existence of the normalized Yamabe flow (which is the re-parametrization of (3.2) to keep the volume fixed) for all time; moreover, in the case when the scalar curvature of the initial metric is negative, he showed the exponential convergence of the flow to a metric of constant scalar curvature.

Since then, a number of works have been established on the convergence of the Yamabe flow on a compact manifold to a metric of constant scalar curvature. Chow [16] showed the convergence of the flow under the conditions that the initial metric is locally conformally flat and of positive Ricci curvature. The convergence of the flow for any locally conformally flat initial metric was shown by Ye [49]. Inspired by this result, Del Pino and Saez [26] proved the convergence to the sphere of a conformally flat metric on $\mathbb{R}^n$ evolving by the Yamabe flow and satisfying a decay condition at infinity.

More recently, Schwetlick and Struwe [46] obtained the convergence of the Yamabe flow on a general compact manifold under a suitable Kazdan-Warner type of condition that rules out the formation of bubbles and that is verified (via the positive mass Theorem) in dimensions $3 \leq n \leq 5$. The convergence result for any general compact manifold was established by Brendle [6] and [7] (up to a technical assumption, in dimensions $n \geq 6$, on the rate of vanishing of Weyl tensor at the points at which it vanishes): starting with any smooth metric on a compact manifold, the normalized Yamabe flow converges to a metric of constant scalar curvature.

Even though the analogue of Perelman's monotonicity formula is still lacking for the Yamabe flow, one expects that Yamabe soliton solutions model finite time singularities. One of results in [23] indicates that in certain cases of Type II singularities one may expect the steady Yamabe solitons to be the singularity models.

Although the Yamabe flow on compact manifolds is well understood, the complete non-compact case is unsettled. In [24] the authors showed that in the conformally flat case and

under certain conditions on the initial data, which in particular imply that the initial metric admits the asymptotic behavior of the cylindrical metric at infinity, complete non-compact solutions to the Yamabe flow develop a finite time singularity and after re-scaling the metric converges to the Barenblatt solution (a certain type of a shrinker, corresponding to the Type I singularity). The general case even when the solution is conformally equivalent to R^n is not well understood. In fact in [24] we show there exist infinitely many shrinking solitons which behave as cylinders at infinity and they are all different than the Barenblatt solution (which unlike other Yamabe shrinkers is given in the explicit form).

All such solutions are prototypes of Type I singularities of the complete non-compact Yamabe flow.

3.1. Classification of Yamabe solitons. One of our results in [24] establishes the rotational symmetry of locally conformally flat Yamabe solitons.

Theorem 3.2 (Rotational symmetry of Yamabe solitons). *All locally conformally flat complete Yamabe gradient solitons with positive sectional curvature have to be rotationally symmetric.*

Our proof of Theorem 3.2 is inspired by the proof of the analogous theorem for complete gradient steady Ricci solitons in [10]. Some time after posting our paper [24], a related work by Cao, Sun and Zhang [11] was posted. Inspired by our work, it was shown in [11] that every complete nontrivial gradient Yamabe soliton admits a special global warped product structure with an one-dimensional base. Consequently, locally conformally flat complete gradient Yamabe solitons with nonnegative Ricci curvature are rotationally symmetric. There is also a related work by Catino, Mantegazza and Mazzieri [15].

One of the goals in [24] was to classify rotationally symmetric Yamabe solitons on R^n. In order to state the theorem which deals with that question lets first state the following proposition.

Proposition 3.3 (PDE formulation of Yamabe solitons). *Let $g_{ij} = u^{\frac{4}{n+2}}\, dx^2$ be a conformally flat rotationally symmetric Yamabe gradient soliton. Then, u is a smooth solution to the elliptic equation*

$$\frac{n-1}{m}\,\Delta u^m + \beta\, x\cdot\nabla u + \gamma\, u = 0, \qquad \text{on } \mathrm{R}^n \tag{3.3}$$

where $\beta \geq 0$ and

$$\gamma = \frac{2\beta+\rho}{1-m}, \qquad m = \frac{n-2}{n+2}.$$

In the case of expanders $\beta > 0$. In addition, any smooth solution to the elliptic equation (3.3) *with β and γ as above defines a gradient Yamabe soliton.*

To simplify the notation, we will assume from now on that $\rho = 1$ in (3.1) in the case of the Yamabe shrinkers, and that $\rho = -1$ in the case of the Yamabe expanders. This can be easily achieved by scaling our metric g. The following result provides the classification of radially symmetric and smooth solutions of the elliptic equation (3.3).

Theorem 3.4 (Classification of radially symmetric Yamabe solitons). *Let $m = (n-2)/(n+2)$. The elliptic equation* (3.3) *admits non-trivial radially symmetric smooth solutions if and only if $\beta \geq 0$ and $\gamma := (2\beta+\rho)/(1-m) > 0$. More precisely, we have:*

(1) **Yamabe shrinkers** $\rho = 1$**:** *For any* $\beta > 0$ *and* $\gamma = (2\beta+1)/(1-m)$*, there exists an one parameter family* u_λ*,* $\lambda > 0$*, of smooth radially symmetric solutions to equation* (3.3) *on* R^n *of slow-decay rate at infinity, namely* $u_\lambda(x) = O(|x|^{-2/(1-m)})$ *as* $|x| \to \infty$*. This asymptotic behavior of* u_λ *gives the asymptotic cylindrical behavior of the corresponding metric* $g_\lambda = u_\lambda^{\frac{4}{n+2}}\, dx^2$*. We will refer to those solutions as to cigar solutions. In the case* $\gamma = \beta n$ *the solutions are given in the closed form*

$$u_\lambda(x) = \left(\frac{C_n}{\lambda^2 + |x|^2}\right)^{\frac{1}{1-m}}, \qquad C_n = (n-2)(n-1) \tag{3.4}$$

and will refer to them as to the Barenblatt solutions. When $\beta = 0$ *and* $\gamma = 1/(1-m)$ *equation* (3.3) *admits the explicit solutions of fast-decay rate*

$$u_\lambda(x) = \left(\frac{C_n\,\lambda}{\lambda^2 + |x|^2}\right)^{\frac{2}{1-m}}, \qquad C_n = (4n(n-1))^{\frac{1}{2}}. \tag{3.5}$$

We will refer to them as to the spheres.

(2) **Yamabe expanders** $\rho = -1$**:** *For any* $\beta > 0$ *and* $\gamma = (2\beta-1)/(1-m) > -1/(1-m)$ *there exists an one parameter family* u_λ*,* $\lambda > 0$*, of smooth radially symmetric solutions to equation* (3.3) *on* R^n*.*

(3) **Yamabe steady solitons** $\rho = 0$**:** *For any* $\beta > 0$ *and* $\gamma = 2\beta/(1-m) > 0$ *there exists an one parameter family* u_λ*,* $\lambda > 0$*, of smooth solutions to equation* (3.3) *on* R^n *which satisfy the asymptotic behavior* $u_\lambda(x) = O((\log|x|/|x|^2)^{1/(1-m)})$*, as* $|x| \to \infty$*. We will refer to them as to logarithmic cigars. For* $\beta = \gamma = 0$*, the solution* u_λ *is a constant, defining the euclidean metric on* $\mathbb{R}^n$*.*

In all of the above cases the solution u_λ *is uniquely determined by its value at the origin.*

Most of the Yamabe solitons we find in Theorem 3.4 have nonnegative sectional curvature.

3.2. Yamabe solitons as singularity models. We consider a complete non-compact metric $g = u^{4/(N+2)}\, dx^2$ which is conformally equivalent to the standard euclidean metric of R^N and evolves by the *Yamabe flow*

$$\frac{\partial g}{\partial t} = -R\, g \tag{3.6}$$

where R denotes the scalar curvature with respect to metric g. Our goal is to study the *singularity formation* of metric g at a singular time T, under the assumption that the initial metric g_0 has *cylindrical behavior* at infinity.

By observing that the conformal metric $g = u^{4/(N+2)}\, dx^2$ has scalar curvature

$$R = -\frac{4(N-1)}{N-2}\, u^{-1}\, \Delta u^{\frac{N-2}{N+2}}$$

it follows that the function u evolves by the fast diffusion equation $u_t = \frac{N-1}{m}\, \Delta u^m$, with exponent $m = (N-2)/(N+2)$. Therefore studying the Yamabe flow equation (3.6) in the conformally flat case is equivalent to studying the fast diffusion equation on R^N.

We would like to relate the singularity profile of conformally flat solutions to the Yamabe flow whose conformal factors have *cylindrical behavior at infinity* with a class of self-similar

shrinking Yamabe solitons that have matched asymptotic behavior at infinity. One special result in this direction was previously shown for example in [24] where the L^1 stability around the explicit Barenblatt profile was established (Barenblatt solution is a complete Yamabe shrinker given by an explicit formula as in Proposition 3.4).

We will assume that the initial metric $g_0 = u_0^{4/(N+2)}\, dx_i dx_j$ is complete, non-compact and has *cylindrical behavior* at infinity, namely

$$u_0(x) = \left(\frac{C^* T}{|x|^2}\right)^{1/n} \big(1 + o(1)\big), \qquad \text{as } |x| \to \infty \tag{3.7}$$

with C^* given by

$$C^* := \frac{2\,\big(((1-m)N - 2\big)}{n}, \qquad n = 1 - m, \quad m = \frac{N-2}{N+2} \tag{3.8}$$

and $T > 0$ any positive constant. In [19] we show that if the initial data $u_0(x)$ satisfies (3.7) then for the solution u we have

$$u(x,t) = \left(\frac{C^*(T-t)}{|x|^2}\right)^{1/n} (1 + o(1)), \qquad \text{as } |x| \to \infty. \tag{3.9}$$

We have seen in [19] that the solution u starting at u_0 that satisfies (3.7) *may or may not become extinct at time* T, depending on the *second order asymptotic behavior*, as $|x| \to \infty$, of the cylindrical tail of the initial data. In either case the metric $g(t) = u^{4/(N+2)}(\cdot, t)\, dx^2$ will develop a singularity at time T. Our goal is to study these singularities. We showed that rescaled limits of solutions u with initial condition satisfying (3.7) behave near a singularity at time T as self-similar shrinking solutions (Yamabe shrinkers). These can be viewed as special solutions of the fast-diffusion equation

$$u_t = \Delta u^{\frac{N-2}{N+2}} \tag{3.10}$$

of the form

$$U(x,t) = (T-t)^{\alpha}\, f(y), \quad y = x\,(T-t)^{\beta}, \qquad \alpha = \frac{1+2\beta}{n}, \;\; \beta > 0, \tag{3.11}$$

where the function f satisfies the elliptic equation

$$\Delta f^{\frac{N-2}{N+2}} + \beta\, y \cdot \nabla f + \alpha\, f = 0. \qquad \text{on } \mathrm{R}^N \tag{3.12}$$

In order to study the singularities of a metric $g = u^{4/(N+2)} dx^2$ evolving by (2.8) and with initial data satisfying (3.7) we need to understand the second order asymptotic behavior at infinity of the self-similar profiles f_λ. We will achieve this by linearizing equation (3.12) around the cylindrical solution.

Let $\gamma_{1,2}$ be the solutions to the characteristic equation of the corresponding linearized equation. They satisfy

$$\gamma^2 + \beta(N-2)\gamma + (N-2) = 0, \tag{3.13}$$

which gives

$$\gamma_{1,2} = \frac{\beta(N-2) \mp \sqrt{\beta^2(N-2)^2 - 4(N-2)}}{2}. \tag{3.14}$$

We see that we need to have $\beta \geq 2/\sqrt{N-2}$ in order for $\gamma_{1,2}$ to be real and the corresponding solution to have non-oscillatory behavior.

One of our results in [19] concerns the second order asymptotics of smooth profiles f on R^N which appear to model the singular behavior of some evolving metrics $g = u^{4/(N+2)}dx^2$ that become extinct at a singular time T. We already know the first order asymptotics is given by a cylindrical behavior at infinity.

Theorem 3.5. *Let $m = (N-2)/(N+2)$, $n = 1-m$, $N \geq 3$, $C^* = 2\left((1-m)N-2\right)/n$, $\beta_0 := 2/\sqrt{N-2}$ and $\beta_1 := 1/(2m)$ (this particular parameter corresponds to having the Barenblatt solution given by* (3.4)*. The following hold:*

- *Let $N \geq 6$ and $\beta > \beta_0$ or $3 \leq N < 6$ and $\beta > \beta_1$: For any $B > 0$ there exists a unique radially symmetric smooth solution $f_{\beta,B}$ of* (3.12) *that satisfies*

$$f_{\beta,B}(y) = \left(\frac{C^*}{|y|^2}\right)^{1/n} \left(1 - B\,|y|^{-\gamma} + o_B(|y|^{-\gamma})\right) \tag{3.15}$$

  with $\gamma = \gamma_1$ given by (3.14)*.*
- *Let $3 \leq N < 6$ and $\beta_0 < \beta < \beta_1$: For any $B < 0$ there exists a unique radially symmetric smooth solution $f_{\beta,B}$ of* (3.12) *that satisfies* (3.15) *with $\gamma = \gamma_1$ given by* (3.14)*.*
- *Let $3 \leq N < 6$ and $\beta = \beta_1$: For any $B < 0$ there exists a unique radially symmetric smooth solution $f_{\beta,B}$ of* (3.12) *that satisfies* (3.15) *with $\gamma = \gamma_2 = 2$ and which is given in closed form by* (3.4)*.*

In all of the above cases we will denote by $U_{\beta,B}$ the self-similar solution of equation (3.10)*. It is given in terms of $f_{\beta,B}$ by* (3.11) *where $f_{\beta,B}$ solves* (3.12)*.*

In describing the asymptotic profile of the solution slightly before time T we will consider the rescaling from the left defined by

$$\bar{u}(y,\tau) := (T-t)^{-\alpha} u(y\,(T-t)^{-\beta}, t)|_{t=T(1-e^{-\tau})}, \qquad (y,\tau) \in \mathbb{R}^N \times (0,\infty). \tag{3.16}$$

Here we discuss only the case when the solution with the cylindrical behavior at infinity becomes extinct at the time T which is the vanishing time of its cylindrical tail as well (in [19] other cases have been considered as well). We will assume in this case that either

- $N \geq 3$ and $\beta \geq \beta_1$ (or equivalently $N\beta \geq \alpha$), or
- $N \geq 6$ and $\beta_0 < \beta < \beta_1$.

The condition $\beta \geq \beta_0 := 2/\sqrt{N-2}$ is imposed so that the self similar solution $U_{\beta,B}$ has non-oscillating behavior as $|x| \to +\infty$. The common feature in both considered cases is that the difference of two self-similar solutions

$$|U_{\beta,B_1} - U_{\beta,B_2}| \notin L^1(\mathbb{R}^N), \qquad \text{if } B_1 \neq B_2.$$

The next theorem generalizes the result proved in [22] in the special case when $\beta = \beta_1$ (see also in [3] for an improvement of the result in [22] shown independently). Our first result is concerned with the case $\beta \geq \beta_1$ in all dimensions $N \geq 3$.

Theorem 3.6. *Let $\beta \geq \beta_1$ and let $u : \mathbb{R}^N \times [0,T) \to \mathbb{R}$ be a solution to* (2.8) *with the initial data u_0 satisfying $0 \leq u_0 \leq U_{\beta,B_1}(\cdot,0)$, for some $B_1 > 0$. Assume in addition that*

$$u_0 - U_{\beta,B} \in L^1(\mathbb{R}^N) \tag{3.17}$$

for some $B > 0$. Then, the rescaled solution $\bar{u}$ given by (3.16) *converges as $\tau \to \infty$ uniformly on compact subsets of R^N to the self-similar solution $U_{\beta,B}$. Moreover, we also have convergence in the $L^1(\mathbb{R}^N)$ norm. If $\beta > \beta_1$ the convergence is exponential.*

In [19] we also consider the case when $\beta < \beta_1$.

References

[1] Angenent, S. and Knopf, D., *An example of neckpinching for Ricci flow on S^{n+1}*, Math. Res. Lett. **11** (2004), 493–518.

[2] Angenet, S., Isenberg, J., and Knopf, D., *Degenerate necpinches in Ricci flow*, arXiv: 1208.4312.

[3] Blanchet, A., Bonforte, M., Dolbeault, J., Grillo, G., and Vaźquez, J.L., *Asymptotics of the fast diffusion equation via entropy estimates*, Arch. Ration. Mech. Anal. **191** (2009), no. 2, 347–385.

[4] C. Böhm and B. Wilking, *Manifolds with positive curvature operators are space forms*, Ann. of Math. (2), **167**(3) (2008), 1079–1097.

[5] Brendle, S., *Rotational symmetry of self-similar solutions to the Ricci flow*, Invent. Math. **194** (2013), no. 3, 731–764.

[6] ———, *Convergence of the Yamabe flow for arbitrary initial energy*, J. Differential Geom. **69** (2005), 217–278.

[7] ———, *Convergence of the Yamabe flow in dimension 6 and higher*, Invent. Math. **170** (2007), 541–576.

[8] Cao, H.-D., *Existence of gradient Kähler-Ricci solitons*, Elliptic and Parabolic Methods in Geometry (Minneapolis, MN, 1994), A K Peters, Wellesley, MA. (1996), 1–16.

[9] ———, *Geometry of complete gradient shrinking Ricci solitons*, Geometry and analysis, No. 1, 227–246, Adv. Lect. Math. (ALM) **17**, Int. Press, Somerville, MA, 2011.

[10] Cao, H.-D. and Chen, Q., *On locally conformally flat gradient steady Ricci solitons*, Trans. Amer. Math. Soc. **364** (2012), no. 5, 2377–2391.

[11] Cao, H.-D., Sun, X., and Zhang, Y., *On the structure of gradient Yamabe solitons*, Math. Res. Lett. **19** (2012), no. 4, 767–774.

[12] Cao, H.-D. and Zhou, D., *On complete gradient shrinking Ricci solitons*, J. Differential Geom. **85** (2010), no. 2, 175–185.

[13] Cao, X., Zhang, and Qi S., *The conjugate heat equation and ancient solutions of the Ricci flow*, Adv. Math. **228** (2011), no. 5, 2891–2919.

[14] Cao, X., Wang, B., and Zhang, Z., *On locally conformally flat gradient shrinking Ricci solitons*, Commun. Contemp. Math. **13** (2011), no. 2, 269–282.

[15] Catino, G., Mantegazza, C., and Mazzieri, L., *On the global structure of conformal gradient solitons with nonnegative Ricci tensor*, Commun. Contemp. Math. **14** (2012), no. 6, 1250045, 12p.

[16] Chow, B., *The Yamabe flow on locally conformally flat manifolds with positive Ricci curvature*, Comm. Pure Appl. Math. **65** (1992), 1003–1014.

[17] Chow, B., Lu, P., and Ni, L., *Hamilton's Ricci flow*, Graduate Studies in Mathematics **77**, American Mathematical Society, Providence, RI; Science Press, New York, 2006.

[18] Chu, S.-C., *Type II ancient solutions to the Ricci flow on surfaces*, Comm. Anal. Geom. **15** (2007), no. 1, 195–215.

[19] Daskalopoulos, P., King, J.R., and Sesum, N., *Extinction profile of complete non-compact solutions to the Yamabe flow*, arXiv:1306.0859.

[20] Daskalopoulos, P. and del Pino M.A., *On a Singular Diffusion Equation*, Comm. in Analysis and Geometry, Vol. 3, 1995, 523–542.

[21] Daskalopoulos, P. and Hamilton, R., *Geometric Estimates for the Logarithmic Fast Diffusion Equation*, Comm. in Analysis and Geometry **12** (2004), 143–164.

[22] Daskalopoulos, P. and Sesum, N., *Eternal solutions to the Ricci flow on* R^2, Int. Math. Res. Not. 2006.

[23] _____, *The classification of locally conformally flat Yamabe solitons*, Adv. Math. **240** (2013), 346–369.

[24] _____, *On the extinction profile of solutions to fast-diffusion*, J. Reine Angew. Math. **622** (2008), 95–119.

[25] Daskalopoulos, P., Hamilton, R., and Sesum, N., *Classification of ancient compact solutions to the Ricci flow on surfaces*, J. Differential Geom. **91** (2012), no. 2, 171–214.

[26] del Pino, M. and Sáez, M., *On the extinction profile for solutions of* $u_t = \Delta u^{(N-2)/(N+2)}$, Indiana Univ. Math. J. **50** (2001), 611–628.

[27] Di Cerbo, L. and Disconzi, M., *Yamabe solitons, determinant of the Laplacian and the uniformization theorem for Riemannian surfaces*, Lett. Math., Phys. **83** (2008), no. 1, 13–18.

[28] Eminenti, M., La Nave, G., and Mantegazza, C., *Ricci solitons: the equation point of view*, Man. Math. **127** (2008), 345-367.

[29] Enders, J., Mller, R., and Topping, P., *On type-I singularities in Ricci flow*, Comm. Anal. Geom. **19** (2011), no. 5, 905–922.

[30] Feldman, M., Ilmanen, T., and Knopf, D., *Rotationally symmetric shrinking and expanding gradient Kähler-Ricci solitons*, J. Diff. Geom., **65** (2003), 169–209.

[31] Gu, H. L. and Zhu, X. P., *The Existence of Type II Singularities for the Ricci Flow on S^{n+1}*, Comm. Anal. geom. (2008), 467–494.

[32] Hamilton, R.S., *The Ricci flow on surfaces*, Contemporary Mathematics **71** (1988), 237–261.

[33] ———, *The Harnack estimate for the Ricci flow*, J. Diff. Geom. **37** (1993), 225–243.

[34] ———, *The formation of singularities in the Ricci flow*, Surveys in Differential Geometry (Cambridge, MA, 1993), 2, 7-136, International Press, Combridge, MA, 1995.

[35] ———, *Lectures on geometric flows* (1989), unpublished.

[36] Hsu, S.-Y., *A note on compact gradient Yamabe solitons*, J. Math. Anal. Appl. **388** (2012), no. 2, 725–726.

[37] Isenberg, J., Knopf, D., and Sesum, N., *Ricci flow neckpinches without rotational symmetry*, arXiv:1312.2933.

[38] Koiso, N., *On rotationally symmmetric Hamilton's equation for Kähler-Einstein metrics*, Recent Topics in Diff. Anal. Geom., Adv. Studies Pure Math., 18-I, Academic Press, Boston, MA (1990), 327–337.

[39] Munteanu, O. and Sesum, N., *On gradient Ricci solitons*, J. Geom. Anal. **23** (2013), no. 2, 539–561.

[40] Munteanu, O. and Wang, J., *Analysis of weighted Laplacian and applications to Ricci solitons*, Comm. Anal. Geom. **20** (2012), no. 1, 55–94.

[41] ———, *Topology of Kähler-Ricci solitons*, arXiv:1312.1318.

[42] Naber, A., *Noncompact shrinking four solitons with nonnegative curvature*, J. Reine Angew. Math. **645** (2010), 125–153.

[43] Ni, L. and Wallach, N., *On a classification of the gradient shrinking solitons*, Math. Res. Lett, **15** (2008), no. 5, 941–955.

[44] Perelman, G., *The entropy formula for the Ricci flow and its geometric applications*, arXiv:math/0211159.

[45] Petersen, P. and Wylie, W., *On the classification of gradient Ricci solitons*, Geom. Topol. **14** (2010), no. 4, 2277–2300.

[46] Schwetlick, H. and Struwe, M., *Convergence of the Yamabe flow for "large" energies*, J. Reine Angew. Math. **562** (2003).

[47] Sesum, N., *Curvature tensor under the Ricci flow*, Amer. J. Math. **127** (2005), no. 6, 1315–1324.

[48] ———, *Convergence of the Ricci flow toward a soliton*, Comm. Anal. Geom. **14** (2006), no. 2, 283–343.

[49] Ye, R., *Global existence and convergence of Yamabe flow*, J. Differential Geom. **39** (1994), 35–50.

[50] Zhang, Z.-H., *Gradient shrinking solitons with vanishing Weyl tensor*, Pacific J. Math. **242** (2009), no. 1, 189–200.

110 Frelinghuysen Road, Room 536, Piscataway, 08854, NJ, USA.
E-mail: natasas@math.rutgers.edu

Extremal Kähler metrics

Gábor Székelyhidi

Abstract. This paper is a survey of some recent progress on the study of Calabi's extremal Kähler metrics. We first discuss the Yau-Tian-Donaldson conjecture relating the existence of extremal metrics to an algebro-geometric stability notion and we give some example settings where this conjecture has been established. We then turn to the question of what one expects when no extremal metric exists.

Mathematics Subject Classification (2010). Primary 58E11; Secondary 32Q20.

Keywords. Extremal metrics, Kähler-Einstein metrics, K-stability.

1. Introduction

A basic problem in differential geometry is to find "best" or "canonical" metrics on smooth manifolds. The most famous example is the classical uniformization theorem, which says that every closed 2-dimensional manifold admits a metric with constant curvature, and moreover this metric is essentially unique in its conformal class. Calabi's introduction of extremal Kähler metrics [10] is an attempt at finding a higher dimensional generalization of this result, in the setting of Kähler geometry.

There are of course other ways in which one could attempt to generalize the uniformization theorem to higher dimensional manifolds. One possibility is the Yamabe problem in the context of conformal geometry. This says [31] that on a closed manifold of arbitrary dimension, every conformal class admits a metric of constant scalar curvature. Moreover this metric is often, but not always, unique up to scaling. A different generalization to 3-dimensional manifolds is given by Thurston's geometrization conjecture, established by Perelman [41]. In this case the goal is to find metrics of constant curvature on a 3-manifold, but this is too ambitious. Instead it turns out that every 3-manifold can be decomposed into pieces each of which admits one of 8 model geometries.

The search for extremal Kähler metrics can be thought of as a complex analogue of the Yamabe problem, where we try to find canonical representatives of a given Kähler class, rather than a conformal class. In both cases the effect of restricting the space of metrics that we allow results in the problems reducing to scalar equations involving the conformal factor, and the Kähler potential, respectively. We will see, however, that in contrast with the Yamabe problem extremal metrics do not always exist, and in these cases one can hope to find a canonical "decomposition" of the manifold into pieces somewhat reminiscent of the geometrization of 3-manifolds.

In order to define extremal metrics, let M be a compact complex manifold of dimension n, equipped with a Kähler class $\Omega \in H^2(M, \mathbf{R})$. Denote by $\mathcal{K}_\Omega$ the set of Kähler metrics in

Proceedings of the International Congress of Mathematicians, Seoul, 2014

the class Ω.

Definition 1.1. An extremal metric is a critical point of the Calabi functional

$$\begin{aligned} \mathrm{Cal} : \mathcal{K}_\Omega &\to \mathbf{R} \\ \omega &\mapsto \int_M (S(\omega) - \underline{S})^2 \, \omega^n, \end{aligned} \tag{1.1}$$

where $S(\omega)$ is the scalar curvature of ω, and $\underline{S}$ is the average of $S(\omega)$ with respect to the volume form ω^n. Note that $\underline{S}$ is independent of the choice of $\omega \in \mathcal{K}_\Omega$.

Calabi [10] has shown that ω is an extremal metric if and only if the gradient $\nabla S(\omega)$ is a holomorphic vector field. Since most complex manifolds do not admit any non-trivial holomorphic vector fields, most extremal metrics are constant scalar curvature Kähler (cscK) metrics. A particularly important special case is when the first Chern class $c_1(M)$ is proportional to the Kähler class Ω. If $c_1(M) = \lambda\Omega$, and $\omega \in \mathcal{K}_\Omega$ is a cscK metric, then it follows that

$$\mathrm{Ric}(\omega) = \lambda\omega, \tag{1.2}$$

and so ω is a Kähler-Einstein metric.

It is known that any two extremal metrics in a fixed Kähler class are isometric (see Chen-Tian [16]), which makes extremal metrics good candidates for being canonical metrics on Kähler manifolds. On the other hand, not every Kähler class admits an extremal metric, the first examples going back to Levine [32] of manifolds which do not admit extremal metrics in any Kähler class. The basic problems are therefore to understand which Kähler classes admit extremal metrics, and what we can say when no extremal metric exists.

The most interesting case of the existence question is when $\Omega = c_1(L)$ is the first Chern class of a line bundle, and consequently M is a projective manifold. In this case the Yau-Tian-Donaldson conjecture predicts that the existence of an extremal metric is related to the stability of the pair (M, L) in the sense of geometric invariant theory. In Section 2 we will discuss two such notions of stability: K-stability, and a slight refinement of it which we call $\widehat{K}$-stability.

As a consequence of work of Tian [54], Donaldson [18, 20], Mabuchi [34], Stoppa [44], Stoppa-Székelyhidi [45], Paul [39], Berman [7], and others, there are now many satisfactory results that show that the existence of an extremal metric implies various notions of stability. The converse direction, however, is largely open. In Section 3 we will discuss two results in this direction. One is the recent breakthrough of Chen-Donaldson-Sun [12] on Kähler-Einstein metrics with positive curvature, and the other is work of the author on extremal metrics on blowups.

Finally in Section 4 we turn to what is to be expected when no extremal metric exists, i.e. when a pair (M, L) is unstable. It is still a natural problem to try minimizing the Calabi functional in a Kähler class, and we will discuss a conjecture due to Donaldson relating this to finding the optimal way to destabilize (M, L). We will give an example where this can be interpreted as the canonical decomposition of the manifold alluded to above.

2. The Yau-Tian-Donaldson conjecture

It is a conjecture going back to Yau (see e.g. [60]) that if M is a Fano manifold, i.e. the anticanonical line bundle K_M^{-1} is ample, then M admits a Kähler-Einstein metric if and

only if M is stable in the sense of geometric invariant theory. Tian [54] introduced the notion of K-stability as a precise candidate of such a stability condition and showed that it is necessary for the existence of a Kähler-Einstein metric. Donaldson [17, 19] generalized the conjecture to pairs (M, L) where $L \to M$ is an ample line bundle, not necessarily equal to the anticanonical bundle. More precisely, Donaldson formulated a more algebraic version of K-stability, and conjectured that K-stability of the pair (M, L) is equivalent to the existence of a cscK metric in the class $c_1(L)$. We start by giving a definition of Donaldson's version of K-stability.

Definition 2.1. A *test-configuration* for (M, L) of exponent r is a $\mathbf{C}^*$-equivariant, flat, polarized family $(\mathcal{M}, \mathcal{L})$ over $\mathbf{C}$, with generic fiber isomorphic to (M, L^r).

The central fiber (M_0, L_0) of a test-configuration has an induced $\mathbf{C}^*$-action, and we write A_k for the infinitesimal generator of this action on $H^0(M_0, L_0^k)$. In other words, the eigenvalues of A_k are the weights of the action. There are expansions

$$\begin{aligned} \dim H^0(M_0, L_0^k) &= a_0 k^n + a_1 k^{n-1} + O(k^{n-2}) \\ \mathrm{Tr}(A_k) &= b_0 k^{n+1} + b_1 k^n + O(k^{n-1}) \\ \mathrm{Tr}(A_k^2) &= c_0 k^{n+2} + O(k^{n+1}). \end{aligned} \tag{2.1}$$

Definition 2.2. Given a test-configuration χ of exponent r for (M, L) as above, its Futaki invariant is defined to be

$$\mathrm{Fut}(\chi) = \frac{a_1 b_0 - a_0 b_1}{a_0^2}. \tag{2.2}$$

The norm of χ is defined by

$$\|\chi\|^2 = r^{-n-2} \left(c_0 - \frac{b_0^2}{a_0} \right), \tag{2.3}$$

where the factor involving r is used to make the norm unchanged if we replace L by a power.

With these preliminaries, we can give a definition of K-stability.

Definition 2.3. The pair (M, L) is K-stable if $\mathrm{Fut}(\chi) > 0$ for all test-configurations χ with $\|\chi\| > 0$.

The condition $\|\chi\| > 0$ is required to rule out certain "trivial" test-configurations. An alternative definition by Li-Xu [33] requires $\mathcal{M}$ to be a normal variety distinct from the product $M \times \mathbf{C}$, but the condition using the norm $\|\chi\|$ will be more natural below, when discussing filtrations.

The central conjecture in the field is the following.

Conjecture 2.4 (Yau-Tian-Donaldson). *Suppose that M has no non-zero holomorphic vector fields. Then M admits a cscK metric in $c_1(L)$ if and only of (M, L) is K-stable.*

When M admits holomorphic vector fields which can be lifted to L, then it is never K-stable according to the previous definition, since in that case one can find test-configurations with total space $\mathcal{M} = M \times \mathbf{C}$, with a non-trivial $\mathbf{C}^*$-action whose Futaki invariant is non-positive. In this case a variant of K-stability, called K-polystability, is used, which rules out such "product test-configurations" and is conjecturally equivalent to the existence of a

cscK metric, even when M admits holomorphic vector fields. A further variant of K-stability, called relative K-stability, was defined by the author [49], and it is conjecturally related to the existence of extremal metrics. In relative K-stability one only considers test-configurations which are orthogonal to a maximal torus of automorphisms of M in a suitable sense.

Example 2.5. Let $(M, L) = (\mathbf{P}^1, \mathcal{O}(1))$. The family of conics $xz = ty^2$ for $t \in \mathbf{C}$ gives a test-configuration χ for (M, L) of exponent 2, degenerating a smooth conic into the union of two lines (see Figure 2.1). A small computation gives $\mathrm{Fut}(\chi) = 1/8$. It is not surprising that this is positive, since the Fubini-Study metric on $\mathbf{P}^1$ has constant scalar curvature.

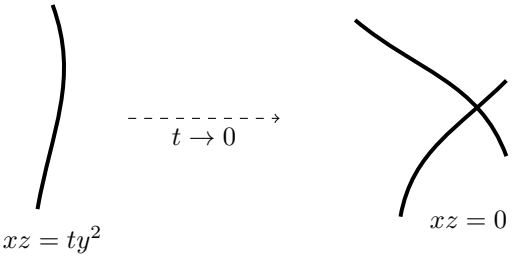


Figure 2.1. A test-configuration degenerating a conic into two lines.

Calculations in Apostolov-Calderbank-Gauduchon-Tønnesen-Friedman [1] suggest that K-stability might not be sufficient to ensure the existence of a cscK metric in general. Indeed they construct examples where the existence of an extremal metric is equivalent to the positivity of a certain function F on an interval (a, b), while relative K-stability only ensures that F is positive at rational points $(a, b) \cap \mathbf{Q}$. It is thus natural to try to work with a completion of the space of test-configurations in a suitable sense in order to detect when this function F vanishes at an irrational point. This motivates the the author's work [47] on filtrations.

Definition 2.6. Let $R = \bigoplus_{k \geqslant 0} H^0(M, L^k)$ denote the homogeneous coordinate ring of (M, L). A filtration of R is a family of subspaces

$$\mathbf{C} = F_0 R \subset F_1 R \subset \ldots \subset R, \tag{2.4}$$

satisfying

1. $(F_i R)(F_j R) \subset F_{i+j} R$,
2. If $s \in F_i R$, and s has degree k piece $s_k \in H^0(M, L^k)$, then $s_k \in F_i R$.
3. $R = \bigcup_{i \geqslant 0} F_i R$.

Witt Nyström [58] showed that every test-configuration for (M, L) gives rise to a filtration. In fact the Rees algebra of the filtration is the coordinate ring of the total space of the test-configuration. On the other hand given any filtration χ of the homogeneous coordinate ring of (M, L), we obtain a flag of subspaces

$$\{0\} = F_0 R_r \subset F_1 R_r \subset \ldots \subset R_r \tag{2.5}$$

of the degree r piece $R_r = H^0(M, L^r)$ for all $r > 0$. In turn, such a flag gives rise to a test-configuration of exponent r for (M, L) – by embedding M into projective space using

a basis of $H^0(M, L^r)$, and acting by a $\mathbf{C}^*$-action whose weight filtration is given by our flag. Therefore any filtration χ induces a sequence of test-configurations $\chi^{(r)}$, where $\chi^{(r)}$ has exponent r. It is natural to think of χ as the limit of the $\chi^{(r)}$, and thus to define

$$\begin{aligned} \mathrm{Fut}(\chi) &= \liminf_{r\to\infty} \mathrm{Fut}(\chi^{(r)}) \\ \|\chi\| &= \lim_{r\to\infty} \|\chi^{(r)}\|, \end{aligned} \tag{2.6}$$

where the limit can be shown to exist. The main difference between filtrations arising from test-configurations, and general filtrations, is that the Rees algebras of the latter need not be finitely generated.

In terms of filtrations we define the following stability notion, which is stronger than K-stability.

Definition 2.7. The pair (M, L) is $\widehat{K}$-stable, if $\mathrm{Fut}(\chi) > 0$ for all filtrations of the homogeneous coordinate ring of (M, L) satisfying $\|\chi\| > 0$.

In view of the examples of Apostolov-et. al. that we mentioned above, it may be that in the Yau-Tian-Donaldson conjecture one should assume $\widehat{K}$-stability instead of K-stability. One direction of this modified conjecture has been established by Boucksom and the author [47].

Theorem 2.8. *Suppose that M has no non-zero holomorphic vector fields. If M admits a cscK metric in $c_1(L)$, then (M, L) is $\widehat{K}$-stable.*

The analogous result for K-stability was shown in Stoppa [44], building on work of Donaldson [20] which we will see in Theorem 4.1, and Arezzo-Pacard [2] which we will discuss in Section 3. In the proof of Theorem 2.8 the main additional ingredient is the use of the Okounkov body [8, 29, 37, 58]. Note that when $L = K_M^{-1}$, then related results were shown by Tian [54] and Paul-Tian [40]. It is likely that a result analogous to Theorem 2.8 can be shown for extremal metrics along the lines of [45].

3. Some existence results

In this section we discuss two special cases, where the Yau-Tian-Donaldson conjecture has been verified.

Kähler-Einstein metrics. We first focus on Kähler-Einstein metrics, i.e. when $c_1(M)$ is proportional to $c_1(L)$. When $c_1(M) = 0$, or $c_1(M) < 0$, then the celebrated work of Yau [59] implies that M admits a Kähler-Einstein metric, and a stability condition does not need to be assumed (see also Aubin [6] for the case when $c_1(M) < 0$).

In the remaining case, when $c_1(M) > 0$, i.e. M is Fano, it was known from early on (see e.g. Matsushima [35]) that a Kähler-Einstein metric does not always exist, and Yau conjectured that the existence is related to stability of M in the sense of geometric invariant theory. Tian [53] found all two-dimensional M which admit a Kähler-Einstein metric, and in [54] he formulated the notion of K-stability, which he conjectured to be equivalent to the existence of a Kähler-Einstein metric. The main difference between Tian's notion of K-stability and the one in Definition 2.3 is that Tian's version of K-stability only requires

$\mathrm{Fut}(\chi) > 0$ for very special types of test-configurations with only mild singularities. In particular their Futaki invariants can be computed differential geometrically using the formula Futaki [24] originally used to define his invariant. By the work of Li-Xu [33] it turns out that in the Fano case Tian's notion of K-stability is equivalent to the a priori stronger condition of Definition 2.3.

Recently, Chen-Donaldson-Sun [12–15] have proved Conjecture 2.4 for Fano manifolds:

Theorem 3.1. *Suppose that M is a Fano manifolds and (M, K_M^{-1}) is K-polystable. Then M admits a Kähler-Einstein metric.*

To construct a Kähler-Einstein metric, the continuity method is used, with a family of equations of the form

$$\mathrm{Ric}(\omega_t) = t\omega_t + \frac{1-t}{m}[D], \tag{3.1}$$

where D is a smooth divisor in the linear system $|mK_M^{-1}|$, and $[D]$ denotes the current of integration. More precisely, a metric ω_t is a solution of (3.1) if $\mathrm{Ric}(\omega_t) = t\omega_t$ on $M \setminus D$, while ω_t has conical singularities along D with cone angle $\frac{2\pi}{m}(1-t)$. One then shows that Equation 3.1 can be solved for $t \in [t_0, T)$ for some $t_0, T > 0$ (see Donaldson [22] for the openness statement), and the question is what happens when $t \to T$.

One of the main results of the work of Chen-Donaldson-Sun is, roughly speaking, that along a subsequence the manifolds (M, ω_{t_i}) have a Gromov-Hausdorff limit W which is a $\mathbf{Q}$-Fano variety, such that the divisor $D \subset M$ converges to a Weil divisor Δ, and W admits a weak Kähler-Einstein metric with conical singularities along Δ (defined in an appropriate sense). Moreover, there are embeddings $\phi_i : M \to \mathbf{P}^N$ and $\phi : W \to \mathbf{P}^N$ into a sufficiently large projective space, such that the pairs $(\phi_i(M), \phi_i(D))$ converge to $(\phi(W), \phi(\Delta))$ in an algebro-geometric sense. In this case either $(\phi(W), \phi(\Delta))$ is in the $SL(N+1, \mathbf{C})$-orbit of the $(\phi_i(M), \phi_i(D))$ in which case we can solve Equation (3.1) for $t = T$, or otherwise we can find a test-configuration for (M, D) with central fiber (W, Δ) to show that (M, K_M^{-1}) is not K-stable. Note that here one needs to extend the theory described in Section 2 to pairs (M, D) resulting in the notion of log K-stability [22].

The fact that a sequence of solutions to Equation (3.1) has a Gromov-Hausdorff limit which is a $\mathbf{Q}$-Fano variety originates in work of Tian [53] on the 2-dimensional case, and it is essentially equivalent to what Tian calls the "partial C^0-estimate" being satisfied by such a sequence of solutions. This partial C^0-estimate was first shown in dimensions greater than 2 by Donaldson-Sun [23] for sequences of Kähler-Einstein metrics, and their method has since been generalized to many other settings: Chen-Donaldson-Sun [14, 15] to solutions of (3.1); Phong-Song-Sturm [42] for sequences of Kähler-Ricci solitons; Tian-Zhang [55] along the Kähler-Ricci flow in dimensions at most 3; the author [48] along Aubin's continuity method; Jiang [28] using only a lower bound for the Ricci curvature, in dimensions at most 3. Note that Tian's original conjecture on the partial C^0-estimate is still open in dimensions greater than 3 – namely we do not yet understand Gromov-Hausdorff limits of Fano manifolds under the assumption of only a positive lower bound on the Ricci curvature.

To close this subsection we mention a possible further result along the lines of Chen-Donaldson-Sun's work. As we described above, if M does not admit a Kähler-Einstein metric, then a sequence of solutions to Equation (3.1) will converge to a weak conical Kähler-Einstein metric on a pair (W, Δ) as $t \to T$. Suppose $T < 1$. We can think of this metric as a suitable weak solution to the equation

$$\mathrm{Ric}(\omega_t) = t\omega_t + \frac{(1-t)}{m}[\Delta] \tag{3.2}$$

for $t = T$ on the space W. Since the pair (W, Δ) necessarily has a non-trivial automorphism group, we cannot expect to solve this equation for $t > T$, however it is reasonable to expect that we can still find weak conical Kähler-Ricci solitons, i.e. we can solve

$$\mathrm{Ric}(\omega_t) + L_{X_t}\omega_t = t\omega_t + \frac{(1-t)}{m}[\Delta], \tag{3.3}$$

for some range of values $t > T$, with suitable vector fields X_t fixing Δ. An extension of Chen-Donaldson-Sun's work to Kähler-Ricci solitons, generalizing Phong-Song-Sturm [42], could then be used to extract a limit (W_1, Δ_1) as $t \to T_1$, with yet another (weak) conical Kähler-Ricci soliton, and so on. Based on this heuristic argument we make the following conjecture.

Conjecture 3.2. *We can solve Equation* (3.3) *up to* $t = 1$ *by passing through finitely many singular times, changing the pair* (W, Δ) *each time. At* $t = 1$ *we obtain a* $\mathbf{Q}$*-Fano variety* W_k *admitting a weak Kähler-Ricci soliton. Moreover, there is a test-configuration for* (M, K_M^{-1}) *with central fiber* W_k.

A further natural expectation would be that the Kähler-Ricci soliton obtained in this way is related to the limiting behavior of the Kähler-Ricci flow on M. Indeed, according to the Hamilton-Tian conjecture (see Tian [54]), the Kähler-Ricci flow is expected to converge to a Kähler-Ricci soliton with mild singularities.

Blow-ups. Beyond the Kähler-Einstein case there are very few general existence results for cscK or extremal metrics. One example is the case of toric surfaces, where Conjecture 2.4 has been established by Donaldson [21], with an extension to extremal metrics by Chen-Li-Sheng [11]. In this section we will discuss a perturbative existence result for cscK metrics on blow-ups.

Suppose that ω is a cscK metric on a compact Kähler manifold M, and choose a point $p \in M$. For all sufficiently small $\epsilon > 0$ the class

$$\Omega_\epsilon = \pi^*[\omega] - \epsilon^2[E] \tag{3.4}$$

is a Kähler class on the blowup $\mathrm{Bl}_p M$, where $\pi : \mathrm{Bl}_p M \to M$ is the blowdown map, and $[E]$ denotes the Poincaré dual of the exceptional divisor. A basic question, going back to work of LeBrun-Singer [30], is whether $\mathrm{Bl}_p M$ admits a cscK (or extremal) metric in the class Ω_ϵ for sufficiently small ϵ. This problem was studied extensively by Arezzo-Pacard [2, 3], and Arezzo-Pacard-Singer [4]. See also Pacard [38] for a survey. The following is the most basic result in this direction.

Theorem 3.3 (Arezzo-Pacard [2]). *Suppose that* M *admits a cscK metric* ω*, and it admits no non-zero holomorphic vector fields. Then there is an* $\epsilon_0 > 0$ *such that* $\mathrm{Bl}_p M$ *admits a cscK metric in the class* Ω_ϵ *for all* $\epsilon \in (0, \epsilon_0)$.

This result not only provides many new examples of cscK metrics, but it is also a key ingredient in the proofs of results such as Theorem 2.8. The construction of cscK metrics on blowups is a typical example of a gluing theorem in geometric analysis. First, one obtains a metric $\omega_\epsilon \in \Omega_\epsilon$ on the blowup, by gluing the metric ω to a scaled down version $\epsilon^2\eta$ of the scalar flat Burns-Simanca [43] metric η on $\mathrm{Bl}_0\mathbf{C}^n$. This is shown in Figure 3.1. In a suitable weighted Hölder space the metric ω_ϵ is sufficiently close to having constant scalar

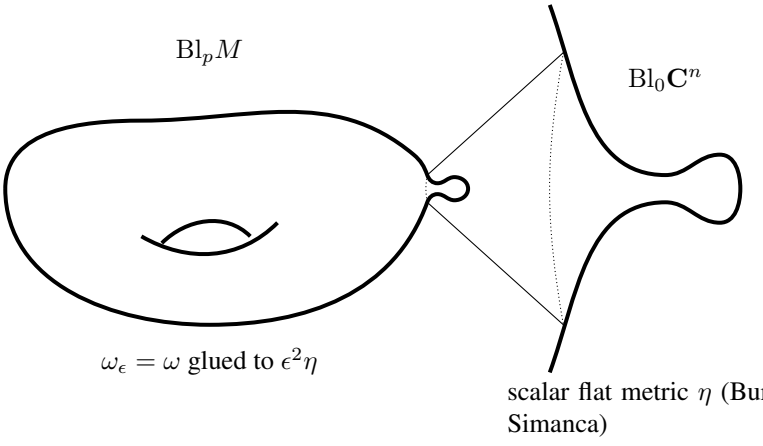


Figure 3.1. The construction of the approximate metric ω_ϵ.

curvature, that one can perturb it to a cscK metric using a contraction mapping argument, for sufficiently small ϵ.

When M admits non-zero holomorphic vector fields, then the problem becomes more subtle, since then $\mathrm{Bl}_p M$ may not admit a cscK (or even extremal) metric for every point p. The problem was addressed by Arezzo-Pacard [3] and Arezzo-Pacard-Singer [4] in the case of extremal metrics, as well as the author [46, 51]. For the case of cscK metrics the sharpest result from [46] is as follows, showing that the Yau-Tian-Donaldson conjecture holds for the pair $(\mathrm{Bl}_p M, \Omega_\epsilon)$ for sufficiently small ϵ.

Theorem 3.4. *Suppose that $\dim M > 2$, M admits a cscK metric ω, and $p \in M$. Then for sufficiently small $\epsilon > 0$, the blowup $\mathrm{Bl}_p M$ admits a cscK metric in the class Ω_ϵ if and only if $(\mathrm{Bl}_p M, \Omega_\epsilon)$ is K-polystable.*

For K-polystability to be defined algebraically, the class Ω_ϵ should be rational, but in fact a very weak version of K-polystability, which can be defined for Kähler manifolds, is sufficient in this theorem. Indeed what we can prove is that if $\mathrm{Bl}_p M$ does not admit a cscK metric in the class Ω_ϵ for sufficiently small ϵ, then there is a $\mathbf{C}^*$-action λ on M such that if we let

$$q = \lim_{t \to 0} \lambda(t) \cdot p, \tag{3.5}$$

then the $\mathbf{C}^*$-action on $\mathrm{Bl}_q M$ induced by λ has non-positive Futaki invariant. In other words when ϵ is sufficiently small, then it is enough to consider test-configurations for $\mathrm{Bl}_p M$ which arise from one-parameter subgroups in the automorphism group of M. While there are also existence results for cscK metrics when $\dim M = 2$, and also for general extremal metrics, in these cases the precise relation with (relative) K-stability has not been established yet.

In the remainder of this section we will give a rough idea of the proof of these existence results. The basic ingredient is the existence of a scalar flat, asymptotically flat metric η on $\mathrm{Bl}_0 \mathbf{C}^n$ due to Burns-Simanca [43], of the form

$$\eta = \sqrt{-1} \partial \overline{\partial} \Big[|w|^2 + \psi(w) \Big] \tag{3.6}$$

on $\mathbf{C}^n \setminus \{0\}$, where

$$\psi(w) = -|w|^{4-2n} + O(|w|^{2-2n}), \quad \text{as } |w| \to \infty \tag{3.7}$$

for $n > 2$. Under the change of variables $w = \epsilon^{-1} z$, we have

$$\epsilon^2 \eta = \sqrt{-1}\partial\overline{\partial}\Big[|z|^2 + \epsilon^2 \psi(\epsilon^{-1} z)\Big]. \tag{3.8}$$

At the same time there are local coordinates near $p \in M$ for which the metric ω is of the form

$$\omega = \sqrt{-1}\partial\overline{\partial}\Big[|z|^2 + \phi(z)\Big], \tag{3.9}$$

where $\phi(z) = O(|z|^4)$. One can then use cutoff functions to glue the metrics ω and $\epsilon^2\eta$ on the level of Kähler potentials on the annular region $r_\epsilon < |z| < 2r_\epsilon$ for some small radius r_ϵ. The result is a metric $\omega_\epsilon \in \Omega_\epsilon$ on $\mathrm{Bl}_p M$, which in a suitable weighted Hölder space is very close to having constant scalar curvature if ϵ is small. It is important here that η is scalar flat, since if it were not, then $\epsilon^2\eta$ would have very large scalar curvature once ϵ is small.

When M has no holomorphic vector fields, then one can show that for sufficiently small ϵ this metric ω_ϵ can be perturbed to a cscK metric in its Kähler class, and this proves Theorem 3.3. Analytically the main ingredient in this proof is to show that the linearization of the scalar curvature operator is invertible, and to control the norm of its inverse in suitable Banach spaces as $\epsilon \to 0$.

The difficulty when M has holomorphic vector fields, or more precisely when when the Hamiltonian isometry group G of (M, ω) is non-trivial, is that the linearized operator will no longer be surjective, since its cokernel can be identified with the Lie algebra $\mathfrak{g}$ of G. One way to overcome this issue is to try to solve a more general equation of the form

$$F(u, \xi) = 0, \tag{3.10}$$

where $\omega_\epsilon + \sqrt{-1}\partial\overline{\partial} u$ is a Kähler metric and $\xi \in \mathfrak{g}$. The operator F is constructed so that if $F(u, \xi) = 0$ and $\xi \in \mathfrak{g}_p$ is in the stabilizer of p, then $\omega_\epsilon + \sqrt{-1}\partial\overline{\partial} u$ is an extremal metric, which has constant scalar curvature if $\xi = 0$. At the same time the linearization of F is surjective. One can then show that for sufficiently small ϵ, for every point $p \in M$ we can find a solution $(u_{\epsilon,p}, \xi_{\epsilon,p})$ of the corresponding equation. The search for cscK metrics is then reduced to the finite dimensional problem of finding zeros of the map

$$\begin{aligned} \mu_\epsilon : M &\to \mathfrak{g} \\ p &\mapsto \xi_{\epsilon,p}, \end{aligned} \tag{3.11}$$

since if $\mu_\epsilon(p) = 0$, then we have found a cscK metric on $\mathrm{Bl}_p M$ in the class Ω_ϵ. More generally to find extremal metrics we need to find p such that $\mu_\epsilon(p) \in \mathfrak{g}_p$.

At this point it becomes important to understand better what the map μ_ϵ is, and for this one needs to construct better approximate solutions than our crude attempt ω_ϵ above. In turn this requires more precise expansions of the metrics $\epsilon^2\eta$ and ω than what we had in Equations (3.8) and (3.9). For the Burns-Simanca metric, according to Gauduchon [25] we have

$$\epsilon^2 \eta = \sqrt{-1}\partial\overline{\partial}\Big[|z|^2 - d_0 \epsilon^{2m-2}|z|^{4-2m} + d_1 \epsilon^{2m}|z|^{2-2m} + O(\epsilon^{4m-4}|z|^{6-4m})\Big], \tag{3.12}$$

where $d_0, d_1 > 0$, while for the metric ω we have

$$\omega = \sqrt{-1}\partial\overline{\partial}\Big[|z|^2 + A_4(z) + A_5(z) + O(|z|^6)\Big], \tag{3.13}$$

where A_4 and A_5 are quartic and quintic expressions. Essentially A_4 is the curvature of ω at p, while A_5 is its covariant derivative at p. The way to obtain better approximate solutions than ω_ϵ is to preserve more terms in these expansions rather than multiplying them all with cutoff functions. In practice this involves modifying the metric ω on the punctured manifold $M \setminus \{p\}$ and $\epsilon^2\eta$ on $\mathrm{Bl}_0\mathbf{C}^n$ to incorporate new terms in their Kähler potentials that are asymptotic to $-d_0\epsilon^{2m-2}|z|^{4-2m} + d_1\epsilon^{2m}|z|^{2-2m}$ and $A_4(z) + A_5(z)$ respectively.

The upshot is that we can obtain an expansion for μ_ϵ which is roughly of the form

$$\mu_\epsilon(p) = \mu(p) + \epsilon^2\Delta\mu(p) + O(\epsilon^\kappa) \tag{3.14}$$

for some $\kappa > 2$, where $\mu : M \to \mathfrak{g}$ is the moment map for the action of G on M, and $\Delta\mu$ is its Laplacian. At this point one can exploit the special structure of moment maps to show that if $\mu(p) + \epsilon^2\Delta\mu(p)$ is in the stabilizer $\mathfrak{g}_p$, and ϵ is sufficiently small, then there is a point $q \in G^c \cdot p$ in the orbit of p under the complexified group such that $\mu_\epsilon(q) \in \mathfrak{g}_q$. Since Bl_pM is biholomorphic to Bl_qM in this case, we end up with an extremal metric on Bl_pM. Under the K-polystability assumption this extremal metric is easily seen to have constant scalar curvature.

Finally, if $\mu(q) + \epsilon^2\Delta\mu(q) \notin \mathfrak{g}_q$ for any $q \in G^c \cdot p$ and sufficiently small ϵ, then the Kempf-Ness principle [36] relating moment maps to GIT stability can be exploited to find a $\mathbf{C}^*$-action on M which induces a destabilizing test-configuration for Bl_pM.

There are several interesting problems which we hope to address in future work.

1. One should extend Theorem 3.4 to the case when $\dim M = 2$ and to general extremal metric. In principle both of these extensions should follow from a more refined expansion of the function μ_ϵ than what we have in Equation (3.14), but it may be more practical to find a different, more direct approach.
2. Can one obtain similar existence results for blow-ups along higher dimensional submanifolds?
3. If M is an arbitrary compact Kähler manifold, is it possible to construct a cscK metric on the blowup of M in a sufficiently large number of points? This would be analogous to Taubes's result [52] on the existence of anti-self-dual metrics on the blowup of a 4-manifold in sufficiently many points. See Tipler [56] for a related result for toric surfaces, where iterated blowups are also allowed.

4. What if no extremal metric exists?

Even if M does not admit an extremal metric in a class $c_1(L)$, it is natural to try minimizing the Calabi functional. That this is closely related to the algebraic geometry of (M, L) is suggested by the following result, analogous to a theorem due to Atiyah-Bott [5] in the case of vector bundles.

Theorem 4.1 (Donaldson [20]). *Given a polarized manifold (M, L), we have*

$$\inf_{\omega \in c_1(L)} \|S(\omega) - \underline{S}\|_{L^2} \geqslant \sup_{\chi} -c_n \frac{\mathrm{Fut}(\chi)}{\|\chi\|}, \tag{4.1}$$

where the supremum runs over all test-configurations for (M, L) with $\|\chi\| > 0$, and c_n is an explicit dimensional constant.

Donaldson also conjectured that in fact equality holds in (4.1). When M admits an extremal metric $\omega_e \in c_1(L)$, then it is easy to check that

$$\|S(\omega_e) - \underline{S}\|_{L^2} = -c_n \frac{\mathrm{Fut}(\chi_e)}{\|\chi_e\|}, \tag{4.2}$$

where χ_e is the product configuration built from the $\mathbf{C}^*$-action induced by $\nabla S(\omega_e)$. In other words equality holds in (4.1) in this case. When (M, L) admits no extremal metric, there is little known, except for the case of a ruled surface [50] where we were able to perform explicit constructions of metrics and test-configurations to realize equality in (4.1). Note that in the case of vector bundles the analogous conjecture is known to hold (i.e. equality in (4.1)) by Atiyah-Bott [5] over Riemann surfaces, and Jacob [27] in higher dimensions.

To describe our result, let Σ be a genus 2 curve, and $M = \mathbf{P}(\mathcal{O} \oplus \mathcal{O}(1))$, where $\mathcal{O}(1)$ denotes any degree one line bundle over Σ. For any real number $m > 0$, we have a Kähler class Ω_m on M, defined by

$$\Omega_m = [F] + m[S_0], \tag{4.3}$$

where $[F], [S_0]$ denote Poincaré duals to the homology classes of a fiber F and the zero section S_0. Up to scaling we obtain all Kähler classes on M in this way. Depending on the value of m, in [50] we observed 3 qualitatively different behaviors of a minimizing sequence for the Calabi functional in Ω_m. There are explicitly computable numbers $0 < k_1 < k_2$ and minimizing sequences ω_i for the Calabi functional in Ω_m with the following properties:

1. When $m < k_1$, then the ω_i converge to the extremal metric in Ω_m whose existence was shown by Tønnesen-Friedman [57].
2. When $k_1 \leqslant m \leqslant k_2$ then suitable pointed limits of the ω_i are complete extremal metrics on $M \setminus S_0$ and $M \setminus S_\infty$. Here S_∞ is the infinity section, and the volumes of the two complete extremal metrics add up to the volumes of the ω_i.
3. When $m > k_2$, then suitable pointed limits of the ω_i are either $\Sigma \times \mathbf{R}$, or complete extremal metrics on $M \setminus S_0$ or $M \setminus S_\infty$. In the first case a circle fibration collapses, and the sum of the volumes of the two complete extremal metrics is strictly less than the volume of the ω_i. Figure 4.1 illustrates the behavior of the metrics ω_i when restricted to a $\mathbf{P}^1$ fiber.

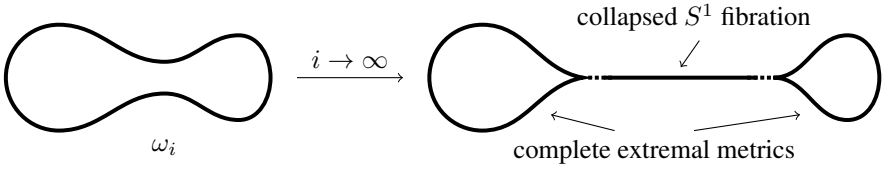


Figure 4.1. The fiber metrics of a minimizing sequence when $m > k_2$.

We interpret cases 2 and 3 as saying that a minimizing sequence breaks the manifold into several pieces. Some of the pieces admit complete extremal metrics, but others display

more complicated collapsing behavior. Having such infinite diameter limits, and possible collapsing is in stark contrast with the case of Fano manifolds that we discussed in Section 3.

The sequences of metrics ω_i above can be written down explicitly using the momentum construction developed in detail by Hwang-Singer [26]. To show that these sequences actually minimize the Calabi energy, one needs to consider the right hand side of (4.1), and construct corresponding sequences of test-configurations χ_i such that

$$\lim_{i\to\infty} \|S(\omega_i) - \underline{S}\|_{L^2} = \lim_{i\to\infty} -c_n \frac{\mathrm{Fut}(\chi_i)}{\|\chi_i\|}. \tag{4.4}$$

For this to make sense we need to assume that m is rational, so that a multiple of Ω_m is an integral class. Such a sequence χ_i can be constructed explicitly, and in the case when $m > k_2$, the exponents of the test-configurations χ_i tend to infinity with i. In other words, we need to embed M into larger and larger projective spaces in order to realize χ_i as a degeneration in projective space. The reason is that the central fiber of χ_i is a normal crossing divisor consisting of a chain of a large number of components isomorphic to M, with the infinity section of each meeting the zero section of the next one. The number of components goes to infinity with i. Figure 4.2 illustrates χ_i restricted to a $\mathbf{P}^1$-fiber of M.

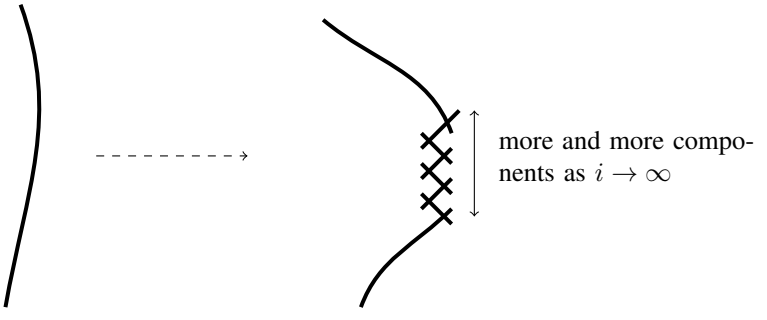


Figure 4.2. The test-configuration χ_i restricted to a $\mathbf{P}^1$ fiber.

From Equation (4.4) together with Theorem 4.1 we obtain the following.

Theorem 4.2. *For the ruled surface M equality holds in Equation 4.1 for any polarization L.*

To conclude this section we point out that already in this example we cannot take a limit of the sequence χ_i in the space of test-configurations, because the exponents go to infinity. However there is a filtration χ, such that χ_i is the induced test-configuration of exponent i, and in this sense the limit of the χ_i exists as a filtration. This filtration achieves the supremum on the right hand side of (4.1) and it is natural to ask whether such a "maximally destabilizing" filtration always exists. In view of the work of Bruasse-Teleman [9] this filtration, if it exists, should be viewed as analogous to the Harder-Narasimhan filtration of an unstable vector bundle.

Acknowledgements. Over the years I have benefited from conversations about extremal metrics with many people. In particular I would like to thank Simon Donaldson, Duong Phong, Julius Ross, Jacopo Stoppa, Richard Thomas and Valentino Tosatti for many useful discussions. The work presented in this survey was partially supported by the NSF.

References

[1] V. Apostolov, D. M. J. Calderbank, P. Gauduchon, and C. W. Tønnesen-Friedman, *Hamiltonian 2-forms in Kähler geometry III, extremal metrics and stability*, Invent. Math. **173** (2008), no. 3, 547–601.

[2] C. Arezzo and F. Pacard, *Blowing up and desingularizing constant scalar curvature Kähler manifolds.*, Acta Math. **196** (2006), no. 2, 179–228.

[3] _____, *Blowing up Kähler manifolds with constant scalar curvature II*, Ann. of Math. (2) **170** (2009), no. 2, 685–738.

[4] C. Arezzo, F. Pacard, and M. A. Singer, *Extremal metrics on blow ups*, Duke Math. J. **157** (2011), no. 1, 1–51.

[5] M. F. Atiyah and R. Bott, *The Yang-Mills equations over Riemann surfaces*, Philos. Trans. Roy. Soc. London Ser. A **308** (1983), 523–615.

[6] T. Aubin, *Équations du type Monge-Ampère sur les varietés kählériennes compactes*, Bull. Sci. Math. (2) **102** (1978), no. 1, 63–95.

[7] R. Berman, *K-polystability of Q-Fano varieties admitting Kähler-Einstein metrics*, arXiv:1205:6214v2.

[8] S. Boucksom and H. Chen, *Okounkov bodies of filtered linear series*, Compos. Math. **147** (2011), 1205–1229.

[9] L. Bruasse and A. Teleman, *Harder-Narasimhan filtrations and optimal destabilizing vectors in complex geometry*, Ann. Inst. Fourier (Grenoble) **55** (2005), no. 3, 1017–1053.

[10] E. Calabi, *Extremal Kähler metrics*, Seminar on Differential Geometry (S. T. Yau, ed.), Princeton, 1982.

[11] B. Chen, An-Min Li, and L. Sheng, *Extremal metrics on toric surfaces*, arXiv:1008.2607

[12] X. X. Chen, S. K. Donaldson, and S. Sun, *Kähler-Einstein metrics and stability*, arXiv:1210.7494.

[13] _____, *Kähler-Einstein metrics on Fano manifolds, I: approximation of metrics with cone singularities*, arXiv:1211.4566.

[14] _____, *Kähler-Einstein metrics on Fano manifolds, II: limits with cone angle less than* 2π, arXiv:1212.4714.

[15] _____, *Kähler-Einstein metrics on Fano manifolds, III: limits as cone angle approaches* 2π *and completion of the main proof*, arXiv:1302.0282.

[16] X. X. Chen and G. Tian, *Geometry of Kähler metrics and foliations by holomorphic discs*, Publ. Math. Inst. Hautes Études Sci. (2008), no. 107, 1–107.

[17] S. K. Donaldson, *Remarks on gauge theory, complex geometry and four-manifold topology*, Fields Medallists' Lectures (Atiyah and Iagolnitzer, eds.), World Scientific, 1997, pp. 384–403.

[18] ———, *Scalar curvature and projective embeddings, I*, J. Differential Geom. **59** (2001), 479–522.

[19] ———, *Scalar curvature and stability of toric varieties*, J. Differential Geom. **62** (2002), 289–349.

[20] ———, *Lower bounds on the Calabi functional*, J. Differential Geom. **70** (2005), no. 3, 453–472.

[21] ———, *Constant scalar curvature metrics on toric surfaces*, Geom. Funct. Anal. **19** (2009), no. 1, 83–136.

[22] ———, *Kähler metrics with cone singularities along a divisor*, Essays in mathematics and its applications, Springer, 2012, pp. 49–79.

[23] S. K. Donaldson and S. Sun, *Gromov-Hausdorff limits of Kähler manifolds and algebraic geometry*, arXiv:1206.2609.

[24] A. Futaki, *An obstruction to the existence of Einstein-Kähler metrics*, Invent. Math. **73** (1983), 437–443.

[25] P. Gauduchon, *Invariant scalar-flat Kähler metrics on* $\mathcal{O}(-l)$, preprint (2012).

[26] A. Hwang and M. A. Singer, *A momentum construction for circle-invariant Kähler metrics*, Trans. Amer. Math. Soc. **354** (2002), no. 6, 2285–2325.

[27] A. Jacob, *The Yang-Mills flow and the Atiyah-Bott formula on compact Kähler manifolds*, arXiv:1109.1550.

[28] W. Jiang, *Bergman kernel along the Kähler Ricci flow and Tian's conjecture*, arXiv: 1311.0428.

[29] R. Lazarsfeld and M. Mustata, *Convex bodies associated to linear series*, Ann. Sci. Éc. Norm. Supér. (4) **42** (2009), no. 5, 783–835.

[30] C. LeBrun and S. R. Simanca, *Extremal Kähler metrics and complex deformation theory*, Geom. and Func. Anal. **4** (1994), no. 3, 298–336.

[31] J. M. Lee and T. H. Parker, *The Yamabe problem*, Bull. Amer. Math. Soc. (N.S.) **17** (1987), no. 1, 37–91.

[32] M. Levine, *A remark on extremal Kähler metrics*, J. Differential Geom. **21** (1985), no. 1, 73–77.

[33] C. Li and C. Xu, *Special test configurations and K-stability of Fano varieties*, arXiv: 1111.5398.

[34] T. Mabuchi, *K-stability of constant scalar curvature polarization*, arXiv:0812.4093.

[35] Y. Matsushima, *Sur la structure du groupe d'homéomorphismes analytiques d'une certaine variété kählérienne*, Nagoya Math. J. **11** (1957), 145–150.

[36] D. Mumford, J. Fogarty, and F. Kirwan, *Geometric invariant theory*, third ed., Ergebnisse der Mathematik und ihrer Grenzgebiete (2) [Results in Mathematics and Related Areas (2)], vol. 34, Springer-Verlag, Berlin, 1994. MR 1304906 (95m:14012)

[37] A. Okounkov, *Brunn-Minkowski inequality for multiplicities*, Invent. Math. **125** (1996), 405–411.

[38] F. Pacard, *Constant scalar curvature and extremal Kähler metrics on blow ups*, Proceedings of the International Congress of Mathematicians. Volume II (New Delhi), Hindustan Book Agency, 2010, pp. 882–898.

[39] S. T. Paul, *Hyperdiscriminant polytopes, Chow polytopes, and Mabuchi energy asymptotics*, Ann. of Math. (2) **175** (2012), no. 1, 255–296.

[40] S. T. Paul and G. Tian, *CM stability and the generalised Futaki invariant II*, Astérisque **328** (2009), 339–354.

[41] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications*, arXiv:math.DG/0211159.

[42] D. H. Phong, J. Song, and J. Sturm, *Degenerations of Kähler-Ricci solitons on Fano manifolds*, arXiv:1211.5849.

[43] S. R. Simanca, *Kähler metrics of constant scalar curvature on bundles over* $C\mathrm{P}_{n-1}$, Math. Ann. **291** (1991), no. 2, 239–246.

[44] J. Stoppa, *K-stability of constant scalar curvature Kähler manifolds*, Adv. Math. **221** (2009), no. 4, 1397–1408.

[45] J. Stoppa and G. Székelyhidi, *Relative K-stability of extremal metrics*, J. Eur. Math. Soc. **13** (2011), no. 4, 899–909.

[46] G. Székelyhidi, *Blowing up extremal Kähler manifolds II*, arXiv:1302.0760.

[47] ———, *Filtrations and test-configurations, with an appendix by S. Boucksom*, arXiv: 1111.4986.

[48] ———, *The partial* C^0*-estimate along the continuity method*, arXiv:1310.8471.

[49] ———, *Extremal metrics and* K*-stability*, Bull. Lond. Math. Soc. **39** (2007), no. 1, 76–84.

[50] ———, *The Calabi functional on a ruled surface*, Ann. Sci. Éc. Norm. Supér. (4) **42** (2009), no. 5, 837–856.

[51] ———, *On blowing up extremal Kähler manifolds*, Duke Math. J. **161** (2012), no. 8, 1411–1453.

[52] C. H. Taubes, *The existence of anti-self-dual conformal structures*, J. Differential Geom. **36** (1992), no. 1, 163–253.

[53] G. Tian, *On Calabi's conjecture for complex surfaces with positive first Chern class*, Invent. Math. **101** (1990), no. 1, 101–172.

[54] ———, *Kähler-Einstein metrics with positive scalar curvature*, Invent. Math. **137** (1997), 1–37.

[55] G. Tian and Z. Zhang, *Regularity of Kähler-Ricci flows on Fano manifolds*, arXiv:1310.5897.

[56] C. Tipler, *A note on blow-ups of toric surfaces and CSC Kähler metrics*, Tohoku Math. J. **66** (2014), no. 2, 15–29.

[57] C. W. Tønnesen-Friedman, *Extremal Kähler metrics on minimal ruled surfaces*, J. Reine Angew. Math. **502** (1998), 175–197.

[58] D. Witt Nyström, *Test configurations and Okounkov bodies*, Compos. Math. **148** (2012), no. 6, 1736–1756.

[59] S.-T. Yau, *On the Ricci curvature of a compact Kähler manifold and the complex Monge-Ampère equation I.*, Comm. Pure Appl. Math. **31** (1978), 339–411.

[60] ———, *Open problems in geometry*, Proc. Symposia Pure Math. **54** (1993), 1–28.

Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556
E-mail: gszekely@nd.edu

Ricci flows with unbounded curvature

Peter M. Topping

Abstract. Until recently, Ricci flow was viewed almost exclusively as a way of deforming Riemannian metrics of bounded curvature. Unfortunately, the bounded curvature hypothesis is unnatural for many applications, but is hard to drop because so many new phenomena can occur in the general case. This article surveys some of the theory from the past few years that has sought to rectify the situation in different ways.

Mathematics Subject Classification (2010). Primary 53C44; Secondary 35K55, 58J35.

Keywords. Ricci flow, well-posedness, unbounded curvature, uniformization, geometrization, flowing beyond singularities.

1. Introduction

Since its inception in 1982 [16], Ricci flow has supported the development of a remarkable and elegant theory. The flow has become well-known as a way of deforming a Riemannian metric in order to improve it, or turn it into a special metric that might satisfy a geometrically rigid condition or simply a natural PDE, and indeed up until now, most applications fit within the following general strategy.

- First we take a space that we do not understand very well, perhaps a Riemannian manifold satisfying a curvature condition, or a metric space with some weak geometric structure.
- Next we deform the space under Ricci flow, keeping track of its properties, for example its topology, its curvature or its conformal structure, until it develops into a very special space, for example sometimes one of constant curvature.
- Such special spaces we can hope to classify, and if the Ricci flow can be sufficiently well understood then we can go back and classify or better understand the space with which we started.

As an example, Hamilton's original insight was that a simply connected three-dimensional closed Riemannian manifold of positive Ricci curvature will flow smoothly under a suitably normalised Ricci flow through a family of manifolds of positive Ricci curvature to a manifold of *constant* curvature, which can then be identified as a round sphere. He deduced that the original manifold of positive Ricci curvature must be diffeomorphic to the three-sphere. Dramatic further development of Ricci flow theory by Hamilton and then Perelman ultimately extended this principle to handle all closed three-manifolds, leading to a resolution of the conjectures of Poincaré and Thurston ([19, 23, 27–29]).

Proceedings of the International Congress of Mathematicians, Seoul, 2014

This general strategy has clearly been very effective, but its scope has nevertheless been severely limited by the Ricci flow existence and asymptotics theory only applying in very special cases, meaning that we are only scratching the surface of the potential applicability of the method. In particular, the vast majority of applications require the underlying manifold to be closed, and without that hypothesis we are not even able in general to start the flow going, even for a short time, without imposing further conditions such as boundedness of the curvature that may damage potential applications.

This article surveys part of the programme to extend the theory of Ricci flow to handle general manifolds or even metric spaces. The central point will be the necessity to handle flows with unbounded curvature; until recently we have had no idea how to even start the flow going starting with a manifold of unbounded curvature, let alone understand its long-time existence and asymptotics or uniqueness. In this unrestricted situation, the flow interacts with itself 'at spatial infinity' in an unfamiliar way that is interesting both geometrically and in terms of the pure PDE questions it raises. Since classical solutions to the Ricci flow are characterised as existing until the curvature becomes unbounded, and we want to consider unbounded curvature from the outset, we now need to understand the issue of long-time existence in much more detail. This also brings to the fore the subtle issue of asymptotics of the flow, and we will witness how Ricci flow organises itself to find a special metric, even when there appear to be obstructions.

2. Why is Ricci flow with unbounded curvature so difficult?

We call a smooth one-parameter family of Riemannian metrics $g(t)$ on a manifold $\mathcal{M}$ a Ricci flow when

$$\frac{\partial g}{\partial t} = -2\,\mathrm{Ric}_{g(t)} \tag{2.1}$$

where Ric is the Ricci tensor (see [16, 38]). One should interpret the right-hand side of the equation as some sort of Laplacian of the metric, and thus interpret this equation as some sort of heat equation, although ultimately this is a nonlinear equation, and is even not quite parabolic, which causes problems when trying to establish existence of solutions even to this day, except in relatively simple situations.

One way of trying to pose this flow is to consider an initial Riemannian metric g_0, and then look for a family $g(t)$, $t \in [0, T)$ satisfying (2.1), with $g(0) = g_0$. In due course, we will see that this is rather naive in general, but it works well in special situations such as on closed manifolds (Hamilton [16]) or more generally when the initial metric is complete and of bounded curvature (Shi [36]). The following hybrid of their work also incorporates a uniqueness assertion of Chen-Zhu [4]; the proofs have been clarified and simplified by DeTurck/Hamilton [10, 19] and Kotschwar [24].

Theorem 2.1 (The Hamilton-Shi flow). *Given a smooth, complete Riemannian manifold* $(\mathcal{M}, g_0)$ *of* bounded curvature*, there exist a unique* $T \in (0, \infty]$ *and Ricci flow* $g(t)$ *for* $t \in [0, T)$ *satisfying the equation* (2.1)*, the initial condition* $g(0) = g_0$*, and the properties that the curvature remains bounded for* $t \in [0, T_0]$*, for any* $T_0 \in [0, T)$*, and that if* $T < \infty$ *then*

$$\sup_{\mathcal{M}} |\mathrm{Rm}_{g(t)}| \to \infty \quad \text{as } t \uparrow T,$$

where $\mathrm{Rm}_{g(t)}$ *is the curvature tensor of* $g(t)$*. Moreover,* $(\mathcal{M}, g(t))$ *is complete for all* $t \in [0, T)$.

We will call the Ricci flow whose existence is asserted by this theorem the *Hamilton-Shi* Ricci flow.

The final assertion here that the Ricci flow is complete is more or less obvious. Indeed completeness of a Riemannian manifold is equivalent to the assertion that the length of any smooth proper curve $\gamma : [0, 1) \to \mathcal{M}$ (i.e. any curve heading off to infinity in $\mathcal{M}$) is infinite, but the boundedness of the curvature in this theorem guarantees that lengths of curves can only grow or decrease at most exponentially (see, for example, [38, Lemma 5.3.2]) and so an infinitely long curve at time $t = 0$ will remain infinitely long at a later time $t \in [0, T_0]$. However, this principle emphatically fails for flows of unbounded curvature. Loosely speaking,

Unbounded curvature allows Ricci flow to feel spatial infinity.

In particular, an unbounded curvature Ricci flow can pull 'points at infinity' to within a finite distance in finite time, as we now illustrate.

Theorem 2.2 (Pulling in points at infinity; Special case of [41]). *There exists a smooth Ricci flow for* $t \in [0, \infty)$*, starting at a smooth, complete Riemannian metric of bounded curvature, that is incomplete for all* $t > 0$.

In particular, if T^2 *is a torus equipped with an arbitrary conformal structure,* $p \in T^2$ *is any point, and we write* h *for the unique complete, conformal, hyperbolic metric on* $T^2 \backslash \{p\}$*, then there exists a smooth Ricci flow* $g(t)$ *on* T^2 *for* $t > 0$ *such that* $g(t) \to h$ *smoothly locally on* $T^2 \backslash \{p\}$ *as* $t \downarrow 0$.

Of course, the Ricci flow $g(t)$ on T^2 can be restricted to $T^2 \backslash \{p\}$ to give the desired Ricci flow that starts complete, but instantaneously becomes incomplete. Intuitively the point p is being pulled in from infinity as t lifts off from zero – see Figure 2.1.

This example also illustrates the subtleties of uniqueness in Ricci flow in the presence of possibly unbounded curvature, since we could have flowed the initial manifold $(T^2 \backslash \{p\}, h)$ using the Hamilton-Shi flow instead. In this example that flow would simply dilate the metric, and could be written explicitly as $g(t) = (1 + 2t)h$.

Examples such as this run somewhat counter to the classical intuition in Ricci flow, and the failure to recognise that they can arise in practice has previously led to errors in important parts of the literature – see the discussion in [42]. We will return later to see how incompleteness can be dealt with, and also how to impose a condition at infinity (analogous to a boundary condition) in order to make the problem well posed.

3. A clear picture of Ricci flow on surfaces

Both the nature of the evolution equation and the existence theory above become crystal clear when the dimension of the underlying manifold $\mathcal{M}$ is two.

In this case, the Ricci tensor can be written in terms of the Gauss curvature K as $\mathrm{Ric} = K.g$, so the flow equation is

$$\frac{\partial}{\partial t} g(t) = -2K.g(t).$$

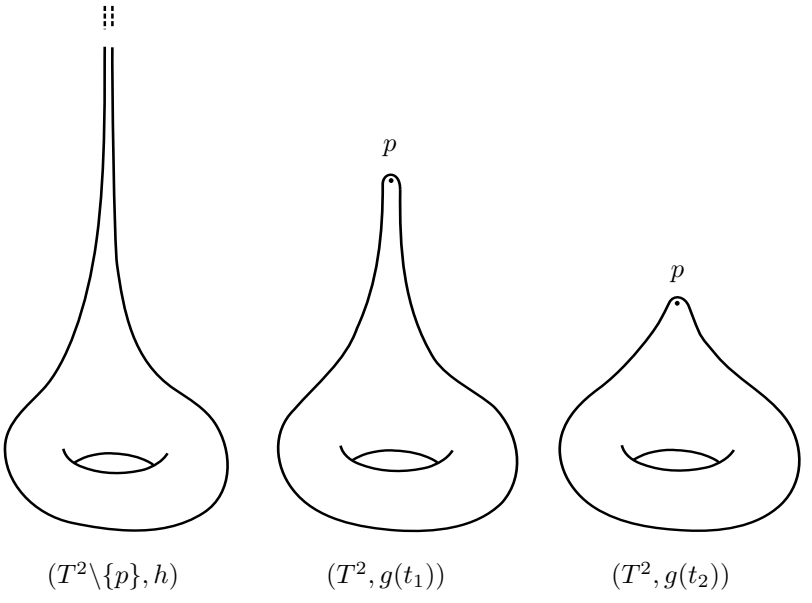


Figure 2.1. Pulling in points at infinity in Theorem 2.2

Thus the flow will always make a conformal deformation of the metric, and we may take local isothermal coordinates x and y, and write the flow $g(t) = e^{2u}|\mathrm{d}z|^2 := e^{2u}(dx^2 + dy^2)$ for some locally-defined, scalar, time-dependent function u, which can then be shown to satisfy the local equation

$$\frac{\partial u}{\partial t} = e^{-2u}\Delta u = -K. \tag{3.1}$$

In particular, in this form, we are dealing with a strictly parabolic PDE. Up to a change of variables, this is the so-called logarithmic fast diffusion equation, which has an extensive literature – see [9, 45] and the references therein.

Theorem 3.1 (Closed surfaces. Hamilton [18], Chow [7]). *Let $\mathcal{M}$ be a closed, oriented surface and g_0 any smooth metric. Define $T = \infty$ unless $\mathcal{M} = S^2$ (topologically) in which case we set $T = \frac{\mathrm{vol}_{g_0}\mathcal{M}}{8\pi}$. Then there exists a unique Ricci flow $g(t)$ on $\mathcal{M}$ for $t \in [0, T)$ so that $g(0) = g_0$. Depending on the genus of $\mathcal{M}$, we have*

$$\begin{aligned}
&\mathcal{M} = S^2: && \frac{g(t)}{2(T-t)} \to G_{+1}, && \textit{a metric of const. curvature } +1, \textit{ as } t \uparrow T.\\
&\mathcal{M} = T^2: && g(t) \to G_0, && \textit{a flat metric, as } t \to \infty.\\
&\mathcal{M} \neq S^2, T^2: && \frac{g(t)}{2t} \to G_{-1}, && \textit{a metric of const. curvature } -1, \textit{ as } t \to \infty.
\end{aligned}$$

The well-posedness theory on closed surfaces was thus completed in the 1980s. An alternative approach to Theorem 3.1 that provides a model for the trickier argument required to prove the Poincaré conjecture can be found in [44].

It is apparent from Theorem 3.1 that Ricci flow *geometrises* a closed surface; more precisely it finds a conformal metric of constant curvature on an arbitrary closed Riemann surface, which is enough to establish the Uniformisation theorem in the restricted case of closed surfaces [6]. It is then a natural question to ask whether Ricci flow performs the same geometrisation task on a general surface. Of course, to be able to even ask this question, we have to be able to start the Ricci flow with a more general metric, and continue it until the flow has had a chance to organise itself into a special metric, whereas the Hamilton-Shi flow from Theorem 2.1 flows restricted metrics, and in general will stop (as we will see) before the flow has achieved anything.

In fact, as we shall demonstrate shortly, Ricci flow is perfectly capable of flowing a completely general surface with possibly unbounded curvature in a uniquely defined way, without even requiring the initial metric to be complete. Before stating the result, we dwell on some issues that such a result must address.

Those unfamiliar with PDE theory often misinterpret existence theory as presumably being obvious. Surely if we are trying to solve an equation $\frac{\partial g}{\partial t} = -2Kg$, then we should simply keep moving g by tiny amounts in the direction $-2Kg$ and a solution should result. The most naive aspect of that suggestion is that it would appear to apply to the problem of solving backwards in time from a given smooth metric, whereas this is certainly not possible. Indeed, Ricci flow has the dramatic smoothing effect of parabolic equations, and immediately makes any metric real analytic [25]. Therefore we could never flow backwards in time starting from a general smooth metric that is not also real analytic.

A more subtle issue that arises once we drop the hypothesis that the underlying manifold is closed, is that we have to worry about boundary conditions. To illustrate this issue, consider simply the problem of starting the Ricci flow on the open disc $D^2 \subset \mathbb{R}^2$ with a metric $g_0 = e^{2u_0}(dx^2 + dy^2)$ that is smooth up to the boundary ∂D^2. From (3.1), this results in the PDE problem

$$\begin{cases} \dfrac{\partial u}{\partial t} = e^{-2u}\Delta u & D^2 \\ u(0) = u_0 & \partial D^2. \end{cases} \tag{3.2}$$

Even amongst solutions that remain continuous up to the boundary at later times, this PDE problem is ill-posed owing to nonuniqueness. Indeed, standard parabolic theory tells us that we are free to specify the restriction of u to ∂D^2 at later times, and only then would we obtain uniqueness. In fact, we will shortly solve the well-posedness problem without resorting to specifying any more data.

A third way we can see that existence theory should be a subtle issue is to consider what it should imply. Geometric flows are typically designed to find special objects, for example constant curvature metrics in the present discussion. On the other hand, it is often possible to give a geometric flow initial data that for some reason cannot be deformed globally and smoothly into a special object – occasionally one does not even exist. Any assertion of global existence of solutions in such cases is also asserting that the geometric flow must organise itself in such a way to resolve this issue, often finding an ingenious way of decomposing the object being flowed into multiple special objects (see [40] and [33, 34] for some other contexts in which this occurs).

We shall shortly see how these issues are resolved in practice by the Ricci flow, but first we give the main well-posedness result in the two-dimensional case. The existence part is joint work with G. Giesen.

Theorem 3.2 (Ricci flows on surfaces, without restriction; [13, 39, 43]). *Let $(\mathcal{M}, g_0)$ be any smooth (connected) Riemannian surface, possibly incomplete and/or with unbounded curvature. Depending on the conformal type, we define $T \in (0, \infty]$ by*[1]

$$T := \begin{cases} \frac{1}{4\pi\chi(\mathcal{M})} \operatorname{vol}_{g_0} \mathcal{M} & \text{if } (\mathcal{M}, g_0) \cong \mathcal{S}^2, \mathbb{C} \text{ or } \mathbb{R}P^2, \\ \infty & \text{otherwise.} \end{cases}$$

Then there exists a smooth Ricci flow $g(t)$ on $\mathcal{M}$, defined for $t \in [0, T)$ such that

(1) *$g(0) = g_0$, and*

(2) *$g(t)$ is instantaneously complete, i.e. complete for all $t \neq 0$ at which it is defined.*

The flow $g(t)$ is unique in the sense that if $\tilde{g}(t)$ is any other smooth Ricci flow, defined now for $t \in [0, \tilde{T})$, satisfying (1) and (2) above, then $\tilde{T} \leq T$ and $g(t) = \tilde{g}(t)$ for all $t \in [0, \tilde{T})$.

In addition, this Ricci flow $g(t)$ is maximally stretched (see Remark 3.3), and the Gauss curvature $K_{g(t)}$ satisfies

$$K_{g(t)} \geq -\frac{1}{2t}$$

for $t \in (0, T)$. If $T < \infty$, then we have

$$\operatorname{vol}_{g(t)} \mathcal{M} = 4\pi\chi(\mathcal{M})(T - t) \ \to \ 0 \quad \text{as } t \uparrow T,$$

and in particular, T is the maximal existence time.

Related results can be found in the literature of the logarithmic fast diffusion literature (e.g. [9, 45]) and the work of Mazzeo, Sesum, Ji and Isenberg [21, 22].

Remark 3.3 ([13] and [15, Remark 1.5]). The *maximally stretched* assertion of the theorem means that $g(t)$ lies 'above' any another Ricci flow with the same or lower initial data. More precisely, if $0 \leq a < b \leq T$ and $\tilde{g}(t)$ is any Ricci flow on $\mathcal{M}$ for $t \in [a, b)$ with $\tilde{g}(a) \leq g(a)$ (with $\tilde{g}(t)$ not necessarily complete or of bounded curvature) then $\tilde{g}(t) \leq g(t)$ for every $t \in [a, b)$.

The mechanism by which this Ricci flow makes an incomplete metric immediately complete is somewhat similar to how the example from Theorem 2.2 made a complete metric immediately incomplete. Unbounded curvature allows points at a finite distance to be sent out to infinity, and vice versa.

Given the unusual nature of these flows, some examples are in order. Although Theorem 3.2 can handle arbitrary initial metrics, with arbitrarily wild behaviour at spatial infinity, we pick the two simplest possible examples [39], both of which start completely flat but have no choice but to have unbounded curvature immediately.

Example 3.4. Let $(\mathcal{M}, g_0)$ be the punctured plane. The metric is incomplete since we can take a curve asymptoting to the origin, of finite length, and the flow will have to do something about this in order to make the metric complete immediately. The flow we construct in Theorem 3.2 deals with this by stretching the metric near the puncture to be asymptotically a hyperbolic cusp, scaled to have Gauss curvature $-\frac{1}{2t}$. See Figure 3.1. (In fact, in this extremely special case, one could compute the flow explicitly as a so-called Ricci soliton, up to the solution of an ODE.)

[1]Note that in the case that $\mathcal{M} = \mathbb{C}$, we set $T = \infty$ if $\operatorname{vol}_{g_0} \mathbb{C} = \infty$. Here $\chi(\mathcal{M})$ is the Euler characteristic.

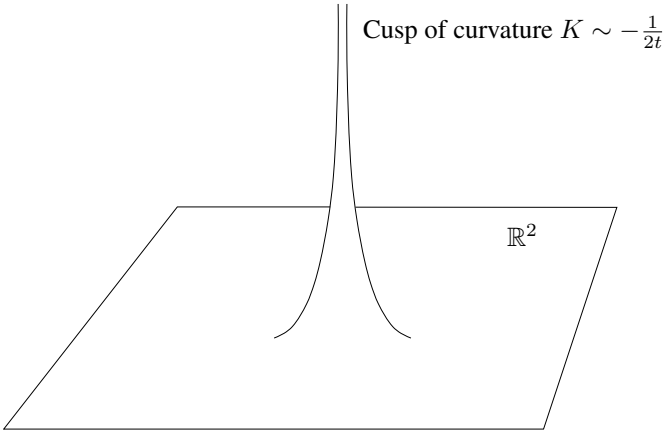


Figure 3.1. Puncture turns into a hyperbolic cusp in Example 3.4

Example 3.5. Now let $(\mathcal{M}, g_0)$ be the Euclidean unit two-dimensional disc. Again, the flow must blow up the metric to make it complete immediately, and it does this by stretching it near to the 'boundary' circle into a Poincaré metric, also scaled to have Gauss curvature $-\frac{1}{2t}$. See Figure 3.2.

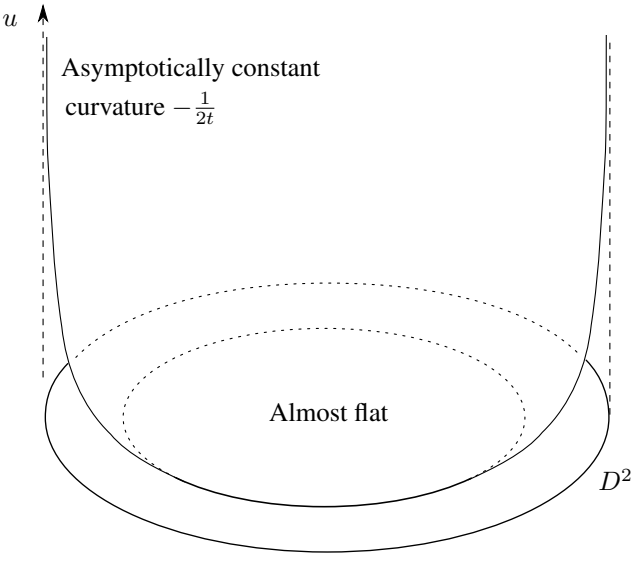


Figure 3.2. Metric stretches at infinity in Example 3.5

This example fits directly into the discussion above on flowing smooth metrics on the disc D^2, where we decided that the addition of boundary data would be the standard way of achieving well-posedness. In Theorem 3.2, the condition on solutions that they are instantaneously complete can be viewed as a substitute for a boundary condition.

Intuitively, the flow from Theorem 3.2 is finding a way of feeding in volume at infinity where the metric is incomplete. The notable feature here is not just that this can be done in order to give a solution, but that it can be done uniquely: If we try to feed in less volume, then the metric will fail to become complete. On the other hand, if we try to feed in more, then the damping within the equation, arising from the e^{-2u} factor in the equation (3.1) is preventing the extra volume from arriving in the interior.

Naively, one might view a common feature of both these examples to be that the conformal factor u in the most obvious, Euclidean coordinates is immediately asymptotically infinity at spatial infinity. However, the conformal factor u depends on the coordinates chosen, and in different coordinates this property will not hold.

Given that Theorem 3.2 makes the metric complete immediately, and also makes the curvature bounded from below, it is reasonable to speculate that maybe the flow also makes the curvature bounded from above immediately, and that therefore after an arbitrarily short time we are in the classical situation of Theorem 2.1 and could make do from then on with the Hamilton-Shi flow. We will see that this suggestion is wrong in two ways. To begin with, we cannot hope ever to be able to construct a complete Ricci flow solution that makes the curvature bounded from above, because our flow is the unique instantaneously complete solution, and carefully chosen initial metrics will have unbounded curvature for all time:

Theorem 3.6 (Ricci flows with unbounded curvature; [14]). *Given any noncompact Riemann surface $\mathcal{M}$, there exists a smooth, complete, conformal metric g_0 such that the unique complete Ricci flow $g(t)$ given by Theorem 3.2 exists for all $t \geq 0$ and $(\mathcal{M}, g(t))$ has unbounded curvature for each $t \geq 0$.*

As we will see in Theorem 4.2, Ricci flows with unbounded curvature at later times were first constructed in higher dimensions in the context of flows with nonnegative complex sectional curvature.

To see a second way in which the classical Hamilton-Shi flow cannot be a substitute for the flow from Theorem 3.2, consider for a moment the restricted situation that $(\mathcal{M}, g_0)$ is both complete and of bounded curvature. We can flow such a metric not only with Theorem 3.2 but also with the Hamilton-Shi flow. By the uniqueness of complete Ricci flows asserted in Theorem 3.2, they must agree – at least while the Hamilton-Shi flow exists. By Theorem 2.1, the Hamilton-Shi flow exists for all time unless the curvature blows up, so one might naively think that this should be when the flow from Theorem 3.2 stops too. However, this more general flow can typically keep going.

Theorem 3.7 (Flowing beyond curvature blow-up; [15]). *There exists a complete surface $(\mathcal{M}, g_0)$ of bounded curvature such that the subsequent Ricci flow $g(t)$ given by Theorem 3.2 exists for all time $t \geq 0$, and that for some times $0 < \tilde{T} < t_0 < \infty$ we have:*

- *The curvature of $g(t)$ is unbounded as $t \uparrow \tilde{T}$, but bounded on $[0, T_0]$ for all $T_0 \in [0, \tilde{T})$.*
- *The curvature is unbounded for each $t \geq t_0$.*

Clearly, the Hamilton-Shi flow will agree with $g(t)$ above until time $\tilde{T}$, when it stops, leaving $g(t)$ to continue alone.

An earlier result of Cabezas-Rivas and Wilking that we state in Theorem 4.2 constructs examples of Ricci flows in dimensions four and higher, with nonnegative complex sectional curvature, with unbounded curvature precisely for $t \geq 1$. In fact, more exotic behaviour can be engineered in which the flow passes back and forth between periods in which the curvature of the flow is unbounded and bounded, such as in the following theorem and Figure 3.3.

Theorem 3.8 (Bursts of unbounded curvature; [15]). *There exists a complete surface $(\mathcal{M}, g_0)$ of bounded curvature such that the subsequent Ricci flow $g(t)$ given by Theorem 3.2 exists for all time $t \geq 0$, and such that for some times $0 < \tilde{T} < t_0 < t_1 < t_2 < \infty$ we have:*

- *The curvature of $g(t)$ is bounded on $[0, T_0]$ for all $T_0 \in [0, \tilde{T})$, but unbounded on $[0, \tilde{T})$.*
- *The curvature is unbounded for each $t \in [t_0, t_1]$.*
- *The curvature is bounded for $t \geq t_2$.*

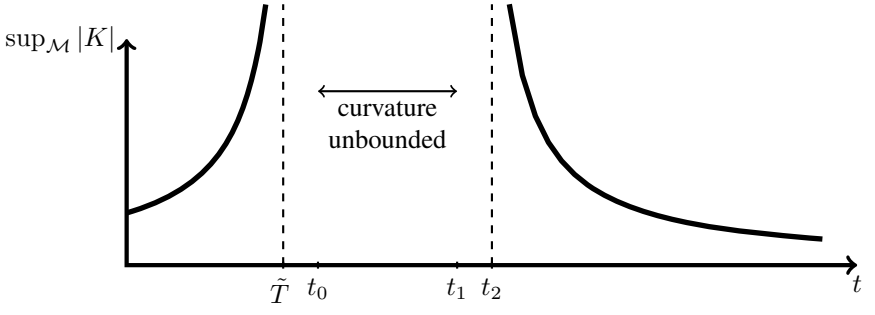


Figure 3.3. The flow in Theorem 3.8 switches back and forth between bounded and unbounded curvature, remaining always complete. The Hamilton-Shi flow ends at $t = \tilde{T}$.

Of course, by uniqueness, the only lever we have to engineer this behaviour is the choice of the initial metric. Once the flow starts, it is determined forever.

Now we have a Ricci flow starting with an arbitrary initial surface, we can return to the question of whether this Ricci flow will geometrise the surface, finding a metric of constant curvature. If we restrict our discussion to initial surfaces $(\mathcal{M}, g_0)$ that are conformal to a hyperbolic metric H, then by definition of T in Theorem 3.2, our Ricci flow must exist for all $t \geq 0$ and we can ask the question as to whether the flow will manage to converge to H. We find that it does:

Theorem 3.9 (Ricci flow geometrises general hyperbolic surfaces; [13]). *If $(\mathcal{M}, g_0)$ has a conformally equivalent, complete, hyperbolic metric H, then the Ricci flow $g(t)$ from Theorem 3.2 exists for all $t \geq 0$ and finds H in the sense that*

$$\frac{g(t)}{2t} \to H$$

smoothly locally as $t \to \infty$.

We view this type of result not as a route to giving an alternative proof of the Uniformisation theorem for general surfaces, but partly as a stepping stone to proving higher-dimensional uniformisation results, partly as a way of comparing the geometry of general metrics to that of conformal constant curvature metrics – for example the determinant of the Laplacian behaves well under Ricci flow [1] – and partly as a means for understanding how Ricci flow organises itself in order to find a special metric. In the latter direction, it is interesting to apply Theorem 3.9 in the case that $\mathcal{M}$ is noncompact (for example, the open disc) and g_0 is a metric supplied from Theorem 3.6. In this case we know on one hand that

$\frac{g(t)}{2t}$ converges to a hyperbolic metric, i.e. one of constant curvature -1, and on the other hand that it remains with unbounded curvature for all time. These two assertions are not contradictory because the convergence to a hyperbolic metric is smooth *local* convergence. In other words, the Ricci flow cannot prevent unbounded curvature, but it does organise itself to force regions of large curvature out towards spatial infinity.

4. Flows with unbounded curvature in higher dimensions

It is interesting to speculate as to how many of the results of the last section generalise to higher dimensions. Certainly the existence theory in Theorem 3.2 cannot possibly be expected to hold in such generality because one can choose smooth, complete, three-dimensional Riemannian manifolds of unbounded curvature that we cannot hope to evolve under Ricci flow, even for a short time. For example, one could take the underlying manifold to be $S^2 \times \mathbb{R}$ and endow it with a warped product metric so that metrically it consists of an infinite chain of three-spheres connected by thinner and thinner (and longer and longer) necks, as in Figure 4.1. However small we take $\varepsilon > 0$, if we pick a neck that is sufficiently thin and long, then the Ricci flow will be inclined to pinch it within time ε. (See [38, §1.3.2] for more about this type of neck-pinch singularity.)

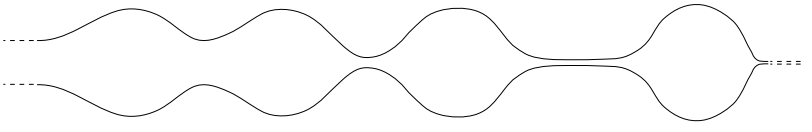


Figure 4.1. $S^2 \times \mathbb{R}$ with thinner and thinner necks

One can, however, hope to flow manifolds of unbounded curvature that also satisfy an appropriate nonnegative curvature condition that can rule out neck-pinch type singularities happening in an arbitrarily short time. An interesting scenario would be to consider complete three-dimensional manifolds of nonnegative Ricci curvature. It is very likely that Ricci flow knows how to flow such manifolds, preserving the nonnegativity of the Ricci curvature, in a unique way.

In 1989, Shi [35] classified complete three manifolds of nonnegative Ricci curvature that have *bounded curvature*. The key step, following Hamilton [17], was to flow forwards using the Ricci flow and prove that either the Ricci curvature becomes strictly positive immediately, or the manifold splits locally as a product. Being able to drop the bounded curvature assumption while still being able to flow even for a short time would have yielded a classification of *all* three manifolds of nonnegative Ricci curvature, which is the most natural class to consider. This classification had to wait over twenty years, and the development of alternative minimal surface techniques by Liu [26]. The natural Ricci flow question of existence with unbounded curvature remains open.

A very natural situation in which the existence side of the theory *has* been successfully developed is the case that we flow manifolds of nonnegative *complex* sectional curvature. This curvature condition implies nonnegative sectional curvature, and is implied by nonnegative curvature operator; it is preserved under Ricci flow [3]. Cabezas-Rivas and Wilking proved existence of solutions in this situation, retaining the curvature condition.

Theorem 4.1 (Cabezas-Rivas and Wilking [3, Theorem 1]). *Given any complete Riemannian manifold $(\mathcal{M}, g_0)$ with nonnegative complex sectional curvature, there exists a Ricci flow $g(t)$ for $t \in [0, \varepsilon)$, some $\varepsilon > 0$, with $g(0) = g_0$ and with $g(t)$ having nonnegative complex sectional curvature for each $t \in [0, \varepsilon)$.*

It is an interesting open question to prove uniqueness of this solution – conceivably there could be many other solutions with the same initial metric. Indeed, *a priori* the Ricci flow could have infinitely many nonunique branches at each instant of time, as is the case for parabolic equations on bounded domains.

Except in low dimensions, the restriction of nonnegative complex sectional curvature does not in itself enforce boundedness of the curvature of a Ricci flow as time advances:

Theorem 4.2 (Cabezas-Rivas and Wilking [3, Theorem 4, Corollary 3]). *There exist Ricci flows with nonnegative complex sectional curvature for $t \in [0, \infty)$ with unbounded curvature for all time, and others with unbounded curvature precisely for $t \geq 1$. On the other hand, if for an n-dimensional Riemannian manifold $(\mathcal{M}, g_0)$ with nonnegative complex sectional curvature there exists $v_0 > 0$ such that*

$$\mathrm{vol}_{g_0}(B_{g_0}(x, 1)) \geq v_0$$

for all $x \in \mathcal{M}$, then we may assume that the curvature of the Ricci flow arising in Theorem 4.1 is bounded above by $\frac{C(n, v_0)}{t}$.

The existence theory above is not explicit about the length of time for which the solution will persist. However, by virtue of the nonnegativity of the complex sectional curvature, a lower bound for the existence time can be read off from the initial geometry of the flow:

Theorem 4.3 (Cabezas-Rivas and Wilking [3, special case of Corollary 5]). *Given an n-dimensional manifold $\mathcal{M}$, there exists $\varepsilon > 0$ depending only on n such that we may assume that a Ricci flow arising from Theorem 4.1 exists for $t \in [0, T)$ where*

$$T = \varepsilon . \sup \left\{ \frac{\mathrm{vol}_{g_0}(B_{g_0}(x, r))}{r^{n-2}} \;\middle|\; x \in \mathcal{M}, r > 0 \right\} \in (0, \infty].$$

Finally we remark that perhaps the most natural context in which one might hope to generalise the results of Section 3 is that of higher dimensional *Kähler* Ricci flow. We leave this discussion for another occasion.

5. Flowing rough metrics or Alexandrov spaces

The entire discussion so far has considered only Ricci flows starting with *smooth* Riemannian metrics. One can also ask whether it is possible to start flowing from a metric that is not smooth, or is possibly not even a Riemannian metric at all, but perhaps a metric space with some basic structure. A Ricci flow arising in this way would generally have unbounded curvature at least in the limit $t \downarrow 0$.

One might naively think that if we can view Ricci flow as a parabolic equation, then it should be irrelevant what the regularity of an initial Riemannian metric is because Ricci flow should instantly smooth it out. However, Ricci flow will *not* smooth in the same quantified

way as the ordinary heat equation. The Harnack inequality, in the sense of [20, Section 1], can be used to prove that positive solutions of the heat equation on, say, Euclidean space must smear out at a certain rate, with the heat kernel being the extreme case. By contrast, the coefficient e^{-2u} in (3.1) can interrupt this behaviour, as is exploited implicitly in the proofs of Theorems 3.6 and 3.8. This effect from a PDE perspective says that a delta-mass for e^{2u} will remain a delta-mass for a while under the evolution equation (3.1), and will not spread out like a heat kernel. See the discussion in [45, Section 8.2].

Nevertheless, there are several situations in which it is possible to start the Ricci flow with a very rough object, perhaps a certain type of metric space. The general principle behind most results of this form, as well as some of the results we have discussed earlier in this survey, is that one approximates the initial data by a sequence of smooth Riemannian manifolds $(\mathcal{M}_i, g_i)$, flows each of these, and then tries to take a limit of these smooth Ricci flows. The real art is to prove uniform estimates on the sequence of smooth Ricci flows and their existence time so that this limit can be taken. A number of estimates of this form could be summarised loosely as

Initial metric g_0 has (i) lower curvature bound, and (ii) noncollapsing hypothesis	$\Longrightarrow$	Curvature decays like $\lvert\mathrm{Rm}_{g(t)}\rvert \leq \frac{C}{t}$

Without the noncollapsing condition (ii), no such uniform estimate can hold as we see by returning to the 'contracting cusp' example of Theorem 2.2 where the curvature decay is expected to be like C/t^2.

Although we do not attempt a complete survey of such results, we do wish to highlight one of the results of M. Simon [37] of this form.

Theorem 5.1 (Special case of [37, Theorem 7.1]). *Given a closed three or two-dimensional manifold $(\mathcal{M}, g_0)$ that satisfies*

(i) $\mathrm{Ric}_{g_0} \geq k \in \mathbb{R}$, *and*

(ii) $\mathrm{vol}_{g_0}(B_{g_0}(x,1)) \geq v_0 > 0$ *for all $x \in \mathcal{M}$,*

there exist constants $T = T(k, v_0) > 0$ and $K = K(k, v_0) > 0$ such that the Ricci flow $g(t)$ with $g(0) = g_0$ exists for $t \in [0, T)$ and satisfies, for each $t \in (0, T)$, the inequalities

(i) $\mathrm{Ric}_{g(t)} \geq -K$,

(ii) $\mathrm{vol}_{g(t)}(B_{g(t)}(x,1)) \geq \frac{v_0}{2}$ *for all $x \in \mathcal{M}$,*

(iii) $\lvert\mathrm{Rm}_{g(t)}\rvert \leq \frac{K}{t}$, *and*

(iv) $e^{K(t-s)} d_{g(s)}(p,q) \geq d_{g(t)}(p,q) \geq d_{g(s)}(p,q) - K(\sqrt{t} - \sqrt{s})$ *for all $s \in (0, t]$ and $p, q \in \mathcal{M}$, where $d_{g(s)}(p,q)$ is the Riemannian distance between p and q with respect to $g(s)$.*

As a consequence, Simon [37] was able to run the Ricci flow for a definite amount of time, starting with any metric space arising as a Gromov-Hausdorff limit of smooth closed three-manifolds satisfying (a) and (b) of Theorem 5.1 uniformly.

Of course, we would like to be able to describe *synthetically* the metric spaces that can be flowed, rather than describe them as limit points under Gromov-Hausdorff convergence. A clean situation in which this can be done was described by T. Richard [31]. The starting

point for his work is the notion of Alexandrov space, which for our purposes is a type of metric space (X, d) that satisfies a very weak notion of curvature bounded below. Such spaces automatically have integral (or infinite) Hausdorff dimension, and we require this dimension to be two. The resulting object turns out to be topologically a surface (possibly with boundary), and we require this surface to be closed. These constraints define the notion of *Alexandrov surface* – see [30] and [2] for a more precise definition and further details.

Following Alexandrov, Richard [30, 31] proved that Alexandrov surfaces can be approximated by smooth Riemannian surfaces satisfying (a) and (b) of Theorem 5.1 uniformly, and hence showed that Ricci flow smooths out an Alexandrov surface. He also proved that it does this in a unique way.

Theorem 5.2 (Flowing Alexandrov surfaces [31]). *Given an Alexandrov surface (X, d) with curvature bounded below by -1, there exists a Ricci flow $g(t)$ on a closed surface $\mathcal{M}$, for $t \in (0, T)$, some $T > 0$, with $K_{g(t)} \geq -1$ and such that the distance function $d_{g(t)} : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ of $g(t)$ converges uniformly to some distance function $d_0 : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ as $t \downarrow 0$, where $(\mathcal{M}, d_0)$ is isometric to (X, d).*

Moreover, the Ricci flow is unique up to isometry in the sense that if $(\hat{\mathcal{M}}, \hat{g}(t))$, defined for $t \in (0, \hat{T})$, is any other Ricci flow with curvature uniformly bounded from below, for which $d_{\hat{g}(t)} \to \hat{d}_0$ as $t \downarrow 0$ with $(\hat{\mathcal{M}}, \hat{d}_0)$ isometric to (X, d), then there exists a smooth map $\varphi : \mathcal{M} \to \hat{\mathcal{M}}$ that is an isometry from $(\mathcal{M}, g(t))$ to $(\hat{\mathcal{M}}, \hat{g}(t))$ for each $t \in (0, \min\{T, \hat{T}\})$.

A curious consequence of this theorem is that Ricci flow knows again how to organise itself, this time to uniquely and instantaneously endow such metric spaces with a conformal structure.

References

[1] P. Albin, C. Aldana, and F. Rochon, *Ricci flow and the determinant of the Laplacian on non-compact surfaces*, Comm. Partial Differential Equations, **38** (2013), 711–749. http://arxiv.org/abs/0909.0807.

[2] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*, Graduate studies in math. **33** AMS 2001.

[3] E. Cabezas-Rivas and B. Wilking, *How to produce a Ricci flow via Cheeger-Gromoll exhaustion*, To appear, J. Eur. Math. Soc.

[4] B.-L. Chen and X.-P. Zhu, *Uniqueness of the Ricci flow on complete noncompact manifolds*, J. Differential Geom. **74** (2006), 119–154.

[5] B.-L. Chen, *Strong uniqueness of the Ricci flow,* J. Differential Geom. **82** (2009), 363–382.

[6] X.-X Chen, P. Lu, and G. Tian, *A note on uniformization of Riemann surfaces by Ricci flow,* Proc. Amer. Math. Soc. **134** (2006), 3391–3393.

[7] B. Chow, *The Ricci flow on the 2-sphere,* J. Differential Geom. **33** (1991), 325–334.

[8] P. Daskalopoulos and M. del Pino, *On a singular diffusion equation,* Communications in Analysis and Geometry **3** (1995), 523–542.

[9] P. Daskalopoulos and C. E. Kenig, *Degenerate Diffusions*, EMS Tracts in Mathematics **1**, 2007.

[10] D. DeTurck, *Deforming metrics in the direction of their Ricci tensors,* In 'Collected papers on Ricci flow.' Edited by H. D. Cao, B. Chow, S. C. Chu, and S. T. Yau. Series in Geometry and Topology, **37**. International Press, 2003.

[11] E. DiBenedetto and D. Diller, *About a Singular Parabolic Equation Arising in Thin Film Dynamics and in the Ricci Flow for Complete* $\mathbb{R}^2$*,* In 'Partial differential equations and applications: collected papers in honor of Carlo Pucci', **177** Lecture notes in pure and applied mathematics, 103–119. CRC Press, 1996.

[12] G. Giesen and P. M. Topping, *Ricci flow of negatively curved incomplete surfaces,* Calc. Var. and PDE, **38** (2010), 357–367. http://arxiv.org/abs/0906.3309.

[13] ———, *Existence of Ricci flows of incomplete surfaces,* Comm. Partial Differential Equations, **36** (2011), 1860–1880. http://arxiv.org/abs/1007.3146.

[14] ———, *Ricci flows with unbounded curvature,* Math. Zeit., 273 (2013), 449–460. http://arxiv.org/abs/1106.2493.

[15] ———, *Ricci flows with bursts of unbounded curvature*, Preprint (2013). http://arxiv.org/abs/1302.5686.

[16] R. S. Hamilton, *Three-manifolds with positive Ricci curvature,* J. Differential Geom. **17** (1982), 255–306.

[17] ———, *Four-manifolds with positive curvature operator,* J. Differential Geom. **24** (1986), 153–179.

[18] ———, *The Ricci flow on surfaces,* Mathematics and general relativity (Santa Cruz, CA, 1986), **71** *Contemporary Mathematics*, 237–262. American Mathematical Society, Providence, RI, 1988.

[19] ———, *The formation of singularities in the Ricci flow,* Surveys in differential geometry, Vol. II (Cambridge, MA, 1993), 7–136, Internat. Press, Cambridge, MA, 1995.

[20] S. Helmensdorfer and P.M. Topping, *The geometry of Differential Harnack Estimates,* Actes du séminaire de Théorie spectrale et géométrie [Grenoble 2011-2012] **30** (2013), 77–89.

[21] J. Isenberg, R. Mazzeo, and N. Sesum, *Ricci flow on asymptotically conical surfaces with nontrivial topology,* Journal für die reine und angewandte Mathematik, **676** (2013), 227–248.

[22] L. Ji, R. Mazzeo, and N. Sesum, *Ricci flow on surfaces with cusps,* Math. Annalen, **345** (2009), 819–834.

[23] B. Kleiner and J. Lott, *Notes on Perelman's papers,* Geometry and Topology, **12** (2008), 2587–2855. [Revised version of February 2013 now available].

[24] B. Kotschwar, *An energy approach to the problem of uniqueness for the Ricci flow,* Preprint (2012) to appear in Comm. Anal. Geom. http://arxiv.org/abs/1206.3225.

[25] B. Kotschwar, *A local version of Bando's theorem on the real-analyticity of solutions to the Ricci flow,* Bull. London Math. Soc. **45** (2013), 153–158.

[26] G. Liu, *3-manifolds with nonnegative Ricci curvature,* Invent. Math. **193** (2013), 367–375.

[27] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications,* http://arXiv.org/abs/math/0211159v1 (2002).

[28] G. Perelman, *Ricci flow with surgery on three-manifolds,* http://arxiv.org/abs/math/0303109v1 (2003).

[29] ———, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds,* http://arxiv.org/abs/math/0307245v1 (2003).

[30] T. Richard, *Flot de Ricci sans bornes supérieures sur la courbure et géométrie de certains espaces métriques,* PhD thesis, University of Grenoble, 2012.

[31] ———, *Canonical smoothing of compact Alexandrov surfaces via Ricci flow,* Preprint (2012). http://arxiv.org/abs/1204.5461.

[32] A. Rodriguez, J.-L. Vázquez, and J.-R. Esteban, *The maximal solution of the logarithmic fast diffusion equation in two space dimensions,* Advances in Differential Equations, **2** (1997), 867–894.

[33] M. Rupflin, P. M. Topping, and M. Zhu, *Asymptotics of the Teichmüller harmonic map flow,* Advances in Math. **244** (2013), 874–893. http://arxiv.org/abs/1209.3783.

[34] M. Rupflin and P. M. Topping, *Teichmüller harmonic map flow into nonpositively curved targets,* http://arxiv.org/abs/1403.3195.

[35] W.-X. Shi, *Complete noncompact three-manifolds with nonnegative Ricci curvature,* J. Differential Geom. **29** (1989), 353–360.

[36] ———, *Deforming the metric on complete Riemannian manifolds,* J. Differential Geom. **30** (1989), 223–301.

[37] M. Simon, *Ricci flow of non-collapsed three manifolds whose Ricci curvature is bounded from below,* J. reine angew. Math. **662** (2012), 59–94.

[38] P. M. Topping, *Lectures on the Ricci flow*, L.M.S. Lecture notes series **325** C.U.P. (2006), http://www.warwick.ac.uk/~maseq/RFnotes.html.

[39] ———, *Ricci flow compactness via pseudolocality, and flows with incomplete initial metrics,* J. Eur. Math. Soc. (JEMS) **12** (2010), 1429–1451.

[40] ———, *Reverse bubbling in geometric flows,* Surveys in Geometric Analysis and Relativity. Advanced Lectures in Mathematics **20**, International Press (2011) eds H. Bray and W.P. Minicozzi II. Volume dedicated to Richard Schoen on the occasion of his sixtieth birthday.

[41] ———, *Uniqueness and nonuniqueness for Ricci flow on surfaces: Reverse cusp singularities,* I.M.R.N., 2012 (2012), 2356–2376. http://arxiv.org/abs/1010.2795.

[42] ———, *Remarks on Hamilton's Compactness Theorem for Ricci flow,* To appear, Journal für die reine und angewandte Mathematik. http://arxiv.org/abs/1110.3714.

[43] ———, *Uniqueness of Instantaneously Complete Ricci flows,* Preprint (2013). http://arxiv.org/abs/1305.1905.

[44] ———, *Applications of Hamilton's Compactness Theorem for Ricci flow,* To appear in IAS/Park City Mathematics Series, AMS.

[45] J.-L. Vázquez, *Smoothing and Decay Estimates for Nonlinear Diffusion Equations*, Oxford Lecture Series in Mathematics and its applications **33**, 2006.

Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK
E-mail: p.m.topping@warwick.ac.uk

Isoperimetric inequalities and asymptotic geometry

Stefan Wenger

Abstract. The m-th isoperimetric or filling volume function of a Riemannian manifold or a more general metric space X measures how difficult it is to fill m-dimensional boundaries in X of a given volume with an $(m+1)$-dimensional surface in X. The asymptotic growth of the isoperimetric functions provides large scale invariants of the underlying space. They have been the subject of intense research in past years in large scale geometry and especially geometric group theory, where the isoperimetric functions appear as Dehn functions of a group.

In this paper and the accompanying talk, I survey relationships between the asymptotic growth of isoperimetric functions and the large scale geometry of the underlying space and, in particular, fine properties of its asymptotic cones. I will furthermore describe recently developed tools from geometric measure theory in metric spaces and explain how these can be used to study the asymptotic growth of the isoperimetric functions.

Mathematics Subject Classification (2010). Primary 53C23; Secondary 49Q15, 20F65.

Keywords. Isoperimetric inequalities, Dehn functions, Gromov hyperbolicity, non-positive curvature, nilpotent groups, Carnot groups, asymptotic cones, currents in metric spaces, compactness theorems.

1. Introduction

The filling area (or first isoperimetric) function of a simply connected Riemannian manifold M measures, roughly speaking, how difficult it is to fill a closed curve in M by a disk-type surface in M. More precisely, it is defined by

$$\mathrm{FA}_0(M,r) := \sup\left\{\mathrm{Fillarea}_0(\gamma) : \gamma \text{ closed curve in } M \text{ of length} \le r\right\},$$

where $\mathrm{Fillarea}_0(\gamma)$ is the least area of a C^1-smooth (or Lipschitz) disk in M bounding γ. The filling area function is closely related to the important Dehn function widely studied in geometric group theory. Indeed, if a finitely presented group Γ acts properly discontinuously and cocompactly by isometries on a simply connected Riemannian manifold M then the Dehn function δ_Γ of Γ and the filling area function FA_0 have the same growth, thus $\delta_\Gamma(n) \simeq \mathrm{FA}_0(M,n)$, as asserted in [37] and proved in [14], [18]. Here, given functions $f, g\colon [0,\infty) \to [0,\infty)$ one writes $f \preceq g$ if there exists $C > 0$ such that

$$f(r) \le Cg(Cr + C) + Cr + C$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

for all $r \geq 0$ and one writes $f \simeq g$ if $f \preceq g$ and $g \preceq f$. The Dehn function of a finitely presented group gives a measure of the complexity of the word problem in a given group. Its growth is a quasi-isometry invariant of groups and thus, under suitable conditions, the same holds true for the filling area function. Higher dimensional filling volume (or isoperimetric) functions in M can be defined analogously. If one uses m-spheres in M as boundaries and $(m+1)$-dimensional disk-type surfaces as fillings, one obtains a homotopical filling function $\delta_M^m(r)$, sometimes called m-th order Dehn function of M. If one uses m-cycles as boundaries and $(m+1)$-chains in M one obtains a homological filling volume function $\mathrm{FV}_{m+1}(M,r)$. See Section 2 below for the precise definitions and relationships between $\delta_M^m(r)$ and $\mathrm{FV}_{m+1}(M,r)$. Like their one-dimensional sibling, the higher dimensional filling functions are quasi-isometry invariants under suitable conditions on the underlying space. They have been intensely studied in large scale geometry and in geometric group theory, where they appear as higher Dehn functions.

The purpose of the present article and the accompanying talk is to survey recent results relating the growth of the filling volume functions to the large scale geometry of the underlying space and to fine properties of its asymptotic cones. In particular, I will explain results from [49, 71–73]. The proofs of these results rely on recently developed tools from geometric measure on metric spaces. I will describe some of these tools, including Ambrosio-Kirchheim's theory [5] of metric currents and the main compactness theorem from [74].

In the rest of this introduction, I briefly describe two results in this direction which can be proved with these techniques. The main results will be explained in Sections 3 and 5.

In the seminal paper [37], Gromov proved a fundamental theorem which shows that linear growth of the filling area function for a Riemannian manifold M (or more generally, for a metric space) is equivalent to M having negative curvature on a large scale, that is, to M being hyperbolic in the sense of Gromov. In fact, Gromov even proved that if $\mathrm{FA}_0(M,r) \leq \lambda r^2$ for all sufficiently large r and a sufficiently small constant $\lambda > 0$ then M is Gromov hyperbolic. This shows, in particular, that the filling area function can not have intermediate growth between quadratic and linear. In [71], I proved the following optimal result.

Theorem 1.1 ([71]). *Let M be a Riemannian manifold. If there exist $r_0, \varepsilon > 0$ such that*

$$\mathrm{FA}_0(M,r) \leq \frac{1-\varepsilon}{4\pi} r^2$$

for all $r \geq r_0$ then M is Gromov hyperbolic and, in particular, $\mathrm{FA}_0(M,r) \simeq r$.

The constant $\frac{1}{4\pi}$ is optimal in view of the classical isoperimetric inequality in the Euclidean plane. Theorem 1.1 actually holds in the complete generality of geodesic metric spaces and an analogous version holds for Cayley graphs, see Theorem 3.2 and Corollary 3.3 below.

A class of groups for which the Dehn function has been particularly well-studied is that of nilpotent groups. While many techniques have been developed to obtain upper bounds for the growth of the Dehn function, only few techniques are known to produce lower bounds. A basic question which had remained open for a long time asks whether the Dehn function δ_Γ of every finitely generated nilpotent group Γ grows exactly polynomially, that is, $\delta_\Gamma(n) \simeq n^\alpha$ for some $\alpha \in \mathbb{N}$. In [72], I answered this question in the negative.

Theorem 1.2 ([72]). *There exists a finitely generated nilpotent group Γ such that*

$$n^2 \varrho(n) \preceq \delta_\Gamma(n) \preceq n^2 \log n$$

for a function ϱ satisfying $\varrho(n) \to \infty$ as $n \to \infty$.

The higher filling volume functions have recently also started to attract attention in relationship with large scale geometry and groups. Interesting questions remain, in particular, concerning the relationship between the growth and the existence of flats in Riemannian manifolds of non-positive curvature and more generally $\mathrm{CAT}(0)$-spaces. They have also recently been studied in the context of geometric group theory for various classes of groups, see Sections 3 and 5 for some results.

The proofs of the results above (and further results which I will describe) rely on powerful tools from geometric measure theory in metric spaces, in particular, on Ambrosio-Kirchheim's theory [5] of currents in metric spaces. This theory provides a suitable notion of chains in the generality of complete metric spaces. One of the main observations used in the proofs of the results relating filling volume and asymptotic geometry is a compactness theorem for a sequence of chains in a sequence of metric spaces, see Theorem 4.1. When applied to a sequence M_n of compact, connected and oriented Riemannian m-manifolds, this theorem says that if the diameters, the volumes and the volumes of the boundaries are uniformly bounded, then there exist a subsequence M_{n_j}, a complete metric space Z, and isometric embeddings $\varphi_j \colon M_{n_j} \hookrightarrow Z$ such that the images $\varphi_j(M_{n_j})$, viewed as integral currents, converge with respect to the flat norm to an integral current in Z. This theorem can for example be used to show that if the $(m+1)$-th filling volume function of a metric space X (satisfying a weak non-positive curvature condition) has growth $\mathrm{FV}_{m+1}(X, r) \simeq r^{\frac{m+1}{m}}$ then there is an asymptotic cone of X which contains a copy of a normed space of dimension $m+1$, see Section 5.

The plan of the paper is as follows. In Section 2, I briefly recall the definition of the Dehn and filling volume functions in a metric space, using Lipschitz maps and Lipschitz chains. Section 3 describes some of the main results which I would like to discuss in this talk, including generalizations of the theorems above. Section 4 outlines some of the analytic tools used in the proofs of the main results. Finally, Section 5 explains how the theory of currents and the compactness theorem mentioned above can be used to prove some relationships between the growth of filling volume functions and fine (Lipschitz) properties of asymptotic cones.

2. Filling volume and Dehn functions

There exist various definitions of filling volume functions in a Riemannian manifold or, more generally, in a metric space X. Depending on the type of boundaries and fillings which are admitted one obtains different notions of filling functions. For example, one may use m-dimensional spheres as boundaries and $(m+1)$-dimensional balls as fillings in order to obtain a homotopical filling function. Or one may uses singular m-cycles and singular $(m+1)$-chains in order to obtain a homological filling function. Here, I will define the filling volume functions using Lipschitz spheres and Lipschitz cycles. If X is a simplicial complex, one may alternatively define the filling volume functions using simplicial maps or simplicial chains.

2.1. Volume of Lipschitz maps. Let X be a metric space and Ω an open subset of an m-dimensional Riemannian manifold (e.g. $\mathbb{R}^m$ or S^m). The volume of a Lipschitz map

$\varphi\colon \Omega \to X$ is defined by

$$\mathrm{Vol}(\varphi) = \int_X \#\{z \in \Omega : \varphi(z) = y\}\, d\mathcal{H}^m(y),$$

where $\mathcal{H}^m$ denotes the Hausdorff m-measure on X. If φ is injective then $\mathrm{Vol}(\varphi)$ is just the Hausdorff m-measure of $\varphi(\Omega)$. If $m = 1$ and Ω is an interval then $\mathrm{Vol}(\varphi) = \mathrm{length}(\varphi)$, the length of the Lipschitz curve φ. If X is a Riemannian manifold or simplicial complex with piecewise Riemannian metric then φ is almost everywhere differentiable by Rademacher's theorem. Thus, by the area formula, $\mathrm{Vol}(\varphi)$ is the integral over Ω of the jacobian of the derivative of φ. If X is a general metric space, then φ is "metrically" differentiable almost everywhere and $\mathrm{Vol}(\varphi)$ is the integral over Ω of the jacobian of the metric derivative of φ by [40].

2.2. Homotopical filling functions. The m-th order Dehn function δ_X^m of a metric space X measures how much volume is needed to fill an m-sphere in X by an $(m+1)$-ball. More precisely, it is defined by

$$\delta_X^m(r) := \sup\{\mathrm{Fillvol}_{0,X}(\varphi) : \varphi\colon S^m \to X \text{ Lipschitz}, \mathrm{Vol}(\varphi) \le r\},$$

where $\mathrm{Fillvol}_{0,X}(\varphi)$ is the smallest volume of a Lipschitz map from the $(m+1)$-dimensional unit ball in $\mathbb{R}^{m+1}$ extending φ; respectively, infinity if no Lipschitz extension exists. If $m = 1$ one often uses the more suggestive notation

$$\mathrm{FA}_0(X, r) := \delta_X^1(r).$$

If X is a Riemannian manifold then this is just the filling area function already defined in Section 1.

2.3. Homological filling functions. The $(m+1)$-th filling volume function $\mathrm{FV}_{m+1}(X, r)$ of a metric space X is defined analogously to the function δ_X^m but employs cycles and chains instead of spheres and balls. For many purposes, a suitable notion of chains is given by the singular Lipschitz chains of [36]. In Section 4, a more involved notion of m-chains will be described which gives rise to powerful analytic tools.

By definition, a (singular) Lipschitz m-chain in X is a formal finite sum $c = \sum_{i=1}^k a_i\varphi_i$, where $a_i \in \mathbb{Z}$ and $\varphi_i\colon \Delta^m \to X$ are Lipschitz maps. Here, Δ^m denotes the standard Euclidean m-simplex. The mass (or volume) of c is defined by

$$\mathbf{M}(c) := \sum_{i=1}^k |a_i| \,\mathrm{Vol}(\varphi_i).$$

The boundary of c is the $(m-1)$-chain given by $\partial c = \sum_{i=1}^k a_i\varphi_i|_{\partial\Delta^m}$. If $\partial c = 0$ then c is called cycle. The $(m+1)$-th filling volume function $\mathrm{FV}_{m+1}(X, r)$ of X is then defined by

$$\mathrm{FV}_{m+1}(X, r) := \sup\left\{\mathrm{Fillvol}_X(z) : z \text{ Lip. } m\text{-cycle in } X \text{ with } \mathbf{M}(z) \le r\right\},$$

where $\mathrm{Fillvol}_X(z)$ is the smallest mass of a Lipschitz $(m+1)$-chain c in X with $\partial c = z$; respectively infinity if no such chain exists. We furthermore define the filling area function for curves by Lipschitz 2-chains by

$$\mathrm{FA}(X, r) := \sup\left\{\mathrm{Fillvol}_X(\gamma) : \gamma\colon S^1 \to X \text{ Lipschitz with } \mathrm{length}(\gamma) \le r\right\}.$$

It follows that $\mathrm{FA}(X,r) \le \mathrm{FV}_2(X,r)$ and $\mathrm{FA}(X,r) \le \mathrm{FA}_0(X,r)$. The notation used here differs from the notation used in some of the existing literature.

Suppose that X has the structure of a finite-dimensional simplicial complex with finitely many isometry types of cells. Then one may equivalently define the $(m+1)$-th filling volume function of X by using simplicial chains instead of Lipschitz chains. This results in a function whose growth is $\simeq$-equivalent to that of $\mathrm{FV}_{m+1}(X,r)$ by the deformation theorem of geometric measure theory [27], see also [26].

If X is an m-connected Riemannian manifold or simplicial complex then the Dehn and filling volume functions are related as follows. One has

$$\delta^2_X(r) \preceq \mathrm{FV}_3(X,r)$$

and

$$\delta^m_X(r) \simeq \mathrm{FV}_{m+1}(X,r)$$

when $m \ge 3$, see [13, 36, 76]. For $m \le 2$ the two functions are different in general, by the results in [1, 77].

2.4. Filling in a larger space. Various notions of coarse filling functions exist in the literature. Their purpose is to make sense of filling functions which are independent of the local structure of a given space. One may take a somewhat different approach to coarse filling functions by allowing fillings in a suitably enlarged space as follows. Let X and Y be metric spaces and $\iota\colon X \hookrightarrow Y$ an isometric (i.e. distance preserving) embedding. The generalized filling area function (by disks) of X with respect to Y is defined by

$$\mathrm{FA}_0(X,Y,r) := \sup\{\mathrm{Fillvol}_{0,Y}(\iota\circ\gamma) : \gamma\colon S^1 \to X \text{ Lipschitz}, \mathrm{length}(\gamma) \le r\}.$$

One may define generalized higher order Dehn functions analogously but they will not appear in this paper. Note that the space X isometrically embeds into the Banach space $\ell^\infty(X)$ of bounded functions on X, endowed with the supremum norm, via a Kuratowski embedding. Since $\ell^\infty(X)$ is an injective metric space it follows that

$$\mathrm{FA}_0(X,\ell^\infty(X),r) \le \mathrm{FA}_0(X,Y,r) \le \mathrm{FA}_0(X,r) \tag{2.1}$$

for every $r \ge 0$ and every metric space Y into which X embeds isometrically. Moreover, one has

$$\mathrm{FA}_0(X,\ell^\infty(X),r) \le \frac{1}{2\pi}r^2 \tag{2.2}$$

for every $r \ge 0$ and every metric space X. Note that X itself need not admit "fillings" of Lipschitz curves. Inequality (2.2) applies, in particular, to the Cayley graph of a finitely presented group. One defines analogously

$$\mathrm{FA}(X,Y,r) := \sup\left\{\mathrm{Fillvol}_Y(\iota\circ\gamma) : \gamma\colon S^1 \to X \text{ Lipschitz with } \mathrm{length}(\gamma) \le r\right\}$$

and obtains $\mathrm{FA}(X,Y,r) \le \mathrm{FA}_0(X,Y,r)$ and the inequalities in (2.1) hold with FA_0 replaced by FA. The generalized higher filling volume functions are defined by

$$\mathrm{FV}_{m+1}(X,Y,r) := \sup\left\{\mathrm{Fillvol}_Y(\iota_\# z) : z \text{ Lip. } m\text{-cycle in } X \text{ with } M(z) \le r\right\},$$

where $\iota_\# z$ is z when viewed as a Lipschitz cycle in Y. One obtains again (2.1) with FA_0 replaced by FV_{m+1}.

2.5. Dehn functions of groups. As already mentioned in the introduction, the filling area function FA_0 is closely related to the Dehn function from geometric group theory. Given a finitely presented group Γ with presentation $\Gamma = \langle a_1, \dots, a_d \mid r_1, \dots, r_s \rangle$, the Dehn function of Γ is defined by

$$\delta_\Gamma(n) := \max \delta(\omega),$$

where the maximum is taken over all words ω in the letters a_i of length at most n representing the identity in Γ. Furthermore, the "area" $\delta(\omega)$ of ω is the smallest number k such that ω can be written as

$$\omega = \prod_{i=1}^{k} g_i r_{j_i}^{\pm 1} g_i^{-1}$$

for some g_i and r_{j_i}, where equality is taken to be in the free group generated by the a_i. A different presentation results in a function whose growth is $\simeq$-equivalent to δ_Γ. If Γ is a group acting properly discontinuously and cocompactly by isometries on a simply connected Riemannian manifold M then $\mathrm{FA}_0(M, n) \simeq \delta_\Gamma(n)$, see [14] and [18]. The Dehn function can also be seen as a generalized filling function $\mathrm{FA}_0(X, Y, r)$ as follows. Let X be the Cayley graph of Γ with respect to the presentation above, endowed with the length metric, and let Y be the universal cover of the 2-presentation complex of Γ. Endow Y with a metric such that each 2-cell is a spherical cap. Then Y is a thickening of X, that is, Y contains X isometrically and is at finite Hausdorff distance from X. The Dehn function of Γ then has the same growth as the generalized filling function $\mathrm{FA}_0(X, Y, r)$.

Higher dimensional analogs of the Dehn function of a group can be defined for groups satisfying certain finiteness properties. The m-th order Dehn function δ_Γ^m can be defined for any group Γ of type $\mathcal{F}_{m+1}$, that is, so that Γ has a $K(\Gamma, 1)$ with finitely many $(m+1)$-cells. Roughly speaking, $\delta_\Gamma^m(n)$ measures how many $(m+1)$-cells are needed to fill (by a cellular $(m+1)$-ball) a cellular m-sphere in a $K(\Gamma, 1)$ made of at most n cells of dimension m. See [4, 14], and [13] for a precise definition and [33, 34] for a comparison with the homotopical and homological filling functions defined above.

3. Growth of filling volume functions

This section describes results which relate the growth of the filling volume and Dehn functions to the large scale geometry of the underlying space. Some more general results in this direction will also be given in Section 5.

3.1. Gromov hyperbolic spaces. The concept of δ-hyperbolicity of a metric space was first introduced and studied by Gromov in his seminal article [37]. It provides a notion of negative curvature on a large scale. A geodesic metric space X is called δ-hyperbolic if every geodesic triangle in X is δ-slim, that is, if each side of the triangle is contained in the δ-neighborhood of the two other sides. A geodesic metric space is called Gromov hyperbolic if it is δ-hyperbolic for some $\delta \geq 0$. Every geodesic metric space of finite diameter is Gromov hyperbolic. So is every simply connected Riemannian manifolds of sectional curvature bounded above by some $\kappa < 0$. Gromov proved in [37] that δ-hyperbolic metric spaces are characterized by having a filling area function which grows linearly. In fact, he proved that if X is a δ-hyperbolic metric space and $\mathrm{FA}_0(X, 100\delta) < \infty$ then $\mathrm{FA}_0(X, r) \simeq r$. Gromov furthermore established a far-reaching converse in [37]. He showed that if the filling area

function of a geodesic metric space X satisfies

$$\mathrm{FA}_0(X,r) \leq \frac{1}{4000} r^2$$

for all sufficiently large r then X is Gromov hyperbolic and thus $\mathrm{FA}_0(X,r) \simeq r$. In particular, there are no geodesic metric spaces with filling area function having growth between linear and quadratic. This result inspired many different proofs, see e.g. [52], [62], [11], [56], [24]. Using methods from geometric measure theory, I improved in [71] Gromov's result to yield the optimal possible constant.

Theorem 3.1 ([71]). *Let X be a geodesic metric space. If there exist $\varepsilon, r_0 > 0$ such that*

$$\mathrm{FA}(X,r) \leq \frac{1-\varepsilon}{4\pi} r^2 \tag{3.1}$$

for every $r \geq r_0$ then X is Gromov hyperbolic.

The constant $\frac{1}{4\pi}$ is optimal as follows from the classical isoperimetric inequality in the Euclidean plane. Previously, the best known constant was $\frac{1}{4000}$ in the case of geodesic metric spaces and $\frac{1}{16\pi}$ for a certain class of Riemannian manifolds, called 'reasonable' in [37]. Theorem 3.1 is thus already new in the case that X is a Riemannian manifold.

Even though optimal, the above theorem is not quite satisfactory in a certain sense. Namely, local deformations of the metric might result in a much bigger isoperimetric constant, destroying (3.1) without changing the large scale structure of X. Moreover, the theorem does not apply to spaces such as Cayley graphs of finitely presented groups. The following more general result remedies some of these deficiencies.

Theorem 3.2 ([71]). *Let X be a geodesic metric space such that $\mathrm{FA}(X,r) \preceq r^2$. If there exist $\varepsilon, r_0 > 0$ such that*

$$\mathrm{FA}(X, \ell^\infty(X), r) \leq \frac{1-\varepsilon}{4\pi} r^2 \tag{3.2}$$

for every $r \geq r_0$ then X is Gromov hyperbolic.

The condition that $\mathrm{FA}(X,r) \preceq r^2$ may be replaced by $\mathrm{FA}(X,Y,r) \preceq r^2$ for some thickening Y of X or by a quadratic bound on a coarse filling area function. Compare (3.2) with the bound (2.2) which holds for every metric space X.

In [71], Theorem 3.2 was stated with $\mathrm{FA}(X, \ell^\infty(X), r)$ replaced by the possibly larger function $\mathrm{FA}_0(X, \ell^\infty(X), r)$. The version above follows from the same arguments as in the proof of [71, Theorem 5.1] together with the fact, recently proven in [17], that the Hausdorff 2-measure is semi-elliptic.

From Theorem 3.2 and (2.2) one obtains the following purely group theoretic result.

Corollary 3.3. *Let $\Gamma = \langle S \mid R \rangle$ be a finitely presented group. Suppose there exist $\varepsilon > 0$ and $n_0 \in \mathbb{N}$ such that the Dehn function of Γ with respect to the presentation $\langle S \mid R \rangle$ satisfies*

$$\delta_\Gamma(n) \leq \frac{1-\varepsilon}{2\pi L} n^2$$

for every $n \geq n_0$, where $L := \max\{|r|^2 : r \in R\}$. Then Γ is Gromov hyperbolic and thus has a linear Dehn function.

It is not clear to me what the best constant is for Gromov hyperbolicity for the Dehn function of finitely presented groups.

As already mentioned, there is no group whose Dehn function grows like r^α for some $\alpha \in (1,2)$. In [12] it was proved that the set of $\alpha \geq 2$ for which there exists a finitely presented group Γ with $\delta_\Gamma(n) \simeq n^\alpha$ is dense in the interval $[2,\infty)$. Thus, there is no other gap than $(1,2)$ in the isoperimetric spectrum for the Dehn function. A similar result for surfaces of revolution can be found in [32]. As regards the higher filling volume functions it is known that geodesic Gromov hyperbolic spaces satisfying suitable conditions on the geometry on small scales have linear filling volume functions $\mathrm{FV}_{m+1}(X,r)$ for all $m \geq 1$. This was shown, in a simplicial setup, in [44]. For filling functions defined with chains with real coefficients in finitely presented groups, this was proved in [51].

3.2. Metric spaces of non-positive curvature. In this section, I describe what is known about the growth of the filling volume functions in $\mathrm{CAT}(0)$-spaces, that is, geodesic metric spaces of non-positive curvature in the sense of Alexandrov. This notion of global non-positive curvature is based on the comparison of geodesic triangles with Euclidean comparison triangles. For precise definitions and an account of the theory of $\mathrm{CAT}(0)$-spaces see e.g. [15] or [16].

In [61], it was proved that if X is a $\mathrm{CAT}(0)$-space then

$$\mathrm{FA}_0(X,r) \leq \frac{1}{4\pi} r^2 \tag{3.3}$$

for every $r \geq 0$. Lytchak and I have recently proved in [49] that (3.3) in fact characterizes $\mathrm{CAT}(0)$-spaces.

Theorem 3.4 ([49]). *Let X be a proper geodesic metric space such that* (3.3) *holds for all $r \geq 0$. Then X is a* $\mathrm{CAT}(0)$*-space.*

The only geodesic metric spaces satisfying (3.3) with $\frac{1}{4\pi}$ replaced by a strictly smaller constant are metric trees, that is, they are such that every geodesic triangle is isometric to a tripod. More precisely, if X is a geodesic metric space satisfying $\mathrm{FA}(X,r) \leq Cr^2$ for some $C < \frac{1}{4\pi}$ and all $r \geq 0$ then X is a metric tree and hence $\mathrm{FA}_0(X,r) = 0$. This follows from the same methods as used in the proof of Theorem 3.1.

We turn to the higher filling volume functions in the setting of $\mathrm{CAT}(0)$-spaces. In his seminal paper [36], Gromov showed that if X is a Hadamard manifold, that is, a complete simply connected Riemannian manifold of non-positive sectional curvature, then for every $m \geq 1$ we have

$$FV_{m+1}(X,r) \leq C_m r^{\frac{m+1}{m}} \tag{3.4}$$

for all $r \geq 0$ and for some constant C_m depending only on m. In [67, 70], I proved that this holds true for general $\mathrm{CAT}(0)$-spaces and, with an appropriate notion of chains, even for metric spaces admitting cone type inequalities, see Section 5 below. The optimal isoperimetric constant in (3.4) is only known in a few cases. In [3] it was found for Euclidean space $X = \mathbb{R}^n$ and all m. For Hadamard manifolds of dimension n, the isoperimetric inequality for domains (i.e. for $m = n-1$) with optimal Euclidean isoperimetric constant was proved in [19] for $n = 4$ and in [41] for $n = 3$.

If X is a $\mathrm{CAT}(\kappa)$-space with $\kappa < 0$, that is, X has a strictly negative upper curvature bound, then it is not difficult to show that

$$\mathrm{FV}_{m+1}(X,r) \leq Cr$$

for all r and for some constant C, see [68]. In general, it is a difficult problem to determine the growth of the filling volume functions for $\mathrm{CAT}(0)$-spaces, even when restricted to Hadamard manifolds. The following conjecture appears somewhat implicitly in [38, p. 128].

Conjecture 3.5. *If X is a proper cocompact $\mathrm{CAT}(0)$-space of Euclidean rank k then for every $m \geq k$ there exists C such that*

$$\mathrm{FV}_{m+1}(X, r) \leq Cr \tag{3.5}$$

for all $r \geq 0$.

Recall that the Euclidean rank of X is the maximal dimension of an isometrically embedded copy of Euclidean space in X. The intuition behind this conjecture comes from a general guiding principle in the theory of non-positively curved spaces which states that above the Euclidean rank a proper cocompact $\mathrm{CAT}(0)$-space should exhibit hyperbolic behavior. Instead of assuming X to be proper, cocompact and of Euclidean rank k, one may more generally formulate the conjecture for the larger class of $\mathrm{CAT}(0)$-spaces all of whose asymptotic cones have geometric dimension at most k, see also Section 5. The conjecture is known to be true in the following cases. If X is a proper cocompact $\mathrm{CAT}(0)$-space X of Euclidean rank $k = 1$ then X is Gromov hyperbolic and thus (3.5) follows from [44] together with the Lipschitz extension results of [46]. As regards the case $k > 1$, the conjecture is known to hold for symmetric spaces of non-compact type. This was asserted in [38], where a proof using projections onto maximal flats was proposed. Recently, a proof which is similar in spirit to the argument outlined in [38] but which uses projections onto suitable horospheres was given in [48]. It seems that the above conjecture remains unanswered for most cases even in the context of Hadamard manifolds.

A consequence of the above conjecture would be that isoperimetric inequalities detect the Euclidean rank. This also follows from [73], where I showed that the $(m+1)$-th filling volume function (defined using metric integral currents) have sub-Euclidean growth $o(r^{\frac{m+1}{m}})$ when m is at least the Euclidean rank. This holds more generally true for metric spaces admitting cone type inequalities, see Theorem 5.4 below. In [58], it was showed that for a simplicial complex X with $H_1(X) = H_2(X) = 0$ and for which every extremal 2-cycle for FV_3 has genus at most some fixed $g \in \mathbb{N}$, the following holds: if $\mathrm{FA}(X, r) \leq Cr^2$ and if $\mathrm{FV}_3(X, r) = o(r^{\frac{3}{2}})$ then for every $\varepsilon > 0$ there exists D_ε such that

$$\mathrm{FV}_3(X, r) \leq D_\varepsilon r^{1+\varepsilon}$$

for every $r \geq 1$. It is unknown at present, whether in $\mathbb{R}^3$, endowed with a non-positively curved metric, extremal 2-cycles have a uniform bound on their genus.

3.3. Nilpotent groups. A class of groups (resp. spaces) for which the Dehn and filling volume functions have been particularly well-studied is that of nilpotent (Lie) groups. While many results on upper bounds have been established, fewer techniques for lower bounds are known. After recalling some well-known results I will focus on the main result of [72], which provided a new lower bound for the Dehn function and helped answering a long-standing open question concerning the possible growth of Dehn functions for nilpotent groups.

Recall that a group G is nilpotent if its lower central series terminates in the trivial subgroup after finitely many steps,

$$G = G_0 > G_1 > \ldots G_c > G_{c+1} = \{e\},$$

where $G_{i+1} = [G, G_i]$ is the subgroup generated by commutators $[g, h]$ with $g \in G$ and $h \in G_i$. The smallest number c such that $G_{c+1} = \{e\}$ is called the class or step of G. A special class of nilpotent Lie groups is given by the so-called homogeneous nilpotent Lie groups or Carnot groups. By definition, a connected and simply connected nilpotent Lie group G of step c is called Carnot group if its Lie algebra $\mathfrak{g}$ admits a decomposition

$$\mathfrak{g} = V_1 \oplus \cdots \oplus V_c$$

such that $[V_1, V_i] = V_{i+1}$ for all $i = 1, \ldots, c-1$ and $[V_1, V_c] = \{0\}$. Here, $[V_1, V_i]$ denotes the smallest subspace of $\mathfrak{g}$ spanned by brackets of the form $[v, w]$ with $v \in V_1$ and $w \in V_i$. The subspace V_i is called the i-th layer of $\mathfrak{g}$. Every connected and simply connected nilpotent Lie group G of step 2 is a Carnot group. When endowed with a left-invariant Riemannian metric d_0, a Carnot group G is locally but not globally biLipschitz homeomorphic to Euclidean space. In fact, on the large scale the metric d_0 behaves like a Carnot-Carathéodory metric by [55]. Carnot groups admit scaling automorphisms. On the basis of the Lie algebra these are given by $s_t(v) = t^i v$ for $v \in V_i$ and $t \geq 0$. When endowed with a left-invariant Riemannian metric in such a way that the subspaces $V_i \subset T_e G$ are pairwise orthogonal, the differential of s_t stretches vectors in the left-translates of V_i by a factor t^i. In what follows, all nilpotent Lie groups will be endowed with a left-invariant Riemannian metric.

Recall that the n-th Heisenberg group $\mathbb{H}^n$ is the nilpotent Lie group $\mathbb{R}^{2n+1} = \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ with multiplication given by

$$(x, y, t) \star (x', y', t') = (x + x', y + y', t + t' + \langle x, y' \rangle).$$

Note that $\mathbb{H}^n$ is of step 2 and has grading of the Lie algebra given by $\mathfrak{h}_n = V_1 \oplus V_2 = \mathbb{R}^{2n} \times \mathbb{R}$. It was proved in [26] and [29] that the first Heisenberg group $\mathbb{H}^1$ has cubic filling area functions:

$$\mathrm{FA}(\mathbb{H}^1, r) \simeq \mathrm{FA}_0(\mathbb{H}^1, r) \simeq r^3.$$

In [31] it was proved that if Γ is a finitely generated nilpotent group of step c then its Dehn function is bounded by $\delta_\Gamma(r) \preceq r^{c+1}$. Previously, it had been shown in [38, $5.A_2'$], [59] that every Carnot group G of step c satisfies $\mathrm{FA}_0(G, r) \preceq r^{c+1}$. This relies essentially on the fact that the curves $t \mapsto s_t(x)$ with $t \in [0, 1]$ have length $\preceq d_0(x, e)^c$, where e is the identity, and stay at bounded distance from each other in the sense that

$$d_0(s_t(x), s_t(x')) \leq t d_0(x, x'),$$

where d_0 is the left-invariant Riemannian metric on G. In [38, $5.A_5$] a strategy has been proposed how to extend this to all simply connected nilpotent Lie groups. The upper growth of r^{c+1} is optimal. Indeed, the Dehn function of every free nilpotent group Γ of step c satisfies $\delta_\Gamma(n) \simeq n^{c+1}$ by [9] and [30]. However, not every nilpotent group of step c has Dehn function growing like r^{c+1}. Indeed, it was proved in [39], and later in [2, 53], that the higher Heisenberg groups $\mathbb{H}^n$, $n \geq 2$, have quadratic filling area functions:

$$\mathrm{FA}(\mathbb{H}^n, r) \simeq \mathrm{FA}_0(\mathbb{H}^n, r) \simeq r^2.$$

In particular, the n-th integer Heisenberg group has quadratic Dehn function whenever $n \geq 2$. Very roughly speaking, the fact that the first layer V_1 of the Lie algebra $\mathfrak{h}_n = V_1 \oplus V_2$ of $\mathbb{H}^n$ contains many 2-dimensional Lie subalgebras is responsible for the quadratic upper

bound. More generally, the n-th central power of any free nilpotent Lie group of step 2 has quadratic filling area functions for $n \geq 2$. This was asserted in [53] and proved in [78]. Recall that if G is a Carnot group of step 2 with grading $\mathfrak{g} = V_1 \oplus V_2$ of its Lie algebra then the n-th central power $G_{Z,n}$ of G is the Carnot group of step 2 whose Lie algebra has grading $\mathfrak{g}_{Z,n} = V_1' \oplus V_2$, where $V_1' := V_1 \oplus \cdots \oplus V_1$ is the n-fold direct sum of copies of V_1, and where the Lie bracket on $\mathfrak{g}_{Z,n}$ is given by

$$[v, w]' = [v_1, w_1] + \cdots + [v_n, w_n]$$

for $v, w \in \mathfrak{g}_{Z,n}$ of the form $v = v_1 + \ldots v_n + \bar{v}$ and $w = w_1 + \ldots w_n + \bar{w}$ with v_i, w_i in the i-th copy of V_1 in V_1' and $\bar{v}, \bar{w} \in V_2$. For example, the n-th Heisenberg group $\mathbb{H}^n$ is the n-th central power of the first Heisenberg group $\mathbb{H}^1$.

A basic question which had remained open for a long time asks whether the Dehn function $\delta_\Gamma(n)$ of every finitely generated nilpotent group Γ has exactly polynomial growth, that is, $\delta_\Gamma(n) \simeq n^\alpha$ for some $\alpha \in \mathbb{N}$. In [72], I answered this question in the negative by proving the following result.

Theorem 3.6 ([72]). *There exists a finitely generated nilpotent group Γ of step 2 with*

$$n^2 \varrho(n) \preceq \delta_\Gamma(n) \preceq n^2 \log n, \tag{3.6}$$

where ϱ is a function satisfying $\varrho(n) \to \infty$ as $n \to \infty$.

One can construct a whole family of such groups and they can be chosen to be lattices of Carnot groups of step 2. The super-quadratic lower bound is proved using methods from geometric measure theory in metric spaces, see Sections 4 and 5. The proof relies among other things on a compactness theorem. Consequently, the precise growth of the function ϱ is not known. The upper bound comes from [53], [78]. More precisely, let G be a Carnot group of step 2 with grading of the Lie algebra $\mathfrak{g} = V_1 \oplus V_2$. Given a subspace $W \subset V_2$ denote by G_W the Carnot group of step 2 whose Lie algebra is $\mathfrak{g}_W = V_1 \oplus (V_2/W)$. If $\mathrm{FA}_0(G, r) \simeq r^2$ then G_W satisfies

$$\mathrm{FA}_0(G_W, r) \preceq r^2 \log r. \tag{3.7}$$

This was asserted in [53] and proved in [78]. Moreover, if W is spanned by elements of the form $[v, v']$ with $v, v' \in V_1$ then $\mathrm{FA}_0(G_W, r) \simeq r^2$ by [78]. In [72], I proved the following lower bound which complements (3.7) and can be used to prove Theorem 3.6.

Theorem 3.7 ([72]). *Let G be a Carnot group of step 2 with grading $\mathfrak{g} = V_1 \oplus V_2$ of its Lie algebra. If $W \subset V_2$ is a non-trivial subspace satisfying*

$$W \cap \{[v, v'] : v, v' \in V_1\} = \{0\} \tag{3.8}$$

then the Carnot group G_W of step 2 whose Lie algebra is $\mathfrak{g}_W = V_1 \oplus (V_2/W)$ satisfies

$$\frac{\mathrm{FA}(G_W, r)}{r^2} \to \infty \quad \text{as } r \to \infty.$$

Note that if G is as in the theorem and if $\dim V_2 \geq 2 \dim V_1$ then there always exists a non-trivial subspace $W \subset V_2$ satisfying (3.8). In [72], the theorem above was only stated for the case that W is 1-dimensional. However, the exact same proof yields the version above.

One can now easily construct finitely generated nilpotent groups Γ satisfying (3.6) as lattices of Carnot groups G_W as above. A concrete example is given as follows. Let $\mathfrak{h} = V_1 \oplus V_2$ be the free nilpotent Lie algebra of step 2 with $\dim V_1 = 6$. Then there exist bases $\{e_1, \dots, e_6\}$ of V_1 and $\{e_{i,j} : 1 \le i < j \le 6\}$ of V_2 such that the Lie bracket on $\mathfrak{h}$ satisfies $[e_i, e_j] = e_{i,j}$ for all $i < j$. Then $u := e_{1,2} + e_{3,4} + e_{5,6}$ satisfies $u \neq [v_1, w_1] + [v_2, w_2]$ for all $v_i, w_i \in V_1$. Denote by H the Carnot group whose Lie algebra is $\mathfrak{h}$ and by G the 2nd central power of H. Then $\mathrm{FA}_0(G, r) \simeq r^2$ and Theorem 3.7 together with (3.7) imply that

$$r^2 \varrho(r) \le \mathrm{FA}(G_{\langle u \rangle}, r) \le \mathrm{FA}_0(G_{\langle u \rangle}, r) \le C r^2 \log r$$

for all sufficiently large r, where $\varrho(r)$ is a function satisfying $\varrho(r) \to \infty$ as $r \to \infty$. Since $G_{\langle u \rangle}$ has rational structure constants it contains a cocompact lattice Γ and hence $\delta_\Gamma(n) \simeq \mathrm{FA}_0(G_{\langle u \rangle}, n)$. Thus Γ satisfies (3.6).

Theorem 3.7 together with the results preceding it moreover yield the following characterization. Let G be a Carnot group of step 2 with grading $\mathfrak{g} = V_1 \oplus V_2$ of its Lie algebra and such that $\mathrm{FA}_0(G, r) \simeq r^2$. Let $W \subset V_2$ be a non-trivial subspace. Then G_W satisfies $\mathrm{FA}_0(G_W, r) \simeq r^2$ if and only if

$$W = \mathrm{span}\{[v, v'] : v, v' \in V_1, [v, v'] \in W\}.$$

Not so much is known yet about the higher filling functions in Carnot groups, except in the top-dimensional case and in the case of the Heisenberg groups. Let G be a Carnot group of step c and of topological dimension n with grading of the Lie algebra given by $\mathfrak{g} = V_1 \oplus \cdots \oplus V_c$. Then $\mathrm{FV}_n(G, r) \simeq r^{\frac{Q}{Q-1}}$, where Q is the homogeneous dimension defined by $Q = \sum_i i \dim V_i$. The upper bound was proved in [54] for $G = \mathbb{H}^1$ and in [66] for general G, see also [20]. The lower bound follows from the volume growth of balls [55].

The growth of the filling volume functions are known for the Heisenberg groups. Indeed, in [78] and [79] it was proved that for $m \le n - 1$ the $(m+1)$-th filling volume function in $\mathbb{H}^n$ is Euclidean,

$$\mathrm{FV}_{m+1}(\mathbb{H}^n, r) \simeq r^{\frac{m+1}{m}}, \tag{3.9}$$

while for $m = n$ it is super-Euclidean,

$$\mathrm{FV}_{n+1}(\mathbb{H}^n, r) \simeq r^{\frac{n+2}{n}},$$

and for $n < m \le 2n$ it is sub-Euclidean,

$$\mathrm{FV}_{m+1}(\mathbb{H}^n, r) \simeq r^{\frac{m+2}{m+1}}.$$

This was previously conjectured in [38]. In [78], it was proved that (3.9) holds more generally for the jet space Carnot groups $J^k(\mathbb{R}^n)$ for $m \le n - 1$ and that $\mathrm{FV}_{n+1}(J^k(\mathbb{R}^n), r) \simeq r^{\frac{n+k+1}{n}}$. For general Carnot groups G of step c the only known upper bound seems to be the following which I proved in [73]:

$$\mathrm{FV}_{m+1}(G, r) \preceq r^{1 + \frac{c^m}{1 + c + \cdots + c^{m-1}}}.$$

It remains a challenging problem to determine the (possible) growth of the filling area and filling volume functions in Carnot groups.

4. Currents in metric spaces and compactness

In this section, I discuss some of the tools used in the proofs of the theorems stated in Section 3 and of the results which will be described in Section 5. One of the main ingredients is a general compactness result, Theorem 4.1, for sequences of m-chains in a sequence of metric spaces. This theorem can be used to show, in particular, that Euclidean growth of the filling volume function $\mathrm{FV}_{m+1}(X,\cdot)$ is reflected in the $(m+1)$-dimensional Lipschitz geometry of the asymptotic cones of X. Singular Lipschitz chains are not suitable to formulate and prove such a compactness theorem since such chains do not enjoy a compactness property even in the setting of a compact ball in Euclidean space.

4.1. Currents in metric spaces. A suitable theory of chains in the generality of complete metric spaces is provided by Ambrosio-Kirchheim's theory [5] of currents in metric space. This theory, based on ideas of De Giorgi [22], generalizes the well-known Federer-Fleming theory [27] of normal and integral currents in Euclidean space to the setting of complete metric spaces. Ambrosio-Kirchheim's theory has been further developed for example in [7, 8, 21, 23, 45, 47, 67, 69, 74].

As is well-known, m-dimensional currents of finite mass in Euclidean space $\mathbb{R}^N$ in the sense of Federer-Fleming [27] are continuous linear functionals on the space $\mathcal{D}^m(\mathbb{R}^N)$ of compactly supported differential m-forms in $\mathbb{R}^N$. In the generality of a complete metric space X, differential forms do not exist. Ambrosio-Kirchheim's theory [5] employs instead $(m+1)$-tuples of real-valued Lipschitz functions on X as a substitute for differential m-forms. More precisely, the space of generalized m-forms on X is defined by

$$\mathcal{D}^m(X) := \{(f,\pi_1,\dots,\pi_m) : f,\pi_i \in \mathrm{Lip}(X),\ f \text{ bounded}\},$$

where $\mathrm{Lip}(X)$ denotes the space of real-valued Lipschitz functions on X. As a rough guiding principle, a tuple $(f,\pi_1,\dots,\pi_m)$ may be thought of as corresponding to the differential form $f d\pi_1 \wedge \cdots \wedge d\pi_m$ whenever $X = \mathbb{R}^N$ and when the f,π_i are smooth. A metric m-current T in the sense of Ambrosio-Kirchheim [5] is a multi-linear functional on $\mathcal{D}^m(X)$ satisfying three conditions. Firstly, a continuity condition requiring T to be continuous along sequences $(f,\pi_1^k,\dots,\pi_m^k)$ such that $\pi_i^k \longrightarrow \pi_i$ pointwise as $k\to\infty$ with bounded Lipschitz constants; secondly, a locality condition which forces $T(f,\pi_1,\dots,\pi_m)$ to depend on the "derivatives" of π_i rather than the π_i themselves; thirdly, a finite mass condition which asks for the existence of a finite Borel measure μ on X such that

$$|T(f,\pi_1,\dots,\pi_m)| \le \int_X |f|\, d\mu$$

for every $(f,\pi_1,\dots,\pi_m) \in \mathcal{D}^m(X)$ such that all π_i are 1-Lipschitz. The smallest such μ is called the mass of T and denoted $\|T\|$. The support of T, denoted $\mathrm{spt}\, T$, is the support of $\|T\|$. One also calls mass of T the number $\mathbf{M}(T) := \|T\|(X)$. An elementary but important example of a metric m-current in $\mathbb{R}^m$, associated to a function $\theta \in L^1(\mathbb{R}^m)$, is given by

$$[\![\theta]\!](f,\pi_1,\dots,\pi_m) := \int_{\mathbb{R}^m} f\theta \det\left(\frac{\partial \pi_i}{\partial x_j}\right) d\mathcal{H}^m.$$

Recall that the boundary and push-forward of a Federer-Fleming current in $\mathbb{R}^N$ are defined by duality with differential forms. In view of the guiding principle above it is thus natural

to define the boundary and push-forward of a metric current as follows. The boundary of a metric m-current T is the functional ∂T on $\mathcal{D}^{m-1}(X)$ defined by

$$(\partial T)(f, \pi_1, \dots, \pi_{m-1}) := T(1, f, \pi_1, \dots, \pi_{m-1}).$$

This functional satisfies all the conditions of a metric current, except maybe the finite mass condition. If ∂T also satisfies the finite mass condition then T is called a normal current. The push-forward of T under a Lipschitz map $\varphi : X \to Y$ is the m-current in Y defined by

$$\varphi_\# T(g, \tau_1, \dots, \tau_m) := T(g \circ \varphi, \tau_1 \circ \varphi, \dots, \tau_m \circ \varphi)$$

for all $(g, \tau_1, \dots, \tau_m) \in \mathcal{D}^m(Y)$.

In general, a normal m-current T in X has little to do with an m-dimensional surface in X, even when $X = \mathbb{R}^N$, since $\|T\|$ may be diffused. The space of integral m-currents in X, denoted $\mathbf{I}_m(X)$, identifies a suitable subclass of normal currents coming from generalized oriented surfaces in some sense. Roughly, a normal m-current T in X is an integral current if it can be written as the countable sum $\sum \varphi_{i\#}[\![\theta_i]\!]$ for some Lipschitz maps $\varphi_i \colon K_i \subset \mathbb{R}^m \to X$ and $\theta_i \in L^1(K_i, \mathbb{Z})$. Thus, T may be thought of as (being induced by) a generalized m-dimensional surface Σ in X with integer multiplicities, where Σ is made of countably many oriented m-dimensional Lipschitz pieces. By the important boundary rectifiability theorem, proved in [5], the boundary of an integral m-current is an integral $(m-1)$-current, that is, can again be written in the form above. Every singular Lipschitz chain $c = \sum_{k=1}^L a_k \varphi_k$ in X induces an integral m-current by $T_c := \sum_{k=1}^L \varphi_{k\#}[\![a_k 1_{\Delta^m}]\!]$. If the φ_k are injective, with pairwise almost disjoint images, then

$$D^{-1}\mathbf{M}(c) \le \mathbf{M}(T_c) \le D\mathbf{M}(c)$$

for some constant $D \ge 1$ only depending on m. The reason for having a constant is that mass measure uses the Gromov mass*-measure rather than the Hausdorff measure used to define $\mathbf{M}(c)$. One of the fundamental results proved in [5] is the closure theorem, which generalizes the corresponding Euclidean result [27] to the setting of general complete metric spaces X. It asserts that if a sequence (T_n) of integral m-currents in X converges weakly (i.e. pointwise) to some normal m-current T and if (T_n) is bounded in the sense that

$$\sup\,[\mathbf{M}(T_n) + \mathbf{M}(\partial T_n)] < \infty \tag{4.1}$$

then T is again an integral current. Another important result proved [5] is a compactness theorem which shows that if (T_n) is a bounded sequence of integral m-currents in a compact metric space X then a subsequence converges weakly to some integral current in X. As a direct consequence of this theorem and the lower semi-continuity of mass under weak convergence, one obtains a solution of the Plateau problem in compact metric spaces X. That is, for every $S \in \mathbf{I}_m(X)$ there exists $S_0 \in \mathbf{I}_m(X)$ such that $\partial S_0 = \partial S$ and

$$\mathbf{M}(S_0) = \inf\{\mathbf{M}(S') : S' \in \mathbf{I}_m(X), \partial S' = \partial S\}.$$

Ambrosio-Kirchheim [5] furthermore solved the Plateau problem for compact boundaries in some infinite dimensional Banach spaces. Using an idea of U. Lang, I solved the Plateau problem for compact boundaries in all Hadamard spaces (i.e. $\mathrm{CAT}(0)$-spaces which are complete) and all dual Banach spaces in [67]. Using Theorem 4.1 below, I extended this in [75] to non-compact boundaries, generalizing recent results in [7].

4.2. Filling volume functions via integral currents. Instead of using singular Lipschitz chains one may use integral currents in order to define the filling volume functions. More precisely, let X and Y be complete metric spaces and $\iota\colon X \hookrightarrow Y$ an isometric embedding. The filling volume in Y of an element $T \in \mathbf{I}_m(Y)$ with $\partial T = 0$ is defined to be

$$\mathrm{Fillvol}^{\mathbf{I}}_Y(T) := \inf\{\mathbf{M}(S) : S \in \mathbf{I}_{m+1}(Y), \partial S = T\}.$$

The filling area function in X with respect to Y using integral currents is defined by

$$\overline{\mathrm{FA}}(X, Y, r) := \sup\left\{\mathrm{Fillvol}^{\mathbf{I}}_Y(\iota_\# T_c) : c \text{ closed Lip. curve in } X, \mathrm{length}(c) \le r\right\}.$$

Here, T_c denotes the integral 1-current in X induced by c, that is, $T_c := c_\#[\![1_{[a,b]}]\!]$ if c is parametrized on $[a, b]$. Then $\overline{\mathrm{FA}}(X, Y, r) \le D\,\mathrm{FA}(X, Y, r)$ for a universal constant $D \ge 1$. If X is a Riemannian manifold then

$$\overline{\mathrm{FA}}(X, r) = \mathrm{FA}(X, r).$$

We furthermore define the $(m + 1)$-th filling volume function using integral currents by

$$\overline{\mathrm{FV}}_{m+1}(X, Y, r) := \sup\left\{\mathrm{Fillvol}^{\mathbf{I}}_Y(\iota_\# T) : T \in \mathbf{I}_m(X), \partial T = 0, \mathbf{M}(T) \le r\right\}$$

for all $r \ge 0$. We abbreviate $\overline{\mathrm{FV}}_{m+1}(X, r) := \overline{\mathrm{FV}}_{m+1}(X, X, r)$. Let X be a Riemannian manifold. Then one can approximate a cycle $T \in \mathbf{I}_m(X)$ by a singular Lipschitz cycle with almost the same mass, see [28, 4.2.19], and hence

$$\overline{\mathrm{FV}}_{m+1}(X, Y, r) \le D\,\mathrm{FV}_{m+1}(X, Y, r)$$

for almost every $r \ge 0$ and some constant D only depending on m; moreover,

$$\overline{\mathrm{FV}}_{m+1}(X, r) = \mathrm{FV}_{m+1}(X, r)$$

for almost every $r \ge 0$.

4.3. A general compactness theorem. The compactness theorem below, which I proved in [74], can be used to obtain relationships between the growth of the filling volume functions in a metric space and the "Lipschitz geometry" of its asymptotic cones. The result combines features of two powerful existing compactness results: Gromov's theorem for uniformly compact sequences of metric spaces and Ambrosio-Kirchheim's compactness result described above.

In order to state the result, recall the definition of the flat norm for currents, which can be thought of as the analog of the filling volume for an integral current with boundary. For given $T \in \mathbf{I}_m(X)$ it is defined by

$$\mathcal{F}_X(T) := \inf\{\mathbf{M}(R) + \mathbf{M}(S) : R \in \mathbf{I}_m(X), S \in \mathbf{I}_{m+1}(X), T = R + \partial S\}. \tag{4.2}$$

If $\partial T = 0$ then $\mathcal{F}_X(T) \le \mathrm{Fillvol}^{\mathbf{I}}_X(T)$. Convergence with respect to the flat norm implies weak (i.e. pointwise) convergence. If the ambient space X admits cone type inequalities in the sense of Definition 5.2 below then the converse is true for bounded sequences (i.e. those satisfying (4.1)) of integral currents, as I showed in [69]. The general compactness theorem alluded to above can be stated as follows.

Theorem 4.1 ([74]). *Let (X_n) be a sequence of complete metric spaces and $T_n \in \mathbf{I}_m(X_n)$ for $n \geq 1$ such that*

$$\sup_n [\operatorname{diam}(\operatorname{spt} T_n) + \mathbf{M}(T_n) + \mathbf{M}(\partial T_n)] < \infty.$$

Then there exist a subsequence (n_j), a complete metric space Z, $T \in \mathbf{I}_m(Z)$, and isometric embeddings $\varphi_j \colon X_{n_j} \hookrightarrow Z$ such that

$$\mathcal{F}_Z(T - \varphi_{j\#} T_{n_j}) \to 0.$$

Moreover, if $\partial T_n = 0$ for all n then $\operatorname{Fillvol}^{\mathbf{I}}_Z(T - \varphi_{j\#} T_{n_j}) \to 0$ as $j \to \infty$.

When applied to a sequence M_n of compact, connected and oriented Riemannian m-manifolds, possibly with boundary, Theorem 4.1 says that if the diameters, the volumes and the volumes of the boundaries are uniformly bounded, then there exist a subsequence M_{n_j}, a complete metric space Z, and isometric embeddings $\varphi_j \colon M_{n_j} \hookrightarrow Z$ such that the images $\varphi_j(M_{n_j})$, viewed as integral currents, converge with respect to the flat norm to an integral current in Z.

In general, it can be shown that the support of the limit T in Theorem 4.1 isometrically embeds into any ultralimit of the sequence (X_{n_j}, x_j), where $x_j \in \operatorname{spt} T_{n_j}$ is an arbitrary point. If the supports $\operatorname{spt} T_{n_j}$ happen to converge in the Gromov-Hausdorff sense to a metric space Y then $\operatorname{spt} T$ may be viewed as an isometric subset of Y. The inclusion $\operatorname{spt} T \subset Y$ may be strict, due to collapsing and cancellation effects. In [63], see also [64], Sormani and I exhibited sufficient conditions on the local "geometry" of the T_n which guarantee that $\operatorname{spt} T = Y$ holds.

Note that, unlike in Ambrosio-Kirchheim's compactness theorem or Gromov's compactness theorem for a sequence of metric spaces, there is no assumption on compactness or bounded "complexity" on the (supports of the) T_n in Theorem 4.1. In the special case that (X_n) is a uniformly compact sequence of metric spaces, Theorem 4.1 is a consequence of Gromov's and Ambrosio-Kirchheim's compactness theorems together with the results in [69]. The rough idea behind the proof of Theorem 4.1 in full generality is to decompose a given current T_n into the sum $T_n = \sum T_n^i$ of integral currents T_n^i such that each T_n^i has a lower bound on the mass of balls up to a certain radius, where the lower bound and the radius bound depend on i but not on n. This kind of thick-thin decomposition is inspired by the arguments in [36] and [67]. The lower bounds on the mass of balls yields uniform compactness of the sequences $(T_n^i)_n$ for fixed i and thus for each of these sequences the compactness theorems of Gromov and of Ambrosio-Kirchheim can be used.

In [47], Lang and I proved a pointed version of Theorem 4.1 for integral currents of finite mass in bounded balls, generalizing the above theorem.

5. Filling volume and asymptotic cones

The theory of metric currents and the compactness theorems mentioned above can be used to obtain relationships between the growth of the filling volume functions in a metric space and fine metric properties of its asymptotic cones. These relationships play a key role in the proofs of the results in Section 3 and of related results described below.

Let (X, d) be a metric space. An asymptotic cone of X is a metric space which one obtains, roughly speaking, by looking at X from infinitely far away. For each choice of a sequence $(p_n) \subset X$ of base points, each sequence (r_n) of scaling factors of the metric with $r_n \to \infty$, as well as each choice of a non-principal ultrafilter ω on $\mathbb{N}$ there is an associated asymptotic cone denoted by $X_\omega = (X, r_n^{-1}d, p_n)_\omega$. Elements of X_ω are equivalence classes of sequences $(x_n) \subset X$ such that the sequence of real numbers $r_n^{-1}d(x_n, p_n)$ is bounded; the distance between two points $[(x_n)]$ and $[(y_n)]$ is the ultralimit, chosen by the non-principal ultra-filter ω, of the bounded sequence of real numbers $r_n^{-1}d(x_n, y_n)$. Asymptotic cones were introduced in [65]. For properties see for example [25]. Asymptotic cones form an ideal tool in the study of the large scale geometry of a space as the local geometry of X disappears in the limit. If X and Y are quasi-isometric metric spaces then, for the same choice of parameters and appropriate choices of base points, the asymptotic cones of X and Y are biLipschitz homeomorphic. A geodesic metric space is Gromov hyperbolic if and only if all its asymptotic cones are metric trees by [37] and [38]. If X is a symmetric space of non-compact type then every asymptotic cone of X is a Euclidean building by [43]. If G is a Carnot group, endowed with left-invariant Riemannian metric d_0, then $(G, r^{-1}d_0, 0)$ converges in the pointed Gromov-Hausdorff distance to (G, d_c), where d_c is the Carnot-Carathéodory metric associated with d_0. This follows from [55]. In particular, (G, d_c) is the unique asymptotic cone of (G, d_0).

The following result, which I proved in [72], shows that admitting a quadratic isoperimetric inequality for curves is preserved in the asymptotic cones.

Theorem 5.1 ([72]). *Let X be a complete length metric space. If $\overline{\mathrm{FA}}(X, r) \preceq r^2$ then every asymptotic cone X_ω of X satisfies*

$$\overline{\mathrm{FA}}(X_\omega, r) \leq C' r^2$$

for all $r \geq 0$ and some constant C'.

The condition $\overline{\mathrm{FA}}(X, r) \preceq r^2$ may be replaced by $\overline{\mathrm{FA}}(X, Y, r) \preceq r^2$ for some thickening Y of X or by a quadratic bound for a coarse filling function. In particular, if G is a finitely presented group with quadratic Dehn function then every asymptotic cone G_ω of G admits a quadratic isoperimetric inequality for integral 1-currents, that is, $\overline{\mathrm{FA}}(G_\omega, r) \leq Cr^2$ for every $r \geq 0$. Previously, it was shown in [57] that every asymptotic cone G_ω of a finitely presented group with quadratic Dehn function is simply connected. Theorem 5.1 does not give simply connectedness but yields strong metric information of G_ω instead. This can be used to prove that certain groups cannot admit a quadratic Dehn function, for example to obtain Theorem 3.6. Note that many groups have simply connected asymptotic cones without having a quadratic Dehn function.

The proofs of Theorems 3.2 and 3.7 rely on Theorem 5.1. The proof of Theorem 3.2 is by condradiction. If X is not Gromov hyperbolic then there exists an asymptotic cone X_ω of X which is not a metric tree and hence contains a non-trivial closed Lipschitz curve c. Since X_ω has quadratic filling area function by Theorem 5.1 the curve c bounds an integral 2-current in X_ω, which is non-zero because the current associated with c is non-zero. Since integral currents are made of Lipschitz pieces it follows that X_ω receives a Lipschitz piece of $\mathbb{R}^2$ with positive Hausdorff 2-measure. By the metric differentiability [40] of Lipschitz maps and after possibly replacing X_ω by a different asymptotic cone of X, one may assume that X_ω contains an isometric copy of a 2-dimensional normed space. Since the isoperimetric constant with respect to the Hausdorff measure for curves in 2-dimensional normed spaces

is $\geq \frac{1}{4}$ and the Hausdorff 2-measure is semi-elliptic one can use (3.2) to produce a contradiction and to complete the proof. As for the proof of Theorem 3.7, one uses (3.8) together with Pansu's differentiability theorem for Lipschitz maps into Carnot groups to show that there exists a closed Lipschitz curve in (G_W, d_c) which does not bound an integral 2-current. Here, d_c denotes the Carnot-Carathéodory metric associated to a left-invariant Riemannian metric d_0 on G_W. In particular, it follows that (G_W, d_c) cannot have a quadratic filling area function and thus, by Theorem 5.1, (G_W, d_0) cannot have a quadratic filling area function either. Consequently, the filling area function of (G_W, d_0) must have super-quadratic growth, completing the proof of Theorem 3.7.

The main construction in the proof of Theorem 5.1 can be described as follows, see [72] for details.

Outline of proof. After possibly replacing X by a suitable thickening of X, one may assume that $\overline{\mathrm{FA}}(X, r) \leq Cr^2$ for all $r \geq 0$ and some constant C. Let $X_\omega = (X, r_n^{-1} d, p_n)_\omega$ be an asymptotic cone of X and let $x^1, \dots, x^k \in X_\omega$ be a chain of points such that $x^k = x^1$. We will construct $S \in \mathbf{I}_2(X_\omega)$ such that ∂S is the current induced by a piece-wise geodesic loop c connecting x^i with x^{i+1}, $i = 1, \dots, k-1$, and such that

$$\mathbf{M}(S) \leq C \operatorname{length}(c)^2.$$

Note first that each x^i comes from a sequence $(x_n^i) \subset X$. Fix a partition $0 = t_0 < t_1 < \cdots < t_k = 1$ and let $c_n : [0,1] \to X$ be such that $c_n|_{[t_i, t_{i+1}]}$ connects x_n^i with x_n^{i+1} by a piece-wise almost geodesic, parametrized proportional to arc-length. For n sufficiently large, there exists $S_n \in \mathbf{I}_2(X)$ with boundary c_n such that $\mathbf{M}(S_n) \leq C \operatorname{length}(c_n)^2$ and such that $\|S_n\|$ satisfies

$$\|S_n\|(B(x, r)) \geq \delta r^2 \tag{5.1}$$

for all $0 < r < \operatorname{dist}(x, c_n)$, where $\delta > 0$ is a constant only depending on C. An area-minimizing integral current with boundary c_n would for example satisfy this. In general, one cannot expect the existence of an area-minimizer. Nevertheless, the completeness of $\mathbf{I}_m(X)$ with respect to the mass norm allows one to prove the existence of a quasi-minimizing integral current and such a current still satisfies (5.1), as already proved in [5]. It follows that the sequence of metric spaces given by

$$Z_n := (\operatorname{spt}(S_n) \cup c_n([0,1]), r_n^{-1} d)$$

is a uniformly compact sequence. Hence, by Gromov's theorem, there exists a compact metric space Z and isometric embeddings $\varphi_n : Z_n \hookrightarrow Z$ for every n. By Ambrosio-Kirchheim's compactness theorem there exists a subsequence (n_j) such that $\varphi_{n_j \#} S_{n_j}$ converges to some $S \in \mathbf{I}_2(Z)$. Furthermore, we may assume that $\varphi_{n_j} \circ c_{n_j}$ converges uniformly to some piecewise geodesic loop c. It follows that $\partial S = c$ and

$$\mathbf{M}(S) \leq \liminf_{j \to \infty} \mathbf{M}(\varphi_{n_j \#} S_{n_j}) \leq C \liminf_{j \to \infty} \operatorname{length}(c_{n_j})^2 = C \operatorname{length}(c)^2.$$

The image of c and the support of S both lie in the ultralimit A of the sequence of subsets $\varphi_n(Z_n)$. Since A isometrically embeds into X_ω it follows that c and S isometrically embed into X_ω, moreover, $c(t_i) = x^i$. Using this construction and a suitable approximation of a given Lipschitz loop in X_ω one proves Theorem 5.1. □

If X is a Carnot group, endowed with a left-invariant Riemannian metric, and if $\overline{\mathrm{FV}}_{m+1}(X,r) \preceq r^{\frac{m+1}{m}}$ then the asymptotic cone X_ω of X satisfies $\overline{\mathrm{FV}}_{m+1}(X_\omega, r) \le Cr^{\frac{m+1}{m}}$ for some constant C and all $r \ge 0$ by [72]. It is not known in which generality Euclidean growth of the filling volume functions passes to the asymptotic cones for more general metric spaces.

For a large class of spaces, Euclidean growth of the filling volume functions results in non-trivial biLipschitz pieces in some of its asymptotic cones. The class of spaces is the following.

Definition 5.2. A complete metric space X is said to admit a cone type inequality for $\mathbf{I}_m(X)$ if there exists $C \ge 0$ such that

$$\mathrm{Fillvol}^{\mathbf{I}}_X(T) \le C \operatorname{diam}(\operatorname{spt} T)\mathbf{M}(T) \tag{5.2}$$

for every $T \in \mathbf{I}_m(X)$ with $\partial T = 0$.

Banach spaces, Hadamard spaces, and complete metric spaces X with a convex metric all admit cone type inequalities for $\mathbf{I}_m(X)$ for all m. This is more generally true for all complete metric spaces X admitting a bounded combing in the following sense. There exist C and $Cd(x,y)$-Lipschitz curves $c_{x,y}\colon [0,1] \to X$ from x to y for all $x, y \in X$ such that

$$d(c_{x,y}(t), c_{x,z}(t)) \le Cd(y,z)$$

for all triples $x, y, z \in X$ and all $t \in [0,1]$. Finally, if X is an m-connected Riemannian manifold or finite dimensional simplicial complex on which a combable group acts properly discontinuously and cocompactly by isometries then X admits a cone type inequality for $\mathbf{I}_m(X)$, see [26]. Admitting a cone type inequality for $\mathbf{I}_1(X)$ is equivalent to admitting a quadratic isoperimteric inequality for $\mathbf{I}_1(X)$, that is,

$$\overline{\mathrm{FA}}(X,r) \le C'r^2$$

for all $r \ge 0$. In [67], I showed that if a complete metric space X satisfies cone type inequalities for $\mathbf{I}_k(X)$ for $k = 1, \dots, m$ then

$$\overline{\mathrm{FV}}_{m+1}(X,r) \le C'r^{\frac{m+1}{m}} \tag{5.3}$$

for every $r \ge 0$, where C' only depends on the constants of the cone type inequalities. For Riemannian manifolds admitting cone type inequalities (for Lipschitz chains) this was previously proved in [36].

For spaces admitting cone type inequalities, the filling volume functions detect the asymptotic rank, defined as follows, see [73].

Definition 5.3. The asymptotic rank of a metric space X, denoted by $\mathrm{Asrk}(X)$, is the supremum over $n \in \mathbb{N}$ such that there exists an asymptotic cone X_ω of X and a biLipschitz embedding $\varphi : K \hookrightarrow X_\omega$ for some compact subset $K \subset \mathbb{R}^n$ with $\mathcal{H}^n(K) > 0$.

The asymptotic rank is a quasi-isometry invariant of metric spaces, see [73]. For a general metric space X one has the bounds

$$\mathrm{Asrk}(X) \le \sup\{\mathrm{Topdim}(C) : C \subset X_\omega \text{ cpt}, X_\omega \text{ an asymptotic cone of } X\}$$

and

$$\mathrm{Asrk}(X) \ge \sup\{n \in \mathbb{N} : \exists \psi : \mathbb{R}^n \to X \text{ quasi-isometric}\}.$$

In the above, $\mathrm{Topdim}(C)$ denotes the topological dimension of C. If X is a Hadamard space then $\mathrm{Asrk}(X)$ is the maximal geometric dimension of an asymptotic cone of X. If X is moreover proper and cocompact then $\mathrm{Asrk}(X)$ coincides with its Euclidean rank. More generally, if X is a proper cocompact length space with a convex metric then

$$\mathrm{Asrk}(X) = \sup\{n \in \mathbb{N} : \exists V \text{ } n\text{-dim. normed space and } \psi : V \to X \text{ isometric}\}.$$

This follows from [42]. If X is a Carnot group, endowed with a left-invariant Riemannian or Carnot-Carathéodory metric, then $\mathrm{Asrk}(X)$ is the highest dimension of a Lie subalgebra contained in the first layer. This follows from [6, 50], and [55]. If X is the mapping class group of a surface of finite type then $\mathrm{Asrk}(X)$ is the maximal rank of an abelian subgroup. This follows from [10] and the inequalities above.

The following theorem, which I proved in [73], shows that the filling volume functions detect the asymptotic rank for a large class of metric spaces. Recall that a metric space (X, d) is said to be quasi-convex if there exists C such that any two points $x, y \in X$ can be joined by a curve of length at most $Cd(x, y)$.

Theorem 5.4 ([73]). *Let X be a complete quasi-convex metric space and $m \geq 1$. Suppose X admits cone type inequalities for $\mathbf{I}_k(X)$ for $k = 1, \dots, m$. Then:*

(i) *If $m < \mathrm{Asrk}(X)$ then there exists $\varepsilon > 0$ such that*

$$\overline{\mathrm{FV}}_{m+1}(X, r) \geq \overline{\mathrm{FV}}_{m+1}(X, \ell^\infty(X), r) \geq \varepsilon r^{\frac{m+1}{m}}$$

for all $r > 0$ large enough.

(ii) *If $m \geq \mathrm{Asrk}(X)$ then*

$$\limsup_{r\to\infty} \frac{\overline{\mathrm{FV}}_{m+1}(X, r)}{r^{\frac{m+1}{m}}} = 0.$$

The proof of part (ii) of Theorem 5.4 is by contradiction and uses the general compactness theorem, Theorem 4.1. The proof can be outlined as follows.

Outline of proof. Suppose there exists a sequence $(T_n) \subset \mathbf{I}_m(X)$ of cycles such that $\mathbf{M}(T_n) \to \infty$ and

$$\liminf_{n\to\infty} \frac{\mathrm{Fillvol}^{\mathbf{I}}_X(T_n)}{\mathbf{M}(T_n)^{\frac{m+1}{m}}} > 0. \tag{5.4}$$

By a thick-thin decomposition procedure, applied to T_n, one obtains a new cycle $T'_n \in \mathbf{I}_m(X)$ such that (5.4) holds for T_n replaced with T'_n and such that

$$\mathrm{diam}(\mathrm{spt}\, T'_n) \leq E\mathbf{M}(T'_n)^{\frac{1}{m}}$$

for every n, where E is a constant independent of n. By the isoperimetric inequality (5.3), there exists $S_n \in \mathbf{I}_{m+1}(X)$ such that $\partial S_n = T'_n$ and such that $\mathbf{M}(S_n) \leq C\mathbf{M}(T'_n)^{\frac{m+1}{m}}$. The S_n can be chosen such that

$$\mathrm{diam}(\mathrm{spt}\, S_n) \leq E'\mathbf{M}(S_n)^{\frac{1}{m+1}}$$

for every n, where E' is a constant independent of n. This can be proved with the same arguments as in the outline of the proof of Theorem 5.1. Define a metric space by $X_n =$

$(X, r_n^{-1}d)$, where $r_n := \mathbf{M}(T'_n)^{\frac{1}{m}}$, and view S_n as a current in X_n. Then (X_n) and S_n satisfy the hypotheses of Theorem 4.1. There thus exists a subsequence (n_j) and a metric space Z such that X_{n_j} embeds isometrically into Z and S_{n_j}, when viewed as a current in Z, converges to some $S \in \mathbf{I}_{m+1}(Z)$ with respect to the flat norm. Since X admits cone type inequalities, the results in [69] and the fact that (5.4) imply that $S \neq 0$. Since $\operatorname{spt} S$ isometrically embeds into some asymptotic cone of X and since S is "made of" biLipschitz pieces from $\mathbb{R}^{m+1}$ it follows that $\operatorname{Asrk}(X) \geq m+1$, a contradiction. □

In view of Conjecture 3.5 it is natural to ask the following question.

Question 5.5. *Let X be a complete quasi-convex metric space admitting cone type inequalities for $\mathbf{I}_k(X)$ for $k = 1, \ldots, m$. Is it true that $\overline{\mathrm{FV}}_{m+1}(X, r)$ grows at most linearly if $m \geq \operatorname{Asrk}(X)$?*

Acknowledgments. I would like to thank Ruth Kellerhals for useful comments and suggestions concerning the presentation of the results.

References

[1] Abrams, A., Brady, N., Pallavi, D., Moon, D., and Young, R., *Homological and homotopical Dehn functions are different*, Proc. Nat. Acad. Sciences, to appear.

[2] Allcock, D., *An isoperimetric inequality for the Heisenberg groups*, Geom. Funct. Anal. **8** (1998), no. 2, 219–233.

[3] Almgren, F., *Optimal isoperimetric inequalities*, Indiana Univ. Math. J. **35** (1986), no. 3, 451–547.

[4] Alonso, J., Wang, X., and Pride, S., *Higher-dimensional isoperimetric (or Dehn) functions of groups*, J. Group Theory **2** (1999), no. 1, 81–112.

[5] Ambrosio, L. and Kirchheim, B., *Currents in metric spaces*, Acta Math. **185** (2000), no. 1, 1–80.

[6] ———, *Rectifiable sets in metric and Banach spaces*, Math. Ann. **318** (2000), no. 3, 527–555.

[7] Ambrosio, L. and Schmidt, T., *Compactness results for normal currents and the Plateau problem in dual Banach spaces*, Proc. Lond. Math. Soc. (3) **106** (2013), no. 5, 1121–1142.

[8] Ambrosio, L. and Wenger, S., *Rectifiability of flat chains in Banach spaces with coefficients in $\mathbb{Z}_p$*, Math. Z. **268** (2011), no. 1–2, 477–506.

[9] Baumslag, G., Miller, C. F., and Short, H., *Isoperimetric inequalities and the homology of groups*, Invent. Math. **113** (1993), no. 3, 531–560.

[10] Behrstock, J. and Minsky, Y., *Dimension and rank for mapping class groups*, Ann. of Math. (2) **167** (2008), no. 3, 1055–1077.

[11] Bowditch, B., *A short proof that a subquadratic isoperimetric inequality implies a linear one*, Michigan Math. J. **42** (1995), no. 1, 103–107.

[12] Brady, N. and Bridson, M., *There is only one gap in the isoperimetric spectrum*, Geom. Funct. Anal. **10** (2000), no. 5, 1053–1070.

[13] Brady, N., Bridson, M., Forester, M., and Shankar, K., *Snowflake groups, Perron-Frobenius eigenvalues and isoperimetric spectra*, Geom. Topol. **13** (2009), no. 1, 141–187.

[14] Bridson, M., *The geometry of the word problem*, Invitations to geometry and topology, 29–91, Oxford Grad. Texts Math., **7**, Oxford Univ. Press, Oxford, 2002.

[15] Bridson, M. and Haefliger, A., *Metric spaces of non-positive curvature*, Grundlehren der Mathematischen Wissenschaften, 319, Springer-Verlag, Berlin, 1999.

[16] Burago, D., Burago, Yu., and Ivanov, S., *A course in metric geometry*, Graduate Studies in Mathematics, 33. American Mathematical Society, Providence, RI, 2001.

[17] Burago, D. and Ivanov, S., *Minimality of planes in normed spaces*, Geom. Funct. Anal. **22** (2012), no. 3, 627–638.

[18] Burillo, J. and Taback, J., *Equivalence of geometric and combinatorial Dehn functions*, New York J. Math. **8** (2002), 169–179.

[19] Croke, Chr., *A sharp four-dimensional isoperimetric inequality*, Comment. Math. Helv. **59** (1984), no. 2, 187–192.

[20] Coulhon, Th. and Saloff-Coste, L., *Isopérimétrie pour les groupes et les variétés*, Rev. Mat. Iberoamericana **9** (1993), no. 2, 293–314.

[21] De Pauw, Th. and Hardt, R., *Rectifiable and flat G chains in a metric space*, Amer. J. Math. **134** (2012), no. 1, 1–69.

[22] De Giorgi, E., *General Plateau problem and geodesic functionals*, Atti Sem. Mat. Fis. Univ. Modena **43** (1995), no. 2, 285–292.

[23] De Lellis, C., *Some fine properties of currents and applications to distributional Jacobians*, Proc. Roy. Soc. Edinburgh Sect. A **132** (2002), no. 4, 815–842.

[24] Drutu, C., *Cônes asymptotiques et invariants de quasi-isométrie pour des espaces métriques hyperboliques*, Ann. Inst. Fourier (Grenoble) **51** (2001), no. 1, 81–97.

[25] _____, *Quasi-isometry invariants and asymptotic cones*, Internat. J. Algebra Comput. **12** (2002), no. 1–2, 99–135.

[26] Epstein, D., Cannon, J., Holt, D., Levy, S., Paterson, M., and Thurston, W., *Word processing in groups*, Jones and Bartlett Publishers, Boston, MA, 1992.

[27] Federer, H. and Fleming, W., *Normal and integral currents*, Ann. of Math. (2) **72** (1960), 458–520.

[28] Federer, H., *Geometric measure theory*, Die Grundlehren der mathematischen Wissenschaften, Band 153 Springer-Verlag New York Inc., New York, 1969.

[29] Gersten, S. M., *Dehn functions and ℓ_1-norms of finite presentations*, Algorithms and classification in combinatorial group theory (Berkeley, CA, 1989), 195–224, Math. Sci. Res. Inst. Publ., 23, Springer, New York, 1992.

[30] ———, *Isoperimetric and isodiametric functions of finite presentations*, Geometric group theory, Vol. 1 (Sussex, 1991), 79–96, London Math. Soc. Lecture Note Ser., 181, Cambridge Univ. Press, Cambridge, 1993.

[31] Gersten, S. M., Holt, D. F., and Riley, T. R., *Isoperimetric inequalities for nilpotent groups*, Geom. Funct. Anal. **13** (2003), no. 4, 795–814.

[32] Grimaldi, R. and Pansu, P., *Remplissage et surfaces de révolution*, J. Math. Pures Appl. (9) **82** (2003), no. 8, 1005–1046.

[33] Groft, Ch., *Generalized Dehn Functions I*, preprint arXiv (2009), arXiv:0901.2303v1.

[34] ———, *Generalized Dehn Functions II*, preprint arXiv (2009), arXiv:0901.2317v1.

[35] Gromov, M., *Groups of polynomial growth and expanding maps*, Inst. Hautes Etudes Sci. Publ. Math. No. 53 (1981), 53–73.

[36] ———, *Filling Riemannian manifolds*, J. Diff. Geom. **18** (1983), no. 1, 1–147.

[37] ———, *Hyperbolic groups*, Essays in group theory, 75–263, Math. Sci. Res. Inst. Publ., 8, Springer, New York, 1987.

[38] ———, *Asymptotic invariants of infinite groups*, Geometric group theory, Vol. 2 (Sussex, 1991), 1–295, London Math. Soc. Lecture Note Ser., 182, Cambridge Univ. Press, Cambridge, 1993.

[39] ———, *Carnot-Carathéodory spaces seen from within*, Sub-Riemannian geometry, 79–323, Progr. Math., 144, Birkhäuser, Basel, 1996.

[40] Kirchheim, B, *Rectifiable metric spaces: local structure and regularity of the Hausdorff measure*, Proc. Amer. Math. Soc. **121** (1994), no. 1, 113–123.

[41] Kleiner, B., *An isoperimetric comparison theorem*, Invent. Math. **108** (1992), no. 1, 37–47.

[42] ———, *The local structure of length spaces with curvature bounded above*, Math. Z. **231** (1999), no. 3, 409–456.

[43] Kleiner, B. and Leeb, B., *Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings*, Inst. Hautes Etudes Sci. Publ. Math. No. 86 (1997), 115–197 (1998).

[44] Lang, U., *Higher-dimensional linear isoperimetric inequalities in hyperbolic groups*, Int. Math. Res. Not. 2000, no. 13, 709–717.

[45] ———, *Local currents in metric spaces*, J. Geom. Anal. **21** (2011), no. 3, 683–742.

[46] Lang, U. and Schlichenmaier, Th., *Nagata dimension, quasisymmetric embeddings, and Lipschitz extensions*, Int. Math. Res. Not. 2005, no. 58, 3625–3655.

[47] Lang, U. and Wenger, S., *The pointed flat compactness theorem for locally integral currents*, Comm. Anal. Geom. **19** (2011), no. 1, 159–189.

[48] Leuzinger, E., *Optimal higher-dimensional Dehn functions for some* $\mathrm{CAT}(0)$ *lattices*, arXiv:1205.4923, preprint 2012.

[49] Lytchak, A. and Wenger, S., in preparation.

[50] Magnani, V., *Unrectifiability and rigidity in stratified groups*, Arch. Math. (Basel) **83** (2004), no. 6, 568–576.

[51] Mineyev, I., *Higher dimensional isoperimetric functions in hyperbolic groups*, Math. Z. **233** (2000), no. 2, 327–345.

[52] Olshanskii, A. Yu., *Hyperbolicity of groups with subquadratic isoperimetric inequality*, Internat. J. Algebra Comput. **1** (1991), no. 3, 281–289.

[53] Olshanskii, A. Yu. and Sapir, M. V., *Quadratic isometric functions of the Heisenberg groups. A combinatorial proof*, Algebra, 11. J. Math. Sci. (New York) **93** (1999), no. 6, 921–927.

[54] Pansu, P., *Une inégalité isopérimétrique sur le groupe de Heisenberg*, C. R. Acad. Sci. Paris Sér. I Math. **295** (1982), no. 2, 127–130.

[55] ———, *Croissance des boules et des géodésiques fermées dans les nilvariétés*, Ergodic Theory Dynam. Systems **3** (1983), no. 3, 415–445.

[56] Papasoglu, P., *On the sub-quadratic isoperimetric inequality*, Geometric group theory (Columbus, OH, 1992), 149–157, Ohio State Univ. Math. Res. Inst. Publ., 3, de Gruyter, Berlin, 1995.

[57] ———, *On the asymptotic cone of groups satisfying a quadratic isoperimetric inequality*, J. Diff. Geom. **44** (1996), no. 4, 789–806.

[58] ———, *Cheeger constants of surfaces and isoperimetric inequalities*, Trans. Amer. Math. Soc. **361** (2009), no. 10, 5139–5162.

[59] Pittet, Ch., *Isoperimetric inequalities for homogeneous nilpotent groups*, Geometric group theory (Columbus, OH, 1992), 159–164, Ohio State Univ. Math. Res. Inst. Publ., 3, de Gruyter, Berlin, 1995.

[60] ———, *Isoperimetric inequalities in nilpotent groups*, J. London Math. Soc. (2) **55** (1997), no. 3, 588–600.

[61] Reshetnyak, Ju. G., *Non-expansive maps in a space of curvature no greater than* K, Sibirsk. Mat. Z. **9** (1968), 918–927.

[62] Short, H. (editor and contributor), *Notes on word hyperbolic groups*, Group theory from a geometrical viewpoint (E. Ghys, A. Haefliger, A. Verjovsky eds.), World Sci. Publ., River Edge, NJ, 1991.

[63] Sormani, C. and Wenger, S., *Weak convergence of currents and cancellation*, With an appendix by Raanan Schul and Wenger, Calc. Var. Partial Differential Equations 38 (2010), no. 1–2, 183–206.

[64] ———, *The intrinsic flat distance between Riemannian manifolds and other integral current spaces*, J. Diff. Geom. **87** (2011), no. 1, 117–199.

[65] van den Dries, L. and Wilkie, A., *Gromov's theorem on groups of polynomial growth and elementary logic*, J. Algebra **89** (1984), no. 2, 349–374.

[66] Varopoulos, N., *Analysis on Lie groups*, J. Funct. Anal. **76** (1988), no. 2, 346–410.

[67] Wenger, S., *Isoperimetric inequalities of Euclidean type in metric spaces*, Geom. Funct. Anal. **15** (2005), no. 2, 534–554.

[68] ———, *Filling invariants at infinity and the Euclidean rank of Hadamard spaces*, Int. Math. Res. Not. 2006, Art. ID 83090, 33 p.

[69] ———, *Flat convergence for integral currents in metric spaces*, Calc. Var. Partial Differential Equations 28 (2007), no. 2, 139–160.

[70] ———, *A short proof of Gromov's filling inequality*, Proc. Amer. Math. Soc. **136** (2008), no. 8, 2937–2941.

[71] ———, *Gromov hyperbolic spaces and the sharp isoperimetric constant*, Invent. Math. **171** (2008), no. 1, 227–255.

[72] ———, *Nilpotent groups without exactly polynomial Dehn function*, J. Topol. **4** (2011), no. 1, 141–160.

[73] ———, *The asymptotic rank of metric spaces*, Comment. Math. Helv. **86** (2011), no. 2, 247–275.

[74] ———, *Compactness for manifolds and integral currents with bounded diameter and volume*, Calc. Var. Partial Differential Equations 40 (2011), no. 3–4, 423–448.

[75] ———, *Plateau's problem for integral currents in locally non-compact metric spaces*, Adv. Calc. Var. **7** (2014), no. 2, 227–240..

[76] White, B., *Mappings that minimize area in their homotopy classes*, J. Diff. Geom. **20** (1984), no. 2, 433–446.

[77] Young, R., *Homological and homotopical higher-order filling functions*, Groups Geom. Dyn. **5** (2011), no. 3, 683–690.

[78] ———, *Filling inequalities for nilpotent groups through approximations*, Groups Geom. Dyn. **7** (2013), no. 4, 977–1011.

[79] ———, *High-dimensional fillings in Heisenberg groups*, preprint arXiv (2010), arXiv:1006.1636v1.

Department of Mathematics,University of Fribourg, Chemin du Musée 23, 1700 Fribourg, Switzerland
E-mail: stefan.wenger@unifr.ch

The cubical route to understanding groups

Daniel T. Wise

Abstract. We survey the methodology and key results used to understand certain groups from a cubical viewpoint, and describe the ideas linking 3-manifolds, cube complexes, and right-angled Artin groups. We close with a collection of problems focused on groups acting on CAT(0) cube complexes.

Mathematics Subject Classification (2010). 20F67, 57M99.

Keywords. CAT(0) cube complexes, right-angled Artin groups, 3-manifolds.

1. Introduction

This survey presents some of the ingredients in a strategy for understanding groups using cube complexes. Developing and applying this has been my research agenda and its details have become increasingly explicit over the past decade. In the hyperbolic case this research program has recently arrived at a major goal: the virtual Haken problem for hyperbolic cube complexes which was completed by Agol, and its associated application towards closed hyperbolic 3-manifolds which was enabled by the closed surfaces of Kahn and Markovic. Some of these developments were precipitated by my use of cubes to understand the underlying structure of hyperbolic groups with a quasiconvex hierarchy, which was in turn partially aimed at the analogous conclusions for cusped hyperbolic manifolds. I describe here some of the tools leading to these developments, mentioning especially work of and with Bergeron, Haglund, Hruska, Hsu, and Sageev. Many of the issues summarized here are treated in a sketchy but intuitive fashion in [67], and in a more dense traditional fashion in [60]. There is still much to do in the cubical realm and we close the survey with a collection of related problems.

Our strategy to understand a group G follows the following scheme:

- Find codimension-1 subgroups and obtain an action of G on the dual cube complex $\widetilde{X}$.
- If the codimension-one subgroups are nice enough then G acts on $\widetilde{X}$ with good finiteness properties.
- Find a finite index subgroup $G' \subset G$ such that $G'\backslash\widetilde{X}$ is very organized in the sense that it is a "special cube complex".
- Obtain an embedding $G' \subset A(\Gamma)$ into a raag, and conclude that G has many nice properties since raags are very nice groups.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

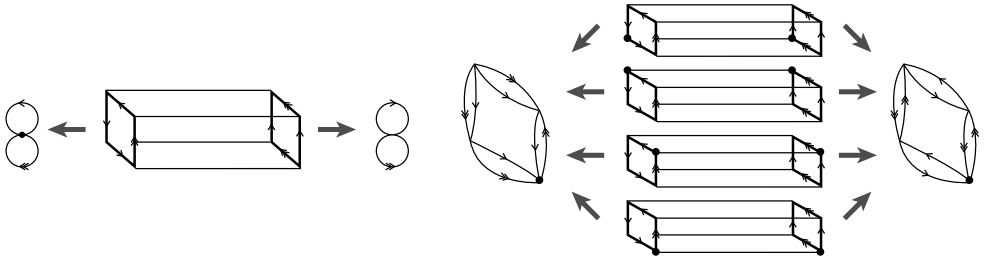


Figure 1.1. A clean cover corresponding to the splitting of $\langle a, b, c, d \mid aaab = cdc^{-1}d^{-1}\rangle$ as an amalgamated free product. The analogous situation for the HNN extension $\langle a, b, c, d, t \mid (aaab)^t = cdc^{-1}d^{-1}\rangle$ is harder.

My trek into this topic began with an interest in residual finiteness that was influenced by the values of the combinatorial group theorists active in the 1960s, and stirred by curiosity about subgroup separability and surfaces in 3-manifolds. In the early 1990s, the simplest examples whose residual finiteness was not understood were cyclic HNN extensions of free groups. These are groups of the form: $\langle a, b, t \mid t^{-1}Ut = V\rangle$ where U, V represent elements of $\langle a, b\rangle$ that do not have conjugate powers. I approached the problem by thinking of covering spaces of a space X that decomposes as a graphs of spaces (see Figure 1.1 for the analogous situation of a free product of two free groups amalgamating a cyclic subgroup). Such a decomposition arises when X is a nonpositively curved square complex with a $\mathcal{VH}$*-structure* in the sense that each of its 1-cells is either "vertical" or "horizontal" and the attaching map of each square alternates between vertical and horizontal edges. The graph of space structure arises from a singular foliation by vertical leaves formed from line segments parallel to the vertical edges. The simplest scenario is where U, V are cyclically reduced words of the same length, in which case the group is the fundamental group of a graph of spaces whose vertex space is a bouquet of circles, and whose edge space is an attached cylinder corresponding to $t^{-1}Ut = V$ that is divided into squares by adding 1-cells parallel to the t edges at the top and bottom. The strategy, as executed in [61], was to first pass to a finite cover $\widehat{X} \to X$ that is *clean* in the sense that the attaching maps of its edge spaces are embeddings. It is straightforward to see that the fundamental groups of these clean graphs of spaces are residually finite. Moreover, there is considerable flexibility in producing finite covering spaces of clean complexes, and this enables one to show that every f.g. subgroup of $\pi_1\widehat{X}$ and hence $\pi_1 X$ is separable. Indeed, this generalizes the ease with which finite covering spaces of graphs can be produced. We refer to Figure 1.1 for a clean cover of the graph of spaces corresponding to the free product of two free groups amalgamating a cyclic subgroup.

These separability ideas were subsequently generalized to arbitrary compact clean $\mathcal{VH}$ square complexes X with $\pi_1 X$ hyperbolic relative to abelian subgroups [64]. A promising application was towards the Dehn complexes of prime alternating links, which are compact nonpositively curved $\mathcal{VH}$ square complexes. A computer experiment showed that many of these Dehn complexes have clean finite covers, and so it became a mission to understand that all Dehn complexes of prime alternating links are virtually clean. In 1997 it seemed that this had introduced a rather new idea towards understanding separability that followed in the footsteps of Hall's work [28], but we now understand in retrospect that it was also an analogue of Scott's right-angled reflection group idea that was used to understand separability

of surface groups [57].

An intriguing early step forward was to show that negatively curved $\mathcal{VH}$-complexes are virtually clean [62]. But although there was an outline of a proof based on the height of an edge group in the splitting, a complete argument that applied to every prime alternating link Dehn complex was unavailable for many years - the malnormal special quotient theorem was the key missing ingredient. I was hopeful that all knot groups were (virtually) fundamental groups of nonpositively curved ($\mathcal{VH}$) square complexes, and went on a fruitless search to generalize the Dehn complex. In another direction, I hoped there was a prospect of generalizing the ideas by showing that there is always a finite index subgroup that acted freely on the product of trees.

Forays into understanding residual finiteness in small-cancellation groups in 1999 led to the study of certain codimension-one subgroups arising from immersed codimension-one graphs. It was then natural to apply Sageev's work [53] on the dual cube complex to cubulate small-cancellation groups [63]. This was an eye-opener for me, as Sageev's construction provided access to a combinatorial geometric structure from ingredients I already had experience with. With Hsu, we cubulated graphs of free groups with cyclic edge groups, and then set out to understand groups with more general splittings [31, 34]. With Hruska, we then made a detailed study of the finiteness properties with papers focusing on the finiteness properties in [29, 30] and generalizing finiteness properties observed in [18, 47, 54, 63].

We began studying raags with Hsu, first showing that they are subgroups of Coxeter groups and hence embed in $SL_n(\mathbb{Z})$ [13, 32], and then examining the separability of the quasi-isometrically embedded subgroups of the raag $\langle\ a, b, c, d \mid [a, b], [b, c], [c, d]\ \rangle$ corresponding to the complement in S^3 of a length 4 chain [33]. Although our proof was not written in that generality, we noticed the separability of complexes that map by a local isometry, but our effort to prove the separability of arbitrary quasi-isometrically embedded subgroups was frustrated.

In 2002, Haglund and I generalized the idea of "clean square complexes" to "special cube complexes" and this dropped the $\mathcal{VH}$ restriction and allowed arbitrary dimensions. The goal of our definition was to obtain separability by extending the notion of canonical completion and retraction to higher dimensions [25]. Surprisingly, this definition magically coincided with the definition of a local isometry to the cube complex of a raag. We immediately set out to generalize the work in [62] to higher dimensions - although it took many years until we finally completed this in [27]. In view of the dual cube complex construction, special cube complexes opened the possibility of reaping algebraic consequences from the hidden cubical geometry of many groups.

2. CAT(0) cube complexes and npc cube complexes

An *n-cube* is a copy of $[-1, 1]^n$. Its *subcubes* are the subspaces obtained by restricting some coordinates to either $+1$ or -1. A *cube complex* is a complex built from cubes glued together along subcubes. The data describing a cube complex is entirely combinatorial; we require that the gluing maps be modeled by (local) isometries where we regard n-cubes as being isometric copies of the corresponding subspaces of $\mathbb{E}^n$.

A *flag complex* S is a simplicial complex with the property that $n + 1$ vertices span an n-simplex if and only if they are pairwise adjacent. Thus a flag complex is determined completely by its 1-skeleton: We can reconstruct S from S^1 by adding an n-simplex for

each complete graph $K(n)$ in S. For instance, a graph S is a flag complex if and only if $\mathsf{girth}(S) \geq 4$. Recall that the *girth* of a graph is the infimum of the lengths of its cycles.

The *link* of a 0-cube v in a cube complex X is the complex built from simplices that corresponds to the ϵ-sphere about v. Specifically, $link(v)$ has an $(n-1)$-simplex for each corner of an n-cube at v. A cube complex X is *nonpositively curved* if $\mathrm{link}(v)$ is a flag complex for each $v \in X^0$. A *CAT(0) cube complex* is a simply-connected nonpositively curved cube complex.

Nonpositively curved cube complexes were introduced by Gromov in [19] as a convenient source of examples of metric spaces with nonpositive curvature. When X is nonpositively curved in the above combinatorial sense, its universal cover $\widetilde{X}$ has a CAT(0) metric such that each n-cube is isometric to the standard Euclidean n-cube. This metric has been constructed in increasing levels of generality by Moussong, Bridson, and Leary [10, 39, 45].

One consequence of a complete CAT(0) metric is that any group acting freely on $\widetilde{X}$ is torsion-free. Some examples of other properties of groups acting on CAT(0) cube complexes are as follows: Let G act cellularly by isometries on the CAT(0) cube complex $\widetilde{X}$. Then G is biautomatic if the action is proper and cocompact [48], G is aTmenable if the action is metrically proper [46, 49], and G satisfies the Tits alternative if $\widetilde{X}$ is finite dimensional and there is an upper bound on the size of stabilizers [56].

3. Raags

Let Γ denote a simplicial graph. The *right-angled Artin group* (*raag*) $A(\Gamma)$ has the following presentation:

$$\langle\ v\ :\ v \in \text{Vertices}(\Gamma)\ \mid\ uv = vu\ :\ (u,v) \in \text{Edges}(\Gamma)\ \rangle$$

Note that the standard 2-complex of a raag is a cube complex with a single 0-cube, a 1-cube for each generator and a 2-cube for each relator. When $\mathsf{girth}(\Gamma) \geq 4$, this 2-complex is a nonpositively curved cube complex. In general, $A(\Gamma)$ is the fundamental group of a nonpositively curved cube complex $R(\Gamma)$ called the *Salvetti Complex* that is obtained by adding an n-cube for each collection of pairwise commutating generators - i.e. each $K(n)$ in Γ. Note that the 2-cubes are already in the standard 2-complex of the above presentation. Some easy examples of raags are: a rank n free group F_n which corresponds to a graph with n vertices and no edges, a rank n free abelian group $\mathbb{Z}^n$ which correspond to a complete graph $K(n)$, and a product $F_n \times F_m$ which corresponds to a complete bipartite graph $K(m,n)$. Every raag is linear, and in fact a subgroup of $SL_n(\mathbb{Z})$ since it is a subgroup of a right-angled Coxeter group [13, 32, 35]. We refer to [12] for more about raags.

4. Special cube complexes

A *midcube* in $[-1,1]^n$ is a connected subspace obtained by restricting exactly one coordinate to 0. Note that an n-cube has n distinct midcubes. A *hyperplane* $\widetilde{Y}$ in a CAT(0) cube complex $\widetilde{X}$ is a connected subspace that intersects each cube in a midcube or in $\emptyset$. As noted by Sageev [53], each midcube lies in a unique hyperplane, and each hyperplane separates $\widetilde{X}$ into two complementary components.

Let $\widetilde{Y}$ be a hyperplane of the universal cover $\widetilde{X}$ of a nonpositively curved cube complex X. Let $H = \mathrm{Stabilizer}(Y)$. Let $Y = H\backslash\widetilde{Y}$, and note that there is a map $Y \to X$. We refer to each such Y as an *immersed hyperplane* of X, and note that each midcube of X extends to a unique immersed hyperplane. When an immersed hyperplane embeds (i.e. its image does not contain distinct midcubes of the same cube of X) then we simply refer to it as a *hyperplane*. For a hyperplane Y that embeds, it is natural to consider the subspace $N^o(Y)$ consisting of the open cubes that it intersects. The open 1-cubes that Y intersects are *dual* to Y. The hyperplane Y is *2-sided* if the space $N^o(Y)$ is isomorphic to $Y \times (-1, 1)$ with $Y \times \{0\}$ corresponding to the subspace Y in the obvious way. When Y is 2-sided we can direct all 1-cubes dual to Y from $Y \times \{-1\}$ to $Y \times \{+1\}$.

We studied the following in [25]:

Definition 4.1. A *special cube complex* is a nonpositively curved cube complex with the following properties:

1. There is no self-crossing hyperplane. That is, each immersed hyperplane of $Y \to X$ is embedded
2. There is no 1-sided hyperplane. That is, each hyperplane of X is 2-sided.
3. There is no hyperplane of X that *self-osculates* in the sense that it is dual to distinct directed 1-cubes with the same initial or terminal 0-cube.
4. There do not exist distinct hyperplanes Y_1, Y_2 of X that *inter-osculate* in the sense that: Y_1, Y_2 intersect, and Y_1, Y_2 are dual to closed 1-cubes that share a 0-cube but do not lie on a common 2-cube.

For instance, any graph is special, any Salvetti complex $R(\Gamma)$ is special, and any CAT(0) cube complex is special. A beautiful source of special cube complexes are the *state complexes* of Abrams and Ghrist [1, 17].

Generalizing separability for a subgroup, we say a double coset HK is *separable* in G if it is closed in the profinite topology of G (i.e. the topology whose basis consist of left cosets of finite-index subgroups). The following criterion for virtual specialness was given in [25]:

Theorem 4.2. *Let X be a compact nonpositively curved cube complex. Suppose that for each immersed hyperplane $A \to X$ we have that $\pi_1 A$ is separable in $\pi_1 X$. Suppose that for each pair of crossing immersed hyperplanes $A \to X$ and $B \to X$, we have $\pi_1 A \pi_1 B$ is separable in $\pi_1 X$. Then X has a finite cover that is special.*

A noncocompact variant of Theorem 4.2 was used in [26] to see that every Coxeter group is virtually special using the cube complex of Theorem 6.1, and it was also used to see that simple-type arithmetic hyperbolic lattices are virtually special in [6]. It is interesting to note that Bergeron had independently engaged with the double hyperplane separability in his studies of totally geodesic submanifolds [5].

Theorem 4.2 has the following consequence:

Corollary 4.3. *Let X be a compact nonpositively curved cube complex. Suppose that $\pi_1 X$ is hyperbolic. If each quasiconvex subgroup of $\pi_1 X$ is separable then X is virtually special.*

A map $\phi : A \to B$ between nonpositively curved cube complexes is a *local isometry* if ϕ is *combinatorial* in the sense that ϕ maps open n-cubes homeomorphically to n-cubes, and ϕ is an *immersion* in the sense that ϕ is locally injective, and finally for each $a \in A^0$ the

induced map $\phi : \mathrm{link}(a) \to \mathrm{link}(\phi(a))$ has the property that if u, v are vertices of $\mathrm{link}(a)$ such that $\phi(u), \phi(v)$ are joined by an edge in $\mathrm{link}(\phi(a))$ then u, v are joined by an edge in $\mathrm{link}(a)$. As observed in [44], if $\phi : A \to B$ is a local isometry then the map $\widetilde{\phi} : \widetilde{A} \to \widetilde{B}$ is an isometric embedding, and moreover $\widetilde{A}$ embeds as a convex subcomplex of $\widetilde{B}$.

The definition of special cube complex was designed to enable the following, and the reader can refer to [25] or [27] for the details:

Lemma 4.4 (Canonical Completion and Retraction). *Let $\phi : A \to B$ be a local isometry of nonpositively curved cube complexes where B is special and A is compact. Then there exists a finite cover $\widehat{B} \to B$ such that $A \to B$ lifts to an embedding $A \hookrightarrow \widehat{B}$. Moreover, there is a retraction $\widehat{B} \to A$.*

Let D be a hyperplane of $\widetilde{X}$. Its *halfspace carriers* are the subcomplexes containing the two components of $\widetilde{X} - D$. Each halfspace carrier is a convex subcomplex since it maps to $\widetilde{X}$ by a local isometry. For a subspace $S \subset \widetilde{X}$ we let $\mathsf{hull}(S)$ denote the intersection of all halfspace carriers of $\widetilde{X}$ containing S. Note that $\mathsf{hull}(S)$ is a convex subcomplex of $\widetilde{X}$ since each halfspace carrier is convex. The following was observed in [23] and [55]. It also has a relatively hyperbolic generalization.

Lemma 4.5. *Let G be a hyperbolic group that acts properly and cocompactly on a CAT(0) cube complex $\widetilde{X}$. Let H be a quasiconvex subgroup of G. Let K be a compact subspace of $\widetilde{X}$. Then $\mathsf{hull}(HK) \subset \mathcal{N}_r(HK)$ for some $r \geq 0$.*

In particular, if X is a compact nonpositively curved cube complex with $\pi_1 X$ hyperbolic, then for each quasiconvex subgroup $H \subset \pi_1 X$, there exists a based local isometry $Y \to X$ such that Y is compact and $\pi_1 Y = H$.

A group G is *residually finite* if the trivial subgroup is the intersection of finite index subgroups of G. A subgroup $H \subset G$ is *separable* if H is the intersection of finite index subgroups of G. Thus G is residually finite precisely if $\{1_G\}$ is separable. Equivalently, H is separable if and only if for each $g \notin H$, there exists a finite quotient $G \to \bar{G}$ such that $\bar{g} \notin \bar{H}$. One can deduce the residual finiteness of $\pi_1 X$ from the specialness of X by applying Lemma 4.5 to a compact convex subcomplex of $\widetilde{X}$ containing $\widetilde{x}, g\widetilde{x}$. Moreover, since G' is residually finite whenever $G' \subset G$, and since $H \subset G$ is separable whenever $[G : G'] < \infty$ and $H \subset G'$ is separable, and since retracts of residually finite groups are separable (they correspond to closed subspaces in the profinite topology) we see from Lemma 4.4 and Lemma 4.5 that quasiconvex subgroups of a special hyperbolic group are separable.

The following theorem [25], connects special cube complexes to raags: Accordingly, we say that a group G is *special* if G is isomorphic to a subgroup of a raag.

Theorem 4.6. *X is special if and only if there is a local isometry $X \to R$ to the Salvetti complex $R = R(\Gamma)$ of a raag.*

Proof. If there is a local isometry $X \to R$, then X is special since an excluded hyperplane pathology would project to a hyperplane pathology in R, which is impossible since R is itself special.

Let Γ be the graph with a vertex for each hyperplane of X and such that two vertices are adjacent if and only if the corresponding hyperplanes cross. Let $R = R(\Gamma)$. As each hyperplane is 2-sided, we can label each 1-cube by the hyperplane it is dual to, and direct all

the dual 1-cubes consistently as above. This provides a map $X^1 \to R^1$ and it is readily seen that this extends to an immersion $X \to R$ because there is no self-osculation, and moreover $X \to R$ is a local isometry because there is no inter-osculation. □

5. Cubical small-cancellation theory

5.a. Classical small-cancellation. Let X denote a graph, and let $\{Y_i \to X\}$ be immersed circles. We summarize this data by $\langle X \mid Y_1, \dots, Y_r\rangle$ and refer to this as a *presentation*. When X is a bouquet of circles, we can regard each Y_i as a word in the generators, and the term presentation nearly corresponds to the usual notion. We let X^* denote the complex associated to the presentation, so X^* is the quotient $X \cup \bigcup_i \mathsf{Cone}(Y_i)$ obtained by attaching a cone with base Y_i to X for each relator Y_i.

A *piece* $P \to X$ is a combinatorial path that factors through lifts to Y_i and Y_j as on the left below, but such that these two lifts are *distinct* in the sense that there is no map $Y_i \to Y_j$ so that the diagram on the right below commutes:

$$\begin{array}{ccc} P & \to & Y_j \\ \downarrow & & \downarrow \\ Y_i & \to & X \end{array} \qquad \begin{array}{ccc} P & \to & Y_j \\ \downarrow & \nearrow & \downarrow \\ Y_i & \to & X \end{array} \tag{5.1}$$

The systole $\|Y_i\|$ is the girth of Y_i. A presentation satisfies the $C'(\frac{1}{n})$ *small-cancellation condition* if $|P| < \frac{1}{n}\|Y_i\|$ whenever P is a piece that lifts to Y_i. Here $|P|$ denotes the length of the immersed path, or equivalently, the distance between the endpoints of the lift of P in $\widetilde{Y}_i \subset \widetilde{X}$.

When $\langle X \mid Y_1, \dots, Y_r\rangle$ satisfies the $C'(\frac{1}{6})$ condition the group $\pi_1 X^* = \pi_1 X / \langle\langle \pi_1 Y_i \rangle\rangle$ is a hyperbolic group many of whose properties are particulary easy to understand. We refer to [41] for a historical account of this theory, and to [43] for a geometric account more consistent with the language below. The main result here is Greendlinger's lemma which can be summarized as follows:

Theorem 5.1. *Any minimal area disk diagram $D \to X^*$ is either*

(1) trivial *in the sense that it consists of a single 0-cell or single closed 2-cell*

(2) *a* ladder *in the sense that it is the union of a sequence of closed 1-cells or 2-cells $R_1 \cup R_2 \cup \cdots \cup R_m$ with $R_i \cap R_j = \emptyset$ for $|i - j| > 1$ and $R_i \cap R_{i+1}$ equal to a (possibly trivial) arc otherwise.*

(3) *D has three or more shells and/or spurs.*

A *shell* is a 2-cell R whose boundary path $\partial_p R = QS$ is the concatenation of two paths, where the *outer path* Q is a subpath of the boundary path $\partial_p D$ and the *inner path* S has the property that $|S| < |Q|$. A *spur* is a 1-cell terminating at a valence 1 vertex in D where $\partial_p D$ backtracks.

5.b. Cubical small-cancellation. A *cubical presentation* $\langle X \mid Y_1, \dots, Y_r\rangle$ consists of a nonpositively curved cube complex X, and a collection of local isometries $Y_i \to X$ of nonpositively curved cube complexes. The group of this presentation is $\pi_1 X / \langle\langle \pi_1 Y_1, \dots, \pi_1 Y_r \rangle\rangle \cong \pi_1 X^*$ where X^* is the complex obtained by attaching cones on the Y_i as above.

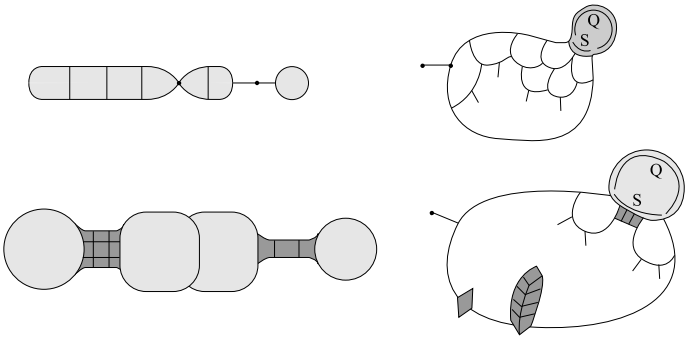


Figure 5.1. Ladders and disk diagrams with spurs, shells, and cornsquares for the classical and cubical cases

In contrast to the classical case, there are now two types of pieces: A *wall-piece* is a path $P \to Y_i$ such that the lift $\widetilde{P} \to \widetilde{Y}_i \subset \widetilde{X}$ has the property that $\widetilde{P}$ is a subpath of $N(V)$ where V is a hyperplane of $\widetilde{X}$ and $N(V)$ is the convex subcomplex consisting of all cubes intersecting V (so $N(V)$ is the closure of $N^o(V)$ discussed earlier). A *cone-piece* is a path $P \to Y_i$ such that $P \to X$ factors as $P \to Y_j$ as on the left of Equation (5.1) and these two paths are distinct in the sense that there is no map $Y_i \to Y_j$ such that the diagram on the right of Equation (5.1). Within $\widetilde{X}$, this means that the lift of P to $\widetilde{Y}_i$ and $\widetilde{Y}_j$ are distinct in the sense that $\widetilde{Y}_i$ and $\widetilde{Y}_j$ look different from the viewpoint of $\widetilde{P}$.

The cubical presentation is $C'(\frac{1}{n})$ if $|P| < \frac{1}{n}\|Y_i\|$ whenever $P \to Y_i$ is a piece. Now $\|Y_i\|$ denotes the infimum of the lengths of the closed essential paths in Y_i, and $|P|$ denotes the distance between the endpoints of the lift of P in $\widetilde{Y}_i \subset \widetilde{X}$. (Here we use the distance in the graph metric on the 1-skeleton.)

Let $P \to X$ be a closed combinatorial path such that $P \to X^*$ is null-homotopic. Note that the 2-cells of X^* are triangles and squares, and the triangles in a disc diagram can be assembled cyclically around 0-cells mapping to conepoints to form *cone-cells* whose interiors are open disks. The boundary path of each cone-cell maps to some Y_i. We are interested in disk diagrams of *minimal area* in the sense that $\big(\#(\text{cone-cells}), \#(\text{squares})\big)$ is minimized in the lexicographical order. Among other things, minimality guarantees that the boundary path of each cone-cell R is essential for otherwise we could replace that cone-cell by a square diagram. A *spur* is as before and a *shell* in D is also as before: it is a cone-cell R such that $\partial_p(R) = QS$ where $|S| < |Q|$ and where Q is a subpath of $P = \partial_p D$. There is now an additional object: A *cornsquare* is a square in D such that (parts of) the two hyperplanes emanating from D end on adjacent 1-cells of $\partial_p D$, and such that they bound a square subdiagram. A *ladder* in X^* is a disk diagram formed from a sequence of cone-cells joined by $n \times m$ square grids with $n, m \geq 0$. See Figure 5.1. The main theorem of cubical small-cancellation theory is very similar to the classical case:

Theorem 5.2. *Suppose that the cubical presentation $\langle X \mid Y_1, \dots, Y_r\rangle$ satisfies the $C'(\frac{1}{12})$ condition. Let $D \to X^*$ be a minimal area diagram. Then either*

(1) *D is a single 0-cell or cone-cell*

(2) *D is a ladder*

(3) *D contains three or more spurs, cornsquares, or shells.*

6. Groups acting on cube complexes

We now turn to the notion of a wallspace introduced by Haglund and Paulin in [24], and the dual cube complex of Sageev [53].

6.a. Wallspaces. A *wall* in a topological space X is a decomposition $X = \overleftarrow{W} \cup \overrightarrow{W}$. It is often simplest to work under the assumption that $\overleftarrow{W} \cap \overrightarrow{W} = \emptyset$. A convenient way to obtain a wall is when a subspace $W \subset X$ has more than one complementary component, in which case we let $\overrightarrow{W}$ be the union of W and some of these components, and we let $\overleftarrow{W} = X - \overrightarrow{W}$. The spaces $\overleftarrow{W}, \overrightarrow{W}$ are the *halfspaces* of the wall. We say p, q are *separated* by the wall $\{\overleftarrow{W}, \overrightarrow{W}\}$ if p, q do not both lie in the same halfspace. A *wallspace* is the space X together with a collection $\mathcal{W}$ of walls such that $\#(p, q) < \infty$ whenever $p, q \in X$. Here $\#(p, q)$ denotes the number of walls separating p, q.

6.b. Codimension-one subgroups. Let G be a f.g. group, and let $\Upsilon = \Upsilon(G, S)$ denote its Cayley graph with respect to a finite generating set S and give Υ the graph metric. A subgroup H of G is *codimension-one* if for some $r > 0$ the complement $\Upsilon - \mathcal{N}_r(H)$ contains more than one H-orbit of component that is *deep* in the sense that it does not lie in $\mathcal{N}_s(H)$ for any $s > 0$. Note that being codimension-one does not depend on the choice of generators although the constant r might be affected. An equivalent formulation is that H is codimension-one if the coset graph $H\backslash\Upsilon$ has more than one end. For instance, any embedding of $\mathbb{Z}^m \subset \mathbb{Z}^n$ is codimension-one precisely when $m = n-1$. Another visual example is any infinite cyclic subgroup of a closed surface group, or indeed, any subgroup $\pi_1 M \subset \pi_1 N$ where M and N are closed aspherical manifolds with $\dim(N) = \dim(M) + 1$.

Given a codimension-one subgroup H of a group G, Sageev produced a wall in the Cayley graph Υ from a codimension-one subgroup H by letting $\overrightarrow{W}$ be HK where K is a deep component, and letting $\overleftarrow{W} = \Upsilon - \overrightarrow{W}$. Note that if $\{\overleftarrow{W}, \overrightarrow{W}\}$ is a wall then $\{g\overleftarrow{W}, g\overrightarrow{W}\}$ is a wall for $g \in G$. By varying $g \in G$, we obtain an infinite set of walls that G will then permute. Performing this procedure for one or more codimension-one subgroups H_i and chosen decompositions of $\Upsilon - \mathcal{N}_{r_i}(H_i)$ we obtain a wallspace that has a *G-action* in the sense that G also acts on the collection of walls. If our collection $\{H_1, \ldots, H_k\}$ is finite then we can be assured that the finiteness condition $\#(p, q) < \infty$ holds so we obtain a genuine wallspace.

6.c. The dual cube complex. Let $(X, \mathcal{W})$ be a wallspace. Sageev defined its *dual CAT(0) cube complex* as follows: Each 0-cube v is a choice of one halfspace from each wall that satisfies the following two conditions:

1. Any two chosen halfspaces of v have nonempty intersection.
2. For some (and hence any) point $x \in X$, all but finitely many of the chosen halfspaces of v contain x.

The 0-cubes u, v are joined by a 1-cube precisely when the u, v choices differ on exactly one wall. For each $n \geq 2$ we add an n-cube whenever its $(n-1)$-skeleton is present. Sageev proved that the dual is CAT(0): the nonpositive curvature condition is immediate, and simple connectivity holds by an argument that uses that every closed path has to pass through 1-cubes corresponding to each wall an even number of times [53].

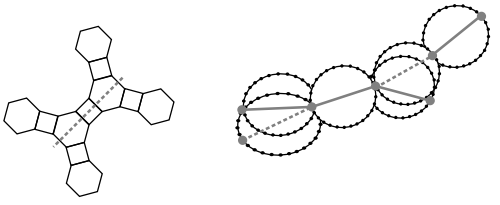


Figure 6.1. A wall for the Coxeter group $\langle a, b, c \mid a^2, b^2, c^2, (ab)^2, (bc)^3 \rangle$, and a wall in an anonymous small-cancellation complex.

We now describe to noteworthy applications of the dual cube complex:

Niblo-Reeves proved the following result in [47]. The walls they used are precisely the reflection walls of the Coxeter group. (See the left of Figure 6.1.)

Theorem 6.1 (Cubulating Coxeter Groups). *Every f.g. Coxeter group acts properly on a finite dimensional CAT(0) cube complex.*

Each $C'(\frac{1}{6})$ 2-complex $\widetilde{X}$ has walls that are trees whose vertices are midpoints of 1-cells of $\widetilde{X}$ and whose edges are arcs in 2-cells joining opposite 1-cells. (See the right of Figure 6.1.)

Theorem 6.2 (Cubulating $C'(\frac{1}{6})$ Groups). *Every f.g. $C'(\frac{1}{6})$ small-cancellation group acts properly and cocompactly on a CAT(0) cube complex [63].*

6.d. Finiteness Properties. The finiteness properties of the dual are related to properties of the walls and their stabilizers. The maximal cubes in the dual (if they exist) correspond to maximal collections of pairwise crossing walls. We obtain cocompactness of the dual when there are finitely many orbits of collections of pairwise crossing walls, as shown by Sageev when the wall stabilizers are quasiconvex and the space X is δ-hyperbolic [54]. We assume now that X is a metric space, as would be the case when X is the Cayley graph of a group or when $X = \mathbb{H}^n$. The group G acts metrically properly on the dual cube complex when there are sufficiently many walls in the sense that $\#(p, q) \to \infty$ when $\mathsf{d}_X(p, q) \to \infty$. A comprehensive examination of the finiteness properties is given in [29]. Producing codimension-one subgroups can be very challenging, and even when they exist, it is often tricky to verify that the group G acts properly on the dual cube complex.

6.e. Cubulating from the boundary. In [7] we gave a criterion to verify that a group G acts properly and cocompactly on a CAT(0) cube complex when there is a rich collection of walls with relatively quasiconvex stabilizers. This result was also obtained by Dufour who applied it to cubulate fibered hyperbolic 3-manifolds in his PhD thesis.

Theorem 6.3. *Let G be a relatively hyperbolic group, and let ∂G denote the Bowditch boundary of G. Suppose that for each pair of points $p, q \in \partial G$ there exists a relatively quasiconvex codimension-one subgroup H such that p, q lie in distinct H-orbits of components of $\partial G - \partial H$. Then G acts properly and relatively cocompactly on a CAT(0) cube complex.*

If the term "relatively" is omitted from Theorem 6.3, the statement applies in the more standard setting where ∂G is the Gromov boundary of a hyperbolic group G, and the meaning of cocompactness in the conclusion is as usual.

In the relatively hyperbolic case, acting *relatively cocompactly* on the dual $\widetilde{C}$ means that there is a collection of convex *parabolic* subcomplexes $\widetilde{C}_i$ that are stabilized by the P_i, and a compact subcomplex K such that $\widetilde{C} = GK \cup \widetilde{G}\widetilde{C}_i$, and these parabolic subcomplexes $\widetilde{C}_i$ are *isolated* in the sense that $g_i\widetilde{C}_i \cap \widetilde{g}_j\widetilde{C}_j \subset GK$ unless $g_i\widetilde{C}_i = g_j\widetilde{C}_j$.

Corollary 6.4. *Let G act freely and cocompactly on $\mathbb{H}^3$. Suppose that for each pair of points $p, q \in \partial\mathbb{H}^3$ there is a quasifuchsian subgroup H such that ∂H is an embedded circle that separates p, q. Then there exists a finite subcollection of quasifuchsian subgroups $\{H_1, \dots, H_k\}$ such that G acts properly and cocompactly on the resulting dual cube complex.*

Kahn and Markovic proved the following in [37]:

Theorem 6.5. *Let M be a closed hyperbolic 3-manifold. For each circle $O \subset \partial\mathbb{H}^3$ there is a sequence of quasifuchsian surfaces $S_i \to M$, such that $\partial\widetilde{S}_i \hookrightarrow \partial\widetilde{M} = \partial\mathbb{H}^3$ converges pointwise to O.*

Kahn and Markovic built these surfaces by amalgamating a family of immersed geodesic pairs of pants, using ergodicity of the geodesic flow to ensure choices so that the surface closes and such that consecutive pants in the gluing nearly followed a fixed $\mathbb{H}^2$ trajectory. In view of Corollary 6.4, we obtain the following consequence of the Kahn-Markovic surfaces:

Corollary 6.6. *Let M be a closed hyperbolic 3-manifold. Then $\pi_1 M$ acts properly and cocompactly on a CAT(0) cube complex.*

Bergeron and I worked this out in 2006 (in the word-hyperbolic case) but the only applications available then were already handled by a direct geometric computation - namely the simple-type arithmetic hyperbolic lattices that were cubulated with Haglund in [6]. Upon hearing of the breakthrough of Kahn-Markovic, we connected it to our work. It appears that there will be further cubulation results following this method, since it avoids certain messy computations. There have not yet been applications using the relatively hyperbolic case, although it is likely that the argument for free-by-cyclic groups in [22] can be generalized in this direction. The case where M is a hyperbolic manifold with nonempty boundary was treated using hierarchies, and is discussed in Section 11.

7. Cubulating Malnormal Amalgams

Definition 7.1. A collection $\{H_1, \dots, H_r\}$ of subgroups of G is *almost malnormal* if $g^{-1}H_i g \cap H_j$ is finite unless $i = j$ and $g \in H_i$. Similarly, the term *malnormal* means that each such intersection of distinct conjugates is trivial.

The following is proven in [31]:

Theorem 7.2. *Let G be a hyperbolic group that splits as a finite graph of groups such that each vertex group is virtually compact special, each edge group is quasiconvex, and the collection of edge groups is almost malnormal. Then G acts properly and cocompactly on a CAT(0) cube complex.*

The idea of the proof of Theorem 7.2 is best motivated by considering the case where all edge groups are infinite cyclic. Let X be a graph of spaces where the vertex spaces are

nonpositively curved cube complexes and the edge spaces are cylinders. Its universal cover, $\widetilde{X}$, is a tree of spaces where the vertex spaces are CAT(0) cube complexes and where the edge spaces are strips attached along geodesics. We make $\widetilde{X}$ into a wallspace as follows: There is a wall for each edge space but the other walls are more complicated. If the attaching geodesics of an edge space have the same combinatorial length on each side, then we choose an equivariant bijection between corresponding hyperplanes in the two vertex spaces and connect corresponding hyperplanes with arcs cutting through the edge space. Otherwise there are more hyperplanes on one side than the other, in which case we choose an injection from the smaller side to the larger side and attach arcs as before. We then equivariantly glue the excess hyperplanes together in pairs using arcs that start and end on excess hyperplanes on the longer side. When these arcs are sufficiently long, and the cyclic subgroup is malnormal, and the vertex space is hyperbolic, then the resulting glued hyperplane-arc-hyperplane objects are genuine walls. Moreover, with careful choices they are quasiconvex and there are sufficiently many so that $\pi_1 X$ acts freely on the resulting dual cube complex.

Remark 7.3. Virtual compact specialness of the vertex groups provide the actual enabling properties of the proof of Theorem 7.2: That the vertex groups act properly and cocompactly on CAT(0) cube complexes, that their edge groups have separable quasiconvex subgroups, and that every quasiconvex wall in an edge group extends to a quasiconvex wall in its vertex group. This last "extension property" is provided by an argument using Lemma 4.5.

Remark 7.4. A relatively hyperbolic variation of Theorem 7.2 is as follows: We assume that G is hyperbolic relative to virtually abelian subgroups $\{P_1, \dots, P_r\}$, but we require that each edge group is relatively quasiconvex and genuinely hyperbolic. We reach the weaker conclusion that G acts *cosparsely* on the dual cube complex $\widetilde{C}$. This means that for each i there is a cube complex $\widetilde{C}_i$ stabilized by P_i and quasi-isometric to some $\mathbb{E}^n$, and there is a compact subcomplex K such that $\widetilde{C} = GK \cup G\widetilde{C}_1 \cup \cdots \cup G\widetilde{C}_r$ and finally $g_i\widetilde{C}_i \cap g_j\widetilde{C}_j \subset GK$ unless $g_i\widetilde{C}_i = g_j\widetilde{C}_j$.

There is more work to be done here: One expects a generalization of Theorem 7.2 to apply to an arbitrary relatively hyperbolic group that splits as a finite graph of virtually compact special vertex groups with edge groups that are relatively quasiconvex and also relatively malnormal in the sense that they intersect their conjugates in finite or parabolic subgroups. (See Problem 13.31.) However, there are groups G that split as a graph of abelian vertex groups with maximal cyclic edge groups, such that G cannot act freely on a CAT(0) cube complex [59].

8. Virtually special cubical Malnormal Amalgams

The following is the main result in [27] which substantially generalizes [62]:

Theorem 8.1 (Virtual Specialness of Malnormal Cubical Amalgams). *Let G be a hyperbolic group that acts properly and cocompactly on a CAT(0) cube complex $\widetilde{X}$. Suppose there is a hyperplane $\widetilde{Y}$ such that $g\widetilde{Y} \cap \widetilde{Y} = \emptyset$ unless $g\overleftarrow{\widetilde{Y}} = \overleftarrow{\widetilde{Y}}$ where $\overleftarrow{\widetilde{Y}}$ denotes the left halfspace of $\widetilde{Y}$ in $\widetilde{X}$. Suppose that* $\mathrm{Stabilizer}(\widetilde{Y})$ *is almost malnormal in G. Let $N^o(\widetilde{Y})$ denote the open cubical neighborhood of $\widetilde{Y}$ in $\widetilde{X}$. Suppose that for each component $\widetilde{X}_*$ of $\widetilde{X} - GN^o(\widetilde{Y})$ the group* $\mathrm{Stabilizer}(\widetilde{X}_*)$ *has a finite index subgroup that acts freely on $\widetilde{X}_*$ with a special quotient. Then G has a finite index subgroup G' that acts freely with a special quotient.*

Corollary 8.2 (Geometric Special Case). *Let X be a compact nonpositively curved cube complex with $\pi_1 X$ hyperbolic. Let $Y \subset X$ be a hyperplane. Suppose that Y is embedded and 2-sided so that $N^o(Y) \cong Y \times (-1, 1)$. Suppose that $\pi_1 Y$ is malnormal in $\pi_1 X$. Suppose that each component of $X - N^o(Y)$ is virtually special. Then X is virtually special.*

To prove virtual specialness, we show that every quasiconvex subgroup is separable and thus apply Corollary 4.3. Let us focus on the statement of Corollary 8.2 in the case where $X - N^o(Y)$ is a disjoint union $L \sqcup R$. The heart of the matter is to be able to show that for any finite covers $\widehat{L}' \to L$ and $\widehat{R}' \to R$, there are finite regular covers $\widehat{L} \to L$ and $\widehat{R} \to R$ that factor through them, so that they induce the same cover $\widehat{Y}$ of Y. We are thus able to build a cover of X from the disjoint union of a collection of covers of copies of $\widehat{L}$ and $\widehat{R}$, by attaching copies of $N(\widehat{Y}) = \widehat{Y} \times [-1, 1]$. The key to producing $\widehat{L}$ and $\widehat{R}$ is to use functorial properties of the canonical completion and retraction of Lemma 4.4. These properties allow one to treat a malnormal quasiconvex subgroup A of a hyperbolic special group B almost as if A were a retract of B.

9. Malnormal quasiconvex hierarchy

Definition 9.1. The group G has a [*malnormal*] *quasiconvex hierarchy terminating in groups of type $\mathcal{T}$* if G can be built from groups of type $\mathcal{T}$ by a finite sequence of finite graphs of groups, such that the edge groups are f.g. and quasi-isometrically embedded [and form an almost malnormal collection] at each stage.

When $\mathcal{T}$ is the class of finite groups, we simply say that G has a *quasiconvex hierarchy*. Our main case of interest is when G is hyperbolic in which case being quasi-isometrically embedded is equivalent to being quasiconvex. When G has a specified relatively hyperbolic structure it is natural to instead require that the edge groups are relatively quasiconvex.

Example 9.2. The fundamental groups of surfaces have hierarchies, and these hierarchies are always quasiconvex, and also often malnormal (except for the torus and klein bottle and when an edge group is generated by a glide reflection). Any Haken 3-manifold M has a (topological) hierarchy which provides a hierarchy for $\pi_1 M$. When M is a fibered hyperbolic 3-manifold this gives a very short hierarchy, but the hierarchy will not be quasiconvex since the fundamental group of the fiber is not quasiconvex. A similar situation occurs for one-relator groups, whose Magnus-Moldavanskii hierarchies (providing the foundation for the theory of one-relator groups), but these hierarchies are typically not quasiconvex.

Combining Theorem 7.2 and Theorem 8.1 we have the following result which functions as a target criterion for virtual specialness.

Theorem 9.3. *Let $\mathcal{S}$ denote the class of hyperbolic virtually compact special groups. Let G be a hyperbolic group with a malnormal quasiconvex hierarchy terminating in groups of type $\mathcal{S}$. Then G is in $\mathcal{S}$.*

10. Special quotient theorem

Theorem 10.1 (Malnormal Special Quotient Theorem). *Let G be hyperbolic and virtually compact special. Let $\{H_1, \ldots, H_r\}$ be an almost malnormal collection of quasiconvex sub-*

groups. There exist finite index subgroups $\{H_i^\circ \subset H_i\}$, such that for any normal finite index subgroups $H_i' \subset H_i$ with $H_i' \subset H_i^\circ$ the quotient $\bar{G} = G/\langle\langle H_1', \dots, H_r' \rangle\rangle$ is hyperbolic and virtually compact special.

Theorem 10.1 is already interesting when G is a rank 2 free group and $\{H_1, \dots, H_r\}$ is a malnormal collection of f.g. subgroups. Among other things we find that $\bar{G}$ is residually finite. In particular, we obtain the following corollary:

Corollary 10.2. *Let $\{W_1, \dots, W_r\}$ be a collection of elements of a free group F with basis $\{a, b\}$, and suppose that W_i, W_j do not have conjugate nontrivial powers for $i \neq j$. Then there exists $k_i \geq 1$ for each i so that the following group is virtually compact special for all $n_i \geq 1$.*

$$\langle a, b \mid W_1^{n_1 k_1}, \dots, W_r^{n_r k_r} \rangle$$

In particular, we can control the relative orders of $\{\bar{W}_1, \dots, \bar{W}_r\}$ in finite quotients $F \to \bar{F}$.

As a finite collection of quasiconvex subgroups has finite *height* which is the maximal number of infinitely intersecting distinct conjugates [18], variations of the following consequence can be obtained by inductively repeatedly applying Theorem 10.1 to an almost malnormal collection of subgroups arising from intersections of conjugates of the $\{H_1, \dots H_r\}$:

Corollary 10.3 (Special Quotient Theorem). *Let G be hyperbolic and virtually compact special. Let $\{H_1, \dots, H_r\}$ be a collection of quasiconvex subgroups. There exist finite index subgroups $H_i' \subset H_i$ such that the quotient $G/\langle\langle H_1', \dots, H_r' \rangle\rangle$ is hyperbolic and virtually compact special.*

We prove Theorem 10.1 by choosing the H_i° so that $\bar{G} = G/\langle\langle H_i' \rangle\rangle$ is arranged to have a finite index subgroup with an almost malnormal quasiconvex hierarchy, and so Theorem 9.3 implies that $\bar{G}$ is virtually compact special. The proof is organized via cubical small-cancellation theory. My original conception for the proof was based on my (thus far unrealized) expectation that $\bar{G}$ would be cubulated (using a generalization of Theorem 6.2), in which case the proof would be completed by using only Theorem 8.1 since the hyperplanes are malnormal. When I was ultimately unable to realize this expectation, I instead applied Theorem 9.3. The required properties of the malnormal quasiconvex hierarchy were verified using cubical small-cancellation theory. In retrospect, much of this could have been done using the theory of Osin or Groves-Manning [20, 50], keeping track of the walls where the virtual splittings would take place. Nevertheless, the cubical small-cancellation theory helped navigate towards the desired virtual hierarchy, and it is another matter to reposition an exposition upon a recognized target.

A relatively hyperbolic version of Theorem 10.1 can also be proven that follows the same argument, although we only implemented this in a limited fashion in [60]. In particular, we showed the following:

Theorem 10.4. *Let G be hyperbolic relative to virtually abelian subgroups $P_1, \dots, P_r$. Suppose that G is virtually compact special. Then there exist finite index subgroups $P_i^\circ \subset P_i$ such that for any normal finite index subgroups $P_i' \subset P_i$ with $P_i' \subset P_i^\circ$, the quotient $\bar{G} = G/\langle\langle P_1', \dots, P_r' \rangle\rangle$ is virtually compact special and hyperbolic.*

One consequence of Theorem 10.4 is that for a closed hyperbolic 3-manifold M with $\partial M = T_1 \sqcup \cdots \sqcup T_r$, there are covers $\widehat{T}_r^\circ \to T_r$ so that for any further covers $\widehat{T}_r' \to \widehat{T}_r^\circ$,

there exists a regular cover $\widehat{M} \to M$ such that each component of the preimage of T_r is isomorphic as a covering space to $\widehat{T}'_r$.

11. Quasiconvex hierarchy

Theorem 11.1. *Let G be a hyperbolic group with a quasiconvex hierarchy. Then G is virtually compact special.*

The proof is staged by applying Theorem 10.1 to the vertex group of an HNN extension $A*_{C^t=D}$ (relative to subgroups related to intersections of various conjugates of C, D) and obtaining a quotient group $\bar{A}*_{\bar{C}^{\bar{t}}=\bar{D}}$ that is closer to being an almost malnormal HNN extension in the sense that $\{\bar{C}, \bar{D}\}$ has lower height in $\bar{G}$ than $\{C, D\}$ in G. The result then follows by induction.

Theorem 11.1 applies to one-relator groups with torsion, since their Magnus-Moldavanskii hierarchy is quasiconvex, and thus resolves Baumslag's conjecture on the residual finiteness of one-relator groups with torsion. It applies to closed hyperbolic 3-manifolds with a geometrically finite incompressible surface since by a result of Thurston's, the Haken hierarchy is quasiconvex provided the first cut is geometrically finite [11].

There are relatively hyperbolic versions of this in [60] which are not powerful enough to handle an arbitrary relatively hyperbolic group with a quasiconvex hierarchy, but they are sufficient to handle fundamental groups of cusped hyperbolic 3-manifolds, and can also handle limit groups which have a hierarchy whose edge groups are cyclic at each stage. A useful step in the relatively hyperbolic direction is the following:

Theorem 11.2. *Let G be hyperbolic relative to $\mathbb{Z}^2$ subgroups. Suppose that G splits as a graph of groups where the vertex groups are virtually compact special and hyperbolic, and the edge groups are relatively quasiconvex. Then G is virtually compact special.*

A hyperbolic 3-manifold M with $\partial M \neq \emptyset$ has a finite cover $\widehat{M}$ with an incompressible geometrically finite surface that cuts all the tori. We thus obtain:

Corollary 11.3. *Let M be a hyperbolic 3-manifold with nonempty boundary. Then $\pi_1 M$ has a finite index subgroup that is isomorphic to π_1 of a compact special cube complex.*

12. Virtually Haken conclusion

A final goal for understanding cubulated groups in the hyperbolic setting was achieved by Agol who proved the following result [3]. I first openly conjectured this and its associated 3-manifold consequence at the 2005 Spring Topology and Dynamics conference and have been working towards it since then.

Theorem 12.1. *Let G be a hyperbolic group that acts properly and cocompactly on a CAT(0) cube complex $\widetilde{X}$. Then G contains a finite index torsion-free subgroup G' such that $G'\backslash\widetilde{X}$ is special.*

In view of Corollary 6.6, this resolved the virtual Haken problem for closed hyperbolic manifolds, and it conclusively impacts geometric group theory, perhaps most obviously since

all f.p. $C'(\frac{1}{6})$ groups are virtually special in view of Theorem 6.2. In a reasonable sense, most groups with comparatively few relators are given by a $C'(\frac{1}{6})$ presentations, so it seems that most groups arising in combinatorial group theory are virtually special.

A quick sketch of the proof of Theorem 12.1 proceeds as follows: A *block* is a compact space that is the quotient $J\backslash\widetilde{Y}$ by a torsion-free subgroup $J \subset G$ where $\widetilde{Y}$ is the intersection of a collection of closed halfspaces of $\widetilde{X}$, and where the hyperplanes of Y embed. The *faces* of $\widetilde{Y}$ are the intersections with hyperplanes that don't separate it, and the images of these faces are *faces* of the block Y. Color all the faces so that intersecting faces have different colors. Repeatedly glue covers of blocks together along faces of the same color to form larger blocks (with fewer types of faces). Here the special quotient theorem is ingeniously applied to produce the covers that ensure that we are gluing along isomorphic faces. With each gluing, there are fewer colors of exposed faces, and one eventually obtains a space corresponding to a quotient of $\widetilde{X}$ by a finite index torsion-free subgroup. With some additional coloring care one can directly arrange that this space already be special. Alternatively, since the hyperplanes embed one can apply Theorem 11.1 and pass to a finite special cover.

We have described that cusped hyperbolic 3-manifolds and closed hyperbolic 3-manifolds have virtually compact special π_1. Liu showed that a graph manifold is virtually special exactly when it admits a metric of nonpositive curvature [40]. To complete the picture, we mention the following obtained with P.Przytycki [52]:

Theorem 12.2. *Let M be a compact irreducible 3-manifold that is neither hyperbolic nor a closed graph manifold. Then M has a finite cover $\widehat{M}$ such that $\pi_1\widehat{M} \cong \pi_1 X$ where X is a special cube complex.*

13. A collection of problems

G is *locally indicable* if every nontrivial finite generated subgroup $H \subset G$ has in infinite cyclic quotient. A hyperbolic group G is *locally quasiconvex* if each f.g. subgroup of G is quasiconvex. Examples of locally indicable groups that are also locally quasiconvex are given in [58]. There are no known torsion-free locally quasiconvex groups that are not also locally indicable. However there are locally indicable hyperbolic groups that are not locally quasiconvex (for instance, hyperbolic free-by-cyclic groups). It is conceivable that the following conjecture holds with only one of the assumptions of locally quasiconvex or locally indicable. There are not yet enough examples understood here.

Conjecture 13.1. Let G be a hyperbolic group that is locally quasiconvex and locally indicable. Then G acts freely and cocompactly on a CAT(0) cube complex.

As described in Lemma 4.5, cyclic subgroups of hyperbolic special groups are virtual retracts. The following proposes that this is essentially a characterization of specialness. It is conceivable that one could find examples of hyperbolic groups where every infinite cyclic subgroup survives in the abelianization of some finite index subgroup.

Conjecture 13.2. A f.g. (hyperbolic) group G is virtually special iff every cyclic subgroup is a virtual retract.

Problem 13.3. Let G be [relatively] hyperbolic. Show that G acts properly on a CAT(0)

cube complex iff every cyclic subgroup is cut by a [relatively] quasiconvex codimension-1 subgroup. (The "if" direction is straightforward).

Problem 13.4. Does every braid group act freely (and cocompactly) on a CAT(0) cube complex? Is every braid group virtually special?

I suspect the answer to the above problems might be no for B_4, although the second problem will be easier to negate. We note that B_4 was shown to be the fundamental group of a compact nonpositively curved space in [9], and other researchers have pushed this to higher B_n, but asking for a nonpositively curved cube complex is considerably more demanding.

Problem 13.5. Characterize when a group that is F_n-by-F_m acts freely on a CAT(0) cube complex. (In particular, find examples that do not act freely.)

Problem 13.6. Let G be the fundamental group of a compact nonpositively curved cube complex. Is G hopfian?

Problem 13.7. Show that the isomorphism problem for fundamental groups of compact nonpositively curved cube complexes is undecidable.

Problem 13.8. Show that not being virtually special is undecidable for compact nonpositively curved cube complexes (even for npc $\mathcal{VH}$-complexes).

The following problem is phrased in the direction that I think it will be resolved:

Problem 13.9. (Gromov) Find a compact nonpositively curved square complex X such that $\pi_1 X$ is not hyperbolic, but $\pi_1 X$ does not contain $\mathbb{Z}^2$ subgroup.

I hope the following variant of a well-known problem in geometric group theory will be approachable for special cube complexes:

Problem 13.10. Let X be a compact special cube complex. Suppose that $\pi_1 X$ is hyperbolic relative to abelian subgroups. Show that $H \subset \pi_1 X$ is quasi-isometrically embedded if H is malnormal. More generally, show that H is quasi-isometrically embedded unless there are infinitely many left cosets $\{Hg_i\}$ such that $g_i^{-1} H g_i \cap H$ is neither parabolic nor finite.

Problem 13.11. Let H be a f.g. quasiisometrically embedded subgroup of a raag. Is H separable? Same question under the stricter hypothesis that H is quasiconvex with respect to the CAT(0) metric.

Problem 13.12. Suppose G is special, so G is a subgroup of a raag. Suppose $G = \pi_1 X$ with X a compact nonpositively curved cube complex. Is X virtually special? Similarly, suppose X, Y are compact nonpositively curved cube complexes with $\pi_1 X \cong \pi_1 Y$. Is X virtually special if and only if Y is virtually special?

Problem 13.13. Find conditions on a 2-complex X that ensure that there are no codimension-one subgroups (á la the Garland spectral gap condition of [4, 68]). Find conditions that control the "direction" of the codimension-one subgroups.

Problem 13.14. Does every infinite f.p. $C(6)$ group have a codimension-1 subgroup? Does every infinite f.p. $C(4) - T(4)$ group have a codimension-1 subgroup?

Problem 13.15. Let G be a hyperbolic group with a hierarchy. Does G act properly and cocompactly on a CAT(0) cube complex? In particular, this should be true for a hyperbolic one-relator group. The case of free-by-cyclic groups is treated in [21, 22]. Another test case to consider are f.g. free-by-free groups.

Problem 13.16. Let H be a malnormal quasiconvex subgroup of a hyperbolic group G that acts properly and cocompactly on a CAT(0) cube complex. Is $G/\langle\langle H'\rangle\rangle$ cubulated when $H' \subset H$ and the systole of H' is sufficient. Does $G/\langle\langle H'\rangle\rangle$ always contain a codimension-1 subgroup? (I guess these have negative answers in general, but they hold when H is cyclic [60]).

Problem 13.17. Give conditions on a cubical presentation $\langle X \mid Y_1, \dots, Y_r\rangle$ that guarantee that the corresponding quotient group $\pi_1 X^*$ acts properly on a CAT(0) cube complex. Some conditions are given when X is 1-dimensional in [60].

Problem 13.18. Generalize the cubical small-cancellation theory in [60] to deal with quotients of groups acting (not freely) on a CAT(0) cube complex $\widetilde{X}$.

Conjecture 13.19. Let X be a compact special cube complex with $\chi(X) \geq 0$. Then either $\pi_1 X$ is abelian, or $\pi_1 X$ contains a f.g. non-abelian subgroup H containing a f.g. normal subgroup N with $H/N \cong \mathbb{Z}$.

Show that if X is 2-dimensional and special and $\chi(X) > 0$, then either $\pi_1 X = 1$ or $\pi_1 X$ contains a f.g. subgroup that is not f.p.

The plan is to apply Bestvina-Brady Morse theory [8] to a suitable immersed subspace. This was my intended route towards virtual fibering before Agol produced his criterion - which is a sophisticated 3-manifold topology argument enabled by a basic property enjoyed by raags [2]. Conjecture 13.19 is examined when G acts properly and cocompactly on a Bourdon building in [66].

A *complete square complex* (CSC) is a nonpositively curved square complex X such that the link of each 0-cube of X is a complete bipartite graph. Equivalently, X is a CSC if $\widetilde{X}$ is isomorphic to the product of two trees.

Problem 13.20. Let X be a compact CSC with $\chi(X) > 0$. Does $\pi_1 X$ contain a f.g. subgroup that is not f.p.?

Problem 13.21. Give novel examples of compact nonpositively curved cube complexes that are not virtually special. Currently, all known nonexamples can be traced to an irreducibe CSC.

Conjecture 13.22. Let X be a compact nonpositively curved cube complex. Then X is virtually special iff for each immersed hyperplane $D \to X$, there are finitely many distinct subgroups of $\pi_1 D$ that are of the form $\pi_1 D \cap g\pi_1 D g^{-1}$.

Motivation for the above conjecture can be found in the realm of CSC's (see [65] and the problems listed there): A CSC is virtually special (equivalently, it is a virtually product) iff one (and hence all) of its hyperplanes satisfy the finiteness condition listed above.

Conjecture 13.23. Let H be a quasiconvex codimension-one subgroup of a hyperbolic group G. Then G has a finite index subgroup that splits along H. This should follow the proof in [3].

The following is a variation on the Algebraic Torus Theorem [16]:

Problem 13.24. Suppose G has a codimension-1 amenable subgroup. Does G virtually splits over an amenable subgroup?

Problem 13.25. (Kropholler-Roller Conjecture [38]) Let H be a f.g. subgroup of a f.g. group G. Suppose there is an H almost-invariant proper subset $A \subset G$ with $A = HAH$. Does G split over a subgroup (presumably related to H)?

Problem 13.26. Let M be a hyperbolic 3-manifold. Does M have a finite cover that is homeomorphic to a nonpositively curved cube complex?

Problem 13.27. Let M be a cusped hyperbolic 3-manifold. Does M have a finite index subgroup that is the fundamental group of a 2-dimensional nonpositively curved cube complex. A $\mathcal{VH}$-complex? Equivalently, does there exist $\widehat{M}$ such that $\pi_1\widehat{M}$ acts freely on the product of two trees?

Problem 13.28. Let M be a closed hyperbolic 3-manifold. Does $\pi_1 M$ have a finite cover that is the fundamental group of a compact 3-dimensional nonpositively curved cube complex. Does $\pi_1 M$ have a finite index subgroup that acts freely on the product of three trees?

Problem 13.29. Let X be a nonpositively curved cube complex that is homeomorphic to a closed manifold. Is X virtually special? (I guess there is an example with an irreducible CSC mapped inside by a local isometry).

Problem 13.30 ("Special" Charney-Davis-Hopf conjecture). Let M be a closed $2n$-manifold homeomorphic to a special cube complex. Show that $(-1)^k\chi(M) \geq 0$.

Problem 13.31. Let G be hyperbolic relative to abelian subgroups. Suppose G splits as a graph of groups where all the vertex (or equivalently all the edge groups) are relatively quasiconvex. Suppose the edge groups are relatively malnormal, and vertex groups are virtually compact special. Show that G acts properly and [relatively cocompactly] on a CAT(0) cube complex. Generalize this further to handle the case where the parabolic subgroups are virtually special (instead of just virtually abelian).

Motivation for the following problem is that it was done in the special case of certain cusped hyperbolic manifolds in [42].

Problem 13.32. Let G be hyperbolic relative to proper subgroups $\{P_i\}$. Suppose G acts properly and cocompactly on a CAT(0) cube complex. Find a codimension-one subgroup that does not contain any infinite order parabolic elements.

Problem 13.33. Let G act properly and cocompactly on a nonpositively curved cube complex. Suppose G is hyperbolic relative to subgroups $\{P_1, \ldots, P_r\}$. Let $P_i' \subset P_i$ be subgroups such that each $P_i/\langle\langle P_i' \rangle\rangle$ acts properly [and cocompactly] on a CAT(0) cube complex. Does $G/\langle\langle P_1', \ldots, P_r' \rangle\rangle$ act properly [and cocompactly] on a CAT(0) cube complex?

One approach to the above problem follows the proof of Theorem 10.1.

Problem 13.34. Let G act properly and cocompactly on a CAT(0) space $\widetilde{X}$. Let H be a quasiconvex subgroup of G (i.e. an orbit Hx is *quasiconvex* in the sense that there exists r such that any geodesic with endpoints in Hx lies in $\mathcal{N}_r(Hx)$). Suppose that W is a wall in

$\widetilde{X}$ whose stabilizer is H. Is the dual cube complex finite dimensional? Equivalently, does H have the bounded packing property (see [30])? The same problem for quasi-isometrically embedded subgroups. Some examples of CAT(0) groups with such walls are the rhombus groups in [36].

Problem 13.35. Suppose a hyperbolic group G acts freely on a CAT(0) cube complex. Does G act freely and cocompactly on a CAT(0) cube complex? I guess there are counterexamples. The work of [14] might be relevant here.

Here are two variants to the above problem: Let G be an aTmenable hyperbolic group. Does G act metrically properly on a CAT(0) cube complex? Let G be a hyperbolic group that is a subgroup of a raag. Is G the fundamental group of a compact nonpositively curved cube complex?

Problem 13.36. Virtual Specialness of hierarchies without relative hyperbolicity. This is very open ended. It should require a finiteness property on the height as in Conjecture 13.22.

Problem 13.37. Suppose that G acts properly and cocompactly on a CAT(0) space. Suppose G splits as a graph of groups where each vertex group acts freely on a CAT(0) cube complex. Does G act freely on a CAT(0) cube complex? (Perhaps B_4 will provide a counterexample.)

Problem 13.38. Let X be a compact nonpositively curved cube complex. Suppose that $\pi_1 D$ is separable in $\pi_1 X$ for each immersed hyperplane $D \to X$. Does it follow that X is virtually special?

Harder: Suppose that all hyperplanes are embedded, 2-sided, and do not self-osculate. Does it follow that X is virtually special?

Problem 13.39. Let G be hyperbolic relative to subgroups $\{P_1, \dots, P_r\}$. Suppose that each P_i is quasi-isometric to a CAT(0) cube complex. Prove that G is quasi-isometric to a CAT(0) cube complex. (This holds when G is hyperbolic [27].)

Problem 13.40. Is every CAT(0) space (with a proper cocompact group action) quasi-isometric to a CAT(0) cube complex?

Problem 13.41. Challenge: Give an example of two groups G_1, G_2 that are quasi-isometric, but where G_1 is the fundamental group of a compact nonpositively curved cube complex, and G_2 has property (T).

Problem 13.42. Characterize the groups acting metrically properly [or freely] on CAT(0) cube complexes that have a quadratic isoperimetric function. (One expects a characterization in terms of properties of a wallspace structure on G.)

Problem 13.43. Let G be hyperbolic relative to abelian groups, and suppose G has a cyclic hierarchy terminating in free-abelian groups (that is, at each step the amalgamated subgroup is is either trivial or cyclic). Is G virtually a limit group? (This is related to special cube complexes through Sela's retractive towers.)

Problem 13.44. Let G act on a CAT(0) cube complex $\widetilde{X}$ with amenable [or even aTmenable] vertex stabilizers, trivial edge stabilizers, and finitely many orbits of hyperplanes. Is G aTmenable? (See [46, 49] for the finite stabilizer case).

Problem 13.45. Let G be a hyperbolic group that splits as a graph of hyperbolic groups with aTmenable vertex groups. Show that G is aTmenable.

Problem 13.46. Let G be an aTmenable hyperbolic group. Let h be an infinite order element. Show that $G/\langle\langle h^n \rangle\rangle$ is aTmenable for sufficiently large n.

Problem 13.47. Does every f.g. raag G admit a faithful discrete affine representation? If not all raags, then which special G admit such representations? See [15] for the case when G is free.

Problem 13.48. Does G satisfy Kaplansky's zero divisor conjecture hold when G acts freely on a CAT(0) cube complex? That is, does the group ring $R[G]$ have no zero-divisors when R is an integral domain? (G need not satisfy the unique product property as Promislow's example [51] acts freely on CAT(0) cube complex.)

Problem 13.49. Let $P(\Gamma)$ be the ring of formal power series with variables equal to the vertices of Γ, and where two such variables commute if the corresponding vertices are adjacent. Let U be the group of units of $P(\Gamma)$. Suppose $G \subset U$ is a f.g. subgroup. Is G (virtually) a subgroup of a raag?

Problem 13.50. Let G be the fundamental group of a special cube complex. Suppose that G is not a surface group nor free. Does G contain an infinite index subgroup that is not free? Does G contain a surface subgroup? (It suffices to prove this when G is not free and G splits as an amalgam of a free group along a malnormal subgroup.)

Problem 13.51. Let G be the fundamental group of a nonpositively curved complex complex X. Let $\mathsf{cd}(G)$ denote the cohomological dimension of G. Prove that for $0 \leq k \leq \mathsf{cd}(G)$ there is a subgroup $H_k \subset G$ with $\mathsf{cd}(H_k) = k$. Arguing by induction on the dimension, this is equivalent to showing that for some hyperplane Y of X, $\mathsf{cd}(\pi_1 Y) \geq \mathsf{cd}(G) - 1$.

Conjecture 13.52. Every lattice in hyperbolic space acts properly and cocompactly on a CAT(0) cube complex. This is still open for uniform arithmetic lattices [6]. (In the nonuniform case one should conclude: either relatively cocompactly or virtually cocompactly.) The same for geometrically finite subgroups.

Problem 13.53. Suppose G is a word-hyperbolic group that acts properly and cocompactly on a CAT(0) cube complex. Prove that G is a discrete subgroup of $\mathrm{Isom}(\mathbb{H}^n)$ for some n.

Problem 13.54. Let M be a hyperbolic m-manifold. Let J be a geometrically finite subgroup of $\pi_1 M$. Prove that for some $n \geq m$ there is (virtually?) a discrete representation $\pi_1 M \to \mathrm{Isom}(\mathbb{H}^n)$ and a hyperplane $\mathbb{H}^p \subset \mathbb{H}^n$ such that J is the preimage of $\mathrm{Stabilizer}(\mathbb{H}^p)$.

Acknowledgement. The author's research is supported by NSERC. I am grateful to Mark Hagen and Daniel Woodhouse for helpful comments and corrections to this text.

References

[1] Aaron Abrams and Robert Ghrist, *Finding topology in a factory: configuration spaces*, Amer. Math. Monthly **109**(2) (2002), 140–150.

[2] Ian Agol, *Criteria for virtual fibering*, J. Topol. **1**(2) (2008), 269–284.

[3] ———, *The virtual haken conjecture*, 2012, With an appendix by Ian Agol, Daniel Groves, and Jason Manning.

[4] W. Ballmann and J. Świątkowski, *On L^2-cohomology and property (T) for automorphism groups of polyhedral cell complexes*, Geom. Funct. Anal., **7**(4) (1997), 615–645.

[5] N. Bergeron, *Cycles géodésiques transverses dans les variétés hyperboliques*, Geom. Funct. Anal., **12**(3) (2002), 437–463.

[6] Nicolas Bergeron, Frédéric Haglund, and Daniel T. Wise, *Hyperplane sections in arithmetic hyperbolic manifolds*, J. Lond. Math. Soc. (2), **83**(2) (2011), 431–448.

[7] Nicolas Bergeron and Daniel T. Wise, *A boundary criterion for cubulation*, Amer. J. Math., **134**(3) (2012), 843–859.

[8] Mladen Bestvina and Noel Brady, *Morse theory and finiteness properties of groups*, Invent. Math., **129**(3) (1997), 445–470.

[9] Thomas Brady, *Artin groups of finite type with three generators*, Michigan Math. J., **47**(2) (2000), 313–324.

[10] Martin R. Bridson and André Haefliger, *Metric spaces of non-positive curvature*, Springer-Verlag, Berlin, 1999.

[11] Richard D. Canary, *Covering theorems for hyperbolic 3-manifolds*, In Low-dimensional topology (Knoxville, TN, 1992), Conf. Proc. Lecture Notes Geom. Topology, III, pages 21–30. Internat. Press, Cambridge, MA, 1994.

[12] Ruth Charney, *An introduction to right-angled Artin groups*, Geom. Dedicata, **125** (2007), 141–158.

[13] Michael W. Davis and Tadeusz Januszkiewicz, *Right-angled Artin groups are commensurable with right-angled Coxeter groups*, J. Pure Appl. Algebra, **153**(3) (2000), 229–235.

[14] Thomas Delzant and Misha Gromov, *Cuts in Kähler groups*, In Infinite groups: geometric, combinatorial and dynamical aspects, volume 248 of Progr. Math., pp 31–55. Birkhäuser, Basel, 2005.

[15] Todd A. Drumm and William M. Goldman, *The geometry of crooked planes*, Topology, **38**(2) (1999), 323–351.

[16] M. J. Dunwoody and E. L. Swenson, *The algebraic torus theorem*, Invent. Math., 140(3):605–637, 2000.

[17] R. Ghrist and V. Peterson, *The geometry and topology of reconfiguration*, Adv. in Appl. Math., **38**(3) (2007), 302–323.

[18] Rita Gitik, Mahan Mitra, Eliyahu Rips, and Michah Sageev, *Widths of subgroups*, Trans. Amer. Math. Soc., **350**(1) (1998), 321–329.

[19] M. Gromov, *Hyperbolic groups*, In Essays in group theory, volume 8 of Math. Sci. Res. Inst. Publ., pp. 75–263. Springer, New York, 1987.

[20] Daniel Groves and Jason Fox Manning, *Dehn filling in relatively hyperbolic groups*, Israel J. Math., **168** (2008), 317–429.

[21] Mark F. Hagen and Daniel T. Wise, *Cubulating hyperbolic free-by-cyclic groups: the irreducible case*, http://arxiv.org/pdf/1311.2084v1, pp. 1–39, 2013, Submitted.

[22] ———, *Cubulating hyperbolic free-by-cyclic groups: the general case*, 2014, In preparation.

[23] Frédéric Haglund, *Finite index subgroups of graph products*, Geom. Dedicata, **135** (2008), 167–209.

[24] *Frédéric Haglund and Frédéric Paulin, Simplicité de groupes d'automorphismes d'espaces à courbure négative*, In The Epstein birthday schrift, pp. 181–248 (electronic). Geom. Topol., Coventry, 1998.

[25] Frédéric Haglund and Daniel T. Wise, *Special cube complexes*, Geom. Funct. Anal., **17**(5) (2008), 1 551–1620.

[26] ———, *Coxeter groups are virtually special*, Adv. Math., **224**(5) (2010), 1890–1903.

[27] ———, *A combination theorem for special cube complexes*, Ann. of Math. (2), **176**(3) (2012), 1427–1482.

[28] Marshall Hall, Jr., *Coset representations in free groups*, Trans. Amer. Math. Soc., **67** (1949), 421–432.

[29] G. Christopher Hruska and Daniel T. Wise, *Finiteness properties of cubulated groups*, Compositio Mathematica, pp. 1–58, to appear.

[30] ———, *Packing subgroups in relatively hyperbolic groups*, Geom. Topol., **13**(4) (2009), 1945–1988.

[31] Tim Hsu and Daniel T. Wise, *Cubulating malnormal amalgams*, Invent. Math. To appear.

[32] ———, *On linear and residual properties of graph products*, Michigan Math. J., **46**(2) (1999), 251–259.

[33] ———, *Separating quasiconvex subgroups of right-angled Artin groups*, Math. Z., **240**(3) (2002), 521–548.

[34] ———, *Cubulating graphs of free groups with cyclic edge groups*, Amer. J. Math., **132**(5) (2010), 1153–1188.

[35] S. P. Humphries, *On representations of Artin groups and the Tits conjecture*, J. Algebra, **169** (1994), 847–862.

[36] David Janzen and Daniel Wise, *Cubulating rhombus groups*, Groups Geom. Dyn., **7**(2) (2013), 419–442.

[37] Jeremy Kahn and Vladimir Markovic, *Immersing almost geodesic surfaces in a closed hyperbolic three manifold*, Ann. of Math. (2), **175**(3) (2012), 1127–1190.

[38] P. H. Kropholler and M. A. Roller, *Relative ends and duality groups*, J. Pure Appl. Algebra, **61**(2) (1989), 197–210.

[39] I.J. Leary, *A metric Kan–Thurston theorem*, Preprint. arXiv:1009.1540.

[40] Yi Liu, *Virtual cubulation of nonpositively curved graph manifolds*, J. Topol., **6**(4) (2013), 793–822.

[41] Roger C. Lyndon and Paul E. Schupp, *Combinatorial group theory*. Springer-Verlag, Berlin, 1977. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 89.

[42] Joseph D. Masters and Xingru Zhang, *Closed quasi-Fuchsian surfaces in hyperbolic knot complements*, Geom. Topol., **12**(4) (2008), 2095–2171.

[43] Jonathan P. McCammond and Daniel T. Wise, *Fans and ladders in small cancellation theory*, Proc. London Math. Soc. (3), **84**(3) (2002), 599–644.

[44] Lee Mosher, *Geometry of cubulated* 3*-manifolds*, Topology, **34**(4) (1995), 789–814.

[45] Gábor Moussong, *Hyperbolic Coxeter Groups*. PhD thesis, Ohio State University, 1988.

[46] G. A. Niblo and L. D. Reeves, *The geometry of cube complexes and the complexity of their fundamental groups*, Topology, **37**(3) (1998), 621–633.

[47] _____, *Coxeter groups act on* $\mathrm{CAT}(0)$ *cube complexes*, J. Group Theory, **6**(3) (2003), 399–413.

[48] Graham Niblo and Lawrence Reeves, *Groups acting on* $\mathrm{CAT}(0)$ *cube complexes*, Geom. Topol., 1:approx. 7 pp. (electronic), 1997.

[49] Graham A. Niblo and Martin A. Roller, *Groups acting on cubes and Kazhdan's property (T)*, Proc. Amer. Math. Soc., **126**(3) (1998), 693–699.

[50] Denis V. Osin, *Peripheral fillings of relatively hyperbolic groups*, Invent. Math., **167**(2) (2007), 295–326.

[51] S. David Promislow, *A simple example of a torsion-free, nonunique product group*, Bull. London Math. Soc., **20**(4) (1988), 302–304.

[52] Piotr Przytycki and Daniel T. Wise, *Mixed 3-manifolds are virtually special*, pages 1–24. Available at arXiv:1205.6742.

[53] Michah Sageev, *Ends of group pairs and non-positively curved cube complexes*, Proc. London Math. Soc. (3), **71**(3) (1995), 585–617.

[54] _____, *Codimension-*1 *subgroups and splittings of groups*, J. Algebra, **189**(2) (1997), 377–389.

[55] Michah Sageev and Daniel T. Wise, *Cores for quasiconvex actions*, Proc. Amer. Math. Soc., To appear.

[56] _____, *The Tits alternative for* $\mathrm{CAT}(0)$ *cubical complexes*, Bull. London Math. Soc., **37**(5) (2005), 706–710.

[57] Peter Scott, *Subgroups of surface groups are almost geometric*, J. London Math. Soc. (2), **17**(3) (1978), 555–565.

[58] D. T. Wise, *Sectional curvature, compact cores, and local quasiconvexity*, Geom. Funct. Anal., **14**(2) (2004), 433–468.

[59] ———, *Cubular tubular groups*, Trans. Amer. Math. Soc. To appear.

[60] ———, *The structure of groups with a quasiconvex hierarchy*, Available at http://www.math.mcgill.ca/wise/papers, pp. 1–189. Submitted.

[61] ———, *Subgroup separability of graphs of free groups with cyclic edge groups*, Q. J. Math., **51**(1) (2000), 107–129.

[62] ———, *The residual finiteness of negatively curved polygons of finite groups*, Invent. Math., **149**(3) (2002), 579–617.

[63] ———, *Cubulating small cancellation groups*, GAFA, Geom. Funct. Anal., **14**(1) (2004), 150–214.

[64] ———, *Subgroup separability of the figure 8 knot group*, Topology, **45**(3) (2006), 421–463.

[65] ———, *Complete square complexes*, Comment. Math. Helv., **82**(4) (2007), 683–724.

[66] ———, *Morse theory, random subgraphs, and incoherent groups*, Bull. Lond. Math. Soc., **43**(5) (2011), 840–848.

[67] ———, *From riches to raags: 3-manifolds, right-angled Artin groups, and cubical geometry*, volume 117 of CBMS Regional Conference Series in Mathematics, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2012.

[68] Andrzej Żuk, *La propriété (T) de Kazhdan pour les groupes agissant sur les polyèdres*. C. R. Acad. Sci. Paris Sér. I Math., **323**(5) (1996), 453–458.

Dept. of Math. & Stats., McGill University, Montreal, QC, Canada H3A 0B9
E-mail: wise@math.mcgill.ca

6. Topology

A guide to (étale) motivic sheaves

Joseph Ayoub

Abstract. We recall the construction, following the method of Morel and Voevodsky, of the triangulated category of étale motivic sheaves over a base scheme. We go through the formalism of Grothendieck's six operations for these categories. We mention the relative rigidity theorem. We discuss some of the tools developed by Voevodsky to analyze motives over a base field. Finally, we discuss some long-standing conjectures.

Mathematics Subject Classification (2010). Primary 14C25, 14F05, 14F20, 14F42; Secondary 18F20.

Keywords. Motives, motivic sheaves, motivic cohomology, Grothendieck's six operations, conservativity conjecture, motivic t-structures.

1. Introduction

The (co)homological invariants associated to an algebraic variety fall into two classes:

(a) the *algebro-geometric invariants* such as higher Chow groups (measuring the complexity of algebraic cycles inside the variety) and Quillen K-theory groups (measuring the complexity of vector bundles over the variety);

(b) the class of *transcendental invariants* such as Betti cohomology (with its mixed Hodge structure) and ℓ-adic cohomology (with its Galois representation).

The distinction between these two classes is extreme.

- The algebro-geometric invariants are abstract Abelian groups, often of infinite rank, carrying no extra structure.[1]They vary chaotically in families and are not computable in any reasonable sense.
- On the other hand, transcendental invariants are concrete groups of finite rank (over some coefficient ring) carrying a rich extra structure. Together with their extra structure, they vary "continuously" in families.

Nevertheless, all these invariants are expected to be shadows of some master invariants, called the *motives* of the algebraic variety. The algebro-geometric invariants are expected to be groups of morphisms, extensions and higher extensions between these motives and other basic ones (such as Tate motives), while each of these motives *realizes* (i.e., gives rise) to a multitude of transcendental invariants of different types that, a priori, look poorly related.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

One of the ultimate goals of the theory of motives is to serve as a bridge *between the above two classes of cohomological invariants.*

Until now, establishing a fully satisfactory theory of motives has defied all attempts. Thinking about it as a bridge between (a) and (b), one can describe the present status of the theory as a broken bridge or, better, as a union of two half-bridges that, for the moment, fail to meet.

- The first half bridge, the one starting from (a), is a theory of motives that gives a satisfactory framework for understanding the algebro-geometric invariants.
- The second half-bridge, the one starting form (b), is a theory of motives that encapsulates the transcendental invariants and endows them with universal extra structures.

Concerning the second half-bridge, we just mention few highlights. In the *pure* case, i.e., for smooth and proper varieties, an approach was pioneered by Grothendieck [20]. Roughly speaking, Grothendieck's idea was to "decompose" smooth and proper varieties into "cohomological atoms" called *pure numerical motives* using certain *algebraic cycles* whose existence would be guaranteed by his (yet unproven) *Standard Conjectures* [12]. Later on, Deligne [11] and then André [2] made Grothendieck's approach unconditional by replacing algebraic cycles with *absolute Hodge cycles* and *motivated cycles* respectively. In the *mixed* case, i.e., for possibly open and singular varieties, an approach was invented by Nori (unpublished, but see [23, §5.3.3] for an account) based on his weak Tannakian reconstruction theorem which is an abstract devise yielding an Abelian category out of a representation of a diagram (aka., quiver). The main geometric ingredient behind most results about Nori's motives is the so-called *Basic Lemma* which can be considered as an enhanced form of the Lefschetz hyperplane theorem. In all these approaches (in the pure and mixed cases), the outcome is a *Tannakian* (and hence *Abelian*) category of motives whose fundamental group is the so-called *motivic Galois group*. It is also important to note here a crucial drawback: except the original construction of Grothendieck which is conditional on the Standard Conjectures, all available unconditional constructions of Abelian categories of motives depend on transcendental data (namely, a Weil cohomology theory such as Betti cohomology or ℓ-adic cohomology). For this reason, the existence of the "true" Abelian category of motives is still considered to be an open question.[2]

The present article is mainly concerned with the first half-bridge, i.e., the one starting from (a). Here the outcome of the theory is a *triangulated* category of motives whose groups of morphisms are blends of the algebro-geometric invariants of algebraic varieties (and more precisely, their higher Chow groups). If the existence of such categories was part of the Grothendieck motivic picture, it was probably Beilinson and Deligne who first expressed the hope that such categories might be easier to construct than their Abelian counterparts. And indeed, three different constructions of triangulated categories of motives appeared in the nineties by Hanamura [13–15], Levine [22] and Voevodsky [29] (see also its precursor [28]).

[1]To avoid confusion, we mention that the kind of extra structures we have in mind are those that can be given by the action of some group of symmetries such as the Galois group of the base field or, more generally, the fundamental group of a Tannakian category such as the category of mixed Hodge structures. It should be mentioned here that higher Chow groups are expected to carry a filtration, the conjectural Bloch–Beilinson filtration, with quite remarkable properties.

[2]Over a field of characteristic zero, it can be shown that if the "true" Abelian category of mixed motives exists, then it must be equivalent to Nori's category, and its subcategory of semi-simple objects must be equivalent to André's category. (The equivalence between André's and Deligne's categories is another story as it would require a weak form of the *Hodge Conjecture*.)

Although, the three categories were found to be equivalent, Voevodsky's construction [29] attracted most attention due to its beauty, simplicity and potential.

Nearly a decade later, it was realized (based on work of Morel and Cisinski–Déglise) that a mild modification of Voevodsky's construction, yields an even simpler (and certainly as beautiful) construction of the same (up to equivalence) triangulated category of motives at least if torsion is neglected or, more precisely, if descent for the étale topology is imposed (which is the right thing to do for many questions concerning integral motives such as the Hodge and Tate conjectures, existence of a motivic t-structure, etc; see §5.2). This simplified construction is more in the spirit of the construction of the Morel–Voevodsky $\mathbb{A}^1$-homotopy category [25] (and more precisely its stabilization that was worked out by Jardine [19]) and has the advantage of giving the correct triangulated categories over any base scheme.[3] These triangulated categories are denoted by $\mathbf{DA}^{\text{ét}}(S;\Lambda)$, where S is the base scheme and Λ is the ring of coefficients, and their objects are called *motivic sheaves* over S or simply *S-motives*;[4] they are the subject of this paper.

The organization is as follows. In §2 we give the details of the construction of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$. We hope to convince the reader that this construction is simple and natural. In §3 we explain the basic operations that one can do on motivic sheaves; the story here is parallel to what one has in the context of étale and ℓ-adic sheaves although the construction of the operations follows a different route. One should consider the formalism of the six operations as a tool to reduce questions about motivic sheaves over general bases to questions about motives over a point (i.e., the spectrum of a field). In order for this formalism to be of any use, one needs information about motives over fields. In §4 we start discussing results about the internal structure of the category of motives over a field. More precisely, we give a concrete description of the group of morphisms between certain motives; such groups are usually called *motivic cohomology*. Here all the results are due to Voevodsky and this is the place where the extra complexity in his original construction pays off. In particular, we recall the original construction of Voevodsky in §4.1 and explain in §4.2 how it permits the computation of motivic cohomology. In §5 we list some of the big open questions concerning motives. It is these conjectures that need to be solved for having a satisfactory theory of motivic sheaves and filling the gap between the two half-bridges discussed above.

2. Construction

In this section, we go through the construction of the categories $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ of *étale motivic sheaves* (or *motivic sheaves* for short) over a base scheme S and with coefficients in a commutative ring Λ. This construction is a slight variation of Voevodsky's original construction of his $\mathbf{DM}^{\text{ét}}(S;\Lambda)$ [24, 29] (see Remark 4.3 for more precisions). In fact, it is really a *simplification* of the latter as sheaves with transfers get replaced by ordinary sheaves. The category $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ should be also considered as the linearized counterpart of the Morel–Voevodsky stable $\mathbb{A}^1$-homotopy category in the étale topology $\mathbf{SH}^{\text{ét}}(S)$ [19, 25]. In fact, both categories $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ and $\mathbf{SH}^{\text{ét}}(S)$ are constructed in a uniform way in [6, Chapitre 4] as special cases of categories $\mathbf{SH}^{T}_{\mathfrak{M}}(S)$ by choosing $\mathfrak{M}$ to be the category of Λ-modules

[3] The original construction of Voevodsky is also known to give the correct triangulated categories when the base scheme is normal. However, the question remains open for more general base schemes (but see Remark 4.6).

[4] It is common to use the terminology "étale motivic sheaves". However, as the main article concerns motives in the étale topology, we use the shorthand "motivic sheaves".

or the category of simplicial symmetric spectra.

In order to keep the technicalities as low as possible, we will be using Verdier localization of triangulated categories [27] instead of the more natural/satisfactory Bousfield localization of model categories [16] which is usually employed in this context. We start by recalling Verdier localization.

2.1. A technical tool: Verdier localization. Recall that a *triangulated category* $\mathcal{T}$ is an additive category endowed with an autoequivalence $A \mapsto A[1]$ and a class of *distinguished triangles* which are diagrams of the form

$$A \xrightarrow{\alpha} B \xrightarrow{\beta} C \xrightarrow{\gamma} A[1] \tag{2.1}$$

satisfying a list of axioms. In particular, given a distinguished triangle as above, one has $\beta \circ \alpha = 0$ and $\gamma \circ \alpha = 0$. Moreover, the distinguished triangle (2.1) is determined by the map $\alpha : A \to B$ up to an isomorphism, which is in general not unique. Nevertheless, it will be sometimes convenient to abuse notation by writing $C = \mathrm{Cone}(\alpha)$ (and thus pretending that C depends canonically on α). Of course, this notation is inspired from topology: one thinks about a distinguished triangle (2.1) as an abstract version of a cofibre sequence. An important fact to keep in mind is the following: *α is an isomorphism if and only if* $\mathrm{Cone}(\alpha)$ *is zero*.

Now, let $\mathcal{T}$ be a triangulated category and $\mathcal{E} \subset \mathcal{T}$ a full subcategory closed under suspensions and desuspensions (i.e., under application of the powers $[n]$, positive and negative, of the autoequivalence $[1]$) and under cones. (Such an $\mathcal{E}$ is called a *triangulated subcategory* of $\mathcal{T}$.) In this situation, we have (see [27, Théorème 2.2.6]):

Proposition 2.1. *There exists a triangulated category $\mathcal{T}/\mathcal{E}$, called the* Verdier quotient *of $\mathcal{T}$ by $\mathcal{E}$, which is universal for the following two properties.*

i) *There is a canonical triangulated functor $\mathcal{T} \to \mathcal{T}/\mathcal{E}$ which is the identity on objects (in particular $\mathcal{T}$ and $\mathcal{T}/\mathcal{E}$ have the same class of objects).*

ii) *For every $A \in \mathcal{E}$, one has $A \simeq 0$ in $\mathcal{T}/\mathcal{E}$.*

Remark 2.2. The construction of $\mathcal{T}/\mathcal{E}$ goes as follows. Consider the class of arrows $S_{\mathcal{E}}$ in $\mathcal{T}$ given by

$$S_{\mathcal{E}} = \{\alpha : A \to B \mid \mathrm{Cone}(\alpha) \in \mathcal{E}\}.$$

The axioms satisfied by the class of distinguished triangles imply that $S_{\mathcal{E}}$ admits a "calculus of fractions". The Verdier quotient is then defined by

$$\mathcal{T}/\mathcal{E} := \mathcal{T}[(S_{\mathcal{E}})^{-1}].$$

In words, $\mathcal{T}/\mathcal{E}$ is the category obtained by formally inverting the arrows in $S_{\mathcal{E}}$.[5] This explains why the Verdier quotient is also called a *localization*.

2.2. An almost correct construction in two steps. The category $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ is obtained from the derived category of étale sheaves on smooth S-schemes by formally forcing two simple properties. In this subsection, we discuss these properties and explain how to force them successively. This yields a slightly naive notion of motivic sheaves. The correct notion will be given in §2.3.

[5] Needless to say that we are ignoring some set-theoretical issues here.

2.2.1. Some notation. From now on, Λ will always denote a commutative ring that we call the *ring of coefficients*. (In practice, Λ is $\mathbb{Z}$, $\mathbb{Q}$, a subring of $\mathbb{Q}$ or a quotient of $\mathbb{Z}$. However, it is sometimes useful to take for Λ a number ring, a number field, a local field, etc.) Given a set E, we denote by $\Lambda \otimes E = \bigoplus_{e\in E} \Lambda \cdot e$ the free Λ-module generated by E.

For simplicity, all schemes will be separated and the reader will not loose much by assuming that all schemes are also Noetherian of finite Krull dimension.

Let S be a base scheme. We denote by Sm/S the category of smooth S-schemes.[6] We endow Sm/S with the étale topology ([3, Exposé VII]) and we denote by $\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$ the category of étale sheaves with values in Λ-modules. If no confusion can arise, objects of $\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$ will be simply called *étale sheaves* on Sm/S. Given a smooth S-scheme X, we denote by $\Lambda_{\text{ét}}(X) := \mathrm{a}_{\text{ét}}(\Lambda \otimes X)$ the étale sheaf associated to the presheaf $U \in \mathrm{Sm}/S \mapsto \Lambda \otimes \mathrm{Hom}_S(U, X)$. This gives a Yoneda functor

$$\Lambda_{\text{ét}} : \mathrm{Sm}/S \to \mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda) \tag{2.2}$$

which one should consider as the first/obvious linearization of the category of smooth S-schemes, a necessary step for passing from S-schemes to S-motives.

The following lemma is left as an exercise and will not be used elsewhere. It shows that étale sheaves on Sm/S have transfers along finite étale covers.

Lemma 2.3. *Let X and U be smooth S-schemes and assume that S is normal. Then $\Lambda_{\text{ét}}(X)(U)$ is the free Λ-module generated by closed integral subschemes $Z \subset U \times_S X$ such that the normalization of Z is étale and finite over U.*

The category $\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$ possesses a monoidal structure. If $\mathcal{M}$ and $\mathcal{N}$ are étale sheaves on Sm/S, then $\mathcal{M} \otimes_\Lambda \mathcal{N}$ is simply the étale sheaf associated to the presheaf $U \in \mathrm{Sm}/S \mapsto \mathcal{M}(U) \otimes_\Lambda \mathcal{N}(U)$. If there is no risk of confusion, we will write $- \otimes -$ instead of $- \otimes_\Lambda -$ for the tensor product of Λ-modules and sheaves of Λ-modules. Given two smooth S-schemes X and Y, it follows readily from the definitions that

$$\Lambda_{\text{ét}}(X) \otimes \Lambda_{\text{ét}}(Y) \simeq \Lambda_{\text{ét}}(X \times_S Y).$$

Said differently, the functor $\Lambda_{\text{ét}}$ is monoidal (when Sm/S is endowed with its Cartesian monoidal structure).

2.2.2. First step: $\mathbb{A}^1$-localization. To motivate what follows, we note that, for a scheme U, the projection $\mathbb{A}^1 \times U \to U$ (where $\mathbb{A}^1 = \mathrm{Spec}(\mathbb{Z}[t])$ is the affine line) induces isomorphisms in most cohomology theories (for instance, in Betti cohomology if $U \in \mathrm{Sm}/\mathbb{C}$, in ℓ-adic cohomology if ℓ is invertible on U, in algebraic K-theory if U is regular, etc). Thus, it is natural to expect the motives of U and $\mathbb{A}^1 \times U$ to be isomorphic.

To impose this in a "homologically correct" manner, we consider the derived category $\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$ of the Abelian category $\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$. Let $\mathcal{T}_{\mathbb{A}^1}$ be the smallest triangulated subcategory of $\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$ which is closed under arbitrary direct sums and containing the 2-terms complexes

$$[\ldots \to 0 \to \Lambda_{\text{ét}}(\mathbb{A}^1 \times U) \to \Lambda_{\text{ét}}(U) \to 0 \to \ldots] \tag{2.3}$$

[6] Recall that *smooth* implies in particular *locally of finite presentation*. One may also restrict to smooth quasi-projective S-schemes and even to smooth quasi-affine S-schemes as these will define equivalent sites for the étale topology.

for all smooth S-schemes U. (In the above complex, the nonzero map is induced by the obvious projection $\mathbb{A}^1 \times U \to U$.) Then define $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$ to be the Verdier quotient of $\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S))$ by $\mathcal{T}_{\mathbb{A}^1}$:

$$\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda) := \mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))/\mathcal{T}_{\mathbb{A}^1}.$$

The categories $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$ and $\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$ have the same objects, that is complexes of étale sheaves on Sm/S; however, a morphism in $\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$ whose cone belongs to $\mathcal{T}_{\mathbb{A}^1}$ gets inverted in $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$. As a matter of fact, the map $\Lambda_{\text{ét}}(\mathbb{A}^1 \times U) \to \Lambda_{\text{ét}}(U)$, whose cone is the complex (2.3), is an isomorphism in $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$.

Definition 2.4. An object of $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$ is called an *effective motivic sheaf* over S (or simply an *effective S-motive*). Given a smooth S-scheme X, then $\Lambda_{\text{ét}}(X)$, viewed as an object of $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$, is called the *effective homological motive* of X and will be denoted by $\mathrm{M}^{\mathrm{eff}}(X)$.

Definition 2.5. Let us denote by $\mathbf{DA}^{\mathrm{eff},\text{ét}}_{\mathrm{ct}}(S;\Lambda)$ the smallest triangulated subcategory of $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$ closed under direct summands and containing the motives $\mathrm{M}^{\mathrm{eff}}(X)$ for $X \in \mathrm{Sm}/S$ of finite presentation. Effective motivic sheaves in $\mathbf{DA}^{\mathrm{eff},\text{ét}}_{\mathrm{ct}}(S;\Lambda)$ are called *constructible*.

Remark 2.6. The category $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$ (as well as $\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$) inherits the monoidal structure of $\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$. If $\mathcal{M}_\bullet$ and $\mathcal{N}_\bullet$ are complexes of étale sheaves on Sm/S (i.e., objets of $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$), then their tensor product $(\mathcal{M}\otimes\mathcal{N})_\bullet$ is the total complex associated to the bi-complex $\mathcal{M}_\bullet \otimes \mathcal{N}_\bullet$.

2.2.3. Second step: naive stabilization. In this subsection, we give a low-tech (and slightly naive) construction yielding the category $\mathbf{DA}^{\text{ét},\,\mathrm{naive}}(S;\Lambda)$ which, nevertheless, captures the essence of the category $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ (see Remark 2.7).

The stabilization here refers to the process of rendering the Tate motive invertible for the tensor product.

To motivate this process, we need to explain another simple fact about the cohomology of algebraic varieties. To fix ideas, we consider ℓ-adic cohomology H^*_ℓ for schemes over an algebraically closed field k in which ℓ is invertible. The reduced cohomology of the pointed (by infinity) projective line $(\mathbb{P}^1_k,\infty)$ is given by

$$H^*_\ell(\mathbb{P}^1_k,\infty) \simeq \mathbb{Z}_\ell(-1)[-2]$$

where, as usual, $\mathbb{Z}_\ell(-1)$ is the dual of the Tate module $\mathbb{Z}_\ell(1) = \mathrm{Lim}_{n\in\mathbb{N}}\ \mu_{\ell^n}(k)$. Hence, seen as an object of the derived category $\mathbf{D}(\mathbb{Z}_\ell)$, the complex $H^*_\ell(\mathbb{P}^1_k,\infty)$ has total rank one and, equivalently, is invertible for the tensor product. It is the latter property that we want to impose on the motivic level.

To this effect, let $L := \Lambda_{\text{ét}}(\mathbb{P}^1_S,\infty_S)$ be the étale sheaf on Sm/S given by the cokernel of the inclusion $\Lambda(\infty_S) \hookrightarrow \Lambda_{\text{ét}}(\mathbb{P}^1_S)$. Seen as an object of $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$, L is the reduced effective homological S-motive of the pointed S-scheme $(\mathbb{P}^1_S,\infty_S)$. We will refer to L as the *Lefschetz motive*; it is the motive that corresponds to the constant complex of ℓ-adic sheaves $\mathbb{Z}_\ell(1)[2]$ over S (for ℓ invertible in $\mathcal{O}_S$).[7] However, it is easy to see that L is not an invertible

[7]This is consistent with what we said before: the ℓ-adic cohomology of $(\mathbb{P}^1_k,\infty)$ is $\mathbb{Z}_\ell(-1)[-2]$ and hence its ℓ-adic homology is $\mathbb{Z}_\ell(1)[2]$; it is the latter that should corresponds to the homological motive of $(\mathbb{P}^1_k,\infty)$.

object of $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$. Therefore, one is lead to invert it formally by considering

$$\mathbf{DA}^{\text{ét},\,\mathrm{naive}}(S;\Lambda) := \mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)[L^{-1}].$$

An objet of $\mathbf{DA}^{\text{ét},\,\mathrm{naive}}(S;\Lambda)$ consists of a pair (M,m) where $M \in \mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$ and $m \in \mathbb{Z}$. The group $\hom_{\mathbf{DA}^{\text{ét},\,\mathrm{naive}}(S;\Lambda)}((M,m),(N,n))$ of morphisms between two such objets is given by

$$\varinjlim_{r\geq -\min(m,n)} \hom_{\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)}(M\otimes L^{r+m}, N\otimes L^{r+n}). \tag{2.4}$$

With this definition, it is easy to see that the endofunctor $-\otimes L$ on $\mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda)$ corresponds to the functor $(M,m)\mapsto(M,m+1)$ on $\mathbf{DA}^{\text{ét},\,\mathrm{naive}}(S;\Lambda)$ which is an equivalence of categories with inverse $(M,m)\mapsto(M,m-1)$.

The formula (2.4) is reminiscent to the formula computing stable homotopy groups ot a topological space. This analogy suggests already that, as in topology, it is technically more convenient to use the formalism of spectra for inverting L. This is indeed the right method and will be explained in §2.3.

Remark 2.7. The category $\mathbf{DA}^{\text{ét},\,\mathrm{naive}}(S;\Lambda)$ suffers many technical defects. For instance, it is not a triangulated category and it doesn't have arbitrary direct sums. However, modulo these technical defects, $\mathbf{DA}^{\text{ét},\,\mathrm{naive}}(S;\Lambda)$ is essentially the right category of S-motives. More precisely, under some technical assumptions,[8] its full subcategory $\mathbf{DA}^{\text{ét},\,\mathrm{naive}}_{\mathrm{ct}}(S;\Lambda)$ consisting of pairs (M,m) with $M \in \mathbf{DA}^{\mathrm{eff},\text{ét}}_{\mathrm{ct}}(S;\Lambda)$, is equivalent to the category $\mathbf{DA}^{\text{ét}}_{\mathrm{ct}}(S;\Lambda)$ of constructible motives (see Definition 2.11 below), which is certainly the most interesting part of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$.

2.3. The definitive construction. This subsection can be skipped by the reader who is satisfied by the almost correct construction explained in §2.2. The goal here is to invert in a "homologically correct" manner the Lefschetz motive $L = \Lambda_{\text{ét}}(\mathbb{P}^1_S,\infty_S)$ for the tensor product. In fact, we will treat the localization (§2.2.2) and the stabilization (§2.2.3) in one single step!

We will borrow the machinery developed by topologists in the context of stable homotopy theory [1, 30] for inverting the (pointed) 1-dimensional sphere S^1 for the smash product. The only difference is that, instead of considering S^1-spectra (for the smash product), we will consider L-spectra (for the tensor product).

Definition 2.8. An *L-spectrum* (of étale sheaves on Sm/S) is a pair

$$\mathcal{E} = ((\mathcal{E}_n)_{n\in\mathbb{N}}, (\gamma_n)_{n\in\mathbb{N}})$$

where $\mathcal{E}_n$ is an étale sheaf on Sm/S and $\gamma_n : L\otimes\mathcal{E}_n \to \mathcal{E}_{n+1}$ is a morphism of sheaves called the n-th *assembly map*. We refer to the sheaf $\mathcal{E}_n$ as the n-th *level* of the L-spectrum $\mathcal{E}$.

A morphism of L-spectra $f : \mathcal{E}\to\mathcal{E}'$ is a collection of morphisms of sheaves $f_n : \mathcal{E}_n \to \mathcal{E}'_n$ that commute with the assembly maps, i.e., such that $f_{n+1}\circ\gamma_n = \gamma'_n\circ(\mathrm{id}_L\otimes f_n)$ for all $n\in\mathbb{N}$. We denote by $\mathrm{Spt}_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$ the category of L-spectra. This is an Abelian category.

[8] Such as S being Noetherian, of finite Krull dimension and of pointwise finite ℓ-cohomological dimension for very prime ℓ which is not invertible in Λ.

Remark 2.9. The functor $\mathrm{Ev}_p : \mathcal{E} \mapsto \mathcal{E}_p$, sending an L-spectrum to its p-th level admits a left adjoint

$$\mathrm{Sus}^p_L : \mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda) \to \mathrm{Spt}_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)).$$

If $\mathcal{F}$ is a complex of sheaves on Sm/S, then $\mathrm{Sus}^p_L\mathcal{F}$ is given by

$$(\mathrm{Sus}^p_L\mathcal{F})_n = \begin{cases} 0 & \text{if} \quad n \leq p-1, \\ L^{\otimes n-p} \otimes \mathcal{F} & \text{if} \quad n \geq p, \end{cases}$$

with the obvious assembly maps. Usually, Sus^0_L is called the *infinite suspension* functor and is denoted by Σ^∞_L.

We will construct $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ as a Verdier localization of the derived category

$$\mathbf{D}(\mathrm{Spt}_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)))$$

of L-spectra over Sm/S. For this, we consider the smallest triangulated subcategory $\mathcal{T}_{\mathbb{A}^1-\mathrm{st}}$ ("st" stands for "stable") of the latter closed under arbitrary direct sums and containing the complexes

$$[\ldots \to 0 \to \mathrm{Sus}^p_L\Lambda_{\text{ét}}(\mathbb{A}^1 \times U) \to \mathrm{Sus}^p_L\Lambda_{\text{ét}}(U) \to 0 \to \ldots] \tag{2.5}$$

$$[\ldots \to 0 \to \mathrm{Sus}^{p+1}_L(L \otimes \Lambda_{\text{ét}}(U)) \to \mathrm{Sus}^p_L\Lambda_{\text{ét}}(U) \to 0 \to \ldots] \tag{2.6}$$

for all smooth S-schemes U and all $p \in \mathbb{N}$. (In the first complex above, the nonzero map is induced by the projection to the second factor; in the second complex above, the nonzero map is the map of L-spectra given by the identity starting from level $p+1$.) We now define a new triangulated category as a Verdier quotient

$$\mathbf{DA}^{\text{ét}}(S;\Lambda) := \mathbf{D}(\mathrm{Spt}_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)))/\mathcal{T}_{\mathbb{A}^1-\mathrm{st}}.$$

Definition 2.10. An object of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ is called a *motivic sheaf* over S (or simply an S-*motive*). Given a smooth S-scheme X, then $\Sigma^\infty_L\Lambda_{\text{ét}}(X)$, viewed as an object of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$, is called the *homological motive* of X and will be denoted by $\mathrm{M}(X)$.

Definition 2.11. Let us denote by $\mathbf{DA}^{\text{ét}}_{\text{ct}}(S;\Lambda)$ the smallest triangulated subcategory of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ closed under direct summands and containing the motives $\mathrm{M}(X)(-p)[-2p] := \mathrm{Sus}^p_L\Lambda_{\text{ét}}(X)$ for $p \in \mathbb{N}$ and $X \in \mathrm{Sm}/S$ of finite presentation. Motivic sheaves in $\mathbf{DA}^{\text{ét}}_{\text{ct}}(S;\Lambda)$ are called *constructible*.

Remark 2.12. It can be shown that $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ is a triangulated category admitting arbitrary direct sums. Therefore, the construction via L-spectra resolves the technical defects of the category $\mathbf{DA}^{\text{ét, naive}}(S;\Lambda)$ constructed in §2.2.3.

Definition 2.13. For $p \in \mathbb{N}$, denote by $\Lambda_S(p)$ (or simply $\Lambda(p)$) the S-motive $\mathrm{Sus}^0_L(L^{\otimes p})[-2p]$ and $\Lambda_S(-p)$ (or simply $\Lambda(-p)$) the S-motive $\mathrm{Sus}^p(\Lambda)[2p]$. These are the Tate motives over S. We also define

$$H^p_{\mathcal{L}}(S;\Lambda(q)) := \hom_{\mathbf{DA}^{\text{ét}}(S;\Lambda)}(\Lambda_S(0), \Lambda_S(q)[p])$$

for p, $q \in \mathbb{Z}$. These groups are called the *étale* (or *Lichtenbaum*) *motivic cohomology* of S (with coefficients in Λ).

2.4. Complements. From Definition 2.10, a motivic sheaf over S is simply a complex of L-spectra on Sm/S, i.e., essentially a sequence of complexes of étale sheaves on Sm/S. This is of course deceiving and slightly misleading. The point is that every complex of L-spectra is *isomorphic* in $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ to a *stably* $\mathbb{A}^1$*-local* complex of L-spectra and it is the latter that deserves better to be called a motivic sheaf. Our goal in this paragraph is to explain this in some detail. We start with the effective case. (Below, $H^i_{\text{ét}}(-;A)$ stands for the étale hyper-cohomology with coefficients in a complex of étale sheaves A.)

Definition 2.14. Let $\mathcal{F}$ be a complex of étale sheaves on Sm/S. We say that $\mathcal{F}$ is $\mathbb{A}^1$*-local* if for all $U \in \mathrm{Sm}/S$ and $i \in \mathbb{Z}$, the map

$$H^i_{\text{ét}}(U;\mathcal{F}) \to H^i_{\text{ét}}(\mathbb{A}^1 \times U;\mathcal{F}),$$

induced by the projection to the second factor, is an isomorphism.

Remark 2.15. $\mathbb{A}^1$-locality is important for the following reason. Let $\mathcal{E}$ and $\mathcal{F}$ be two complexes of étale sheaves on Sm/S. Then, if $\mathcal{F}$ is $\mathbb{A}^1$-local, the natural homomorphism

$$\hom_{\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))}(\mathcal{E},\mathcal{F}) \to \hom_{\mathbf{DA}^{\text{eff},\,\text{ét}}(S;\Lambda)}(\mathcal{E},\mathcal{F})$$

is an isomorphism. In words, computing morphisms between effective motivic sheaves can be performed in the more familiar derived category of étale sheaves when the target is $\mathbb{A}^1$-local. The next result gives, in theory, a way to reduce to this favorable case.

Lemma 2.16. *There is, up to a unique isomorphism, a triangulated endofunctor* $\mathrm{Loc}_{\mathbb{A}^1}$ *of* $\mathbf{D}(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$ *endowed with a natural transformation* $\mathrm{id} \to \mathrm{Loc}_{\mathbb{A}^1}$ *such that the following two properties are satisfied for every complex* $\mathcal{F}$ *of étale sheaves on* Sm/S*:*

- $\mathrm{Loc}_{\mathbb{A}^1}(\mathcal{F})$ *is* $\mathbb{A}^1$*-local, and*
- $\mathcal{F} \to \mathrm{Loc}_{\mathbb{A}^1}(\mathcal{F})$ *is an* $\mathbb{A}^1$-weak equivalence (*i.e., becomes an isomorphism in the category* $\mathbf{DA}^{\text{eff},\,\text{ét}}(S;\Lambda)$).

$\mathrm{Loc}_{\mathbb{A}^1}$ *is called the* $\mathbb{A}^1$*-localization functor.*

Remark 2.17. If one adopts the convention that an "effective S-motive" is an $\mathbb{A}^1$-local complex of sheaves on Sm/S, then the effective motive of a smooth S-scheme X would be given by $\mathrm{Loc}_{\mathbb{A}^1}(\Lambda_{\text{ét}}(X))$. Therefore, understanding the $\mathbb{A}^1$-localization functor is of utmost importance in the theory of motives!

Remark 2.18. One of the drawback of the abstract construction is that it gives no information about the $\mathbb{A}^1$-localization functor. We will explain in §4.2 how Voevodsky is able to overcome this crucial difficulty (sadly, only when S is the spectrum of a field) using his theory of *homotopy invariant presheaves with transfers*.

We now turn to the stable setting.

Definition 2.19. Let $\mathcal{K} = ((\mathcal{K}_n)_{n\in\mathbb{N}}, (\gamma_n)_{n\in\mathbb{N}})$ be a complex of L-spectra of étale sheaves on Sm/S. We say that $\mathcal{K}$ is *stably* $\mathbb{A}^1$*-local* if the following two properties are satisfied for all $U \in \mathrm{Sm}/S$, $i \in \mathbb{Z}$ and $n \in \mathbb{N}$:

i) the map

$$H^i_{\text{ét}}(U;\mathcal{K}_n) \to H^i_{\text{ét}}(\mathbb{A}^1 \times U;\mathcal{K}_n),$$

induced by the projection to the second factor, is an isomorphism;

ii) the map

$$H^i_{\text{ét}}(U;\mathcal{K}_n) \to H^{i+2}_{\text{ét}}((\mathbb{P}^1,\infty)\times U;\mathcal{K}_{n+1}),$$

induced by the n-th assembly map, is an isomorphism.

Remark 2.20. Stably $\mathbb{A}^1$-local complexes of L-spectra are important for the same reason as the one explained in Remark 2.15.

Remark 2.21. Let $\mathcal{K}$ be a stably $\mathbb{A}^1$-local complex of L-spectra. Writing $\mathcal{K}(n)$ for the complex $\mathcal{K}_n[-2n]$, the two properties in Definition 2.19 gives the familiar isomorphisms:

i) $H^*_{\text{ét}}(\mathbb{A}^1\times U;\mathcal{K}(n)) = H^*_{\text{ét}}(U;\mathcal{K}(n))$;

ii) $H^*_{\text{ét}}((\mathbb{A}^1\smallsetminus 0)\times U;\mathcal{K}(n)) \simeq H^*_{\text{ét}}(U;\mathcal{K}(n)) \oplus H^{*-1}_{\text{ét}}(U;\mathcal{K}(n-1))$.

Lemma 2.22. *There is, up to a unique isomorphism, a triangulated endofunctor* $\mathrm{Loc}_{\mathbb{A}^1\text{-st}}$ *of* $\mathbf{D}(\mathrm{Spt}_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)))$ *endowed with a natural transformation* $\mathrm{id}\to\mathrm{Loc}_{\mathbb{A}^1\text{-st}}$ *such that the following two properties are satisfied for every complex of L-spectra $\mathcal{K}$:*

- $\mathrm{Loc}_{\mathbb{A}^1\text{-st}}(\mathcal{K})$ *is stably $\mathbb{A}^1$-local, and*
- $\mathcal{K}\to\mathrm{Loc}_{\mathbb{A}^1\text{-st}}(\mathcal{K})$ *is a* stable $\mathbb{A}^1$-weak equivalence *(i.e., becomes an isomorphism in* $\mathbf{DA}^{\text{ét}}(S;\Lambda)$*).*

Remark 2.23. As in the effective case, if one adopts the convention that an "S-motive" is a stably $\mathbb{A}^1$-local complex of L-spectra, then the motive of a smooth S-scheme X would be given by $\mathrm{Loc}_{\mathbb{A}^1\text{-st}}(\Sigma^\infty_T\Lambda_{\text{ét}}(X))$.

2.5. Relative rigidity theorem. When the characteristic of Λ is non-zero, the category $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ has a very simple description. Indeed, one has the following (see [9, Théorème 4.1]):

Theorem 2.24. *Let $n\in\mathbb{N}\smallsetminus\{0\}$ be an integer invertible in $\mathcal{O}(S)$. If Λ is a $\mathbb{Z}/n\mathbb{Z}$-algebra (and S satisfies some mild technical hypothesis*[9]*), then there is an equivalence of categories*

$$\mathbf{DA}^{\text{ét}}(S;\Lambda)\simeq\mathbf{D}(S_{\text{ét}};\Lambda)$$

where $\mathbf{D}(S_{\text{ét}};\Lambda)$ is the derived category of étale sheaves on $S_{\text{ét}}$ (the small étale site of S).

Remark 2.25. Theorem 2.24 is a relative version of a well-known result of Suslin–Voevodsky [29, Proposition 3.3.3 of Chapter 5] stating the same conclusion for the category $\mathbf{DM}^{\text{ét}}(S;\Lambda)$ when S is a field.

Remark 2.26. From a certain perspective, Theorem 2.24 is disappointing. Indeed, its shows that the categories $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ are too simple to capture the complexity of the torsion in Chow groups. This is not so surprising as it is well-known that higher Chow groups do not satisfy étale descent. A way around this is to replace in the construction "étale" by "Nisnevich" which yields the categories $\mathbf{DA}(S;\Lambda)$. The latter "see" the higher Chow groups integrally (but also other things like oriented Chow groups).

[9] These hypothesis are satisfied when S is excellent.

Remark 2.27. From another perspective, Theorem 2.24 is encouraging. Indeed, it is also well-known that integrality in Chow groups is chaotic in general. For instance, there are famous counterexamples (the first ones by Atiyah–Hirzebruch [4, Theorem 6.5] and Kollár [21, page 134–135]) to the integral Hodge and Tate conjectures. Imposing étale descent forces a better organization in the integral structure of higher Chow groups. As a matter of fact, it has been shown recently by Rosenschon–Srinivas [26] that the Hodge and Tate conjectures can be "corrected" integrally by replacing the Chow groups by their étale version.[10] See also Remark 5.7 below for another (but related) reason to be happy about Theorem 2.24.

3. Operations on motivic sheaves

In this section, we review the functorialities of the categories of motivic sheaves. As for the classical "cohomological coefficients" (in the sense of Grothendieck), one has for motivic sheaves the Grothendieck six operations formalism and Verdier's duality. One also has the nearby cycles formalism, but this will not be discussed here (see [6, Chapitre 4] and [9]).

3.1. Operations associated to morphisms of schemes. In this subsection, we will recall the construction of the formalism of the four operations f^*, f_*, $f_!$ and $f^!$, associated to a morphism of schemes f, in the context of motivic sheaves.

3.1.1. Ordinary inverse and direct images. Let $f : T \to S$ be a morphism of schemes. Then f induces a pair of adjoint functors:

$$f^* : \mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda) \rightleftarrows \mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/T;\Lambda) : f_*. \tag{3.1}$$

The functor f_* is easy to understand; given an étale sheaf $\mathcal{G}$ over Sm/T, one has $f_*\mathcal{G}(U) := \mathcal{G}(T \times_S U)$ for all $U \in \mathrm{Sm}/S$. The functor f^* is characterized by its property of commuting with arbitrary colimits and by the formula

$$f^*\Lambda_{\text{ét}}(U) \simeq \Lambda_{\text{ét}}(T \times_S U) \tag{3.2}$$

for all $U \in \mathrm{Sm}/S$.

The adjunction (3.1) can be derived yielding an adjunction on the level of effective motivic sheaves

$$\mathrm{L}f^* : \mathbf{DA}^{\mathrm{eff},\,\text{ét}}(S;\Lambda) \rightleftarrows \mathbf{DA}^{\mathrm{eff},\,\text{ét}}(T;\Lambda) : \mathrm{R}f_*. \tag{3.3}$$

It can also be extended to L-spectra and then derived yielding an adjunction on the level of motivic sheaves

$$\mathrm{L}f^* : \mathbf{DA}^{\text{ét}}(S;\Lambda) \rightleftarrows \mathbf{DA}^{\text{ét}}(T;\Lambda) : \mathrm{R}f_*. \tag{3.4}$$

These functors are triangulated.

[10] For a smooth algebraic variety X over a field k, the étale Chow groups of X can be defined by the formula (see Definition 2.13)

$$\mathrm{CH}^n_{\text{ét}}(X) := H^{2n}_{\mathcal{L}}(X;\mathbb{Z}(n)) = \hom_{\mathbf{DA}^{\text{ét}}(k;\mathbb{Z})}(\mathrm{M}(X),\mathbb{Z}(n)[2n])$$

(or, equivalently, using $\mathbf{DM}^{\text{ét}}(k;\mathbb{Z})$ instead of $\mathbf{DA}^{\text{ét}}(k;\mathbb{Z})$). When $k = \mathbb{C}$, Rosenschon and Srinivas construct in [26] a cycle map $\mathrm{CH}^n_{\text{ét}}(X) \to H^{2n}(X(\mathbb{C}),\mathbb{Z})$ and show that if the Hodge conjecture holds for the rational Chow groups (i.e., for $\mathrm{CH}^n_{\mathbb{Q}}(X) := \mathrm{CH}^n(X) \otimes \mathbb{Q}$) then it also holds integrally for the étale Chow groups. They also show a similar statement for the Tate conjecture.

Remark 3.1. The formula (3.2) still holds for the left derived functors $\mathrm{L}f^*$ in (3.3) and (3.4). In words, $\mathrm{L}f^*$ takes the homological motive of an S-scheme U to the homological motive of the T-scheme $T \times_S U$ (in the effective and non-effective settings).

Lemma 3.2. *Assume that f is smooth. Then, the functor f^* admits a left adjoint*

$$f_\sharp : \mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/T; \Lambda) \to \mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S; \Lambda).$$

If $V \in \mathrm{Sm}/T$, then $f_\sharp \Lambda_{\text{ét}}(V/T) = \Lambda_{\text{ét}}(V/S)$. Moreover, $f_\sharp$ can be left derived yielding left adjoints to $\mathrm{L}f^*$ on the level of motivic sheaves:

$$\mathrm{L}f_\sharp : \mathbf{DA}^{\mathrm{eff},\text{ét}}(T; \Lambda) \to \mathbf{DA}^{\mathrm{eff},\text{ét}}(S; \Lambda) \text{ and } \mathrm{L}f_\sharp : \mathbf{DA}^{\text{ét}}(T; \Lambda) \to \mathbf{DA}^{\text{ét}}(S; \Lambda).$$

Remark 3.3. The existence of a left adjoint to f^*, when f is smooth, is part of the formalism of the six operations of Grothendieck. However, in the classical setting, this property is one of the deepest, whereas for motivic sheaves one has it for free!

3.1.2. A list of axioms. From now on, we will drop the "L" and "R" when dealing with the operations $\mathrm{L}f^*$, $\mathrm{L}f_\sharp$ and $\mathrm{R}f_*$.

Let SCH be the category of all schemes and $\mathfrak{TR}$ the 2-category of triangulated categories. Then, the 2-functor

$$\begin{array}{rccc} \mathbf{DA}^{\text{ét}}(-;\Lambda) \ : & \mathrm{SCH} & \to & \mathfrak{TR} \\ & f & \mapsto & f^* \end{array}$$

satisfies the following list of axioms. (Only one of these axioms fails to hold for $\mathbf{DA}^{\mathrm{eff},\text{ét}}(-,\Lambda)$, namely the sixth!)

1. $\mathbf{DA}^{\text{ét}}(\emptyset;\Lambda)$ is equivalent to the zero triangulated category.
2. For every morphism of schemes $f : T \to S$, the functor $f^* : \mathbf{DA}^{\text{ét}}(S;\Lambda) \to \mathbf{DA}^{\text{ét}}(T;\Lambda)$ admits a right adjoint f_*.
3. For every *smooth* morphism $f : T \to S$, the functor $f^* : \mathbf{DA}^{\text{ét}}(S;\Lambda) \to \mathbf{DA}^{\text{ét}}(T;\Lambda)$ admits a left adjoint $f_\sharp$. Moreover, given a cartesian square

$$\begin{array}{ccc} T' & \xrightarrow{g'} & T \\ \downarrow{\scriptstyle f'} & & \downarrow{\scriptstyle f} \\ S' & \xrightarrow{g} & S, \end{array}$$

   the natural exchange morphism $f'_\sharp \circ g'^* \to g^* \circ f_\sharp$ is an isomorphism.
4. For every closed immersion i with complementary open immersion j, the pair (i^*, j^*) is conservative (i.e., if a motive M satisfies $i^*M \simeq 0$ and $j^*M \simeq 0$, then $M \simeq 0$). Moreover, the counit of the adjunction $i^* \circ i_* \to \mathrm{id}$ is an isomorphism.
5. If $p : V \to S$ is the projection of a vector bundle, then the unit of the adjunction $\mathrm{id} \to p_* p^*$ is an isomorphism.
6. If $f : T \to S$ is smooth and $s : S \to T$ is a section of f (i.e., $f \circ s = \mathrm{id}_S$), then the functor $f_\sharp \circ s_*$ is an autoequivalence of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$.

We will call such a 2-functor an *extended stable homotopical* 2*-functor*.

Remark 3.4. Except the fourth axiom, all these axioms follow readily from the construction. For instance, the fifth axiom is a consequence of the $\mathbb{A}^1$-localization and the sixth axiom follows from inverting the Lefschetz motive (for the tensor product).

The fourth axiom (aka., the locality axiom) is due to Morel–Voevodsky [25, Theorem 2.21 of §3.2]. (In loc. cit., only the non-Abelian setting is considered but their proof can be adapted to the additive setting without much difficulties; see [6, §4.5.3].) It is the proof of this axiom that dictates some of the choices that were made by Morel–Voevodsky (and repeated in §2) such as considering sheaves on *smooth* S-schemes instead of sheaves on larger categories of S-schemes.

Remark 3.5. That these axioms suffices to derive the full formalism of the four operations is due to Voevodsky (unpublished). The details of the verifications were carried on in [5, Chapitre 1].

For later use, we make the following definition.

Definition 3.6. Given an $\mathcal{O}_S$-module $\mathcal{M}$ on a scheme S, we set $\mathrm{Th}(\mathcal{M}) = p_\sharp \circ s_*$ where $p : \mathrm{V}(\mathcal{M}) \to S$ is the projection of the associated vector bundle and s is its zero section. By the sixth axiom, $\mathrm{Th}(\mathcal{M})$ is an autoequivalence of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$, called the *Thom equivalence*. Its inverse is denoted by $\mathrm{Th}^{-1}(\mathcal{M})$.

Remark 3.7. It is customary to denote $\mathrm{Th}(\mathcal{O}_S^{\oplus r})(-)[-2r]$ by $(-)(r)$ and to call it the r-th *Tate twist* (extended to negative integers in the usual way).

If $\mathcal{M}$ has constant rank r, it can be shown that $\mathrm{Th}(\mathcal{M})[-2r]$ is canonically equivalent to $(-)(r)$ (see [9, Remarque 11.3]). This is a special property of $\mathbf{DA}^{\text{ét}}(-;\Lambda)$ called *orientation*.

3.1.3. The proper base change theorem. One of the most surprising fact here is that the axioms of §3.1.2 imply quite formally the so-called *proper base change theorem*. (All the axioms are used in the proof of this theorem; as a matter of fact, this theorem fails for the categories $\mathbf{DA}^{\text{eff},\text{ét}}(-;\Lambda)$.)

Theorem 3.8. *Given a cartesian square*

$$\begin{array}{ccc} Y' & \xrightarrow{g'} & Y \\ \downarrow{\scriptstyle f'} & & \downarrow{\scriptstyle f} \\ X' & \xrightarrow{g} & X \end{array}$$

with f proper, the exchange morphism $g^* \circ f_*(\mathcal{M}) \to f'_* \circ g'^*(\mathcal{M})$ is an isomorphism for every motivic sheaf $\mathcal{M} \in \mathbf{DA}^{\text{ét}}(Y;\Lambda)$.

To prove Theorem 3.8, it is enough to treat the case where f is the projection $p_n : \mathbb{P}^n_X \to X$. (This reduction is easy and classical; it appears for example in [3, Exposé XII].) To treat the case of p_n, one needs a completely different approach than the one used in [3, Exposés XII et XIII]. Here is a sketch of the proof following [5, Chapitre 1]:

Proof. In contrast with the étale formalism, here we define the extraordinary push-forward functors $f_!$ before knowing the validity of the proper base change theorem. That this can be done relies on the (easy) existence of a left adjoint $h_\sharp$ to h^* when h is smooth. Indeed,

assuming that f is smoothable, i.e., can be written as $f = h \circ i$ with h smooth and i a closed immersion, one sets

$$f_! := h_\sharp \circ \mathrm{Th}^{-1}(\Omega_h) \circ i_* \qquad \text{(and dually} \quad f^! := i^! \circ \mathrm{Th}(\Omega_h) \circ h^*).$$

A big deal of effort in [5, Chapitre 1] is devoted to showing that these definitions are independent (up to natural isomorphisms) of the choice of the factorization $f = h \circ i$ and that there are coherent choices of isomorphisms $(f \circ f')_! \simeq f_! \circ f'_!$, for composable smoothable morphisms, etc. Assuming this is granted, it is then easy to explain the strategy of the proof of Theorem 3.8.

From the third axiom in §3.1.2 and the definition of the extraordinary direct image, it is quite easy to see that one has an exchange isomorphism $g^* \circ f_! \simeq f'_! \circ g'^*$ (without any condition on f beside being smoothable).

On the other hand, one can construct a natural transformation $\alpha_f : f_! \to f_*$ (which is reminiscent to the obvious morphism from cohomology with support to ordinary cohomology). It is defined as follows. Consider the commutative diagram

$$\begin{array}{ccc} Y & & \\ & \searrow^{\Delta} & \\ & Y \times_X Y & \xrightarrow[pr_1]{} Y \\ & \downarrow{\scriptstyle pr_2} & \downarrow{\scriptstyle f} \\ Y & \xrightarrow[f]{} & X. \end{array}$$

From the square, one gets a natural exchange morphism $f_! \circ pr_{1*} \to f_* \circ pr_{2!}$ (deduced by adjunction from the exchange isomorphism given by the third axiom of §3.1.2). Applying this to $\Delta_* = \Delta_!$ and using the identifications $pr_{1*} \circ \Delta_* = \mathrm{id}$ and $pr_{2!} \circ \Delta_! = \mathrm{id}$, one gets the promised natural transformation.

This is said, we are left to showing that $p_{n!} \to p_{n*}$ is an isomorphism for $p_n : \mathbb{P}^n \times X \to X$. This is done by induction on n using a rather tricky argument. The point is to realize that it suffices to show that

$$p_{n!} \circ p_n^* \to p_{n*} \circ p_n^* \qquad \text{and} \qquad p_{n!} \circ p_n^! \to p_{n*} \circ p_n^!$$

are both isomorphisms. Indeed, assuming this, one can then define two maps $p_n^* \to p_n^!$ by the compositions of

$$p_{n*} \xrightarrow{\eta} p_{n*} \circ p_n^* \circ p_{n*} \simeq p_{n!} \circ p_n^* \circ p_{n*} \xrightarrow{\delta} p_{n!}$$

$$p_{n*} \xrightarrow{\eta} p_{n*} \circ p_n^! \circ p_{n!} \simeq p_{n!} \circ p_n^! \circ p_{n!} \xrightarrow{\delta} p_{n!}$$

A direct computation shows that these morphisms give respectively left and right inverses to the canonical morphism $p_{n!} \to p_{n*}$. See [5, §1.7.2] for the complete proof. □

3.1.4. Extraordinary direct and inverse images. As said in the sketch of the proof of Theorem 3.8, one has, for f smoothable (and, in particular, for f quasi-projective), two extraordinary operations $f^!$ and $f_!$.

Once the proper base change theorem is established, it is possible to extend the extraordinary operations to the case where f is of finite presentation (but not necessarily smoothable)

following the receipt of [3, Exposé XVII]. Indeed, by Nagata's compactification, we may factor $f = \overline{f} \circ j$ where $\overline{f}$ is proper and j is an open immersion. Then, one sets $f_! := \overline{f}_* \circ j_\sharp$. The proper base change theorem implies that this is independent of the choice of the compactification.[11]

In any case, one has an adjunction $(f_!, f^!)$ for every finite type separated morphism. (The existence of $f^!$ is local over the source of f and hence, one may reduce to the case where f is quasi-projective.)

Theorem 3.9. *For every cartesian square*

$$\begin{array}{ccc} Y' & \xrightarrow{g'} & Y \\ \downarrow{\scriptstyle f'} & & \downarrow{\scriptstyle f} \\ X' & \xrightarrow{g} & X \end{array}$$

with f of finite type and g arbitrary, one has exchange isomorphisms

$$g^* f_! \simeq f'_! g'^* \qquad \text{and} \qquad f^! g_* \simeq g'_* f'^!.$$

3.2. Closed monoidal structures and Verdier duality. As constructed in §2.3, the category $\mathbf{DA}^{\text{ét}}(S;\Lambda)$, possesses a monoidal structure. However, as it is the case for the smash product of spectra in topology, it is not possible to define the tensor product directly on the category $\mathrm{Spt}_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$ of L-spectra. Different ways around this difficulty have been developed in topology. One of these ways is via the notion of *symmetric spectra* [18] that had been greatly generalized in [17].

More specifically, one considers the Abelian category $\mathrm{Spt}^\Sigma_L(\mathrm{Shv}_{\text{ét}}(S;\Lambda))$ of symmetric L-spectra of étale sheaves on Sm/S. A symmetric L-spectrum is an L-spectrum $\mathcal{E}$ endowed with an action of the n-th symmetric group Σ_n on its n-th level $\mathcal{E}_n$ and such that the assembly maps are equivariant in an appropriate sense.

The point is that the extra symmetry that symmetric L-spectra possess permits to define a symmetric and associative tensor product on $\mathrm{Spt}^\Sigma_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))$. The latter induces a tensor product on $\mathbf{D}(\mathrm{Spt}^\Sigma_L(\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)))$ and its localization with respect to its triangulated subcategory $\mathcal{T}^\Sigma_{\mathbb{A}^1-\mathrm{st}}$ defined similarly as in §2.3. Finally, one can show that this localization yields an equivalent category to $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ inducing a monoidal structure on the latter.

Unfortunately, the details of this story are quite technical and boring. We refer the interested reader to [6, Chapitre 4] for a complete (and self-contained) account (using however the language of model categories).

Theorem 3.10. *The categories $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ are symmetric monoidal and closed (i.e., $A \otimes -$ admits a right adjoint $\underline{Hom}(A,-)$ for every S-motive A). The operations f^* are monoidal functors. One also has the usual formulas*

$$f_!(-) \otimes - \simeq f_!(- \otimes f^*(-)), \quad f^! \underline{Hom}(-,-) \simeq \underline{Hom}(f^*(-), f^!(-)),$$

$$f_* \underline{Hom}(f^*(-),-) \simeq \underline{Hom}(-, f_*(-)), \quad \underline{Hom}(f_!(-),-) \simeq f_* \underline{Hom}(-, f^!(-)), \text{ etc.}$$

[11] It is worth noting here that checking that $\overline{f}_* \circ j_\sharp$ is independent of the factorization $f = \overline{f} \circ j$ is easier than checking that $h_\sharp \circ \mathrm{Th}(\Omega_h) \circ i_*$ is independent of the factorization $f = h \circ i$. The reason for this is that "the category of compactifications" is filtered whereas the "category of smoothifications" is not.

Finally, assuming that S is of finite type over a characteristic zero field k and denoting π_S to projection to the point, there is a dualizable objet in $\mathbf{DA}^{\text{ét}}_{\text{ct}}(S;\Lambda)$ given by $\pi_S^!\Lambda(0)$.

Another important result to mention here is:

Theorem 3.11. *If X is a proper and smooth S-scheme of pure relative dimension d, then $\mathrm{M}(X)$ admits a strong dual given by $\mathrm{M}(X)(-d)[-2d]$.*

Proof. This follows from Theorem 3.10 using that

$$\mathrm{M}(X)\simeq(\pi_X)_!(\pi_X)^!\Lambda_S(0)\qquad\text{and}\qquad \mathrm{M}(X)(-d)[-2d]\simeq(\pi_X)_*(\pi_X)^*\Lambda_S(0)$$

where $\pi_X : X\to S$ is the structural morphism. □

4. Motives over a base field

The formalism of Grothendieck's six operations is a powerful tool for reducing questions about general sheaves to questions about lisse sheaves and, ultimately, to questions about (germs of) sheaves on generic points of varieties. For this formalism to be of any use in the context of motivic sheaves, one needs informations about motives over fields.

In this section we list some of what is known concerning motives over a field; everything here is essentially due to Voevodsky. When dealing with Voevodsky's motives, we mostly work over a base field k except for the construction §4.1.1 and the comparison theorem §4.1.2 where this restriction is irrelevant. The use of the étale topology results in inverting automatically the exponent-characteristic of k.[12] Therefore, there is no need in assuming k perfect in quoting [24, 29].

4.1. Voevodsky's motives. Many theorems about motives over a field and morphisms between them are obtained by using a slightly more complicated construction than the one explained in §2. The extra complication is the requirement of having transfers and is the key for many concrete computations.

4.1.1. The construction. The construction of Voevodsky's category $\mathbf{DM}^{\text{ét}}(k;\Lambda)$ follows exactly the same pattern as the construction given in §2 with only one difference: one uses the Abelian category of étale *sheaves with transfers* instead of the Abelian category of ordinary étale sheaves. To expand on this, we need some notation.

Let S be a base scheme that we assume to be Noetherian. In [29, Chapter 2], a category of finite correspondences SmCor/S was constructed. This is an additive category whose objects are smooth S-schemes. Given two smooth S-schemes U and V, the group of morphisms from U to V in SmCor/S is denoted by $\mathrm{Cor}_S(U,V)$. When S is regular, this group is freely generated by integral and closed subschemes $Z\subset U\times_S V$ such that the projection

[12] This is well-known and easy. Indeed, if $k=\mathbb{F}_p$, then the Artin–Schreier exact sequence of étale sheaves on $\mathrm{Sm}/\mathbb{F}_p$:

$$0\to\mathbb{Z}/p\mathbb{Z}\to\mathcal{O}\overset{(-)^p}{\to}\mathcal{O}\to 0,$$

and the fact that $\mathcal{O}$ is $\mathbb{A}^1$-contractible, show that the constant étale sheaf $\mathbb{Z}/p\mathbb{Z}$ is also $\mathbb{A}^1$-contractible. From this, it is easy to deduce that multiplication by p is invertible in $\mathbf{DA}^{\text{ét}}(\mathbb{F}_p;\Lambda)$ and more generally in $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ for every $\mathbb{F}_p$-scheme S. The same holds true for $\mathbf{DM}^{\text{ét}}(\mathbb{F}_p;\Lambda)$ and $\mathbf{DM}^{\text{ét}}(S;\Lambda)$.

$Z \to U$ is finite and surjective over a connected component of U. Moreover, the composition of finite correspondences is then given by the usual formula involving Serre's multiplicities.

Definition 4.1. A *presheaf with transfers* on Sm/S is a contravariant additive functor from SmCor/S to the category of Λ-modules. An *étale sheaf with transfers* is a presheaf with transfers Sm/S which is, after forgetting transfers, a sheaf for the étale topology. Étale sheaves with transfers form an Abelian category that we denote by $\mathrm{Str}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$.

Example 4.2. For a smooth S-scheme X, we denote by $\Lambda_{\mathrm{tr}}(X)$ the presheaf with transfers on Sm/S represented by X, i.e., given by $\Lambda_{\mathrm{tr}}(X)(U) = \mathrm{Cor}_S(U,X)\otimes_{\mathbb{Z}}\Lambda$ for all $U \in \mathrm{Sm}/S$. In fact, $\Lambda_{\mathrm{tr}}(X)$ is an étale sheaf with transfers on Sm/S. After forgetting transfers, one has an inclusion of étale sheaves $\Lambda_{\text{ét}}(X) \subset \Lambda_{\mathrm{tr}}(X)$.

As said before, replacing everywhere "$\mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$" by "$\mathrm{Str}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)$" in §2 yields Voevodsky's triangulated categories of S-motives. More precisely, one obtains two versions.

- The category of *effective Voevodsky S-motives* given by
$$\mathbf{DM}^{\mathrm{eff},\,\text{ét}}(S;\Lambda) := \mathbf{D}(\mathrm{Str}_{\text{ét}}(\mathrm{Sm}/S;\Lambda))/\mathcal{T}^{\mathrm{tr}}_{\mathbb{A}^1}$$
where $\mathcal{T}^{\mathrm{tr}}_{\mathbb{A}^1}$ is defined similarly as $\mathcal{T}_{\mathbb{A}^1}$ in §2.2.2 (writing "Λ_{tr}" instead of "$\Lambda_{\text{ét}}$" in (2.3)).
- The category of (non-effective) *Voevodsky S-motives* given by
$$\mathbf{DM}^{\text{ét}}(S;\Lambda) := \mathbf{D}(\mathrm{Spt}_{L_{\mathrm{tr}}}(\mathrm{Str}_{\text{ét}}(\mathrm{Sm}/S;\Lambda)))/\mathcal{T}^{\mathrm{tr}}_{\mathbb{A}^1-\mathrm{st}}$$
where $L_{\mathrm{tr}} = \Lambda_{\mathrm{tr}}(\mathbb{P}^1_S,\infty_S)$ and $\mathcal{T}^{\mathrm{tr}}_{\mathbb{A}^1-\mathrm{st}}$ is defined similarly as $\mathcal{T}_{\mathbb{A}^1-\mathrm{st}}$ in §2.3 (writing "Λ_{tr}" and "L_{tr}" instead of "$\Lambda_{\text{ét}}$" and "L" in (2.5) and (2.6)).

Remark 4.3. Strictly speaking, Voevodsky [24] considered categories
$$\mathbf{DM}^{\mathrm{eff},\,\text{ét}}_{-}(S;\Lambda) \qquad \text{and} \qquad \mathbf{DM}^{\text{ét}}_{\mathrm{gm}}(S;\Lambda)$$
for S the spectrum of a perfect field (with finite cohomological dimension). The category $\mathbf{DM}^{\mathrm{eff},\,\text{ét}}_{-}(S;\Lambda)$ is the triangulated subcategory of $\mathbf{DM}^{\mathrm{eff},\,\text{ét}}(S;\Lambda)$ consisting of complexes that are bounded on the right. The category $\mathbf{DM}^{\mathrm{eff},\,\text{ét}}_{\mathrm{gm}}(S;\Lambda)$ is the triangulated subcategory of $\mathbf{DM}^{\mathrm{eff},\,\text{ét}}(S;\Lambda)$ generated by $\Lambda_{\mathrm{tr}}(X)$ for $X \in \mathrm{Sm}/S$ of finite type. Finally, $\mathbf{DM}^{\text{ét}}_{\mathrm{gm}}(S;\Lambda)$ is obtained from $\mathbf{DM}^{\mathrm{eff},\,\text{ét}}_{\mathrm{gm}}(S;\Lambda)$ by formally inverting tensoring by the Lefschetz motive L_{tr} (i.e., using the naive construction as in §2.2.3); it is also the triangulated subcategory of $\mathbf{DM}^{\text{ét}}(S;\Lambda)$ generated by S-motives of finite type smooth S-schemes and their negative Tate twists.

4.1.2. The comparison theorem. There is a pair of adjoint functors:
$$\mathrm{a}_{\mathrm{tr}} : \mathrm{Shv}_{\text{ét}}(\mathrm{Sm}/S;\Lambda) \rightleftarrows \mathrm{Str}_{\text{ét}}(\mathrm{Sm}/S;\Lambda) : \mathrm{o}_{\mathrm{tr}}. \tag{4.1}$$
The functor o_{tr} is a forgetful functor: it takes an étale sheaf with transfers to its underlying étale sheaf. The functor a_{tr} is characterized by its property of commuting with arbitrary colimits and by the formula
$$\mathrm{a}_{\mathrm{tr}}(\Lambda_{\text{ét}}(U)) \simeq \Lambda_{\mathrm{tr}}(U)$$

for all $U \in \mathrm{Sm}/S$. The adjunction (4.1) can be derived yielding an adjunction on the level of effective S-motives:

$$\mathrm{La}_{\mathrm{tr}} : \mathbf{DA}^{\mathrm{eff},\text{ét}}(S;\Lambda) \rightleftarrows \mathbf{DM}^{\mathrm{eff},\text{ét}}(S;\Lambda) : \mathrm{Ro}_{\mathrm{tr}}. \tag{4.2}$$

It can also be extended to spectra and then derived yielding an adjunction on the level of (non-effective) S-motives:

$$\mathrm{La}_{\mathrm{tr}} : \mathbf{DA}^{\text{ét}}(S;\Lambda) \rightleftarrows \mathbf{DM}^{\text{ét}}(S;\Lambda) : \mathrm{Ro}_{\mathrm{tr}}. \tag{4.3}$$

Theorem 4.4. *If S is normal (and some technical assumptions are satisfied), the functors in* (4.3) *are equivalences of categories.*

Proof. When Λ is a $\mathbb{Q}$-algebra, Theorem 4.4 was proved by Morel, for S the spectrum of a field, and was generalized later by Cisinski–Déglise.[13] In [9, Annexe B], we simplified the proof of Cisinski–Déglise and extended their result to more general coefficient rings using Theorem 2.24. □

Remark 4.5. If the normal scheme S has characteristic zero and if Λ is a $\mathbb{Q}$-algebra, then the functors in (4.2) are also known to be equivalences of categories by [8, Théorème B.1]. (This is indeed a stronger statement!)

Remark 4.6. It is unknown if Theorem 4.4 holds for general base schemes (e.g., reducible). This is because the theory of finite correspondences over non-normal schemes is quite complicated. A related (and probably equivalent) open question is to know if the 2-functor $\mathbf{DM}^{\text{ét}}(-;\Lambda)$ satisfies the localization axiom (i.e., the fourth axiom in §3.1.2). In fact, this is the only missing property that prevents one to promote $\mathbf{DM}^{\text{ét}}(-;\Lambda)$ into an extended stable homotopical 2-functor. But, in our opinion, these questions have minor impact for the following reasons:

1. A stable homotopical 2-functor H, say over quasi-projective S-schemes with S regular, is essentially determined by its values on smooth S-schemes. Indeed, if X is a quasi-projective S-scheme, one can choose an embedding $i : X \hookrightarrow Y$ with Y a smooth S-scheme. Then, thanks to the locality axiom, $\mathsf{H}(X)$ can be described as the subcategory of $\mathsf{H}(Y)$ consisting of those objects supported on X, i.e., those objects that vanish when pulled back along the complement of i. Therefore, Theorem 4.4 tells that $\mathbf{DA}^{\text{ét}}(-;\Lambda)$ is, up to an equivalence, the unique stable homotopical 2-functor that extends Voevodsky's category of motives over regular bases.
2. A stressed before, the construction of $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ is really simpler than $\mathbf{DM}^{\text{ét}}(S;\Lambda)$. Moreover, the advantage of using transfers in defining motivic sheaves disappears when the base scheme S has dimension ≥ 1. Indeed, all the results that will be explained in §4.2 require the base to be a field.

Remark 4.7. The reader might wonder which construction of categories of motives is better. The answer is that both $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ and $\mathbf{DM}^{\text{ét}}(S;\Lambda)$ have their advantages and disadvantages.

[13] In fact, Morel and Cisinski–Déglise prove a stronger result where the étale topology is replaced by the Nisnevich topology. Indeed, they prove that $\mathbf{DM}(k;\mathbb{Q})$ is equivalent to a direct summand $\mathbf{DA}(S;\mathbb{Q})_+$ of $\mathbf{DA}(S;\mathbb{Q})$ whose complement vanishes when étale descent is imposed.

- $\mathbf{DA}^{\text{ét}}(S;\Lambda)$ is simpler[14] and is the correct category of motivic sheaves for any S. On the other hand, one does not have a concrete model for the $\mathbb{A}^1$-localization functor when S is the spectrum of a field.
- Over a field, one has the theory of homotopy invariant presheaves with transfers which is a powerful tool to study the category $\mathbf{DM}^{\text{ét}}(k;\Lambda)$. However, over a curve and higher dimensional bases, this advantage disappears as the theory of homotopy invariant presheaves with transfers breaks down completely. Moreover, it is unclear if $\mathbf{DM}^{\text{ét}}(S;\Lambda)$ is the correct category when S is not normal.

4.2. Homotopy invariant presheaves with transfers. Let $\mathcal{F}$ be a presheaf on Sm/k. We say that $\mathcal{F}$ is homotopy invariant if $\mathcal{F}(U) \to \mathcal{F}(\mathbb{A}^1 \times U)$ is an isomorphism for all $U \in \mathrm{Sm}/k$. For simplicity, we assume that the exponent characteristic of k is invertible in Λ. A basic theorem of Voevodsky [29, Chapter 3] states the following.[15]

Theorem 4.8. *Let $\mathcal{F}$ be a homotopy invariant presheaf with transfers on Sm/k (with values in Λ-modules). Then $\mathrm{a}_{\text{ét}}(\mathcal{F})$, the étale sheaf associated to $\mathcal{F}$, is an $\mathbb{A}^1$-local object of $\mathbf{D}(\mathrm{Str}_{\text{ét}}(\mathrm{Sm}/k;\Lambda))$. More concretely,*

$$H^i_{\text{ét}}(U;\mathrm{a}_{\text{ét}}(\mathcal{F})) \to H^i_{\text{ét}}(\mathbb{A}^1 \times U;\mathrm{a}_{\text{ét}}(\mathcal{F}))$$

is an isomorphism for all $i \in \mathbb{N}$ and $U \in \mathrm{Sm}/k$.

Remark 4.9. All the hypothesis in this theorem are necessary. For instance, the theorem is wrong for presheaves without transfers. It is also wrong if k is replaced by a curve or a higher dimensional base.

One reason why this theorem is important is that it enables one to construct very easily the $\mathbb{A}^1$-localization of any complex of étale sheaves with transfers. To explain this, we need some notation.

Definition 4.10. For $n \in \mathbb{N}$, set

$$\Delta^n = \mathrm{Spec}(\mathbb{Z}[t_0,\dots,t_n]/(t_0+\dots+t_n-1)).$$

These schemes form a cosimplicial scheme $\Delta^\bullet$. Given a complex of presheaves with transfers $\mathcal{K}_\bullet$, we define $\mathrm{Sing}^{\mathbb{A}^1}(\mathcal{K})$ to be the total complex of the double complex $\underline{\hom}(\Delta^\bullet;\mathcal{K}_\bullet)$. (Recall that $\underline{\hom}(\Delta^n,\mathcal{F})(U) = \mathcal{F}(\Delta^n \times U)$ for any presheaf $\mathcal{F}$ and any $U \in \mathrm{Sm}/k$.) The functor $\mathrm{Sing}^{\mathbb{A}^1}$ is called the *Suslin–Voevodsky construction*.

Corollary 4.11. *Let $\mathcal{K}$ be a complex of étale sheaves with transfers. Then $\mathrm{Loc}_{\mathbb{A}^1}(\mathcal{K})$ is given by the Suslin–Voevodsky construction $\mathrm{Sing}^{\mathbb{A}^1}(\mathcal{K})$.*

Proof. It follows formally from the construction that the canonical map $\mathcal{K} \to \mathrm{Sing}^{\mathbb{A}^1}(\mathcal{K})$ is an isomorphism in $\mathbf{DM}^{\text{eff},\text{ét}}(k;\Lambda)$. It remains to show that $\mathrm{Sing}^{\mathbb{A}^1}(\mathcal{K})$ is $\mathbb{A}^1$-local. But again, it follows formally form the construction that the homology presheaves of the complex

[14] For instance, it is very convenient not to have to worry about transfers when discussing realizations!

[15] In loc. cit., the result is established for the Nisnevich topology. However, it is an exercise to deduce the result for the étale topology using Suslin's rigidity theorem [24, Theorem 7.20] and the homotopy invariance of étale cohomology with values in $\Lambda/n\Lambda$ for n prime to the exponent-characteristic of k.

$\mathrm{Sing}^{\mathbb{A}^1}(\mathcal{K})$ are homotopy invariant (and admits transfers). Applying Theorem 4.8 to these and using a spectral sequence, one deduces that the maps

$$H^i_{\text{ét}}(U, \mathrm{Sing}^{\mathbb{A}^1}(\mathcal{K})) \to H^i_{\text{ét}}(\mathbb{A}^1 \times U; \mathrm{Sing}^{\mathbb{A}^1}(\mathcal{K}))$$

are isomorphisms. □

4.3. Application: morphisms between motivic sheaves. A basic question about motivic sheaves is the following.

Question. *Given two motivic sheaves $\mathcal{M}$ and $\mathcal{N}$ over a base scheme S, how to compute the group* $\hom_{\mathbf{DA}^{\text{ét}}(S;\Lambda)}(\mathcal{M}, \mathcal{N})$*?*

As said before, in theory, the formalism of the six operations reduces the above question to computing some groups of morphisms (usually many) in $\mathbf{DA}^{\text{ét}}(k;\Lambda) \simeq \mathbf{DM}^{\text{ét}}(k;\Lambda)$ (for various fields k). Therefore, it is important to have a solution of this question when the base is a field.

Let k be a field and assume that the exponent-characteristic of k is invertible in Λ. We will explain the solution of the above question in the case where $\mathcal{M}$ and $\mathcal{N}$ are the motives of smooth k-varieties X and Y respectively. Hence, we concentrate on the groups

$$\hom_{\mathbf{DM}^{\text{ét}}(k;\Lambda)}(\mathrm{M}(X); \mathrm{M}(Y)[n]).$$

For simplicity, we assume that Y is proper of pure dimension d_Y. By Theorem 3.11, we know that $\mathrm{M}(Y)$ has a strong dual given by $\mathrm{M}(Y)^\vee = \mathrm{M}(Y)(-d_Y)[-2d_Y]$. Hence, we are left to compute the étale motivic cohomology groups

$$H^p_{\mathcal{L}}(Z; \Lambda(q)) := \hom_{\mathbf{DM}^{\text{ét}}(k;\Lambda)}(\mathrm{M}(Z); \Lambda(q)[p])$$

(for $Z = X \times_k Y$ and $q = d_Y$ and $p = n + 2d_Y$). The answer is as follows.

Theorem 4.12. *Let X be a smooth k-variety. Then there is a canonical isomorphism*

$$\hom_{\mathbf{DM}^{\text{ét}}(k;\Lambda)}(\mathrm{M}(X); \Lambda(q)[p]) \simeq H^{p-2q}_{\text{ét}}(X; \mathrm{Sing}^{\mathbb{A}^1}\Lambda_{\mathrm{tr}}(\mathbb{P}^1_k, \infty_k)^{\wedge q}) \tag{4.4}$$

where the right-hand side is the étale hypercohomology of X with values in the complex of étale sheaves $\mathrm{Sing}^{\mathbb{A}^1}\Lambda_{\mathrm{tr}}(\mathbb{P}^1_k, \infty_k)^{\wedge q}$.

Remark 4.13. Theorem 4.12 is an immediate consequence of Theorem 4.8. Another theorem of Voevodsky asserts that the complex $\mathrm{Sing}^{\mathbb{A}^1}\Lambda_{\mathrm{tr}}(\mathbb{P}^1_k, \infty_k)^{\wedge q}$ satisfies Nisnevich descent. Therefore, if Λ is a $\mathbb{Q}$-algebra (or when "étale" is replaced by "Nisnevich"), the right hand side in (4.4) is simply the cohomology of a concrete complex of cycles, namely $\mathrm{Cor}_k(\Delta^\bullet \times X, (\mathbb{P}^1_k, \infty_k)^{\wedge q}) \otimes \Lambda$.

5. Conjectures

There are many outstanding conjectures concerning motives and algebraic cycles. Some of these seem desperately out of reach such as the Hodge and Tate conjectures (that already made an appearance in Remark 2.27) or the Grothendieck and Kontsevich–Zagier conjectures on periods.

In this section we will discuss two other conjectures that, in comparison with the previous ones, seem more approachable. These two conjectures (as well as the previous ones) predict relations between algebro-geometric objects and transcendental objects, and each one of these conjectures fills some part of the gap between the two half-bridges discussed in the Introduction.

5.1. The conservativity conjecture. Let k be a field of characteristic zero and let $\sigma : k \hookrightarrow \mathbb{C}$ be a complex embedding. Given a finite type k-scheme X, denote by X_{an} the set $X(\mathbb{C})$ endowed with its analytic topology. One has a Betti realization functor [7]

$$\mathrm{B}^*_X : \mathbf{DA}^{\text{ét}}(X;\Lambda) \to \mathbf{D}(X_{\mathrm{an}};\Lambda) \tag{5.1}$$

where $\mathbf{D}(X_{\mathrm{an}};\Lambda)$ is the derived category of sheaves of Λ-modules on X_{an}. A central conjecture concerning motives states the following.

Conjecture 5.1 (Conservativity Conjecture). *The functor B^*_X, restricted to the subcategory $\mathbf{DA}^{\text{ét}}_{\mathrm{ct}}(X;\Lambda)$, is conservative. Said differently, if $\mathcal{M}$ is a constructible motivic sheaf on X such that $\mathrm{B}^*_X(\mathcal{M}) \simeq 0$, then necessarily $\mathcal{M} \simeq 0$.*

Lemma 5.2. *It suffices to prove Conjecture 5.1 for $X = \mathrm{Spec}(k)$ and $\Lambda = \mathbb{Q}$.*

Proof. The reduction to the case $\Lambda = \mathbb{Q}$ follows from Theorem 2.24. The reduction to the case $X = \mathrm{Spec}(k)$ is a consequence of the compatibility of the Betti realization with inverse images. □

Conjectures such as the Hodge and Tate Conjectures concern existence of algebraic cycles (and hence elements in motivic cohomology). On the contrary, Conjecture 5.1 concerns motives which makes it look more approachable. However, the next remark suggests that this hope might be too naive.

Remark 5.3. It is well-known that the category of Chow motives with rational coefficients embeds fully faithfully inside $\mathbf{DM}^{\text{ét}}(k;\mathbb{Q})$. Applying Conjecture 5.1 to Chow motives one obtains the following particular case. *Let X and Y be smooth and projective varieties over k of pure dimension d. Let $\gamma \in \mathrm{CH}^d_{\mathbb{Q}}(X \times_k Y)$ be an algebraic cycle inducing an isomorphism in cohomology $\gamma : H^*(Y(\mathbb{C});\mathbb{Q}) \xrightarrow{\sim} H^*(X(\mathbb{C});\mathbb{Q})$. Then, there exists an algebraic cycle $\delta \in \mathrm{CH}^d_{\mathbb{Q}}(Y \times_k X)$ such that $\delta \circ \gamma = [\Delta_X]$ and $\gamma \circ \delta = [\Delta_Y]$.* This reveals a strong analogy/connexion between the Conservativity Conjecture and the Standard Conjecture of Lefschetz type [12].

Remark 5.4. On a more optimistic note, we mention that we formulated in [8, Conjecture B of §2.4] a concrete (although very complicated) conjecture that would implies Conjecture 5.1. We like to think that this is a non trivial step (although, probably, a very small one) towards a potential solution of the Conservativity Conjecture.

5.2. Existence of a motivic *t*-structure. Keep the notation as in §5.1.

Conjecture 5.5 (t-Structure Conjecture). *The category $\mathbf{DA}^{\text{ét}}_{\mathrm{ct}}(X;\Lambda)$ carries a t-structure, called the* motivic t-structure*, making B^*_X exact. (Said differently, if $\mathcal{M}$ is a constructible X-motive which belongs to the heart of the motivic t-structure, then $\mathrm{B}^*_X(\mathcal{M})$ is concentrated in degree zero, i.e., is isomorphic to a constructible sheaf on X_{an}.) Moreover, this t-structure is independent of the choice of the complex embedding σ.*

Remark 5.6. Conjecture 5.5 can be reduced to the case where $X = \mathrm{Spec}(k)$ using gluing techniques. Moreover, these gluing techniques can also be used to define perverse motivic t-structures assuming the existence of the usual motivic t-structure.

Remark 5.7. It is important to note that we do not assume Λ to be a $\mathbb{Q}$-algebra in Conjecture 5.5. Indeed, the t-Structure Conjecture is expected to hold *integrally* for $\mathbf{DA}^{\text{ét}}_{\text{ct}}(X;\Lambda)$; in fact, assuming that $\mathbf{DA}^{\text{ét}}_{\text{ct}}(S;\mathbb{Q})$ admits a motivic t-structure, it is easy to construct a motivic t-structure on $\mathbf{DA}^{\text{ét}}_{\text{ct}}(X;\mathbb{Z})$ using Theorem 2.24.

This is particularly significant as it is well-known that $\mathbf{DM}_{\text{gm}}(k;\Lambda)$ (The "Nisnevich" variant of $\mathbf{DM}^{\text{ét}}_{\text{gm}}(k;\Lambda)$) cannot admit a motivic t-structure unless Λ is a $\mathbb{Q}$-algebra. (A simple explanation for this was given by Voevodsky [29, Remark on page 217].) This indicates that, in view of a future theory of *Abelian motivic sheaves*, it is more natural to impose étale descent.

Remark 5.8. In [8, Conjecture A of §2.4] we formulated a very concrete conjecture that, together with Conjecture B of loc. cit., should imply Conjecture 5.5 and more. (By "more", we have in mind the property that $\mathbf{DA}^{\text{ét}}_{\text{ct}}(S;\Lambda)$ is equivalent to the derived category of the heart of its motivic t-structure.)

Remark 5.9. As a measure of the deepness of Conjectures 5.1 and 5.5, we mention that they imply the Standard Conjectures in characteristic zero (as explained by Beilinson [10]). They imply many other well-established conjectures such as the Bloch Conjecture for surfaces and its generalizations, Kimura finiteness for Chow motives, the existence of the Bloch–Beilinson filtration on Chow groups, etc.

Acknowledgements. The author was supported in part by the Swiss National Science Foundation, project no. 200021-144372/1.

References

[1] J. F. Adams, *Stable homotopy and generalised homology*, University of Chicago Press, Chicago, Ill., 1974, Chicago Lectures in Mathematics.

[2] Yves André, *Pour une théorie inconditionnelle des motifs,* Inst. Hautes Études Sci. Publ. Math., (83):5–49, 1996.

[3] Michael Artin, Alexandre Grothendieck, and Verdier Jean-Louis, *Théorie des Topos et Cohomologie Étale des Schémas,* Lecture Notes in Mathematics, Vol. 269, 270 and 305. Springer-Verlag, Berlin; New York, 1972. Séminaire de Géométrie Algébrique du Bois-Marie SGA 4, Avec la collaboration de N. Bourbaki, P. Deligne and B. Saint-Donat.

[4] M. F. Atiyah and F. Hirzebruch, *Analytic cycles on complex manifolds,* Topology, **1**: 25–45, 1962.

[5] Joseph Ayoub, *Les six opérations de Grothendieck et le formalisme des cycles évanescents dans le monde motivique. I,* Astérisque, (314):x+466 pp. (2008), 2007.

[6] ———, *Les six opérations de Grothendieck et le formalisme des cycles évanescents dans le monde motivique. II,* Astérisque, (315):vi+364 pp. (2008), 2007.

[7] ———, *Note sur les opérations de Grothendieck et la réalisation de Betti,* J. Inst. Math. Jussieu, **9**(2) (2010), 225–263.

[8] ———, *L'algèbre de Hopf et le groupe de Galois motiviques d'un corps de caractéristique nulle, I,* J. reine angew. Math., Ahead of Print, 2013.

[9] ———, *La réalisation étale et les opérations de Grothendieck,* Ann. Sci. Éc. Norm. Supér. (4), **47**(1) (2014), 1–141.

[10] Alexander Beilinson, *Remarks on Grothendieck's standard conjectures,* Preprint, 2010.

[11] Pierre Deligne, James S. Milne, Arthur Ogus, and Kuang-yen Shih, *Hodge cycles, motives, and Shimura varieties*, volume 900 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1982.

[12] Alexandre Grothendieck, *Standard conjectures on algebraic cycles,* In Algebraic Geometry (Internat. Colloq., Tata Inst. Fund. Res., Bombay, 1968), pages 193–199. Oxford Univ. Press, London, 1969.

[13] Masaki Hanamura, *Mixed motives and algebraic cycles. I,* Math. Res. Lett., **2**(6):811–821, 1995.

[14] ———, *Mixed motives and algebraic cycles. III,* Math. Res. Lett., **6**(1) (1999), 61–82.

[15] ———, *Mixed motives and algebraic cycles. II,* Invent. Math., **158**(1) (2004), 105–179.

[16] Philip S. Hirschhorn, *Model categories and their localizations*, volume 99 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 2003.

[17] Mark Hovey, *Spectra and symmetric spectra in general model categories,* J. Pure Appl. Algebra, **165**(1) (2001), 63–127.

[18] Mark Hovey, Brooke Shipley, and Jeff Smith, *Symmetric spectra,* J. Amer. Math. Soc., **13**(1) (2000), 149–208.

[19] J. F. Jardine, *Motivic symmetric spectra,* Doc. Math., **5**:445–553 (electronic), 2000.

[20] Steven L. Kleiman, *Motives In Algebraic geometry, Oslo 1970* (Proc. Fifth Nordic Summer-School in Math.), pages 53–82. Wolters-Noordhoff, Groningen, 1972.

[21] János Kollár, *Trento examples In Classification of irregular varieties,* (Trento, 1990), volume 1515 of Lecture Notes in Math., pages 134–139. Springer, Berlin, 1992.

[22] Marc Levine, *Mixed motives*, volume 57 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 1998.

[23] ———, *Mixed motives,* In Handbook of K-theory. Vol. 1, 2, pages 429–521, Springer, Berlin, 2005.

[24] Carlo Mazza, Vladimir Voevodsky, and Charles Weibel, *Lecture notes on motivic cohomology*, volume 2 of Clay Mathematics Monographs, American Mathematical Society, Providence, RI, 2006.

[25] Fabien Morel and Vladimir Voevodsky, $\mathbf{A}^1$*-homotopy theory of schemes,* Inst. Hautes Études Sci. Publ. Math., (90) (2001), 45–143, 1999.

[26] Andreas Rosenschon and Vasudevan Srinivas, *Étale motivic cohomology and algebraic cycles,* Preprint, 2014.

[27] Jean-Louis Verdier, *Des catégories dérivées des catégories abéliennes,* Astérisque, (239):xii+253 pp. (1997), 1996, With a preface by Luc Illusie, Edited and with a note by Georges Maltsiniotis.

[28] Vladimir Voevodsky, *Homology of schemes,* Selecta Math. (N.S.), 2(1):111–153, 1996.

[29] Vladimir Voevodsky, Andrei Suslin, and Eric M. Friedlander, *Cycles, transfers, and motivic homology theories*, volume 143 of Annals of Mathematics Studies. Princeton University Press, Princeton, NJ, 2000.

[30] Rainer Vogt, *Boardman's stable homotopy category*, Lecture Notes Series, No. 21. Matematisk Institut, Aarhus Universitet, Aarhus, 1970.

Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland
E-mail: joseph.ayoub@math.uzh.ch

Isogenies, power operations, and homotopy theory

Charles Rezk

Abstract. The modern understanding of the homotopy theory of spaces and spectra is organized by the chromatic philosophy, which relates phenomena in homotopy theory with the moduli of one-dimensional formal groups. In this paper, we describe how certain phenomena $K(n)$-local homotopy can be computed from knowledge of isogenies of deformations of formal groups of height n.

Mathematics Subject Classification (2010). Primary 55S25; Secondary 55N34, 55P43.

Keywords. Homotopy theory, formal groups, power operations.

1. Introduction

A sweeping theme in the study of homotopy theory over the past several decades is the *chromatic viewpoint*. In this philosophy, phenomena in homotopy theory are associated to phenomena in the theory of one-dimensional formal groups. This program was instigated by Quillen's observation of the connection between complex bordism and formal group laws [32].

The chromatic picture is best described in terms of localization at a chosen prime p. After one localizes at a prime p, the moduli of formal groups admits a descending filtration, called the *height* filtration. According the chromatic philosophy, this filtration is mirrored by a sequence of successive approximations to homotopy theory. The difference between adjacent approximation is the *nth chromatic layer*, which is associated by the chromatic picture to formal groups of height n. Phenomena in the nth chromatic layer may be detected using cohomology theories called *Morava K-theories* and *Morava E-theories*, which are typically (and unimaginatively) denoted $K(n)$ and E_n. A good recent introduction to this point of view is is [17].

In this paper I will describe a particular manifestation of the chromatic picture, which relates "$K(n)$-local homotopy theory" (i.e., one manifestation of the nth chromatic layer in homotopy theory), to *isogenies* of formal groups.

Some of the results describe here are joint work with others, including some not-yet-published work with Matt Ando, Mark Behrens, and Mike Hopkins.

2. Formal groups and localized homotopy theory

We briefly recall the role and significance of formal groups in homotopy theory.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2.1. Complex orientation and formal groups. Recall that a generalized cohomology theory E^* is said to be **complex orientable** if (i) it takes values in graded rings, and (ii) if there exists an element $x \in \widetilde{E}^*\mathbb{CP}^\infty$ which restricts to the fundamental class in $\widetilde{E}^*\mathbb{CP}^1$. To such a theory is associated a (one-dimensional, commutative) **formal group** $\mathbb{G}_E$, which is the formal scheme over the ring E^* with coefficient ring $\mathcal{O}_{\mathbb{G}_E} = E^*\mathbb{CP}^\infty$.

Remark 2.1. For a complex orientable theory, a choice of element x as in (ii) gives rise to an "Euler class" on complex line bundles, defined by $e_x(L \to X) := f^*(x) \in E^*(X)$ where $f\colon X \to \mathbb{CP}^\infty$ classifies L, together with a power series $F_x(t_1, t_2) \in E^*[\![t_1, t_2]\!]$ expressing the Euler class of a tensor product of lines:

$$e_x(L_1 \otimes L_2) = F_x(e_x(L_1), e_x(L_2)),$$

which is an example of a **formal group law** on E^*. Both e_x and F_x depend on the choice of "coordinate" $x \in \mathcal{O}_{\mathbb{G}_E} = E^*\mathbb{CP}^\infty$. The formal group $\mathbb{G}_E$ is a coordinate-free expression of the collection of formal group laws associated to E, and depends only on the cohomology theory E itself.

Example 2.1. For a ring R, let HR^* denote ordinary cohomology with coefficients in R. For any R, the theory HR^* is complex orientable, and the resulting formal group $\mathbb{G}_R$ is the *additive* formal group. In fact, if we take $x \in HR^2(\mathbb{CP}^\infty)$ to be any generator, we have that $F_x(t_1, t_2) = t_1 + t_2$, and recover the classical addition formula for first Chern classes of complex line bundles.

Example 2.2. Complex bordism MU^* is a complex oriented theory, which comes with a tautological choice of coordinate $x \in MU^2\mathbb{CP}^\infty$. Quillen [32] identified the resulting formal group F_x as the *universal* formal group law. In coordinate-free language, we may say that the formal group $\mathbb{G}_{MU}$ of complex bordism is the universal example of a formal group equipped with the data of a choice of coordinate.

All formal groups over a given field of characteristic 0 are isomorphic to the additive formal group. For a formal groups $\mathbb{G}$ over fields k of characteristic p, there is an isomorphism invariant called the **height** of $\mathbb{G}$, which is an element $n \in \mathbb{Z}_{\geq 1} \cup \{\infty\}$. For separably closed k, the height is a complete invariant.

Example 2.3. Fix a prime p. For any height $n \in \mathbb{Z}_{\geq 1}$, there exist complex cohomology theories whose formal group is one of height n. The standard examples are the **Morava K-theories** $K(n)$, whose coefficient ring is $K(n)^* = \mathbb{F}_p[v_n^{\pm}]$ with $v_n \in K(n)^{-2(p^n-1)}$, and whose formal group is the **Honda formal group** of height n.

Example 2.4. For any formal group G_0 of height $1 \leq n < \infty$ over a perfect field k, Lubin and Tate constructed its *universal deformation*, which is a formal group G defined over the complete local ring $\mathbb{W}_p[\![u_1, \dots, u_{n-1}]\!]$, whose restriction at the special fiber is G_0. There is a corresponding cohomology theory called **Morava E-theory**, which will play an important role in our story; see §4.2 below.

Formal groups of infinite height over fields of characteristic p are isomorphic to the additive formal group. Ordinary homology $H\mathbb{F}_p$ with mod p coefficients is an example of a complex oriented theory whose formal group has infinite height.

It is conventional to say that any formal group over a field of characteristic zero has height 0. Ordinary homology $H\mathbb{Q}$ with rational coefficients is an example of a theory with such a formal group.

2.2. Localized homotopy theory. Associated to any homology theory E is a corresponding **localization functor**, first constructed in full generality by Bousfield [10]. Say that a based space Y is E**-local** if for any based CW-complex K such that $\widetilde{E}_*(K) \approx 0$, the space $\mathrm{Map}_*(K, Y)$ of based maps is weakly contractible. Bousfield showed that for any space X, there exists a space X_E and a map $\eta_E\colon X \to X_E$, called the E**-localization** of X, such that (i) the map η_E induces an isomorphism in E-homology, and (ii) X_E is E-local. Furthermore, the operation $X \mapsto X_E$ can be realized functorially.

Example 2.5. For ordinary homology theories $E = HR$, localization of spaces is well-behaved in the absence of fundamental groups. For instance, $\pi_*(X_{H\mathbb{Q}}) \approx \pi_* X \otimes \mathbb{Q}$ if X is simply connected, and $\pi_*(X_{H\mathbb{F}_p}) \approx (\pi_* X)^\wedge_p$ if X is simply connected and finite type.

There is an analogous localization construction for spectra. In what follows we will be most interested in localization with respect to Morava K-theory spectra. In particular, for every prime p and $n \geq 1$, there is a localization functor

$$X \mapsto X_{K(n)}\colon h\mathrm{Spectra} \to h\mathrm{Spectra}_{K(n)} \subset h\mathrm{Spectra}$$

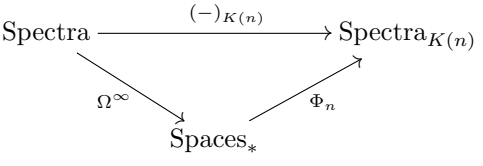
from the homotopy category of spectra to the full subcategory of $K(n)$-local spectra.

2.3. The functor of Bousfield and Kuhn. It is a remarkable observation of Bousfield[11] and Kuhn[22] that localization functors on spectra with respect to certain homology theories (such as Morava K-theories) actually factor through the underlying space.

Fix a prime p and an integer $n \geq 1$. There exists a functor

$$\Phi_n\colon \mathrm{Spaces}_* \to \mathrm{Spectra}_{K(n)} \subset \mathrm{Spectra}$$

from pointed spaces to $K(n)$-local spectra which makes the following diagram commute up to natural weak equivalence.



The functor Φ_n is constructed using the existence of periodic phenomena in stable homotopy theory [20]. Observe that given any space Y and map $g\colon Y \to \Omega^d Y$ with $d \geq 1$, we can obtain a spectrum E by setting $\underline{E}_{kd} = Y$ using the g as the structure map $\underline{E}_{kd} \to \Omega^d \underline{E}_{kd+d}$ (much as periodic K-theory is obtained from the Bott periodicity map $U \to \Omega^2 U$, though in our case g need not be an equivalence). Given a based finite CW-complex K and a map $f\colon \Sigma^d K \to K$ for some $d \geq 1$, define a functor $\Phi_{K,f}\colon \mathrm{Spaces}_* \to \mathrm{Spectra}$ by associating to a based space X the map

$$f^*\colon \mathrm{Map}_*(K, X) \xrightarrow{f} \mathrm{Map}_*(\Sigma^d K, X) \approx \Omega^d \mathrm{Map}_*(K, X),$$

from which we obtain a spectrum $\Phi_{K,f}(X)$ as above. The functor $\Phi_{K,f}(X)$ has the properties:

- $\Phi_{K,f}(\Omega X) \approx \Omega \Phi_{K,f}(X)$,

- If $X = \Omega^\infty Y$ is the 0-space of a spectrum Y, then

$$\Phi_{K,f}(\Omega^\infty Y) \approx \operatorname{hocolim}_k \Sigma^{-kd}\underline{\mathrm{Hom}}(\Sigma^\infty K, Y) \approx (\Sigma^\infty f)^{-1}\underline{\mathrm{Hom}}(\Sigma^\infty K, Y),$$

  the "telescope" of the function spectrum $\underline{\mathrm{Hom}}(\Sigma^\infty K, Y)$ induced by the map $\Sigma^\infty f$.

Non-trivial examples are provided by a v_n-self map, i.e., a pair (K, f) such that K is a finite CW complex with $K(n)_*K \neq 0$ and $f\colon \Sigma^d K \to K$ with $d \geq 1$ such that $K(n)_*f$ is an isomorphism. (This implies that for any spectrum Y, the map $\underline{\mathrm{Hom}}(K, Y) \to f^{-1}\underline{\mathrm{Hom}}(K, Y)$ induces an isomorphism on $K(n)_*$-homology, and thus $\Phi_{K,f}$ is seen to be non-trivial.)

Such v_n-self maps are plentiful by the *periodicity theorem* of Hopkins-Smith [20]. Using this theory, [22] constructs Φ_n^f as a homotopy inverse limit of a suitably chosen family of functors $\Sigma^{-q_i}\Phi_{K_i,f_i}$; then $\Phi_n(X) := \Phi_n^f(X)_{K(n)}$.

Remark 2.2. Given a v_n-self map f of K, the homotopy groups

$$\pi_*\Phi_{K,f}(X) \approx f^{-1}\pi_* \operatorname{Map}(K, X)$$

are the v_n**-periodic homotopy groups** of the space X, usually denoted $v_n^{-1}\pi_*(X; K)$ (they depend on K, but not on the choice of v_n-self map f). As a result, the spectrum $\Phi_n(X)$ contains information about the v_n-periodic homotopy groups of the space X. The extent to which this information is captured depends in part on the status of the *telescope conjecture*, which if true would imply that $\Phi_n = \Phi_n^f$; see the discussion in [23].

Bousfield has developed a theory to effectively compute invariants of $\Phi_1(X)$ for certain spaces X, such as spheres and finite H-spaces [13, 14]. In §5, we will outline an approach to generalize Bousfield's results to the case of $n \geq 2$.

2.4. The Bousfield-Kuhn idempotent. Given a basepoint preserving unstable map $f\colon \Omega^\infty F \to \Omega^\infty E$ where E is a $K(n)$-local spectrum, the $K(n)$-local Bousfield-Kuhn functor gives rise to a map of spectra

$$F \xrightarrow{\iota} F_{K(n)} \approx \Phi_n\Omega^\infty F \xrightarrow{\Phi_n(f)} \Omega^\infty E \approx E,$$

and hence an infinite loop space map

$$\phi_n(f) := \Omega^\infty(\Phi_n(f) \circ \iota)\colon \Omega^\infty F \to \Omega^\infty E.$$

If $f = \Omega^\infty g$ for a map $g\colon F \to E$ of spectra, then $\Phi_n(f) \circ \iota = g$. Thus, the function

$$\phi_n\colon h\mathrm{Spaces}_*(\Omega^\infty F, \Omega^\infty E) \xrightarrow{\Phi} h\mathrm{Spectra}(F, E) \xrightarrow{\Omega^\infty} h\mathrm{Spaces}_*(\Omega^\infty F, \Omega^\infty E)$$

is *idempotent*, with image precisely the set of homotopy classes of maps $\Omega^\infty F \to \Omega^\infty E$ which are infinite loop maps.

It turns out to be possible to compute something about the map $\phi_n(f)$ (as an *unstable* map), when E is a complex oriented cohomology theory to which the character theory of Hopkins-Kuhn-Ravenel [18] can be applied, such as Morava E-theory. For the purposes of stating a result, we recall that the Hopkins-Kuhn-Ravenel theory provides a natural ring homomorphism

$$\chi\colon E^0(X \times BG) \to \prod_{g \in G_{n,p}} D \otimes_{E^0} E^0(X),$$

for any finite group G, where: X is a finite CW-complex, $G_{n,p}$ is the set of conjugacy classes of homomorphism $\Lambda_N = (\mathbb{Z}/p^N)^n \to G$ (for N sufficiently large, depending on G), and D_N is a certain faithfully flat extension of E^0.

Let $f\colon \Omega^\infty F \to \Omega^\infty E$ be an H-map; i.e., the induced operation $F^0(-) \to E^0(-)$ is a group homomorphism. The following computes the operation $\phi_n(f)$.

Theorem 2.3. *Let E be a $K(n)$-local complex orientable theory such that π_*E is a complete local ring, and let $f\colon \Omega^\infty F \to \Omega^\infty E$ be an H-space map. Then for any finite CW-complex X, the map*

$$\phi_n(f)\colon F^0(X) \to E^0(X)$$

is given (modulo torsion in $E^0(X)$) by

$$(\phi_n f)(x) = \sum_{k=0}^{n} (-1)^k p^{\binom{k}{2}} \left(p^{-k} \sum_{A \subseteq \Lambda_1^*,\ |A|=p^k} \chi(f(\widetilde{x} \wedge t_A))(g_A). \right) \tag{2.1}$$

The inner sum is taken over all subgroups of $\Lambda_1^* = \mathrm{Hom}(\Lambda_1, U(1))$ of given order; the map $g_A\colon \Lambda_1 \to A^* = \mathrm{Hom}(A, U(1))$ is the dual homomorphism to the inclusion $A \subseteq \Lambda_1^*$; $t_A\colon \Sigma_+^\infty BA^* \to S^0$ is the stable transfer map, and $\widetilde{x}\colon \Sigma_+^\infty X \to F$ is the map of spectra representing $x \in F^0(X)$.

It turns out that the inner sum of (2.1) in fact takes values in $E^0(X) \subset D_1 \otimes_{E^0} E^0(X)$, and furthermore this value is divisible by p^k, so that the bracketed expression in (2.1) gives a well-defined element of $E^0(X)$ modulo torsion. The statement (2.3) has not appeared in print elsewhere, but it is a consequence of the methods of [36].

Example 2.6. Let $E = K_p$ be p-complete K-theory. Then the formula of (2.3) becomes

$$(\Phi f)(x) = f(x) - p^{-1}\chi(f(\widetilde{x} \wedge t_{Z/p}))(g_{\mathbb{Z}/p}).$$

In particular, if $f\colon \Omega^\infty F \to \Omega^\infty K_p$ is an H-map such that $\chi(f(\widetilde{x} \wedge t_{\mathbb{Z}/p}))(g_{\mathbb{Z}/p}) = 0$, then f admits the structure of an infinite loop map. This immediately recovers a well-known theorem of Madsen-Snaith-Tornehave [31].

There is a generalization of (2.1) for f which is not necessarily an H-map, though it is too cumbersome to give it here. Formulas of this nature, where the target spectrum is $E = K(n)$, are obtained in [42].

3. Units and orientations

The Bousfield-Kuhn idempotent can be usefully applied to the study of the units spectrum of a commutative S-algebra.

3.1. The units of a commutative ring spectrum. Let R be a homotopy associative ring spectrum. The **units space** of R is called $GL_1(R)$; it is defined by the pullback square of spaces

$$\begin{array}{ccc} GL_1(R) & \longrightarrow & \Omega^\infty R \\ \downarrow & & \downarrow \\ (\pi_0 R)^\times & \rightarrowtail & \pi_0 R \end{array}$$

For a space X, we have $h\mathrm{Spaces}(X, GL_1(R)) \approx R^0(X)^\times \subseteq R^0(X)$.

When R is a commutative S-algebra, then $GL_1(R)$ admits a canonical structure of a grouplike E_∞-space, induced by the multiplicative structure of R. Write $\mathrm{gl}_1(R)$ for the (-1)-connected spectrum which is the infinite delooping of $GL_1(R)$, called the **units spectrum** of R.

The units spectrum carries the obstruction to constructing orientations of commutative S-algebras, as shown by May, Quinn, Ray, and Tornehave in [26]; see [27] and [1] for recent treatments of this theory. Let $f\colon g \to o$ be a map of (-1)-connective spectra, where o is the infinite delooping of BO, the classifying space of the infinite dimensional orthogonal group. Let $BG = \Omega^\infty g$ denote the infinite delooping of g, and let MG denote the Thom spectrum of the virtual vector bundle classified by $B(\Omega^\infty f)\colon BG \to BO$; the spectrum MG admits (up to weak equivalence) the structure of a commutative S-algebra. Then one can show that space of commutative S-algebra maps $MG \to R$ is weakly equivalent to the space of null-homotopies of the composite map

$$g \xrightarrow{f} o \xrightarrow{j} \mathrm{gl}_1(S) \to \mathrm{gl}_1(R).$$

Thus, understanding the homotopy type of the spectrum $\mathrm{gl}_1(R)$ is essential to understanding G-orientations of R which are realized by maps of commutative S-algebras.

3.2. A "logarithmic" operation. We use the Bousfield-Kuhn functor to obtain information about the $K(n)$-localization of $\mathrm{gl}_1(R)$. To do this, we consider the "shift" map

$$s\colon GL_1(R) \xrightarrow{x \mapsto x-1} \Omega^\infty R.$$

This shift map is a based map between infinite loop spaces, and thus we may apply the idempotent operator of §2.4 to it. If R is a $K(n)$-local commutative S-algebra, we obtain in this way from this cohomology operation of the form

$$\ell_n = \phi_n(s)\colon R^0(X)^\times \to R^0(X),$$

which is "logarithmic", in the sense that $\ell_n(xy) = \ell_n(x) + \ell_n(y)$. The operation ℓ_n is represented by a map of spectra $\mathrm{gl}_1(R) \to R$.

Example 3.1. To get a sense of what such an operation provides, consider the following analogous situation, where E is a *rational* commutative S-algebra. For any pointed and *connected* space X, we can define a group homomorphism

$$\ell_{\mathbb{Q}}\colon (1 + \overline{E}^0(X))^\times \to E^0(X) \qquad \text{by} \qquad \ell_{\mathbb{Q}}(x) = -\textstyle\sum_{m\geq 1}(1-x)^m/m = \log(x).$$

The series defining $\ell_{\mathbb{Q}}$ is converges: because X is connected, $1 - x$ is nilpotent when restricted to any connected finite CW-complex mapping to X. The operation $\ell_{\mathbb{Q}}$ is in fact stable: it is represented by a map of spectra $(\mathrm{gl}_1 E)_{\geq 1} \to E$ (where $Z_{\geq n}$ denotes the $(n-1)$-connected cover of a spectrum Z).

The above theory applies in this case to give the following.

Theorem 3.1 ([36]). *Let E be a Morava E-theory* (2.4)*, associated to the Lubin-Tate universal deformation of a height n-formal group. Then its logarithmic operation is given (modulo*

torsion) by the formula

$$\ell_n(x) = \frac{1}{p}\log\left(\prod_{k=0}^{n}\Big(\prod_{A\subset\Lambda^*,\ |A|=p^k}\psi_A(x)\Big)^{(-1)^k p^{\binom{k}{2}-k+1}}\right). \tag{3.1}$$

The functions $\psi_A\colon E^0(X)\to D\otimes_{E^0}E^0(X)$ are certain natural additive and multiplicative cohomology operations (described below (4.1)*), and* $\log(x) = -\sum_{m\geq 1}(1-x)^m/m$.

The expression inside "log" in (3.1) is a multiplicative analog of (2.1), with the role of $x\mapsto\chi(f(\widetilde{x})\wedge t_A, g_A)$ in (2.1) replaced by the operation ψ_A. It turns out that the expression inside log in (3.1) is in fact contained in $E^0(X)\subseteq D\otimes_{E^0}E^0(X)$, and is congruent to 1 modulo p; thus evaluating the formal expansion of log at this expression converges p-adically to an element of $E^0(X)$.

Example 3.2. An example of a Morava E-theory spectrum at height 1 is KU_p, the p-completion of complex K-theory. In fact, it is possible to generalize (3.1) in the case of $n=1$ to any $K(1)$-local commutative S-algebra E, so we will describe the result in this case [36, Thm. 1.9]. The formula (3.1) takes the form

$$\ell_1(x) = \frac{1}{p}\log\big(x^p/\psi^p(x)\big) = -\sum_{m\geq 1}\frac{1}{pm}\big(1-x^p/\psi^p(x)\big)^m. \tag{3.2}$$

If E is KU_p or KO_p, the operation $\psi^p\colon E^0(-)\to E^0(-)$ is the usual pth Adams operation on p-complete real-or-complex K-theory.

We can do a little better in this case: the operation ψ^p on $E^0(-)$ satisfies a "Frobenius congruence" $\psi^p(x)\equiv x^p \mod pE^0(-)$; therefore the infinite series of (3.2) converges p-adically. The Frobenius congruence is "witnessed" by a cohomology operation $\theta^p\colon E^0(-)\to E^0(-)$, satisfying the identity $\psi^p(x) = x^p + p\,\theta^p(x)$. Thus we can write

$$\ell_1(x) = \sum_{m\geq 1}(-1)^m\frac{p^{m-1}}{m}(\theta^p(x)/x^p)^m, \tag{3.3}$$

and (3.3) in fact holds on the nose (i.e., not merely modulo torsion [36, Thm. 1.9]). The right-hand side of (3.3) recovers the **Artin-Hasse logarithm** of tom Dieck [43], who originally realized this operation as as spectrum maps $\mathrm{gl}_1(KU_p)\to KU_p$ and $\mathrm{gl}_1(KO_p)\to KO_p$ without reference to the Bousfield-Kuhn functor.

We can use (3.3), to compute the map ℓ_1 on homotopy groups, and we thus recover the well-known equivalences of connected covers $\mathrm{gl}_1(KU_p)_{\geq 3}\xrightarrow{\sim}(KU_p)_{\geq 3}$ and $\mathrm{gl}_1(KO_p)_{\geq 2}\xrightarrow{\sim}(KO_p)_{\geq 2}$.

To understand (3.1) in the general case, we can formally pull the operations ψ_A (which are ring homomorphisms) out of the logarithmic series, obtaining

$$\ell_n(x) = \sum_{k=0}^{n}(-1)^k p^{\binom{k}{2}}T_j(\log x) \qquad \text{where} \qquad T_j := p^{-k}\sum_{A\subseteq\Lambda^*,\ |A|=p^k}\psi_A. \tag{3.4}$$

For $x = 1+y\in E^0(X)^\times$ such that $y^2=0$, this becomes

$$\ell_n(x) = \sum_{k=0}^{n}(-1)^k p^{\binom{k}{2}}T_j(y). \tag{3.5}$$

In particular, taking X to be a sphere, we obtain a formula which computes the effect of $\ell_n\colon \mathrm{gl}_1(E) \to E$ on homotopy groups (up to torsion).

To understand how we can compute these operations, we must discuss "power operations" for $K(n)$-local commutative rings (such as Morava E-theory). The short answer is that such operations are controlled by certain isogenies of the formal group associated to the theory, and in particular the operators T_j are "Hecke operators" for the theory. We will come back to this in §4.

3.3. Application of the logarithm to orientation problems.

Example 3.3. Orientations of K-theory. Consider the composite

$$spin \to o \xrightarrow{j} \mathrm{gl}_1(S) \to \mathrm{gl}_1(KO_p) \xrightarrow{\ell_1} KO_p. \tag{3.6}$$

It is a standard calculation that all maps $spin = \Sigma^{-1}(KO_{\geq 4}) \to KO_p$ are null homotopic. As ℓ_1 is here an equivalence on 2-connected covers (3.2), we immediately see that the composite of (3.6) is null-homotopic, and thus there must exist a map $MSpin \to KO_p$ of commutative S-algebras. It can be shown that the Atiyah-Bott-Shapiro orientation can be realized by one such map; see [19, §6.1] for a sketch.

3.4. Application to the string orientation of tmf. With Matt Ando and Mike Hopkins, we have shown that tmf, the spectrum of **topological modular forms**, admits a commutative S-algebra map $MString \to \mathrm{tmf}$ which realizes the Witten genus. Our proof only exists in preprint form, though the result was announced in [19, §6], to which the reader is referred for background. Here we will only note the way in which the logarithmic operation enters into the proof.

The key point is to understand the homotopy type of $\mathrm{gl}_1(\mathrm{tmf}_p)$, where tmf_p is the completion of tmf at a prime p. General results localizations show that there is a commutative square of spectra

$$\begin{array}{ccc} \mathrm{gl}_1(\mathrm{tmf}_p) & \xrightarrow{\ell_2} & \mathrm{tmf}_{K(2)} \\ {\scriptstyle \ell_1}\downarrow & & \downarrow{\scriptstyle \iota_{K(2)}} \\ \mathrm{tmf}_{K(1)} & \xrightarrow{\gamma} & (\mathrm{tmf}_{K(2)})_{K(1)} \end{array}$$

which, after taking 3-connected covers, becomes a homotopy pullback. Both $\mathrm{tmf}_{K(2)}$ and $\mathrm{tmf}_{K(1)}$ are relatively well-understood objects: $\mathrm{tmf}_{K(2)}$ is closely related to Morava E-theory spectra at height 2, while $\mathrm{tmf}_{K(1)}$ is related to the theory of p-adic modular forms. To understand the homotopy type of $\mathrm{gl}_1(\mathrm{tmf}_p)$, we must get our hands on the map γ. It can be shown that maps $\mathrm{tmf}_{K(1)} \to (\mathrm{tmf}_{K(2)})_{K(1)}$ are characterized (up to homotopy) by their effect on homotopy groups. Thus, the key is to compute the effect of γ on homotopy groups.

Recall that there is a map $\pi_*\mathrm{tmf} \to MF_*$ to the ring of modular forms (with integer coefficients), which is an isomorphism up to torsion. Given an element in $\pi_{2k}\mathrm{tmf}$ corresponding to a modular form f of weight k, we use (3.2) to obtain

$$\ell_1(f) = f^*(q) := f(q) - p^{k-1}f(q^p),$$

where the result is stated in terms of the q-expansion of f. The series $f^*(q)$ is the q-expansion of a p-adic modular form, and thus corresponds to an element of $\pi_{2k}\mathrm{tmf}_{K(1)}$.

When we evaluate ℓ_2 at an element of $\pi_{2k}\mathrm{tmf}$ associated to a modular form f, the result turns out to be again a modular form; i.e., the image of $\ell_2\colon \pi_*\mathrm{tmf}_p \to \pi_*\mathrm{tmf}_{K(2)}$ is contained in the image of $\iota_{K(2)}\colon \pi_*\mathrm{tmf}_p \to \pi_*\mathrm{tmf}_{K(2)}$. In fact, (3.5) implies

$$\ell_2(f) = (1 - T_1 + p^{k-1})f,$$

where T_1 is a classical Hecke operator on modular forms (usually written $T(p)$ in this context).

Using these calculations, one can deduce that

$$\gamma = (\iota_{K(2)})_{K(1)} \circ (\mathrm{id} - U),$$

where $U\colon \mathrm{tmf}_{K(1)} \to \mathrm{tmf}_{K(1)}$ is topological lift of the **Atkin operator** on p-adic modular forms; see [5] for a construction of this topological lift. This calculation provides enough control on the homotopy type of $\mathrm{gl}_1(\mathrm{tmf}_p)$ to study the set of string-orientations of tmf.

Remark 3.2. These ideas actually allow one to construct ("by hand") a spectrum map $\ell_{\mathrm{tmf}_p}\colon\ \mathrm{gl}_1(\mathrm{tmf}_p) \to \mathrm{tmf}_p$, so that $\iota_{K(2)} \circ \ell_{\mathrm{tmf}_p} = \ell_2$ and $\iota_{K(1)} \circ \ell_{\mathrm{tmf}_p} = (\mathrm{id} - U) \circ \ell_1$. This fact is in need of a more natural explanation.

Remark 3.3. If $f = \sum a_n q^n$ is an *eigenform* of weight k, then $(1 - T_1 + p^{k-1})f = (1 - a_p + p^{k-1})f$. In particular, if f is an *Eisenstein series*, then $\ell_2(f) = 0$, an observation which is is key to realizing the Witten genus as a string-orientation.

We note in passing that for an eigenform which is a cusp form and normalized so that $a_1 = 1$, the expression $L(f, s) = \prod_p (1 - a_p p^{-s} + p^{k-1-2s})^{-1}$ is precisely the L-series associated to the form. The significance of this to homotopy theory remains unclear.

4. Power operations

The notion of a power operation originated in Steenrod's construction of the eponymously named operations in ordinary cohomology with coefficients in $\mathbb{F}_p$. A convenient modern formulation is in terms of structured commutative ring spectra. There are various equivalent models of such; I will not distinguish among them here, and I will call them **commutative S-algebras**; see [15] and [28] for introductions to some of these models.

A (generalized) cohomology theory $X \mapsto E^*(X)$ is represented by a spectrum E. If E is equipped with the structure of a commutative monoid in the homotopy category of spectra, then $X \mapsto E^*(X)$ takes values in graded commutative rings. For any $m \geq 0$, there is a resulting cohomology operation $x \mapsto x^m\colon E^0(-) \to E^0(-)$ defined by taking mth powers with respect to the product.

If E is equipped with the structure of a commutative S-algebra, then the mth power map admits a refinement to a "total mth power operation" of the form

$$P_m\colon E^0(X) \to E^0(X \times B\Sigma_m),$$

where $B\Sigma_m$ is the classifying space of the symmetric group on m letters. The function P_m is a multiplicative (but non-additive) natural transformation $E^0(-) \to E^0(- \times B\Sigma_m)$ of cohomology groups.

It is convenient to regard power operations as operations on the homotopy groups of commutative S-algebras. For a commutative S-algebra R, the power construction determines gives a function

$$P_m\colon \pi_0 R \to \pi_0 R^{B\Sigma_m^+}, \tag{4.1}$$

where $R^X := \underline{\mathrm{Hom}}(\Sigma^\infty X, R)$, the spectrum of maps from the suspension spectrum of a pointed space X to R. (Operations in dimensions other 0 can also be obtained, by replacing $B\Sigma_m^+$ with a suitable Thom spectrum. In the discussion below, I will concentrate on operations in dimension 0 for simplicity.) Cohomology operations for $E^*(X)$ (such as Steenrod's for ordinary cohomology) can be obtained by setting $R = E^{X_+}$, to be thought of as the ring of E-valued cochains on X.

Let us fix a commutative S-algebra E. To calculate power operations for commutative E-algebras R, one must understand the functor $R \mapsto \pi_0(R^{B\Sigma_m^+}) = R^0(B\Sigma_m)$, which in practice can be non-trivial. The best case scenario is to have a natural "Künneth isomorphism"

$$\pi_* R^{B\Sigma_m^+} \approx \pi_* R \ \otimes_{\pi_* E}\ \pi_* E^{B\Sigma_m^+}, \tag{4.2}$$

together with calculational control of the rings $\pi_* E^{B\Sigma_m^+} \approx E^{-*} B\Sigma_m$, the E-cohomology rings of symmetric groups.

This best case scenario is in fact relatively rare. It does hold for HF-algebras, where F is a field [9]. It holds also for $K(n)$-local commutative E-algebras, where E is a Morava E-theory spectrum.

4.1. Power operations in the $K(n)$-local setting. In 1993, Hopkins and Miller perceived that a Morava E-theory spectrum must admit an essentially unique commutative S-algebra structure; the proof is result is in [16]. Therefore Morava E-theories admit a theory of power operations; such operations were first originally by Ando [4][1].

Example 4.1. The operations ψ_A appearing in (3.1) are obtained as power operations for Morava E-theory, namely as the composite

$$E^0(X) \xrightarrow{P_{p^k}} E^0(X \times B\Sigma_{p^k}) \xrightarrow{(\mathrm{id}\times Bi)^*} E^0(X \times BA^*) \xrightarrow{\chi(-,g_A)} D \otimes_{E^0} E^0(X),$$

where P_{p^k} is the total power operation for the Morava E-theory, $i\colon A^* \to \Sigma_{p^k}$ is the inclusion defined by the left-action of A^* on its underlying set, where $p^k = |A^*|$. These are examples of the operations constructed in [4].

The theory of power operations for commutative algebras over Morava E-theories is now very largely understood, based mainly on work by Ando, Hopkins, and Strickland, who were motivated by the problem of rigidifying the Witten genus to a map of spectra [2], along with some contributions by the author [37].

Our goal in this section is two-fold: to show (i) the homotopy groups $\pi_* R$ of a $K(n)$-local commutative E-algebra R take values in a category $\mathrm{QCoh}(\mathrm{Def})$ of sheaves on a moduli problem of "formal groups and isogenies", and (ii) the category $\mathrm{QCoh}(\mathrm{Def})$ can in practice be described using a small amount of data, and in fact at small heights can be described completely explicitly. In addition to the references given below, the material in this section is developed in detail in the preprint [34].

[1] In fact, Ando did not make use of the commutative S-algebra structure of E, which was unavailable at the time, though he does show that the operations he constructs are the same as those obtained from any commutative S-algebra structure which might exist on E.

4.2. Deformations and Morava E-theory. We fix a height n formal group (commutative, one-dimensional) G_0 over a perfect field k.

Given a formal group G over a complete local ring B, a **(G_0-)deformation structure** on G/B is a pair (i, α) consisting of an inclusion $i\colon k \to B/\mathfrak{m}$ of fields and an isomorphism $\alpha\colon i^*G_0 \xrightarrow{\sim} G_{B/\mathfrak{m}}$ of formal groups over $B/\mathfrak{m}$. We write $\mathcal{D}(G/A) = \mathcal{D}_{G_0}(G/A)$ for the set of deformation structures on G/A. Note that if $g\colon A \to A'$ is a local homomorphism, there is map $g^*\colon \mathcal{D}(G/A) \to \mathcal{D}(g^*G/A')$ induced by base change.

An **isogeny** of formal groups over A is a homomorphism $f\colon G \to G'$ such that the induced map $\mathcal{O}_{G'} \to \mathcal{O}_G$ on function rings is finite and locally free; we write $\deg(f)$ for the rank of $\mathcal{O}_G$ as an $\mathcal{O}_{G'}$-module. Given such an isogeny, there is an induced pushforward map $f_!\colon \mathcal{D}(G/A) \to \mathcal{D}(G'/A)$ on sets of deformation structures, so that

$$f_*(i, \alpha) := (i \circ \phi^r, \alpha'),$$

where $\phi(a) = a^p$ is the absolute Frobenius on rings, $p^r = \deg f$, and α' is the unique isomorphism such that $\alpha' \circ F^r = f_{B/\mathfrak{m}} \circ \alpha$, where

$$F^r\colon G \to (\phi^r)^*G$$

denotes the **Frobenius isogeny**, i.e., the relative p^rth power Frobenius defined for any G over an $\mathbb{F}_p$-algebra.

Remark 4.1. An easy exercise shows that, when $\mathbb{F}_p \subseteq A$, there is an identity $F_* = \phi^*$ of maps $\mathcal{D}(G/A) \to \mathcal{D}(\phi^*G/A)$.

Given a complete local ring B, let $\mathrm{Def}^0(B) = \mathrm{Def}^0_{G_0}(B)$ denote the groupoid so that

- *objects* are pairs $(G, (i, \alpha) \in \mathcal{D}(G/B))$,
- *morphisms* are isomorphims $f\colon G \to G'$ such that $f_*(i, \alpha) = (i', \alpha')$.

Proposition 4.2 (Lubin-Tate [24]). *All automorphisms in* $\mathrm{Def}^0(B)$ *are identity maps (i.e.,* $\mathrm{Def}(B)$ *is equivalent to a discrete groupoid). There exists a ring* A_0 *and a natural bijection*

$$\{\textit{local homomorphisms } A_0 \to B\} \longleftrightarrow \{\textit{iso. classes of objects in } \mathrm{Def}^0(B)\}.$$

There is a (non-canonical) isomorphism $A_0 \approx \mathbb{W}_p k[\![a_1, \ldots, a_{n-1}]\!]$.

The tautological object of $\mathrm{Def}^0(A_0)$ is the **universal deformation** of G_0. It is the formal group of Morava E-theory, whose existence follows from the following.

Theorem 4.3 (Morava [30], Goerss-Hopkins-Miller [16]). *Given a formal group* G_0 *of height* n *over a perfect field, there exists an essentially unique commutative* S*-algebra* E*, which is a complex oriented cohomology theory with* $\pi_*E \approx A_0[u, u^{-1}]$ *with* $|u| = 2$*, whose formal group is the universal deformation of* G_0*.*

4.3. The "pile" of deformation structures. We enlarge the groupoid $\mathrm{Def}^0(B)$ to a category $\mathrm{Def}(B)$, with the same objects, but so that

- *morphisms* are isogenies $f\colon G \to G'$ such that $f_*(i, \alpha) = (i', \alpha')$.

To each continuous homomorphism $g\colon B \to B'$ there is an associated pullback functor $g^*\colon \mathrm{Def}(B') \to \mathrm{Def}(B)$. Thus, Def defines a (pseudo)functor

$$\{\text{complete local rings}\}^{\mathrm{op}} \to \{\text{categories}\}.$$

A *stack* is a (kind of) presheaf of groupoids. The functor Def_{G_0} gives rise to a more general kind of object, namely a presheaf of categories on the opposite category of complete local rings. This more general concept demands a new name; thus, we will speak of Def_{G_0} as the **pile**[2] of deformations of G_0 and its Frobenius isogenies.

There is a category $\mathrm{QCoh}(\mathrm{Def})$ of **quasicoherent (pre)sheaves** of modules over the structure sheaf $\mathcal{O}_{\mathrm{Def}}$ of the pile Def. An object of $\mathrm{QCoh}(\mathrm{Def})$ amounts to a choice of data $\{M_B, M_g\}$ consisting of

- for each complete local ring B, a functor $M_B\colon \mathrm{Def}(B)^{\mathrm{op}} \to \mathrm{Mod}_B$, and
- for each for each local homomorphism $g\colon B \to B'$ a natural isomorphism

$$M_g\colon B' \otimes_B M_B \Longrightarrow M_{B'} \circ g^*\colon \mathrm{Def}(B)^{\mathrm{op}} \to \mathrm{Mod}_{B'},$$

where $g^*\colon \mathrm{Def}(B) \to \mathrm{Def}(B')$ is the functor induced by base change along g,

together with coherence data equating $M_{g'g}$ with a composition of $M_{g'}$ and M_g.

Example 4.2. Given G/B, let $\omega_B(G/B)$ denote the B-module of invariant 1-forms on G. Because 1-forms can be pulled back along isogenies, this defines an object $\omega \in \mathrm{QCoh}(\mathrm{Def})$.

Example 4.3. Given G/B, let $\deg_B(G/B) := B$. To an isogeny $f\colon G \to G'$, we associate the map $\deg_B(f)\colon \deg_B(G'/B) \to \deg_B(G/B)$ induced by multiplication by the integer $\deg(f)$. This defines an object $\deg \in \mathrm{QCoh}(\mathrm{Def})$, the **degree sheaf**.

4.4. Power operations and $\mathbf{QCoh(Def)}$. Let E denote the Morava E-theory associated to our fixed formal group G_0/k.

Recall (4.1) the total power operation $P_m\colon \pi_0 R \to \pi_0 R \otimes_{\pi_0 E} E^0 B\Sigma_m$ defined for $K(n)$-local commutative E-algebras R. The function P_m is multiplicative (i.e., $P_m(ab) = P_m(a)P_m(b)$), but not additive. We may obtain a ring homomorphism by passing to the quotient $A_m = E^0 B\Sigma_m / I_{\mathrm{tr}}$ of $E^0 B\Sigma_m$ by the ideal $I_{\mathrm{tr}} \subseteq E^0 B\Sigma_m$ generated by the image of all transfers maps from inclusions of the form $\Sigma_i \times \Sigma_{m-i} \subset \Sigma_m$ with $0 < i < m$. The composite map

$$\overline{P}_m\colon \pi_0 R \xrightarrow{P_m} \pi_0 R^{B\Sigma_m^+} \approx \pi_0 R \otimes_{\pi_0 E} E^0 B\Sigma_m \to \pi_0 R \otimes_{\pi_0 E} E^0 B\Sigma_m / I_{\mathrm{tr}} \tag{4.3}$$

is a ring homomorphism.

The key to understanding power operations are the following result due to Strickland. To state it, it is useful to note that we can form a quotient category $\mathrm{Def}(B)/\sim$ of $\mathrm{Def}(B)$ by formally identifying isomorphic objects (possible exactly because there are no non-trivial automorphisms in this category), and the projection functor is an equivalence of categories.

We write $\mathrm{Def}^r(B)/\sim$ for the set of morphisms which correspond to isogenies of degree p^r. It is straightforward to show that elements of $\mathrm{Def}^r(B)/\sim$ are in bijective correspondence with pairs (G, H), where G is an object of $\mathrm{Def}^0(B)/\sim$ and $H \leq G$ is a finite subgroup scheme of rank p^k; the correspondence sends an isogeny to its kernel.

[2]This term was suggested by Matt Ando.

Theorem 4.4 (Strickland [39, 40]). *There exist complete local rings A_r (for $r \geq 0$), finite and free as A_0-modules, so that*

$$\{\textit{local homomorphisms } A_r \to B\} \longleftrightarrow \mathrm{Def}^r(B)/\sim .$$

Furthermore, there is a natural identification of rings

$$E^0 B\Sigma_{p^r}/I_{\mathrm{tr}} \approx A_r.$$

As a consequence, the functor $B \mapsto \mathrm{Def}(B)$ from complete local rings to (graded) categories is represented by a graded affine category object $\{A_r\}$ in $(\text{complete local rings})^{\mathrm{op}}$. Thus, $\mathrm{QCoh}(\mathrm{Def})$ is equivalent to a category of *comodules*, whose objects are A_0-modules M equipped with module maps

$$\psi_r \colon M \to {}^t A_r{}^s \otimes_{A_0} M,$$

which satisfy an evident coassociativity property. (There are ring maps $s, t \colon A_0 \to A_r$ corresponding to "source" and "target" in the graded category; we use superscripts to indicate the corresponding A_0-module structures on A_r.) Furthermore, the power operation maps $\overline{P}_{p^r}$ of (4.3) make $\pi_0 R$ into a comodule; i.e., (4.4) refines π_0 to a functor

$$\underline{\pi}_0 \colon h\mathrm{Com}(E)_{K(n)} \to \mathrm{QCoh}(\mathrm{Def})$$

from the homotopy category of $K(n)$-local commutative E-algebra spectra, to the category of quasi-coherent sheaves of modules on Def, so that the value of $\underline{\pi}_0(R)$ at the universal deformation in $\mathrm{Def}(A_0)$ precisely the ring $\pi_0 R$.

Remark 4.5. The existence of the functor $\underline{\pi}_0$ is essentially an observation of Ando, Hopkins, and Strickland (see [2]). A construction is given in [37].

4.5. Additional structure. The functor π_0 to sheaves on Def admits several additional refinements, which we pass over quickly. (Most are discussed in [37]; see [7] for a treatment of completion.)

- $\mathrm{QCoh}(\mathrm{Def})$ is a symmetric monoidal category, and $\underline{\pi}_0$ naturally takes values in $\mathrm{QCoh}(\mathrm{Def}, \mathrm{Com})$, the category of sheaves of commutative rings in $\mathrm{QCoh}(\mathrm{Def})$.
- There is an extension to a functor $\underline{\pi}_* \colon h\mathrm{Com}(E)_{K(n)} \to \mathrm{QCoh}^*(\mathrm{Def}, \mathrm{Com})$, where the target is a category of *graded* ring objects in $\mathrm{QCoh}(\mathrm{Def})$.
- The output of π_* is (in a suitable sense) complete with respect to the maximal ideal of A_0.
- The rings $\underline{\pi}_0 R$ satisfy a **Frobenius congruence**. We say that an object $M \in \mathrm{QCoh}(\mathrm{Def}, \mathrm{Com})$ satisfies this condition if, for all formal groups G/B with $\mathbb{F}_p \subseteq B$, the map

  $$B^{\phi} \otimes_B M_B(G, d) \approx M_B(\phi^* G, \phi^*(d)) = M_B(\phi^* G, M_*(d)) \xrightarrow{F^*} M_B(G, d)$$

  coincides with the relative pth power map on the ring $M_B(G, d)$, for any $d \in \mathcal{D}(G/B)$.

  In terms of the comodule formulation of $\mathrm{QCoh}(\mathrm{Def}, \mathrm{Com})$, this amounts to saying that the composite

  $$M \xrightarrow{\psi_1} A_1{}^s \otimes_{A_0} M \xrightarrow{\gamma \otimes \mathrm{id}} (A_0/p) \otimes_{A_0} M = M/p$$

is the pth power map on M, where $\gamma\colon A_1 \to A_0/p$ is the map representing $\mathrm{Def}^0 \to \mathrm{Def}^1$ sending $G \mapsto (F\colon G \to \phi^* G)$.

- The Frobenius congruence for $\underline{\pi}_0 R$ is *witnessed* by a non-additive operation on π_0. To state this, note that there is A_0-module homomorphism $\epsilon\colon A_1 \to A_0$ lifting $\gamma\colon A_1 \to A_0/p$. Using this, we define a homomorphism of abelian groups

$$Q\colon \pi_0 R \xrightarrow{\psi_1} A_1 \otimes_{A_0} \pi_0 R \xrightarrow{\epsilon\otimes \mathrm{id}} A_0 \otimes_{A_0} \pi_0 R = \pi_0 R,$$

which satisfies $Q(x) \equiv x^p \mod p$. The *witness* is a (non-additive) natural operation $\theta\colon \pi_0 R \to \pi_0 R$ satisfying $Q(x) = x^p + p\theta(x)$.

Remark 4.6. Mathew, Naumann, and Noel have observed [29] that the mere existence of a witness for the Frobenius congruence allows one to show that any p^r-torsion element in the homotopy of a $K(n)$-local commutative E-algebra is nilpotent. Using this together with the nilpotence theorem of Devinatz-Hopkins-Smith, they prove a conjecture of May: for any commutative S-algebra R the kernel of the Hurewicz map $\pi_* R \to H_*(R, \mathbb{Z})$ consists of nilpotent elements.

The outcome of the additional structure outlined above is that there exists a refinement of the homotopy functor $\pi_*\colon h\mathrm{Com}(E)_{K(n)} \to \mathrm{Mod}(E_*)$ to a functor

$$\underline{\pi}_*\colon h\mathrm{Com}(E_{G_0})_{K(n)} \to \mathcal{T}_{G_0}$$

to a certain algebraically defined category $\mathcal{T}_{G_0}$, whose construction depends only on the formal group G_0/k.

Example 4.4. If $G_0/k = G_m/\mathbb{F}_p$, the formal multiplicative group, then $E_{G_0} = KU_p$ is p-complete complex K-theory. The category $\mathcal{T}_{G_0}$ is the category of p-complete $\mathbb{Z}/2$-graded θ^p-rings described by Bousfield [12].

4.6. The quadratic nature of QCoh(Def). Remarkably, making calculations about objects in $\mathrm{QCoh}(\mathrm{Def})$ is far more tractable than the above suggests. This is because the representing coalgebra $\{A_r\}$ is "quadratic". This means the following: an object of $\mathrm{QCoh}(\mathrm{Def})$ is determined, up to canonical isomorphism, by its underlying module M and the structure map $\psi_1\colon M \to A_1 \otimes_{A_0} M$, which is subject to a single relation, namely that there *exists* a dotted arrow in the following diagram of $\pi_0 E$-modules:

$$\begin{array}{ccc} M & \xrightarrow{\quad\psi_1\quad} & {}^t A_1{}^s \otimes_{A_0} M \\ \vdots & & \downarrow{\scriptstyle \mathrm{id}\otimes\psi_1} \\ {}^t A_2{}^s \otimes_{A_0} M & \xrightarrow{\nabla\otimes\mathrm{id}} & {}^t A_1{}^s \otimes_{A_0} {}^t A_1{}^s \otimes_{A_0} M \end{array} \tag{4.4}$$

where ∇ encodes composition of two morphisms of degree p in Def. In particular, the category $\mathrm{QCoh}(\mathrm{Def})$ can be reconstructed using only knowledge of the rings $\pi_0 E$, A_1, and A_2, and the ring homomorphisms s, t, and ∇.

Example 4.5. *Multiplicative group.* For $G_0/k = G_m/\mathbb{F}_p$, the rings $A_r \approx \mathbb{Z}_p$ for all r. An object in $\mathrm{QCoh}(\mathrm{Def})$ amounts to a $\mathbb{Z}_p$-module M equipped with an endomorphism $\psi\colon M \to M$.

Example 4.6. *Height 2.* When E is associated to a formal group of height 2, it is possible to give a completely explicit description of $\mathrm{QCoh}(\mathrm{Def})$, by using explicit formulas obtained from the theory of elliptic curves. For instance, let G_0 be the formal completion of the supersingular elliptic curve over $\mathbb{F}_2$. In this case, $A_0 = \mathbb{Z}_2[\![a]\!]$, $A_1 = \pi_0 E[d]/(d^3 - ad - 2)$, and the ring homomorphisms $s, t\colon A_0 \to A_1$ are given by $s(a) = a$ and $t(a) = a^2 + 3d - ad^2$. The ring A_2 is the pullback of

$$A_1 \overset{t}{\otimes}_{A_0} \overset{s}{A_1} \xrightarrow{w\otimes \mathrm{id}} A_1 \xleftarrow{s} A_0$$

where $w\colon A_1 \to A_1$ sends $w(a) = t(a)$ and $w(d) = a - d^2$. (The map w classifies the operation of sending a p-isogeny of elliptic curves to its dual isogeny.) The map Δ is the evident inclusion map. The above description is outlined (admittedly in very rough form) in [33]. Zhu [44] has calculated a similar example at the prime 3.

Example 4.7. *Height 2, modulo p.* It is possible to give a uniform description of this structure at height 2, if we work modulo the prime. Fix a supersingular elliptic curve C_0 over $\mathbb{F}_p$ whose Frobenius isogeny satisfies $F^2 = -p$ (such always exist), and let G_0 be its formal completion. Then, following an observation of [21, 13.4.6], we see that $A_1/p \approx \mathbb{F}_p[\![a_1, a_2]\!]/(\,(a_1^p - a_2)(a_1 - a_2^p)\,)$, so that $s, t\colon A_0/p = \mathbb{F}_p[\![a]\!] \to A_1/p$ send $s(a) = a_1$ and $t(a) = a_2$; the rings A_r/p can be described similarly. See [38], especially §3.

In the general case, the quadraticity of $\mathrm{QCoh}(\mathrm{Def})$ is a consequence of a stronger theorem: that the algebra of power operations for Morava E-theory is Koszul.

4.7. The ring of power operations is Koszul. We observe that $\mathrm{QCoh}(\mathrm{Def})$ is equivalent to a category of modules over an associative ring $\Gamma := \bigoplus \mathrm{Hom}_{A_0}(A_r, A_0)$. In particular, it is an abelian category with enough projectives and injectives. In their work, Ando, Hopkins, and Strickland perceived that the ring Γ should have finite homological dimension (see discussion at the end of §14 in [39]). That this is so is a consequence of the following theorem.

Theorem 4.7 ([35]). *The ring Γ is Koszul, and thus objects of $\mathrm{QCoh}(\mathrm{Def}_{G_0})$ admit a functorial resolution by a "Koszul complex". More precisely, there is a functor $\mathcal{C}\colon \mathrm{QCoh}(\mathrm{Def}) \to \mathrm{Ch}(\mathrm{QCoh}(\mathrm{Def}))$ together with a natural augmentation $\epsilon\colon \mathcal{C}(M) \to M$ which is a quasi-isomorphism if M is projective as a $\pi_0 E$-module. Furthermore,*

$$\mathcal{C}_k(M) = \Gamma \otimes_{\pi_0 E} C_k \otimes_{\pi_0 E} M,$$

where C_k is a $\pi_0 E$-module which is (i) free and finitely generated as a right $\pi_0 E$-module, and (ii) $C_k = 0$ if $k > n$, where n is the height of the formal group G_0.

As a consequence, Γ has global dimension $2n$, where n is the height of the formal group.

Remark 4.8. The proof of (4.7) given in [35] is purely topological, making no reference to the interpretation of $\mathrm{QCoh}(\mathrm{Def})$ in terms of isogenies of formal groups. The proof is inspired by the theory of the Goodwillie calculus of the identity tower, and in particular by the work of Arone-Mahowald [3] on the $K(n)$-local homotopy type of the layers of the tower of the identity functor evaluated at odd spheres. They show that the $K(n)$-local homotopy type of an odd sphere is concentrated purely at Goodwillie layers p^k for $0 \le k \le n$.

Remark 4.9. The statement of (4.7) is purely a statement about deformations about formal groups and their isogenies, and thus should in particular admit a proof which does not use topology. I do not know such a proof in general, but such a purely algebro-geometric proof exists in the cases $n = 1$ and $n = 2$; see [38] for the height 2 case.

Remark 4.10. Ando, Hopkins, and Strickland originally conjectured a particular form for a finite complex such as that in terms of (4.7), in terms of a "Tits building" associated to subgroups of $\mathbb{G}_E[p]$, the p-torsion subgroup of the formal group $\mathbb{G}_E$. Their original complex in the height 2 case can be constructed using the arguments of [38]. Recently, Jacob Lurie has shown how (4.7) can be used to recover the original proposal of Ando, Hopkins, and Strickland.

4.8. Computing maps of $K(n)$-local E-algebras. We can use the theory described above to describe the E_2-term of a spectral sequence computing the space of maps between commutative E-algebras. We describe how this works in a special case.

Let R, F be two $K(n)$-local commutative E-algebras equipped with an augmentation to E. There is a spectral sequence

$$E_2^{s,t} \Longrightarrow \pi_{t-s}\mathrm{Com}(E)^{\mathrm{aug}}_{K(n)}(R, F) \tag{4.5}$$

computing homotopy groups of the derived space of maps in the category of augmented $K(n)$-local commutative E-algebras. When π_*R is *smooth* over π_*E, the E_2-term takes the form

$$E_2^{s,t} \approx \begin{cases} \mathcal{T}_{/E_*}(\pi_*R, \pi_*F) & \text{if } (s,t) = (0,0), \\ \mathrm{Ext}^s_{\mathrm{QCoh(Def)}}(\omega^{-1/2} \otimes \widehat{Q}(\pi_*R), \omega^{(t-1)/2} \otimes \pi_*\overline{F}) & \text{otherwise.} \end{cases} \tag{4.6}$$

Here $\widehat{Q}(\pi_*R)$ is the module of indecomposables of the augmented ring π_*R (completed with respect to the maximal ideal of π_*E), $\pi_*\overline{F}$ is the augmentation ideal of π_*F, and ω is the module of invariant 1-forms (4.2). In this situation, the spectral sequence is strongly convergent, and is non-zero for only finitely many values of s.

5. Tangent spaces to cochains and $\Phi_n(S^{2d-1})$

5.1. Derived indecomposables of commutative ring spectra. Fix a commutative ring A, and consider an augmented commutative A-algebra R; i.e., an A-algebra equipped with an A-algebra map $\pi\colon R \to A$. We can consider the A-module $T^*_{A,\pi}(R) := I/I^2$ of **indecomposables** with respect to the augmentation, where $I = \mathrm{Ker}(\pi)$. In geometrical terms, $\mathrm{Indec}(R)$ is the cotangent space to $\mathrm{Spec}(R)$ at the point corresponding to π. The dual module $T_\pi(R) := \underline{\mathrm{Hom}}_A(T^*_{A,\pi}(R), A)$ can be viewed as a tangent space at π.

This cotangent space construction admits a derived generalization to commutative ring spectra. Given a commutative S-algebra A, and an augmented commutative A-algebra R, there is an A-module of **derived indecomposables** constructed by Basterra [6], and which we will also denote by $T^*_A(R)$ (taking the map π to be understood).[3] We write $T_A(R) := \underline{\mathrm{Hom}}_A(T^*_A(R), A)$ for the corresponding "tangent space".

[3]This functor is also called "reduced topological Andre-Quillen cohomology".

We note two alternate descriptions of these constructions, which follow from work of Basterra and Mandell ([8], especially §2).

- The cotangent functor $T^*_A\colon \mathrm{Com}^{\mathrm{aug}}_A \to \mathrm{Mod}_A$ is a kind of stabilization functor. It is most conveniently expressed in terms of an equivalence $\mathrm{Com}^{\mathrm{aug}}_A \approx \mathrm{Com}^{\mathrm{nu}}_A$ between augmented and non-unital algebras, so that
$$T^*_A(R) \approx \mathrm{hocolim}_n\, \Omega^n_{\mathrm{nu}}\Sigma^n_{\mathrm{nu}}(I),$$
where I is the homotopy fiber of the augmentation $R \to A$ viewed as a non-unital algebra, and Σ_{nu} and Ω_{nu} are loop and suspension functors in the homotopy theory of $\mathrm{Com}^{\mathrm{nu}}_A$.
- The tangent space module can also be described as "functions to the dual numbers". That is, the underlying spaces of the spectrum $T_A(R)$ can be identified as
$$\Omega^{\infty+t}T_A(R) \approx \mathrm{Com}^{\mathrm{aug}}_A(R, A \ltimes \Omega^t A),$$
where $A \ltimes \Omega^t A$ is a split square-zero extension of A by a shift of A.

Given a based space X and a commutative S-algebra A, we can apply these constructions to the cochain algebra $A^{X_+} := \underline{\mathrm{Hom}}(\Sigma^\infty(X_+), A)$, which is a commutative A-algebra equipped with an augmentation corresponding to the basepoint of X.

Example 5.1. Take $A = H\mathbb{Q}$, the rational Eilenberg-Mac Lane spectrum. For spaces X which are simply connected and of finite-type, we have a natural isomorphism
$$\pi_* T_{H\mathbb{Q}}(H\mathbb{Q}^{X_+}) \approx \pi_*(X) \otimes \mathbb{Q}$$
from the homotopy of the tangent space spectrum to the rational homotopy groups of X. This is a modern restatement of a well-known fact of rational homotopy theory (e.g., [41, Thm. 10.1].

Example 5.2. Take $A = H\overline{\mathbb{F}}_p$, the Eilenberg-Mac Lane spectrum of the algebraic closure of $\mathbb{F}_p$. Then $T_{H\overline{\mathbb{F}}_p}(H\overline{\mathbb{F}}_p^{X_+}) \approx 0$ by [25, Prop. 3.4]. Mandell's work shows that the cochain spectrum $H\overline{\mathbb{F}}_p^{X_+}$ contains complete information about the mod p homotopy type of simply connected finite-type X, but this information cannot be extracted from the tangent space.

It turns out that for $K(n)$-local rings, the (co)tangent space behaves more like rational homology than mod p-homology, where the role of rational homotopy groups is replaced with the Bousfield-Kuhn functor.

5.2. The tangent space to cochains for $K(n)$-local rings. Let A be a $K(n)$-local commutative S-algebra. For a based space X, we can construct **comparison maps** which relate the A cohomology/homology of the spectrum $\Phi_n X$ with the tangent/cotangent space of the cochain ring A^{X_+}. These take the form
$$c^*_X\colon T^*_A(A^{X_+}) \to \underline{\mathrm{Hom}}_A(\Phi_n X, A) \quad \text{and} \quad c_X\colon A \wedge \Phi_n X \to T_A(A^{X_+}).$$

Remark 5.1. Here is an idea of how to build c^*_X (the map c_X is obtained by taking A-linear duals). Given a space X, apply Φ_n to the tautological map $u\colon X \to \Omega^\infty\Sigma^\infty X$, obtaining
$$\Phi_n X \xrightarrow{\Phi_n(u)} \Phi(\Omega^\infty\Sigma^\infty X) \approx (\Sigma^\infty X)_{K(n)}.$$

Taking functions into a $K(n)$-local ring A gives

$$\kappa_X \colon A^X \approx \underline{\mathrm{Hom}}((\Sigma^\infty X)_{K(n)}, A) \to \underline{\mathrm{Hom}}(\Phi_n X, A).$$

The object A^X is the augmentation ideal of A^{X_+}, and its stabilization as a non-unital A-algebra is $T^*_A(A^{X_+})$. The map c^*_X is constructed as the limit of the collection of maps

$$\Sigma^n_{\mathrm{nu}}(A^X) \to A^{\Omega^n X} \xrightarrow{\kappa_X} \underline{\mathrm{Hom}}_A(\Phi_n(\Omega^n X), A) \approx \Omega^{-n}\underline{\mathrm{Hom}}_A(\Phi_n X, A),$$

where we have used the fact that Φ_n commutes with Ω up to weak equivalence.

Mark Behrens and I have recently proved the following result, which shows the comparison map is an isomorphism for odd-dimensional spheres.

Theorem 5.2. *For any $K(n)$-local commutative ring A, and $X = S^{2d-1}$, the maps c_X and c^*_X induce isomorphisms in $K(n)$-homology.*

Remark 5.3. A consequence of the proof is a natural identification of E_*-modules

$$C_k \approx E_*^\wedge \partial_{p^k}(S^1)_{K(n)},$$

where C_k is the module in the Koszul resolution of (4.7), with the E-homology of the p^kth layer of the Goodwillie tower of the identity functor, evaluated at the circle S^1.

Remark 5.4. Combining (5.2) with remarks from §5.1, we see that we can use the spectral sequence (4.5) to compute $\pi_*(E \wedge \Phi_n(S^{2d+1})$. By (4.6), the E_2-term is

$$E_2^{s,t} \approx \mathrm{Ext}^s_{\mathrm{QCoh(Def)}}(\omega^{d-1}, \omega^{(t-1)/2} \otimes \mathrm{nul}) \Longrightarrow E^\wedge_{t-s}\Phi_n S^{2d-1},$$

where $\mathrm{nul} \in \mathrm{QCoh(Def)}$ is the object corresponding to the comodule $M = A_0$ whose coaction maps $\psi_r \colon M \to A_r \otimes_{A_0} M$ are identically 0.

Explicit calculations show that, for $n = 1, 2$, the only non-vanishing groups are when $s = n$, and thus this gives a complete calculation in that case. For $n = 1$, this recovers calculations of Bousfield [13]. Details are provided in [34].

Acknowledgements. The author's work described in this paper was supported by the NSF under grants DMS-0505056 and DMS-1006054, and was written while the author was in residence at the Mathematical Sciences Research Institute in Berkeley, California, during the Spring 2014 semester supported by NSF grant 093207800.

References

[1] M. Ando, A. J. Blumberg, D. Gepner, M. J. Hopkins, and C. Rezk, *Units of ring spectra, orientations, and thom spectra via rigid infinite loop space theory*, To appear in Journal of Topology.

[2] M. Ando, M. J. Hopkins, and N. P. Strickland, *The sigma orientation is an H_∞ map*, Amer. J. Math., **126** (2004), no. 2, 247–334.

[3] G. Arone and M. E. Mahowald, *The Goodwillie tower of the identity functor and the unstable periodic homotopy of spheres*, Invent. Math., **135** (1999), no. 3, 743–788.

[4] M. Ando, *Isogenies of formal group laws and power operations in the cohomology theories E_n*, Duke Math., J. **79** (1995), no. 2, 423–485.

[5] A. J. Baker, *Elliptic cohomology, p-adic modular forms and Atkin's operator U_p*, Algebraic topology (Evanston, IL, 1988), Contemp. Math., vol. 96, Amer. Math., Soc., Providence, RI, 1989, pp. 33–38.

[6] M. Basterra, *André-Quillen cohomology of commutative S-algebras*, J. Pure Appl. Algebra, **144** (1999), no. 2, 111–143.

[7] T. Barthel and M. Frankland, *Completed power operations for morava E-theory*, arXiv: 1311.7123.

[8] M. Basterra and M. A. Mandell, *Homology and cohomology of E_∞ ring spectra*, Math. Z., **249** (2005), no. 4, 903–944.

[9] R. R. Bruner, J. P. May, J. E. McClure, and M. Steinberger, *H_∞ ring spectra and their applications*, Lecture Notes in Mathematics, vol. 1176, Springer-Verlag, Berlin, 1986.

[10] A. K. Bousfield, *The localization of spaces with respect to homology*, Topology, **14** (1975), 133–150.

[11] ———, *Uniqueness of infinite deloopings for K-theoretic spaces*, Pacific J. Math., **129** (1987), no. 1, 1–31.

[12] ———, *On λ-rings and the K-theory of infinite loop spaces*, K-Theory, **10** (1996), no. 1, 1–30.

[13] ———, *The K-theory localizations and v_1-periodic homotopy groups of H-spaces*, Topology, **38** (1999), no. 6, 1239–1264.

[14] ———, *On the 2-primary v_1-periodic homotopy groups of spaces*, Topology, **44** (2005), no. 2, 381–413.

[15] A. D. Elmendorf, I. Kriz, M. A. Mandell, and J. P. May, *Rings, modules, and algebras in stable homotopy theory*, Mathematical Surveys and Monographs, vol. 47, American Mathematical Society, Providence, RI, 1997, With an appendix by M. Cole.

[16] P. G. Goerss and M. J. Hopkins, *Moduli spaces of commutative ring spectra*, Structured ring spectra, London Math. Soc. Lecture Note Ser., vol. 315, Cambridge Univ. Press, Cambridge, 2004, pp. 151–200.

[17] P. G. Goerss, *The Adams-Novikov spectral sequence and the homotopy groups of spheres*, arXiv:0802.1006.

[18] M. J. Hopkins, N. J. Kuhn, and D. C. Ravenel, *Generalized group characters and complex oriented cohomology theories*, J. Amer. Math. Soc., **13** (2000), no. 3, 553–594.

[19] M. J. Hopkins, *Algebraic topology and modular forms*, Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002) (Beijing), Higher Ed. Press, 2002, pp. 291–317.

[20] Michael J. Hopkins and Jeffrey H. Smith, *Nilpotence and stable homotopy theory. II*, Ann. of Math., (2) **148** (1998), no. 1, 1–49.

[21] N. M. Katz and B. Mazur, *Arithmetic moduli of elliptic curves*, Annals of Mathematics Studies, vol. 108, Princeton University Press, Princeton, NJ, 1985.

[22] N. J. Kuhn, *Morava K-theories and infinite loop spaces*, Algebraic topology (Arcata, CA, 1986) (Berlin), Lecture Notes in Math., vol. 1370, Springer, 1989, pp. 243–257.

[23] ———, *A guide to telescopic functors*, Homology, Homotopy Appl., **10** (2008), no. 3, 291–319.

[24] Jonathan Lubin and John Tate, *Formal moduli for one-parameter formal Lie groups*, Bull. Soc. Math., France, **94** (1966), 49–59.

[25] M. A. Mandell, *Cochains and homotopy type*, Publ. Math., Inst. Hautes Études Sci. (2006), no. 103, 213–246.

[26] J. P. May, *E_∞ ring spaces and E_∞ ring spectra*, Lecture Notes in Mathematics, Vol. 577, Springer-Verlag, Berlin-New York, 1977, With contributions by Frank Quinn, Nigel Ray, and Jørgen Tornehave.

[27] ———, *What are E_∞ ring spaces good for?*, New topological contexts for Galois theory and algebraic geometry (BIRS 2008), Geom. Topol. Monogr., vol. 16, Geom. Topol. Publ., Coventry, 2009, pp. 331–365.

[28] M. A. Mandell, J. P. May, S. Schwede, and B. Shipley, *Model categories of diagram spectra*, Proc. London Math. Soc., (3) **82** (2001), no. 2, 441–512.

[29] A. Mathew, N. Naumann, and J. Noel, *On a nilpotence conjecture of J. P. may*, arXiv: 1403.2023.

[30] J. Morava, *Completions of complex cobordism*, Geometric applications of homotopy theory (Proc. Conf., Evanston, Ill., 1977), II, Lecture Notes in Math., vol. 658, Springer, Berlin, 1978, pp. 349–361.

[31] I. Madsen, V. Snaith, and J. Tornehave, *Infinite loop maps in geometric topology*, Math. Proc. Cambridge Philos. Soc., **81** (1977), no. 3, 399–430.

[32] D. G. Quillen, *On the formal group laws of unoriented and complex cobordism theory*, Bull. Amer. Math. Soc., **75** (1969), 1293–1298.

[33] C. Rezk, *Power operations for Morava E-theory of height 2 at the prime 2*, arXiv: 0812.1320.

[34] ———, *Power operations in Morava E-theory: structure and calculations*, http:// www.math.uiuc.edu/~rezk/power-ops-ht-2.pdf.

[35] ———, *Rings of power operations for morava E-theories are Koszul*, arXiv:1204. 4831.

[36] ———, *The units of a ring spectrum and a logarithmic cohomology operation*, J. Amer. Math. Soc., **19** (2006), no. 4, 969–1014.

[37] ———, *The congruence criterion for power operations in Morava E-theory*, Homology, Homotopy Appl., **11** (2009), no. 2, 327–379.

[38] ———, *Modular isogeny complexes*, Algebr. Geom. Topol., **12** (2012), no. 3, 1373–1403.

[39] N. P. Strickland, *Finite subgroups of formal groups*, J. Pure Appl. Algebra, **121** (1997), no. 2, 161–208.

[40] ———, *Morava E-theory of symmetric groups*, Topology, **37** (1998), no. 4, 757–779.

[41] D. Sullivan, *Infinitesimal computations in topology*, Inst. Hautes Études Sci. Publ. Math., (1977), no. 47, 269–331 (1978).

[42] A. Stacey and S. Whitehouse, *Stable and unstable operations in mod p cohomology theories*, Algebr. Geom. Topol., **8** (2008), no. 2, 1059–1091.

[43] T. tom Dieck, *The Artin-Hasse logarithm for λ-rings*, Algebraic topology (Arcata, CA, 1986), Lecture Notes in Math., vol. 1370, Springer, Berlin, 1989, pp. 409–415.

[44] Y. Zhu, *The power operation structure on Morava E-theory of height 2 and prime 3*, Algebr. Geom. Topol., **14** (2014), no. 2, 953–977.

University of Illinois, Urbana, Illinois, USA
E-mail: rezk@math.uiuc.edu

Quasi-morphisms and quasi-states in symplectic topology

Michael Entov

Abstract. We discuss certain "almost homomorphisms" and "almost linea" functionals that have appeared in symplectic topology and their applications concerning Hamiltonian dynamics, functional-theoretic properties of Poisson brackets and algebraic and metric properties of symplectomorphism groups.

Mathematics Subject Classification (2010). Primary 53D35, 53D40, 53D45; Secondary 17B99, 20F69, 46L30.

Keywords. Symplectic manifold, Poisson brackets, Hamiltonian symplectomorphism, quantum homology, quasi-morphism, quasi-state, symplectic rigidity.

1. Introduction

Symplectic manifolds carry several interesting mathematical structures of different flavors, coming from algebra, geometry, topology, dynamics and analysis. In this survey we discuss certain "almost homomorphisms" (called *Calabi quasi-morphisms*) on groups of symplectomorphisms of symplectic manifolds and certain "almost linear" functionals (called *symplectic quasi-states*) on the spaces of smooth functions on symplectic manifolds that have been useful in finding new relations between these structures. In particular, we describe applications of these new tools to Hamiltonian dynamics, functional-theoretic properties of Poisson brackets as well as algebraic and metric properties of the groups of symplectomorphisms. We also briefly discuss a relation between the symplectic quasi-states and von Neumann's mathematical foundations of quantum mechanics. We end the survey with a discussion on the function theory approach to symplectic topology.

A detailed introduction to the subject can be found in the forthcoming book [58] by L.Polterovich and D.Rosen.

2. Quasi-morphisms and quasi-states - generalities

2.1. Quasi-morphisms. Let G be a group. A function $\mu : G \to \mathbb{R}$ is called a *quasi-morphism*, if there exists a constant $C > 0$ so that $|\mu(xy) - \mu(x) - \mu(y)| \leq C$ for any $x, y \in G$. We say that a quasi-morphism $\mu : G \to \mathbb{R}$ is *homogeneous*[1], if $\mu(x^k) = k\mu(x)$ for any $x \in G$, $k \in \mathbb{Z}$.

[1] Proceedings of the International Congress of Mathematicians, Seoul, 2014

Clearly, any $\mathbb{R}$-valued homomorphism on G is a homogeneous quasi-morphism but finding homogeneous quasi-morphisms that are not homomorphisms is usually a non-trivial task. Let us also note that any homogeneous quasi-morphism μ is conjugacy-invariant and satisfies $\mu(xy) = \mu(x) + \mu(y)$ for any commuting elements x, y (in particular, any homogeneous quasi-morphism on an abelian group is a homomorphism). For more on quasi-morphisms see e.g. [16].

2.2. Quasi-states and quantum mechanics. Roughly speaking, quasi-states are "almost linear" functionals on algebras of a certain kind. The term "quasi-state" comes from the work of Aarnes (see [1] and the references to the Aarnes' previous work therein) but its history goes back to the mathematical model of quantum mechanics suggested by von Neumann [72]. A basic object of this model is a real Lie algebra of observables that will be denoted by $\mathcal{A}_q$ (q for quantum): its elements (in the simplest version of the theory) are Hermitian operators on a finite-dimensional complex Hilbert space H and the Lie bracket is given by $[A, B]_\hbar = \frac{i}{\hbar}(AB - BA)$, where $\hbar$ is the Planck constant. Observables represent physical quantities such as energy, position, momentum etc. In von Neumann's model a state of a quantum system is given by a functional $\zeta : \mathcal{A}_q \to \mathbb{R}$ which satisfies the following axioms:

Additivity: $\zeta(A + B) = \zeta(A) + \zeta(B)$ for all $A, B \in \mathcal{A}_q$.

Homogeneity: $\zeta(cA) = c\zeta(A)$ for all $c \in \mathbb{R}$ and $A \in \mathcal{A}_q$.

Positivity: $\zeta(A) \geq 0$ provided $A \geq 0$..

Normalization: $\zeta(Id) = 1$.

As a consequence of these axioms von Neumann proved that for every state ζ there exists a non-negative Hermitian operator U_ζ with trace 1 so that $\zeta(A) = \mathrm{tr}(U_\zeta A)$ for all $A \in \mathcal{A}_q$. An easy consequence of this formula is that for every state ζ there exists an observable A such that

$$\zeta(A^2) - (\zeta(A))^2 > 0 \, . \tag{2.1}$$

In his book [72] von Neumann adopted a statistical interpretation of quantum mechanics according to which the value $\zeta(A)$ is considered as the expectation of a physical quantity represented by A in the state ζ. In this interpretation the equation (2.1) says that there are no dispersion-free states. This result led von Neumann to a conclusion which can be roughly described as the impossibility to present random quantum-mechanical phenomena as an observable part of some "hidden" underlying deterministic mechanism. This conclusion caused a major discussion among physicists (see e.g. [6]) some of whom disagreed with the additivity axiom of a quantum state. Their reasoning was that the formula $\zeta(A+B) = \zeta(A)+\zeta(B)$ makes sense *a priori* only if observables A and B are simultaneously measurable, that is, commute: $[A, B]_\hbar = 0$.

In 1957 Gleason [36] proved his famous theorem which can be viewed as an additional argument in favor of von Neumann's additivity axiom. Recall that two Hermitian operators on a finite-dimensional Hilbert space commute if and only if they can be written as polynomials of the same Hermitian operator. Let us define a *quasi-state* on $\mathcal{A}_q$ as an $\mathbb{R}$-valued functional which satisfies the homogeneity, positivity and normalization axioms above, while the additivity axiom is replaced by one of the two *equivalent* axioms:

[1]Sometimes homogeneous quasi-morphisms are also called *pseudo-characters* – see e.g. [66].

Quasi-additivity-I: $\zeta(A+B) = \zeta(A) + \zeta(B)$, provided A and B commute: $[A, B]_\hbar = 0$;

Quasi-additivity-II: $\zeta(A + B) = \zeta(A) + \zeta(B)$, provided A and B belong to a singly generated subalgebra of $\mathcal{A}_q$.

According to the Gleason theorem, every quasi-state ζ on $\mathcal{A}_q$ is linear, that is, a state, provided the complex dimension of the Hilbert space H is at least 3 (it is an easy exercise to show that in the two-dimensional case there are plenty of non-linear quasi-states).

Let us turn now to the mathematical model of classical mechanics. Here the algebra $\mathcal{A}_c$ of observables (c for classical) is the space $C^\infty(M)$ of smooth functions on a symplectic manifold M. The space $C^\infty(M)$ carries two structures. On one hand, it is a Lie algebra with respect to the Poisson bracket (see Section 3.1). On the other hand, it is a dense subset (in the uniform norm) of the commutative Banach algebra $C(M)$ of continuous functions on M. For both frameworks one can define its own version of the notion of a quasi-state adapting, respectively, the first or the second definition of quasi-additivity – as a result one gets the so-called *Lie quasi-states* and *topological quasi-states* (see Section 2.3).

Symplectic quasi-states that appear in symplectic topology and will be discussed below in Section 3 belong to both of these worlds – they are simultaneously Lie and topological quasi-states[2]. Note that for the Lie algebra $C^\infty(M)$ the first definition of quasi-additivity fits in with the physical Correspondence Principle according to which the bracket $[\ ,\ ]_\hbar$ corresponds to the Poisson bracket $\{\ ,\ \}$ in the classical limit $\hbar \to 0$. The existence of non-linear symplectic quasi-states on certain symplectic manifolds (see Section 3) can be viewed as an "anti-Gleason phenomenon" in classical mechanics. Interestingly, at least for $M = S^2$, the symplectic quasi-state that we construct is dispersion-free (see Example 3.3), unlike states in von Neumann's model of quantum mechanics. For more information on the connection of symplectic quasi-states to physics see [29] and Remark 4.21 below.

2.3. Lie and topological quasi-states. Here is the precise definition of a Lie quasi-state. Let $\mathfrak{g}$ be a (possibly infinite-dimensional) Lie algebra over $\mathbb{R}$ and let $W \subset \mathfrak{g}$ be a vector subspace. A function $\zeta : W \to \mathbb{R}$ is called a *Lie quasi-state*, if it is linear on every abelian subalgebra of $\mathfrak{g}$ contained in W.

Finding non-linear Lie quasi-states is, in general, a non-trivial task: for instance, the difficult Gleason theorem mentioned above is essentially equivalent – in the finite-dimensional setting – to the claim that any Lie quasi-state on the unitary Lie algebra $\mathfrak{u}(n)$, $n \geq 3$, which is bounded on a neighborhood of zero, has to be linear [24]. Choosing an appropriate regularity class of Lie quasi-states is essential for this kind of results: if $\mathfrak{g}$ is finite-dimensional, then any Lie quasi-state on $\mathfrak{g}$ which is differentiable at 0 is automatically linear while the space of all, not necessarily continuous, Lie quasi-states on $\mathfrak{g}$ might be infinite-dimensional [24].

Another source of interest to Lie quasi-states lies in their connection to quasi-morphisms on Lie groups: if $\mathfrak{g}$ is the Lie algebra of a Lie group G and $\mu : G \to \mathbb{R}$ is a homogeneous quasi-morphism continuous on 1-parametric subgroups, then *the derivative of* μ, that is, the composition of μ with the exponential map, is a Lie quasi-state on $\mathfrak{g}$, invariant under the adjoint action of G on $\mathfrak{g}$. Symplectic quasi-states that we will discuss below appear as a particular case of this construction.

Unfortunately, rather little is known about non-linear (continuous) Lie quasi-states and

[2]Interestingly, symplectic quasi-states had appeared in an infinite-dimensional setting in symplectic topology before Lie quasi-states were properly studied in the finite-dimensional setting.

their connections to quasi-morphisms in general – almost all known facts (in particular, a non-trivial description of the space of all non-linear continuous Lie quasi-states on $\mathfrak{sp}(2n, \mathbb{R})$, $n \geq 3$) and some basic open questions on the subject can be found in [24].

Let us now define the notion of a topological quasi-state – it is due to Aarnes [1] (who called it just a "quasi-state"). Let X be a compact Hausdorff topological space and let $C(X)$ be the space of continuous functions on X equipped with the uniform norm. For a function $F \in C(X)$ denote by $\mathcal{A}_F$ the closure in $C(X)$ of the set of functions of the form $p \circ F$, where p is a real polynomial. A functional $\zeta : C(X) \to \mathbb{R}$ is called a *topological quasi-state* [1], if it satisfies the following axioms:

Quasi-linearity: ζ is linear on $\mathcal{A}_F$ for every $F \in C(X)$ (in particular, ζ is homogeneous).

Monotonicity: $\zeta(F) \leq \zeta(G)$ for $F \leq G$.

Normalization: $\zeta(1) = 1$.

A linear topological quasi-state is called a *state* (similarly to states in von Neumann's model of quantum mechanics – see Section 2.2). The existence of non-linear topological quasi-states was first proved by Aarnes [1].

By the classical Riesz representation theorem, states on $C(X)$ are in one-to-one correspondence with regular Borel probability measures on X. In [1] Aarnes proved a generalized Riesz representation theorem that associates to each topological quasi-state ζ a *quasi-measure*[3] τ on X which is defined only on sets that are either open or closed and is finitely additive but not necessarily sub-additive. The relation between ζ and τ extends the relation between states and measures given by the Riesz representation theorem. In particular, if A is closed, $\tau(A)$ can be thought of as the "value" of ζ on the (discontinuous) characteristic function of A.

3. Calabi quasi-morphisms, symplectic quasi-states

3.1. Symplectic preliminaries. Referring the reader to [47] for the foundations of symplectic geometry we briefly recall the basic notions needed for the further discussion.

Let M^{2n} be a closed connected manifold equipped with a symplectic form ω, that is, a closed and non-degenerate differential 2-form. In terms of classical mechanics, M can be viewed as the phase space of a mechanical system and smooth functions on M (possibly depending smoothly on an additional parameter, viewed as time) are called Hamiltonians. Whenever we consider a time-dependent Hamiltonian we assume that it is 1-periodic in time, i.e. has the form $F : M \times S^1 \to \mathbb{R}$. Set $F_t := F(\cdot, t)$. The support of F is defined as $\operatorname{supp} F := \cup_{t \in S^1} \operatorname{supp} F_t \subset M$. We say that F is *normalized*, if $\int_M F_t \omega^n = 0$ for any $t \in S^1$.

We denote by $C(M)$ (respectively, $C^\infty(M)$), the space of continuous (respectively, smooth) functions on M and by $\|\cdot\|$ the uniform norm on these spaces: $\|F\| := \max_M |F|$.

Given a (time-dependent) Hamiltonian F, define its *Hamiltonian vector field* X_{F_t} by $\omega(\cdot, X_{F_t}) = dF_t(\cdot)$. Denote the flow of X_{F_t} by ϕ_F^t – it preserves ω and is called the *Hamiltonian flow of* F. Symplectomorphisms of M (that is, diffeomorphisms of M preserving ω) that can be included in such a flow are called *Hamiltonian symplectomorphisms* and form a

[3] Quasi-measures are sometimes also called *topological measures*.

group $Ham(M)$ which is a subgroup of the identity component $Symp_0(M)$ of the full symplectomorphism group $Symp(M)$. Its universal cover is denoted by $\widetilde{Ham}(M)$. We say that $\phi_F := \phi_F^1$ is *the Hamiltonian symplectomorphism generated by* F. The Hamiltonian F also generates an element of $\widetilde{Ham}(M)$ that will be denoted by $\widetilde{\phi}_F$: it is given by the homotopy class (with the fixed end-points) of the path ϕ_F^t, $0 \leq t \leq 1$, in $Ham(M)$.

The space of smooth functions on M will be denoted by $C^\infty(M)$. Given $F, G \in C^\infty(M)$, define the *Poisson bracket* $\{F, G\}$ by $\{F, G\} := \omega(X_G, X_F)$. Together with the Poisson bracket $C^\infty(M)$ becomes a Lie algebra whose center is $\mathbb{R}$ (the constant functions).

It is instructive to view $Ham(M)$ and $\widetilde{Ham}(M)$ as infinite-dimensional Lie groups whose Lie algebra (the algebra of time-independent Hamiltonian vector fields on M) is naturally isomorphic to $C^\infty(M)/\mathbb{R}$, or to the subalgebra of $C^\infty(M)$ formed by normalized functions, with the map $F \mapsto \phi_F$ being viewed as the exponential map and the natural action of $Ham(M)$ on $C^\infty(M)$ being viewed as the adjoint action.

Similarly to the closed case, for an open symplectic manifold (U^{2n}, ω) one can define $Ham(U)$ as the group formed by Hamiltonian symplectomorphisms generated by (time-dependent) Hamiltonians supported in U and $\widetilde{Ham}(U)$ as its universal cover. The group $\widetilde{Ham}(U)$ admits the *Calabi homomorphism* $\widetilde{Cal}_U : \widetilde{Ham}(U) \to \mathbb{R}$ defined by $\widetilde{Cal}_U(\widetilde{\phi}_F) := \int_{S^1} \int_M F_t \omega^n dt$, where $\operatorname{supp} F \subset U$. If ω is exact, $\widetilde{Cal}_U$ descends to a homomorphism $Cal_U : Ham(U) \to \mathbb{R}$. If U is an open subset of M, then there are natural inclusion homomorphisms $Ham(U) \to Ham(M)$, $\widetilde{Ham}(U) \to \widetilde{Ham}(M)$, whose images will be denoted by G_U and $\widetilde{G}_U$[4].

Let U be an open subset of M. Each $\phi \in Ham(M)$ (respectively, $\widetilde{\phi} \in \widetilde{Ham}(M)$) can be represented as a product of elements of the form $\psi\theta\psi^{-1}$ with θ lying in G_U (respectively, $\widetilde{G}_U$). Moreover, assuming that $\widetilde{Cal}_U$ descends to a homomorphism Cal_U on $Ham(U)$, one can make sure that each such θ satisfies $Cal_U(\theta) = 0$. This follows from Banyaga's fragmentation lemma [5]. Denote the minimal number of factors in such a product by $\|\phi\|_U$ (respectively, $\|\widetilde{\phi}\|_U$), if there is no condition on $Cal_U(\theta)$, and $\|\phi\|_{U,0}$, if the condition $Cal_U(\theta) = 0$ is imposed. All the norms are defined as 0 on the identity elements.

Let T^k be a torus. A *Hamiltonian T^k-action* on M is a homomorphism $T^k \to Ham(M)$. (We will always assume that such an action is effective). In such a case the action of the i-th S^1-factor of $T^k = S^1 \times \ldots \times S^1$, $i = 1, \ldots, k$, is a Hamiltonian flow generated by a Hamiltonian H_i. The Hamiltonians $H_1, \ldots, H_k$ commute with respect to the Poisson bracket. The map $\Phi = (H_1, \ldots, H_k) : M \to \mathbb{R}^k$ is called the *moment map* of the Hamiltonian T^k-action. If all H_i are normalized, we say that Φ is the normalized moment map.

A submanifold L of (M^{2n}, ω) is called *Lagrangian*, if $\dim L = n$ and $\omega|_L \equiv 0$.

A (closed) symplectic manifold (M, ω) admits a preferred class of almost complex structures compatible in a certain sense with ω. All these almost complex structures have the same first Chern class c_1, called the first Chern class of M. A closed symplectic manifold (M, ω) is called *monotone*, if $[\omega]$ and c_1 are positively proportional on spherical homology classes and *symplectically aspherical*, if $[\omega]$ vanishes on such classes.

Finally, we say that a subset $X \subset M$ is *displaceable from* $Y \subset M$ *by a group* G (where G is either $Ham(M)$, or $Symp_0(M)$, or $Symp(M)$), if there exists $\phi \in G$ such that $\phi(X) \cap \overline{Y} = \emptyset$. If X can be displaced from itself by G, we say that it is *displaceable by* G

[4]Note that the homomorphism $\widetilde{Ham}(U) \to \widetilde{Ham}(M)$ does not have to be injective and, accordingly, $\widetilde{G}_U$ does not have to be the preimage of G_U under the universal cover $\widetilde{Ham}(M) \to Ham(M)$ – see [40].

or just *displaceable*, if $G = Ham(M)$.

Consider $T^*S^1 = \mathbb{R} \times S^1$ with the coordinates (r, θ) and the symplectic form $dr \wedge d\theta$. We say that $X \subset M$ is *stably displaceable*, if $X \times \{r = 0\}$ is displaceable in $M \times T^*S^1$ equipped with the split symplectic form $\omega \oplus (dr \wedge d\theta)$. Any displaceable set is stably displaceable (but not necessarily vice versa).

3.2. Quantum homology and spectral numbers. A closed symplectic manifold M carries a rich algebraic structure called the *quantum homology* of M: additively it is just the singular homology of M with coefficients in a certain ring, while multiplicatively the quantum product is a deformation of the classical intersection product on homology which is defined using a count of certain pseudo-holomorphic spheres[5] in M. In fact, there are several possible algebraic setups for this structure – we refer the reader to [48], as well as [22, 70], for the precise definitions and more details. In any case, the resulting algebraic object is a ring with unity given by the fundamental class $[M]$. Passing, if needed (depending on M and the algebraic setup of the construction), to an appropriate subring with unity one gets a finite-dimensional commutative algebra with unity over a certain field that we will denote by $\mathcal{K}$. Typically, $\mathcal{K}$ is the field of semi-infinite (Laurent-type) power series with coefficients in a base field $\mathcal{F}$, where $\mathcal{F}$ is one of the fields $\mathbb{Z}_p$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. Abusing the terminology we will denote the latter finite-dimensional commutative algebra by $QH(M)$ and still call it the quantum homology of M.

Let us also mention the constructions of Usher [70] and Fukaya-Oh-Ohta-Ono [31] (the so-called *deformed quantum homology*) that, roughly speaking, use certain homology classes of M for an additional deformation of the quantum homology product and sometimes allow to obtain *different* finite-dimensional commutative algebras as above for a given M – abusing the terminology we will still call any of these different algebras the quantum homology of M and denote it by $QH(M)$ and emphasize the difference between them only when needed.

Given a non-zero $a \in QH(M)$ and a Hamiltonian $F : M \times S^1 \to \mathbb{R}$, one can define the *spectral number* $c(a, F)$ [53, 61] (see [51, 52, 71] for earlier versions of the construction and [68, 69] for additional important properties of the spectral numbers). It generalizes the following classical minimax quantity: given a singular non-zero (rational) homology class a of M and a continuous function F on M, consider the smallest value c of F so that a can be realized by a cycle lying in $\{F \leq c\}$ – for a smooth Morse function F this definition can be reformulated in terms of the Morse homology of F. The construction of the spectral number $c(a, F)$ is based on the same concept, where the singular homology is replaced by the quantum homology of M and the Morse homology of F is replaced by its *Floer homology*. The latter can be viewed as an infinite-dimensional version of the Morse homology for a certain functional, associated with F, on a covering of the space of free contractible loops in M, with the critical points of the functional being pre-images of contractible 1-periodic orbits of the Hamiltonian flow of F under the covering (see e.g. [48] for a detailed introduction to the subject).

If $F, G : M \times S^1 \to \mathbb{R}$ are normalized and $\widetilde{\phi}_F = \widetilde{\phi}_G$, then $c(a, F) = c(a, G)$. Thus, given $\widetilde{\phi} \in \widetilde{Ham}(M)$, one can define $c(a, \widetilde{\phi}) := c(a, F)$ for any normalized F generating $\widetilde{\phi}$.

[5] Pseudo-holomorphic spheres are (j, J)-holomorphic maps $(\mathbb{C}P^1, j) \to (M, J)$ for the standard complex structure j on $\mathbb{C}P^1$ and an almost complex structure J on M compatible with the symplectic form.

3.3. The main theorem. Assume $a \in QH(M)$ is an idempotent (for instance, $a = [M]$). **Here and further on, whenever we mention an idempotent, we assume that it is non-zero.** Define $\mu_a : \widetilde{Ham}(M) \to \mathbb{R}$ by

$$\mu_a(\widetilde{\phi}) := -\mathrm{vol}(M) \lim_{k\to+\infty} \frac{c(a, \widetilde{\phi}^k)}{k},$$

where $\mathrm{vol}(M) := \int_M \omega^n$, and $\zeta_a : C^\infty(M) \to \mathbb{R}$ by

$$\zeta_a(F) := \lim_{k\to+\infty} \frac{c(a, kF)}{k}.$$

One can check [21] that the limits exist and

$$\zeta_a(F) = \frac{\int_M F\omega^n - \mu_a(\widetilde{\phi}_F)}{\mathrm{vol}(M)}.$$

The next theorem shows that under certain conditions on $QH(M)$ and a the function μ_a is a homogeneous quasi-morphism and accordingly ζ_a is a Lie quasi-state invariant under the adjoint action of $\widetilde{Ham}(M)$ on $C^\infty(M)$, since, up to a scaling factor and an addition of a linear map invariant under the adjoint action of $\widetilde{Ham}(M)$, it is the derivative of μ_a (see Section 2.3).

We will say that $QH(M)$ is *field-split*, if it can be represented, in the category of $\mathcal{K}$-algebras, as a direct sum of two subalgebras at least one of which is a field. Such a field will be called *a field factor* of $QH(M)$.

Theorem 3.1. *Assume $QH(M)$ is field-split and a is the unity in a field factor of $QH(M)$. Then μ_a satisfies the following properties:*

(A) **(Stability)** $\int_0^1 \min_M(F_t - G_t)\, dt \le \mu_a(\widetilde{\phi}_G) - \mu_a(\widetilde{\phi}_F) \le \int_0^1 \max_M(F_t - G_t)\, dt$.

(B) *The function $\mu_a : \widetilde{Ham}(M) \to \mathbb{R}$ is a homogeneous quasi-morphism, that is,*

(B1) **(Homogeneity)** $\mu_a(\widetilde{\phi}^k) = k\mu_a(\widetilde{\phi})$ *for any $\widetilde{\phi} \in \widetilde{Ham}(M)$ and $k \in \mathbb{Z}$.*

(B2) **(Quasi-additivity)** *There exists $C > 0$ such that $|\mu_a(\widetilde{\phi}\widetilde{\psi}) - \mu_a(\widetilde{\phi}) - \mu_a(\widetilde{\psi})| \le C$ for any $\widetilde{\phi}, \widetilde{\psi} \in \widetilde{Ham}(M)$.*

(C) **(Calabi property)** *If $U \subset M$ is stably displaceable and $\operatorname{supp} F \subset U$, then $\mu_a(\widetilde{\phi}_F) = \int_0^1 \int_U F_t \omega^n dt$. In other words, the Calabi homomorphism $\widetilde{Cal}_U$ descends from $\widetilde{Ham}(U)$ to $\widetilde{G}_U \subset \widetilde{Ham}(M)$[6] and $\mu|_{\widetilde{G}_U} = \widetilde{Cal}_U$.*

At the same time, ζ_a satisfies the following properties:

(a) **(Monotonicity)** $\min_M(F - G) \le \zeta_a(F) - \zeta_a(G) \le \max_M(F - G)$ *for any $F, G \in C^\infty(M)$ and, in particular, if $F \le G$, then $\zeta_a(F) \le \zeta_a(G)$. Hence, ζ_a is 1-Lipschitz with respect to the uniform norm and extends to a functional on $C(M)$ that we will still denote by ζ_a.*

[6] This was tacitly assumed in [20].

(b) *The functional ζ_a is a Lie quasi-state, that is*

(b1) **(Homogeneity)** $\zeta_a(\lambda F) = \lambda \zeta_a(F)$ *for any* $F \in C(M)$ *and* $\lambda \in \mathbb{R}$.

(b2) **(Strong quasi-additivity)** *If* $F, G \in C^\infty(M)$ *and* $\{F, G\} = 0$, *then* $\zeta_a(F+G) = \zeta_a(F) + \zeta_a(G)$. *In fact,* ζ_a *satisfies a stronger property: for any* $F, G \in C^\infty(M)$ *one has*

$$|\zeta_a(F+G) - \zeta_a(F) - \zeta_a(G)| \leq \sqrt{2C||\{F,G\}||},$$

where $C > 0$ *is the constant from (*B2).

(c) **(Vanishing property)** *If* $U \subset M$ *is stably displaceable and* $F \in C(M)$ *with* $\operatorname{supp} F \subset U$, *then* $\zeta_a(F) = 0$.

(d) **(Normalization)** $\zeta_a(1) = 1$.

(e) **(Invariance)** $\zeta_a : C(M) \to \mathbb{R}$ *is invariant under the action of* $Symp_0(M)$ *on* $C(M)$.

The functionals $\mu_a : \widetilde{Ham}(M) \to \mathbb{R}$ and $\zeta_a : C(M) \to \mathbb{R}$ satisfying the properties listed in Theorem 3.1 are called, respectively, a *Calabi quasi-morphism* and a *symplectic quasi-state*. In particular, the restriction of a symplectic quasi-state to $C^\infty(M)$ is always a Lie quasi-state. Moreover, one can readily check that any symplectic quasi-state is also a topological quasi-state. The converse is true only if $\dim M = 2$ [21]. The quasi-measure associated to a symplectic quasi-state is $Symp_0(M)$-invariant and vanishes on stably displaceable open sets.

For an arbitrary M and an arbitrary idempotent $a \in QH(M)$ (for instance, $a = [M]$) one gets a weaker set of properties of μ_a and ζ_a.

Theorem 3.2. *Assume* $a \in QH(M)$ *is an arbitrary idempotent. Then* μ_a *satisfies the properties* (A) *and* (C) *from Theorem 3.1 and a weaker version of the properties* (B1) *and* (B2)*:*

(B1') **(Partial homogeneity)** $\mu_a(\widetilde{\phi}^k) = k\mu_a(\widetilde{\phi})$ *for any* $\widetilde{\phi} \in \widetilde{Ham}(M)$ *and* $k \in \mathbb{Z}_{\geq 0}$.

(B2') **(Partial quasi-additivity)** *Given a displaceable open set* $U \subset M$, *there exists* $C = C(\mu_a, U) \geq 0$, *so that* $|\mu_a(\widetilde{\phi}\widetilde{\psi}) - \mu_a(\widetilde{\phi}) - \mu_a(\widetilde{\psi})| \leq C \min\{\|\widetilde{\phi}\|_U, \|\widetilde{\psi}\|_U\}$ *for any* $\widetilde{\phi}, \widetilde{\psi} \in \widetilde{Ham}(M)$.

At the same time, ζ_a *satisfies the properties* (a), (c), (d), (e) *from Theorem 3.1 and a weaker version of the properties* (b1) *and* (b2)*:*

(b1') **(Partial homogeneity)** $\zeta_a(\lambda F) = \lambda \zeta_a(F)$ *for any* $F \in C(M)$ *and* $\lambda \in \mathbb{R}_{\geq 0}$.

(b2') **(Partial strong quasi-additivity)** *If* $F, G \in C^\infty(M)$ *and* $\{F, G\} = 0$ *and either* $\operatorname{supp} G$ *is displaceable or* G *is constant, then* $\zeta_a(F+G) = \zeta_a(F) + \zeta_a(G)$. *In fact,* ζ_a *satisfies a stronger property: for any* $F, G \in C^\infty(M)$ *one has*

$$|\zeta_a(F+G) - \zeta_a(F) - \zeta_a(G)| \leq \sqrt{2C||\{F,G\}||},$$

where $C > 0$ *is a constant depending on* ζ_a *and on the supports of* F *and* G.

The functionals μ_a and ζ_a satisfying the properties listed in Theorem 3.2 are called, respectively, a *partial Calabi quasi-morphism* and a *partial symplectic quasi-state*. Clearly, a genuine Calabi quasi-morphism or a genuine symplectic quasi-state is also a partial one.

Theorems 3.1 and 3.2 were proved under various additional restrictions on M in [20], [21]. In [45, 54, 68, 69] (see also [22, 23]) the restrictions were removed and the theorems were proved for new classes of closed symplectic manifolds. The stronger part of the property (b2) in Theorems 3.1 was proved in [28]. The stronger part of the property (b2') in Theorems 3.2 was proved in [57]. The Calabi and vanishing properties (C) and (c) in Theorems 3.1, 3.2 were originally proved for a displaceable U – it was later observed by Borman [10] that the displaceability assumption on U can be weakened to stable displaceability.

Examples of closed manifolds with a field-split $QH(M)$ (for an appropriate algebraic setup of $QH(M)$) include complex projective spaces (or, more generally, complex Grassmanians and symplectic toric manifolds), as well as blow-ups of symplectic manifolds – all with appropriate (and, in a certain sense, generic [70]) symplectic structures [20, 22, 31, 32, 44, 45, 54, 55, 70]. The direct product of symplectic manifolds with field-split quantum homology algebras also has this property – possibly under some additional assumptions on the manifolds, depending on the algebraic setup of the quantum homology [22]. As an example of M whose quantum homology $QH(M)$ is *not* field-split one can take any symplectically aspherical manifold – in such a case there are no pseudo-holomorphic spheres in M and hence there is no difference between quantum and singular homology. Let us emphasize that, in general, the question whether $QH(M)$ is field-split may depend not only on M but also on the algebraic setup of the quantum homology (and on a choice of the deformation in case of the deformed quantum homology).

Example 3.3 ([20]). Assume $M = S^2$ and $a = [S^2]$. Then $\zeta_a : C(S^2) \to \mathbb{R}$ is a symplectic quasi-state and its restriction to the set of smooth Morse functions on S^2 (which is dense in $C(S^2)$ in the uniform norm) can be described in combinatorial terms.

Namely, assume that the symplectic (that is, area) form ω on S^2 is normalized so that the area of S^2 is 1 and let F be a smooth Morse function on S^2. Consider the space Δ of connected components of the level sets of F as a quotient space of S^2. As a topological space Δ is homeomorphic to a tree. The function F descends to Δ and the push-forward of the measure defined by ω on S^2 yields a non-atomic Borel probability measure on Δ. There exists a unique point $x \in \Delta$ such that each connected component of $\Delta \setminus x$ has measure $\leq 1/2$ (such a point x is called *the median* of the measured tree Δ). Then $\zeta_a(F) = F(x)$.

Let us note that the symplectic quasi-state $\zeta_a : C(S^2) \to \mathbb{R}$ in Example 3.3 is *dispersion-free*, that is, satisfies $\zeta_a(F^2) = (\zeta_a(F))^2$. Equivalently, the corresponding quasi-measure takes only values 0 and 1. The following open question is of utmost importance for the study of symplectic quasi-states:

Question 3.4. *Is it true that the (partial) symplectic quasi-states constructed in Theorems 3.1 and 3.2 are always dispersion-free?*

Remark 3.5. Sometimes the (partial) Calabi quasi-morphism $\mu_a : \widetilde{Ham}(M) \to \mathbb{R}$ descends to $Ham(M)$ (that is, vanishes on $\pi_1 Ham(M)$). Abusing the notation we will denote the resulting (partial) Calabi quasi-morphism on $Ham(M)$ also by μ_a. The list of manifolds for which μ_a is known to descend to $Ham(M)$ for *all* a includes symplectically aspherical manifolds [61], complex projective spaces [20] and their monotone products [13, 20], a monotone blow-up of $\mathbb{C}P^2$ at three points and the complex Grassmannian $Gr(2, 4)$ [13].

The list of manifolds for which it is known that μ_a does *not* descend to $\mathit{Ham}(M)$ at least for *some* a includes various symplectic toric manifolds and, in particular, the monotone blow-ups of $\mathbb{C}P^2$ at one or two points [23, 54].

Let us note that if (M,ω) is monotone, the restriction of μ_a to $\pi_1\mathit{Ham}(M)$ does not depend on the choice of a (for a fixed algebraic setup of $QH(M)$) [23]. This is not necessarily true if M is not monotone [55].

Remark 3.6. For a closed connected M the group $\mathit{Ham}(M)$ is simple and the group $\widetilde{\mathit{Ham}}(M)$ is perfect [5]. Hence, these groups do not admit non-trivial homomorphisms to $\mathbb{R}$ and therefore partial Calabi quasi-morphisms on $\widetilde{\mathit{Ham}}(M)$ and $\mathit{Ham}(M)$ are never homomorphisms (they are non-trivial because of the Calabi property). Also, partial symplectic quasi-states are never linear (use a partition of unity with displaceable supports to check it). Moreover, in certain cases one can verify that a partial symplectic quasi-state ζ_a is not a genuine quasi-state (and, accordingly, μ_a is not a genuine quasi-morphism) – see Section 4.1.

Remark 3.7. Denote by $\mathcal{E}$ the collection of all open displaceable $U \subset (M,\omega)$ such that $\omega|_U$ is exact. For any $U \in \mathcal{E}$ the Calabi homomorphism $\mathit{Cal}_U : G_U \to \mathbb{R}$ is well-defined and, by Banyaga's theorem [5], the group $\operatorname{Ker}\mathit{Cal}_U$ is simple, meaning that, up to a scalar factor, Cal_U is the unique non-trivial $\mathbb{R}$-valued homomorphism on G_U continuous on 1-parametric subgroups. If $U, V \in \mathcal{E}$ and $U \subset V$, then $G_U \subset G_V$ and $\mathit{Cal}_U = \mathit{Cal}_V$ on G_U. Thus, if a partial Calabi quasi-morphism μ_a descends to $\mathit{Ham}(M)$, we get the following picture: there is a family $\mathcal{E}$ of subgroups of $\mathit{Ham}(M)$, with each subgroup $G_U \in \mathcal{E}$ carrying the unique $\mathbb{R}$-valued homomorphism Cal_U (continuous on 1-parametric subgroups), and while it is impossible to patch up all these homomorphisms into an $\mathbb{R}$-valued homomorphism on $\mathit{Ham}(M)$, it is possible to patch them up into a partial Calabi quasi-morphism μ_a (which may be non-unique).

Now let us discuss the existence and uniqueness of genuine Calabi quasi-morphisms and symplectic quasi-states on a given symplectic manifold and, in particular, the dependence of μ_a and ζ_a on a and the algebraic setup of $QH(M)$.

The set of idempotents in $QH(M)$ carries a partial order: namely, given idempotents $a, b \in QH(M)$, we write $a \succeq b$ if $ab = b$. Clearly, $[M] \geq b$ for any idempotent $b \in QH(M)$. If $a \succeq b$, then $a - b$ is also an idempotent and $a \succeq a - b$. Conversely, if $b, b' \in QH(M)$ are two idempotents such that $bb' = 0$, then $b + b'$ is also an idempotent and $b + b' \succeq b, b'$.

The following theorem follows from basic properties of spectral numbers (cf. [23], Theorem 1.5).

Theorem 3.8. *Assume $a, b \in QH(M)$ are idempotents, so that $a \succeq b$. Then*

(a) $\mu_a \leq \mu_b$, $\zeta_a \geq \zeta_b$.

(b) *If μ_a is a genuine (i.e. not only partial) Calabi quasi-morphism, then $\mu_a = \mu_b$, $\zeta_a = \zeta_b$ and thus μ_b is also a genuine Calabi quasi-morphism and ζ_b is a genuine symplectic quasi-state.*

At the same time it is possible that $a \succeq b$, μ_a is a partial Calabi quasi-morphism while μ_b is a genuine Calabi quasi-morphism – see Examples 4.13, 4.14.

In fact, different idempotents may define linearly independent Calabi quasi-morphisms and symplectic quasi-states. For instance, if M is a blow-up of $\mathbb{C}P^2$ at one point with an appropriate non-monotone symplectic structure, one can find Calabi quasi-morphisms μ_a and

μ_b on $\widetilde{Ham}(M)$ (for some idempotents $a, b \in QH(M)$) that have linearly independent restrictions to $\pi_1 Ham(M)$ [55], and if M is $S^2 \times S^2$ with a monotone symplectic structure, one can find linearly independent[7] symplectic quasi-states ζ_a, ζ_b on $C(M)$ – see Example 4.14.

Moreover, a change of the algebraic setup of $QH(M)$ may yield new Calabi quasi-morphisms and symplectic quasi-states on the same M. For instance, if $M = \mathbb{C}P^n$, one can choose algebraic setups of $QH(\mathbb{C}P^n)$, both for $\mathcal{F} = \mathbb{Z}_2$ and $\mathcal{F} = \mathbb{C}$, so that *in both cases* $QH(\mathbb{C}P^n)$ is a field and thus the *only* idempotent in $QH(\mathbb{C}P^n)$ is the unity $a = [\mathbb{C}P^n]$. By Theorem 3.1, in both cases $a = [\mathbb{C}P^n]$ defines a Calabi quasi-morphism μ_a on $\widetilde{Ham}(\mathbb{C}P^n)$ that descends to $Ham(\mathbb{C}P^n)$ (for $\mathcal{F} = \mathbb{C}$ this is proved in [20], the same proof works also for $\mathcal{F} = \mathbb{Z}_2$). However, it follows from [73] for $n = 2$ and from [50] for $n = 3$ that the symplectic quasi-states defined by $[\mathbb{C}P^n]$ in both cases are *different* (see Example 4.12).

Remark 3.9. Let us note that historically the first Calabi quasi-morphism on $\widetilde{Ham}(\mathbb{C}P^n)$ was implicitly constructed by Givental in [34, 35] in a completely different way; the fact that it is indeed a Calabi quasi-morphism was proved by Ben Simon [7] (the stability property of the quasi-morphism is proved in [12]). Givental's Calabi quasi-morphism descends from $\widetilde{Ham}(\mathbb{C}P^n)$ to $Ham(\mathbb{C}P^n)$ [64]. It would be interesting to find out whether this quasi-morphism on $Ham(\mathbb{C}P^n)$ can be expressed as μ_a for some $a \in QH(\mathbb{C}P^n)$ (for some algebraic setup of $QH(\mathbb{C}P^n)$).

Question 3.10. *Is the quasi-morphism μ_a for $a = [\mathbb{C}P^1]$ the only Calabi quasi-morphism on $Ham(\mathbb{C}P^1)$?*

Let us note that in the case of $\mathbb{C}P^1$ the symplectic quasi-state ζ_a for $a = [\mathbb{C}P^1]$, described in Example 3.3, is known to be the unique symplectic quasi-state on $C(S^2)$ [20].

In some cases the non-uniqueness can be prove by using the different algebras $QH(M)$ appearing in the deformed quantum homology construction – see [31] for examples of symplectic manifolds with infinitely many linearly independent Calabi quasi-morphisms and symplectic quasi-states constructed in this way.

Let us also mention a construction due to Borman [10, 11] (also see [12]) that allows to use a Calabi quasi-morphism on $\widetilde{Ham}(N)$ in order to build a Calabi quasi-morphism on $\widetilde{Ham}(M)$ if M is obtained from N by a symplectic reduction or if M is a symplectic submanifold of N. Using different presentations of a symplectic manifold as a symplectic reduction one can construct examples of M with infinitely many linearly independent Calabi quasi-morphisms on $\widetilde{Ham}(M)$ (the corresponding symplectic quasi-states are also linearly independent).

Let us also note that apart from the Calabi quasi-morphisms from Theorem 3.1, the Givental quasi-morphism mentioned in Remark 3.9 and the quasi-morphisms produced from them by Borman's reduction method, there are no known examples of homogeneous quasi-morphisms on $\widetilde{Ham}(M)$ satisfying the stability property (A) from Theorem 3.1, though otherwise there are many homogeneous quasi-morphisms on $\widetilde{Ham}(M)$ (that sometimes descend to $Ham(M)$) – see [19, 33, 59, 60, 65]. Let us also note that for symplectic manifolds of dimension greater than 2 the constructions above are the only currently known constructions of partial symplectic quasi-states on closed manifolds. (As it was mentioned above,

[7] Note that Calabi quasi-morphisms and symplectic quasi-states on a given manifold form convex sets respectively in the vector spaces of all homogeneous quasi-morphisms and all Lie quasi-states and their linear dependence is considered in these vector spaces.

in dimension 2 any topological quasi-state is also symplectic. There is a number of ways to construct topological quasi-states in any dimension – see e.g. [2, 41]). The basic open case for the existence of a (stable) Calabi quasi-morphism on $\widetilde{Ham}(M)$ is $M = T^2$ and for symplectic quasi-states it is $M = T^4$. It is also unknown whether any symplectic quasi-state has to come from a Calabi quasi-morphism and whether different Calabi quasi-morphisms may define the same symplectic quasi-state.

Remark 3.11. There is a straightforward extension of the notion of a (partial) Calabi quasi-morphism to the case of an open symplectic manifold M and there are several constructions of such quasi-morphisms.

First, one can consider a conformally symplectic embedding[8] of M in a closed symplectic manifold N carrying a Calabi quasi-morphism defined on $Ham(N)$ and use the homomorphism $Ham(M) \to Ham(N)$ induced by the embedding to pullback the quasi-morphism from $Ham(N)$ to $Ham(M)$. (Of course, one then has to prove that the resulting quasi-morphism on $Ham(M)$ is non-trivial). In this way one can, for instance, use conformally symplectic embeddings of a standard round ball B^{2n} in $\mathbb{C}P^n$ in order to construct a continuum of linearly independent Calabi quasi-morphisms on $Ham(B^{2n})$ [9].

There are also *intrinsic* constructions of (partial) Calabi quasi-morphisms for certain open symplectic manifolds following the lines of the construction presented above – see [42, 49] for more details as well as for extensions of the notion of a (partial) symplectic quasi-state to the open case. These constructions allow to extend many of the results mentioned in this survey to the open case.

4. Applications

4.1. Quasi-states and rigidity of symplectic intersections. A key phenomenon in symplectic topology is rigidity of intersections of subsets of symplectic manifolds: namely, sometimes a subset X of a symplectic manifold M cannot be displaced from a subset Y by $Ham(M)$ (or $Symp_0(M)$, or $Symp(M)$), even though X can be displaced from Y by a smooth isotopy. A central role in the applications of partial symplectic quasi-states is played by their connection to this phenomenon. Namely, to each partial symplectic quasi-state, and, in particular, to each idempotent $a \in QH(M)$, one can associate a certain hierarchy of non-displaceable sets in M. The interplay between the hierarchies associated to different a is an interesting geometric phenomenon in itself.

The key definitions describing the hierarchy are as follows [23]. Let $\zeta : C(M) \to \mathbb{R}$ be a partial symplectic quasi-state. We say that a closed subset $X \subset M$ is *heavy with respect to* ζ, if $\zeta(F) \geq \inf_X F$ for all $F \in C(M)$, and *superheavy with respect to* ζ, if $\zeta(F) \leq \sup_X F$ for all $F \in C(M)$. Equivalently, X is superheavy with respect to ζ, if $\zeta(F) = F(X)$ for any $F \in C(M)$ which is constant on X. If $\zeta = \zeta_a$ for an idempotent $a \in QH(M)$ (and a prefixed algebraic setup of $QH(M)$), we use the terms *a-heavy* and *a-superheavy* for the heavy and superheavy sets with respect to ζ_a. Clearly, a closed set containing a heavy/superheavy subset is itself heavy/superheavy. The basic properties of heavy and superheavy sets are summarized in the following theorems.

[8] A map $f : (M, \omega) \to (N, \Omega)$ between symplectic manifolds is called *conformally symplectic* if $f^*\Omega = c\omega$ for some non-zero constant c.

Theorem 4.1 ([23]). *Heavy and superheavy sets with respect to a fixed partial symplectic quasi-state ζ satisfy the following properties:*

(a) *Every superheavy set is heavy, but, in general, not vice versa.*

(b) *The classes of heavy and superheavy sets are Symp$_0(M)$-invariant.*

(c) *Every superheavy set has to intersect every heavy set. Therefore, in view of (b), any superheavy set cannot be displaced from any heavy set by Symp$_0(M)$. In particular, any superheavy set is non-displaceable by Symp$_0(M)$. On the other hand, two heavy sets may be disjoint.*

(d) *Every heavy subset is stably non-displaceable. However, it may be displaceable by Symp$_0(M)$.*

(e) *If ζ is a genuine (that is, not partial) symplectic quasi-state, then the classes of heavy and superheavy sets are identical: they coincide with the class of closed sets of full quasi-measure (that is, of quasi-measure 1) for the quasi-measure on M associated with ζ.*

Theorem 4.2 ([23]). *Assume that X_i is an a_i-heavy (resp. a_i-superheavy) subset of a closed connected symplectic manifold M_i for some idempotent $a_i \in QH(M_i)$, $i = 1, 2$. Then the product $X_1 \times X_2 \subset M_1 \times M_2$ is $a_1 \otimes a_2$-heavy (resp. $a_1 \otimes a_2$-superheavy)*[9].

Changing an idempotent a or changing the algebraic setup of $QH(M)$ may completely change the heaviness/superheaviness property of a set: there are examples of disjoint sets that are superheavy with respect to different idempotents in $QH(M)$ (see Example 4.14) and there is an example of a set that is $[M]$-superheavy, if $QH(M)$ is set up over $\mathcal{F} = \mathbb{Z}_2$, and is disjoint from an $[M]$-superheavy set, if $QH(M)$ is set up over $\mathcal{F} = \mathbb{C}$ (see Example 4.12).

The partial order on the set of idempotents mentioned in Section 3 yields the following relation between the corresponding collections of heavy and superheavy sets which follows immediately from Theorem 3.8 and the examples below.

Theorem 4.3. *Assume $a, b \in QH(M)$ are idempotents and $a \succeq b$. Then*

(a) *Every a-superheavy set is also b-superheavy (but not necessarily vice versa).*

(b) *Every b-heavy set is also a-heavy (but not necessarily vice versa).*

Remark 4.4. There is a natural action of *Symp* (M) on $QH(M)$. The subgroup *Symp*$_0(M)$ $\subset$ *Symp* (M) acts on $QH(M)$ trivially and this explains the *Symp*$_0(M)$-invariance of the partial symplectic quasi-states ζ_a and, accordingly, of the classes of a-heavy and a-superheavy sets. If an idempotent $a \in QH(M)$ is invariant under the action of the full group *Symp* (M) (e.g. if $a = [M]$), then *Symp*$_0(M)$ can be replaced by *Symp* (M) everywhere in Theorems 3.1, 3.2, 4.1. In particular, *any* $[M]$-superheavy set cannot be displaced from *any* a-heavy set (for *any* idempotent $a \in QH(M)$) by *Symp* (M).

Remark 4.5. The reason why every superheavy set X (with respect to a partial quasi-state ζ) must intersect every heavy set Y is very simple: If $X \cap Y = \emptyset$, pick a function F so that $F|_X \equiv 0$ and $F|_Y \equiv 1$. Then, by the definition of heavy and superheavy sets, $\zeta(F) = 0$ and $\zeta(F) \geq 1$, which is impossible.

[9] There is an analogue of the Künneth formula for quantum homology – in particular, to each pair of idempotents $a_1 \in QH(M_1)$, $a_2 \in QH(M_2)$, one can associate an idempotent $a_1 \otimes a_2 \in QH(M_1 \times M_2)$. Let us note that even if ζ_{a_1} and ζ_{a_2} are genuine symplectic quasi-states, $\zeta_{a_1 \otimes a_2}$ may be only a partial one – see Example 4.14.

Remark 4.6. For certain Lagrangian submanifolds X and Y of M one can prove the non-displaceability of X from Y by means of the Lagrangian Floer homology of the pair (X, Y) (see e.g. [30]). The advantage of this method is that, unlike Theorem 4.1, it gives a non-trivial lower bound on the number of transverse intersection points of $\phi(X)$, $\phi \in \mathit{Ham}\,(M)$, and Y. On the other hand, unlike the Lagrangian Floer theory, symplectic quasi-states allow to prove non-displaceability results for singular sets (see the examples below).

The most basic examples of heavy and superheavy sets are an equator in S^2 (that is, an embedded circle dividing S^2 into two parts of equal area) which is $[S^2]$-superheavy and a meridian in T^2 which is $[T^2]$-heavy but not $[T^2]$-superheavy, since it is displaceable by $\mathit{Symp}_0(T^2)$ – in particular, it implies that the partial symplectic quasi-state on $C(T^2)$ defined by $[T^2]$ is not a genuine symplectic quasi-state [23] (cf. Remark 3.6). The union of a meridian and a parallel in T^2 is $[T^2]$-superheavy [38]. More complicated examples come from the following constructions.

Let $\mathbb{A} \subset C^\infty(M)$ be a finite-dimensional Poisson-commutative vector subspace (meaning that $\{F, G\} = 0$ for any $F, G \in \mathbb{A}$) . The map $\Phi : M \to \mathbb{A}^*$ defined by $\langle \Phi(x), F\rangle = F(x)$ is called *the moment map* of $\mathbb{A}$. As an example of such a moment map one can consider the moment map of a Hamiltonian torus action on M, or a map $M \to \mathbb{R}^N$ whose components have disjoint supports.

A non-empty fiber $\Phi^{-1}(p)$ is called a *stem* of $\mathbb{A}$ (see [21]), if all non-empty fibers $\Phi^{-1}(q)$ with $q \neq p$ are displaceable, and a *stable stem*, if they are stably displaceable. If a subset of M is a (stable) stem of *some* finite-dimensional Poisson-commutative subspace of $C^\infty(M)$, it will be called just *a (stable) stem*. Any stem is a stable stem but possibly not vice versa [10].

Theorem 4.7 ([21, 23]). *A stable stem is superheavy with respect to any partial symplectic quasi-state ζ on $C(M)$.*

Using the partial symplectic quasi-state $\zeta_{[M]}$ we get

Corollary 4.8 ([21]). *For any finite-dimensional Poisson-commutative subspace of $C^\infty(M)$ its moment map Φ has at least one non-displaceable fiber.*

The following question is closely related to Question 3.4.

Question 4.9. *Is it true that for any finite-dimensional Poisson-commutative subspace of $C^\infty(M)$ its moment map Φ has at least one heavy fiber (at least with respect to some symplectic quasi-state ζ on $C(M)$)?*

Remark 4.10. If ζ a genuine symplectic quasi-state on $C(M)$, then Theorem 4.7 can be proved using the quasi-measure τ associated to ζ [21]. Namely, the push-forward of τ to $\mathbb{A}^*$ by the moment map Φ of a Poisson commutative subspace $\mathbb{A}$ is a Borel probability measure ν on $\mathbb{A}^*$ [21]. As we already noted above, the vanishing property of ζ implies that τ vanishes on stably displaceable open subsets of M. Therefore if a fiber $\Phi^{-1}(p)$ of Φ is a stable stem, the support of ν has to be concentrated at p, meaning that $\tau(\Phi^{-1}(p)) = 1$ or, in other words (see Theorem 4.1), $\Phi^{-1}(p)$ is superheavy with respect to ζ.

Here are a few examples of (stable) stems [21, 23]. The Lagrangian Clifford torus in $\mathbb{C}P^n$, defined as $L = \{[z_0 : \ldots : z_n] \in \mathbb{C}P^n \mid |z_0| = \ldots = |z_n|\}$, is a stem, hence $[\mathbb{C}P^n]$-superheavy (this generalizes the example of an equator in S^2 mentioned above). The

[10] There are no known examples of a stable stem that is not a stem, i.e. not a stem of *any* $\mathbb{A}$.

codimension-1 skeleton of a triangulation of a closed symplectic manifold M^{2n} all of whose $2n$-dimensional simplices are displaceable is a stem. A fiber $\Phi^{-1}(0)$ of the normalized moment map Φ of a compressible[11] Hamiltonian torus action on M is a stable stem[12]. A direct product of (stable) stems is a (stable) stem and that the image of a (stable) stem under *any* symplectomorphism is again a (stable) stem.

In case when the Hamiltonian torus action is not compressible, much less is known about (stable) displaceability of fibers of the moment map of the action (aside from the case of symplectic toric manifolds where many results have been obtained in recent years by different authors). If (M, ω) is monotone, one can explicitly find a so-called *special fiber* of the moment map Φ of the action which is a-superheavy for *any* idempotent $a \in QH(M)$ [23]. For a Hamiltonian T^n-action torus on a monotone (M^{2n}, ω) (that is, for a monotone symplectic toric manifold) the special fiber (which in this case is a Lagrangian torus) can be described in simple combinatorial terms involving the moment polytope (that is, the image of the moment map which is a convex polytope) – see [23]. It is not known whether in the latter case the special fiber is always a stem – see [46] for a detailed investigation of this question. Interestingly enough, the question whether the special fiber of the *normalized* moment map for a monotone symplectic toric manifold M coincides with the fiber over zero is related to the existence of a Kähler-Einstein metric on M – see [23, 64].

Finally, heaviness/superheaviness of Lagrangian submanifolds can be proved using various versions of Lagrangian Floer homology. Namely, to certain Lagrangian submanifolds L of M one can associate the *quantum homology* (or the *Lagrangian Floer homology*) $QH(L)$ that comes with an *open-closed map* $i_L : QH(L) \to QH(M)$ [3, 8, 30, 31]. If $i_L(x)$ is non-zero for certain $x \in QH(L)$, then L is $[M]$-heavy[13], and if $i_L(x)$ divides an idempotent $a \in QH(M)$, then L is a-superheavy – this is shown in [23] in the monotone case, cf. [8, 30, 31].

Here are a few examples where this method can be applied. Let us emphasize that the applications to specific L do depend on a proper choice of the algebraic setup for the quantum homology in each case – see e.g. Example 4.12; we will ignore this issue in the other examples below and refer the reader to [8, 23, 31] for details.

Example 4.11 ([23]). Assume that $L \subset M$ is a Lagrangian submanifold and $\pi_2(M, L) = 0$. Then L is $[M]$-heavy. Note that in this case heaviness may not be improved to superheaviness: the meridian in T^2 is $[T^2]$-heavy but not $[T^2]$-superheavy.

Example 4.12. The real projective space $\mathbb{R}P^n$, which is a Lagrangian submanifold of $\mathbb{C}P^n$, is $[\mathbb{C}P^n]$-superheavy, *as long as* $QH(\mathbb{C}P^n)$ *is set up over* $\mathcal{F} = \mathbb{Z}_2$ [8, 23]. In particular, this implies that $\mathbb{R}P^n$ is not displaceable by *Symp* $(\mathbb{C}P^n)$ from the Clifford torus (see [4, 67] for other proofs of this fact).

On the other hand, $\mathbb{R}P^n$ may not be $[\mathbb{C}P^n]$-superheavy, if $QH(\mathbb{C}P^n)$ is set up over $\mathcal{F} = \mathbb{C}$ – for $n = 2$ this follows from [73] and for $n = 3$ from [50]. This implies that the symplectic quasi-states defined by $[\mathbb{C}P^n]$ for the setups of $QH(M)$ over $\mathcal{F} = \mathbb{Z}_2$ and $\mathcal{F} = \mathbb{C}$ are different for $n = 2, 3$.

---

[11] An effective Hamiltonian T^k-action on (M, ω) is called *compressible* if the image of the homomorphism $\pi_1(T^k) \to \pi_1(Ham\,(M))$, induced by the action, is a finite group.

[12] Stable stems appearing in this way potentially may not be genuine stems.

[13] The a-heaviness for an arbitrary idempotent $a \in QH(M)$ can also be proved by the same method under a stronger non-vanishing assumption.

Example 4.13 ([23]). Consider the torus T^{2n} equipped with the standard symplectic structure $\omega = dp \wedge dq$. Let $M^{2n} = T^{2n} \sharp \overline{\mathbb{C}P^n}$ be a symplectic blow-up of T^{2n} at one point (the blow-up is performed in a small ball B around the point). Assume that the Lagrangian torus $L \subset T^{2n}$ given by $q = 0$ does not intersect B.

Then the proper transform of L is a Lagrangian submanifold of M which is $[M]$-heavy but *not* a-heavy for some other idempotent $a \in QH(M)$ (that, roughly speaking, depends on the exceptional divisor of the blow-up). In this case the functional $\zeta_{[M]}$ is a partial (but not genuine) symplectic quasi-state while ζ_a is a genuine symplectic quasi-state.

Example 4.14 ([18, 23]). Let S^2 be the standard unit sphere in $\mathbb{R}^3$ with the standard area form σ. Let $M := S^2 \times S^2$ be equipped with the symplectic form $\sigma \oplus \sigma$. Denote by x_1, y_1, z_1 and x_2, y_2, z_2 the Euclidean coordinates on the two S^2-factors.

Consider the following three Lagrangian submanifolds of M: the anti-diagonal $\Delta := \{(u, v) \in S^2 \times S^2 \ : \ u = -v\}$, the Clifford torus $L = \{z_1 = z_2 = 0\}$ and the torus $K = \{z_1 + z_2 = 0, \ x_1x_2 + y_1y_2 + z_1z_2 = -1/2\}$. Clearly, L intersects both K and Δ, while $K \cap \Delta = \emptyset$.

For a certain algebraic setup of $QH(M)$ (with $\mathcal{F} = \mathbb{C}$) the algebra $QH(M)$ is a direct sum of two fields whose unities will be denoted by a_- and a_+ (in particular, $a_- + a_+ = [M]$). The idempotents a_- and a_+ define symplectic quasi-states ζ_{a_-}, ζ_{a_+}. At the same time $\zeta_{[M]}$ is only a partial, not genuine, symplectic quasi-state. The submanifolds Δ and L are a_--superheavy, while K is not. At the same time K and L are a_+-superheavy, while Δ is not. All three sets Δ, K, L are $[M]$-heavy but L is the only one of them that is $[M]$-superheavy.

In particular, the Lagrangian tori L and K cannot be mapped into each other by any symplectomorphism of M. See [50] and the references therein for more results on the Lagrangian torus K.

For more examples of heavy and superheavy Lagrangian submanifolds obtained by means of an open-closed map see [23, 31].

4.2. Quasi-states and connecting trajectories of Hamiltonian flows. Here is an application of symplectic quasi-states to Hamiltonian dynamics. As above, we assume that M is a closed connected symplectic manifold.

Theorem 4.15 ([14]). *Let $X_0, X_1, Y_0, Y_1 \subset M$ be a quadruple of closed sets so that $X_0 \cap X_1 = Y_0 \cap Y_1 = \emptyset$ and the sets $X_0 \cup Y_0, Y_0 \cup X_1, X_1 \cup Y_1, Y_1 \cup X_0$ are all a-superheavy for an idempotent $a \in QH(M)$. Let $G \in C^\infty(M \times S^1)$ be a 1-periodic Hamiltonian with $G_t|_{Y_0} \leq 0$, $G_t|_{Y_1} \geq 1$ for all $t \in S^1$.*

Then there exists a point $x \in M$ and time moments $t_0, t_1 \in \mathbb{R}$ so that $\phi_G^{t_0}(x) \in X_0$ and $\phi_G^{t_1}(x) \in X_1$. Furthermore, $|t_0 - t_1|$ can be bounded from above by a constant depending only on a, if G is time-independent, and both on a and the oscillation $\max_{M \times S^1} G - \min_{M \times S^1} G$ of G, if G is time-dependent.

Remark 4.16. The proof of Theorem 4.15 uses the following important notion [14]: Given a quadruple X_0, Y_0, X_1, Y_1 of compact subsets of a (possibly open) symplectic manifold such that $X_0 \cap X_1 = Y_0 \cap Y_1 = \emptyset$, define $pb_4(X_0, Y_0, X_1, Y_1)$ (where pb stands for the "Poisson brackets" and 4 for the number of sets) by $pb_4(X_0, Y_0, X_1, Y_1) = \inf_{F,G} ||\{F, G\}||$, where the infimum is taken over all compactly supported smooth F, G satisfying $F|_{X_0} \leq 0$, $F|_{X_1} \geq 1$, $G|_{Y_0} \leq 0$, $G|_{Y_1} \geq 1$.

Given a quadruple X_0, Y_0, X_1, Y_1 as in Theorem 4.15, one can use the strong form of the property (b2) in Theorem 3.1 to prove the positivity of pb_4 for certain stabilizations [14]

of the four sets. The existence of a connecting trajectory of the Hamiltonian flow is then deduced from the positivity of pb_4 using an averaging argument [14].

Example 4.17 ([14]). Consider an open tubular neighborhood U of the zero-section T^n in T^*T^n. Pick $q_0, q_1 \in T^n$ and consider the open cotangent disks $D_i = T^*_{q_i}T^n \cap U$, $i = 0, 1$. Let $M = S^2 \times \ldots \times S^2$ be the product of n copies of S^2 equipped with the split symplectic structure $\omega = \sigma \oplus \ldots \oplus \sigma$. Let $Y_1 \subset M$ be the product of equators in the S^2 factors. It is a Lagrangian torus. If $\int_{S^2} \sigma$ is sufficiently large, then, by the Weinstein neighborhood theorem (see [47]), U can be symplectically identified with a tubular neighborhood of Y_1 in M. Using the identification we consider $X_i := \overline{D}_i$, $i = 0, 1$, Y_1 and U as subsets of M. Set $Y_0 := M \setminus U$.

If $\int_{S^2} \sigma$ is sufficiently large, the quadruple $X_0, Y_0, X_1, Y_1 \subset M$ satisfies the assumptions of Theorem 4.15 [14]. In particular, let $G : T^*T^n \times S^1 \to \mathbb{R}$ be a Hamiltonian supported in $U \subset T^*T^n$ which is ≥ 1 on $T^n \times S^1$. Theorem 4.15 implies the existence of a trajectory of the Hamiltonian flow of G passing through D_0 and D_1.

Switching the pair X_0, X_1 with the pair Y_0, Y_1 and applying Theorem 4.15 to the switched pairs one can show in a similar way that if $F : T^*T^n \times S^1 \to \mathbb{R}$ is a compactly supported Hamiltonian such that $F|_{D_0 \times S^1} \leq 0$, $F|_{D_1 \times S^1} \geq 1$, then there exists a trajectory of the Hamiltonian flow of F connecting the zero-section of T^*T^n with ∂U. It would be interesting to find out whether this fact can be related to the well-known Arnold diffusion phenomenon in Hamiltonian dynamics that concerns trajectories of a similar kind.

4.3. Quasi-states and C^0-rigidity of Poisson brackets. In this section we still assume that (M, ω) is a closed connected symplectic manifold.

The Poisson brackets of two smooth functions on (M, ω) depend on their first derivatives. Nevertheless, as it was first discovered in [17], the Poisson bracket displays a certain rigidity with respect to the *uniform* norm of the functions. This rigidity is best expressed in terms of the *profile function* defined as follows [14].

Equip the space $\Pi := C^\infty(M) \times C^\infty(M)$ with the product uniform metric:

$$d((F, G), (H, K)) = ||F - H|| + ||G - K||.$$

For each $s \geq 0$ define $\Pi_s := \{(H, K) \in \Pi \mid ||\{H, K\}|| \leq s\}$. In particular, Π_0 is the set of Poisson-commuting pairs. Given a pair $(F, G) \in \Pi$, define the *profile function* $\rho_{F,G} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ by $\rho_{F,G}(s) = d((F, G), \Pi_s)$.

Question 4.18. *Given a pair $(F, G) \in \Pi$, what can be said of $\rho_{F,G}(0)$? In other words, how well can (F, G) be approximated with respect to d by a Poisson-commuting pair?*

Let us note that similar approximation questions have been extensively studied for matrices – see e.g. [37] and the references therein. It follows from [25] that the sets Π_s, $s \geq 0$, are closed with respect to d and therefore $\rho_{F,G}(s) > 0$ for $s \in [0, ||\{F, G\}||)$. Symplectic quasi-states help to give a more precise answer in certain cases.

Theorem 4.19 ([14]). *Let $\zeta : C(M) \to \mathbb{R}$ be a symplectic quasi-state.*

(a) *Assume $X, Y, Z \subset M$ are closed sets that are superheavy with respect to ζ and satisfy $X \cap Y \cap Z = \emptyset$. Assume $F|_X \leq 0$, $G|_Y \leq 0$, $(F + G)|_Z \geq 1$ and at least one of the functions F, G has its range in $[0, 1]$. Then $\rho_{F,G}(0) = 1/2$ and for some positive*

constant C, independent of F, G, and for all $s \in [0; ||\{F, G\}||]$

$$\frac{1}{2} - C\sqrt{s} \leq \rho_{F,G}(s) \leq \frac{1}{2} - \frac{s}{2||\{F, G\}||}.$$

(b) *Let $X_0, X_1, Y_0, Y_1 \subset M$ be closed sets so that $X_0 \cap X_1 = Y_0 \cap Y_1 = \emptyset$ and the sets $X_0 \cup Y_0, Y_0 \cup X_1, X_1 \cup Y_1, Y_1 \cup X_0$ are all superheavy with respect to ζ. Assume $F, G \in C^\infty(M)$, $F|_{X_0} \leq 0$, $F|_{X_1} \geq 1$, $G|_{Y_0} \leq 0$, $G|_{Y_1} \geq 1$ and at least one of the functions F, G has its range in $[0, 1]$. Then $\rho_{F,G}(0) = 1/2$ and for some positive constant C, independent of F, G, and for all $s \in [0; ||\{F, G\}||]$*

$$\frac{1}{2} - Cs \leq \rho_{F,G}(s) \leq \frac{1}{2} - \frac{s}{2||\{F, G\}||}.$$

The proof of part (b) uses the fact that $pb_4(X_0, Y_0, X_1, Y_1) > 0$ (see Remark 4.16) and part (a) is based on the positivity of a similar Poisson bracket invariant $pb_3(X, Y, Z)$ – see [14] for more details, as well as for examples where the theorem can be applied, including an example where the lower bound in part (a) is asymptotically sharp. For a version of Theorem 4.19 for *iterated* Poisson brackets see [27].

Here is another fact concerning the C^0-rigidity of Poisson brackets whose proof uses partial symplectic quasi-states and the strong version of their partial quasi-additivity (see Theorem 3.2). Let $\mathcal{U} = \{U_1, \dots, U_N\}$ be a finite cover of M by displaceable open sets. Given a partition of unity $\vec{F} = \{F_1, \dots, F_N\}$ subordinated to $\mathcal{U}$ (that is, $\operatorname{supp} F_i \subset U_i$ for every i), consider the following measure of its Poisson non-commutativity:

$$\kappa(\vec{F}) := \inf_{x,y \in [-1,1]^N} \Big\| \Big\{ \sum_{i=1}^N x_i F_i, \sum_{j=1}^N y_j F_j \Big\} \Big\|,$$

where the infimum is taken over all $x = (x_1, \dots, x_N), y = (y_1, \dots, y_N) \in [-1, 1]^N$. Set $pb(\mathcal{U}) := \inf_{\vec{F}} \kappa(\vec{F})$, where the infimum is taken over all partitions of unity $\vec{F}$ subordinated to $\mathcal{U}$. We say that $\mathcal{U}$ is *dominated by an open set* $U \subset M$ if for each $i = 1, \dots, N$ there exists $\phi_i \in Ham\,(M)$ so that $U_i \subset \phi_i(U)$.

Theorem 4.20 ([57]). *Assume $\mathcal{U}$ is dominated by a displaceable open set U. Then there exists a constant $C = C(U) > 0$ so that*

$$pb(\mathcal{U}) \geq C/N^2. \tag{4.1}$$

The theorem strengthens a similar result proved previously in [28]. It is not clear whether the inequality (4.1) can be improved – see [57] for a discussion.

Remark 4.21. Amazingly, Theorem 4.20 that belongs to the mathematical formalism of classical mechanics can be used to prove results about mathematical objects of quantum nature appearing in the Berezin-Toeplitz quantization of a symplectic manifold – see [56–58].

4.4. Quasi-morphisms and metric properties of *Ham* (M). The group *Ham* (M) carries various interesting metrics. Here we will discuss how Calabi quasi-morphisms can be used to study these metrics.

The most remarkable metric on *Ham* (M) is the Hofer metric. Namely, define the *Hofer norm* $||\phi||_H$ of $\phi \in$ *Ham* (M) as $||\phi||_H = \inf_F \int_0^1 (\max_M F_t - \min_M F_t)dt$, where the infimum is taken over all (time-dependent) Hamiltonians F generating ϕ (if M is open, F is also required to be compactly supported). The *Hofer metric* is defined by $\varrho(\phi, \psi) = ||\phi\psi^{-1}||_H$. It is a deep result of symplectic topology that ϱ is a bi-invariant metric – see e.g. [47] and the references therein.

Assume *Ham* (M) admits a partial Calabi quasi-morphism μ. Then the stability property of μ (see Theorem 3.1) implies that μ is Lipschitz with respect to ϱ and therefore the diameter of *Ham* (M) with respect to ϱ is infinite [20]. Moreover, the Lipschitz property of μ with respect to ϱ allows to obtain the following result on the growth of 1-parametric subgroups of *Ham* (M) with respect to the Hofer norm.

Theorem 4.22 ([58], cf. [20]). *Assume Ham* (M) *admits a partial Calabi quasi-morphism. Then there exists a set* $\Xi \subset C^\infty(M)$ *which is* C^0*-open and* C^∞*-dense in* $C^\infty(M)$ *so that* $\lim_{t\to+\infty} \frac{||\phi_F^t||_H}{t||F||} > 0$ *for any* $F \in \Xi$.

Calabi quasi-morphisms can be also applied to the study of the metric induced by ϱ on certain spaces of Lagrangian submanifolds of M – see [39, 63].

The group *Ham* (M) also carries the C^0-topology: equip M with a distance function d, given by a Riemannian metric on M, and define the C^0-topology on *Ham* (M) as the one induced by the metric $dist(\phi, \psi) = \max_{x\in M} d(\phi(x), \psi(x))$. The relation between the C^0-topology and the Hofer metric on *Ham* (M) is rather delicate (for instance, the C^0-metric is never continuous with respect to the Hofer metric). One can use the Calabi quasi-morphisms on *Ham* (B^{2n}) (see Remark 3.11) in order to construct infinitely many linearly independent homogeneous quasi-morphisms on *Ham* (B^{2n}) that are both Lipschitz with respect to the Hofer metric and continuous in the C^0-topology [26]. This yields the following corollary answering a question of Le Roux [43]:

Corollary 4.23 ([26]). *For any* $c \in \mathbb{R}$ *the set* $\{\phi \in$ *Ham* $(B^{2n}) \mid ||\phi||_H \geq c\}$ *has a non-empty interior in the* C^0*-topology.*

See [62] for an extension of this result to a wider class of open symplectic manifolds.

(Partial) Calabi quasi-morphisms can be also applied to the study of the norms $\|\phi\|_U$ and $\|\phi\|_{U,0}$ on *Ham* (M) (see Section 3.1) and the metrics defined by them. Namely, let $U \subset M$ be a displaceable open set. Assume *Ham* (M) admits a Calabi quasi-morphism μ. A standard fact about quasi-morphisms bounded on a generating set yields that there exists a constant $C = C(\mu) > 0$ so that $\|\phi\|_{U,0} \geq C|\mu(\phi)|$ for any $\phi \in$ *Ham* (M). In particular, *Ham* (M) is unbounded with respect to the norm $\|\cdot\|_{U,0}$.

On the other hand, if $\mu :$ *Ham* $(M) \to \mathbb{R}$ is only a partial *but not genuine* Calabi quasi-morphism, it can be used to show the unboundedness of *Ham* (M) with respect to the norm $\|\cdot\|_U$ [15]. Namely, assume $X, Y = \varphi(X) \subset M$, $\varphi \in$ *Symp* $_0(M)$, are heavy[14] disjoint[15]

[14]With respect to the partial symplectic quasi-state ζ associated to μ. If X is heavy with respect to ζ, then so is $\varphi(X)$, since ζ is *Symp* $_0(M)$-invariant.

[15]This is possible only if μ is *not* a genuine Calabi quasi-morphism, since otherwise each heavy set (with respect to ζ) is also superheavy and therefore must intersect any other heavy set.

closed subsets of M and V, W are their disjoint open neighborhoods (for instance, X and Y can be two meridians on a standard symplectic torus). Let F, G be smooth functions supported, respectively, in V, W so that $F|_X \equiv 1 = \max_M F$, $G|_Y \equiv 1 = \max_M G$. One can easily show that

$$k = |\mu(\phi_F^k \phi_G^k) - \mu(\phi_F^k) - \mu(\phi_G^k)| \leq C\|\phi_F^k\|_U$$

for any $k \in \mathbb{N}$ and a constant $C > 0$ depending only on μ and U. Thus in such a case $\|\phi_F^k\|_U$ grows asymptotically linearly with k and the norm $\|\cdot\|_U$ is unbounded on $Ham\,(M)$.

Question 4.24. *Does there exist a closed symplectic manifold M for which $\|\cdot\|_U$ is bounded on $Ham\,(M)$?*

4.5. First steps of symplectic function theory – discussion. A smooth manifold M and various geometric structures on it can be described in terms of the function space $C^\infty(M)$ – for instance, subsets of M correspond to ideals in $C^\infty(M)$, tangent vectors to derivations on $C^\infty(M)$ etc. In particular, a symplectic structure on M is completely determined by the corresponding Poisson brackets on $C^\infty(M)$ which means that, in principle, any symplectic phenomenon has a counterpart in the *symplectic function theory*, that is, the function theory of the Poisson brackets. The key feature of symplectic topology is C^0-rigidity appearing in various forms for smooth objects on symplectic manifolds. Its counterpart in the symplectic function theory is the rigidity of the Poisson brackets with respect to the C^0-norm on functions – see e.g. [58] for a deduction of the foundational Eliashberg-Gromov theorem on the C^0-closedness of $Symp(M)$ from the C^0-rigidity of the Poisson brackets.

The results and methods presented in this survey show that thinking about symplectic phenomena in terms of the function theory may have a number of advantages. First, it allows to use the "Lie group - Lie algebra" connection between $\widetilde{Ham}\,(M)$ and $C^\infty(M)/\mathbb{R}$: most of the properties of symplectic quasi-states are proved using the properties of Calabi quasi-morphisms. Second, it allows to deal with singular sets (see Remark 4.6). Third, it allows to apply functional methods, like averaging, to Hamiltonian dynamics (see Remark 4.16). Fourth, it helps to find connections between symplectic topology and quantum mechanics since it is the Poisson algebra $C^\infty(M)$ that is being quantized in various quantization constructions (see e.g. Remark 4.21). Moreover, one may hope to discover new geometric and dynamical phenomena by studying the function theory of the Poisson bracket. For instance, the behavior of the profile function $\rho(t)$ as $t \to 0$ (see Section 4.3 and [14]) is obviously of interest in symplectic function theory but its geometric or dynamical implications are absolutely unclear.

Acknowledgements. Most of the material presented in this survey is based on our joint papers with Leonid Polterovich – I express my deep gratitude to Leonid for the long and enjoyable collaboration. Some of the results were obtained jointly with Paul Biran, Lev Buhovsky, Frol Zapolsky, Pierre Py and Daniel Rosen – I thank them all. The author was partially supported by the Israel Science Foundation grant # 723/10.

References

[1] Aarnes, J.F., *Quasi-states and quasi-measures*, Adv. Math. **86** (1991), 41–67.

[2] _____, *Construction of non-subadditive measures and discretization of Borel measures*, Fund. Math. **147** (1995), 213–237.

[3] Albers, P., *On the extrinsic topology of Lagrangian submanifolds*, Int. Math. Res. Not. **38** (2005), 2341–2371. Erratum: Int. Math. Res. Not. (2010), 1363–1369.

[4] Alston, G., *Lagrangian Floer homology of the Clifford torus and real projective space in odd dimensions*, J. Sympl. Geom. **9** (2011), 83–106.

[5] Banyaga, A., *Sur la structure du groupe des difféomorphismes qui préservent une forme symplectique*, Comm. Math. Helv. **53** (1978), 174–227.

[6] Bell, J.S., *On the problem of hidden variables in quantum mechanics*, Rev. Modern Phys. **38** (1966), 447–452.

[7] Ben Simon, G., *The nonlinear Maslov index and the Calabi homomorphism*, Comm. Contemp. Math. **9** (2007), 769–780.

[8] Biran, P. and Cornea, O., *Rigidity and uniruling for Lagrangian submanifolds*, Geom. Topol. **13** (2009), 2881–2989.

[9] Biran, P., Entov, M., and Polterovich, L., *Calabi quasimorphisms for the symplectic ball*, Comm. Contemp. Math. **6** (2004), 793–802.

[10] Borman, M.S., *Symplectic reduction of quasi-morphisms and quasi-states*, J. Sympl. Geom. **10** (2012), 225–246.

[11] _____, *Quasi-states, quasi-morphisms, and the moment map*, Int. Math. Res. Not. (2013), 2497–2533.

[12] Borman, M.S. and Zapolsky, F., *Quasi-morphisms on contactomorphism groups and contact rigidity*, preprint, arXiv:1308.3224, 2013.

[13] Branson, M., *Symplectic manifolds with vanishing action-Maslov homomorphism*, Algebr. Geom. Topol. **11** (2011), 1077–1096.

[14] Buhovsky, L., Entov, M., and Polterovich, L., *Poisson brackets and symplectic invariants*, Selecta Math. (N.S.) **18** (2012), 89–157.

[15] Burago, D., Ivanov, S., and Polterovich, L., *Conjugation-invariant norms on groups of geometric origin, in Groups of Diffeomorphisms*: In Honor of Shigeyuki Morita on the Occasion of His 60th Birthday. Adv. Studies in Pure Math. 52, Math. Soc. of Japan, Tokyo, 2008.

[16] Calegari, D., *scl*. MSJ Memoirs 20, Math. Soc. of Japan, Tokyo, 2009.

[17] Cardin, F. and Viterbo, C., *Commuting Hamiltonians and Hamilton-Jacobi multi-time equations*, Duke Math. J. **144** (2008), 235–284.

[18] Eliashberg, Y. and Polterovich, L., *Symplectic quasi-states on the quadric surface and Lagrangian submanifolds*, preprint, arXiv:1006.2501, 2010.

[19] Entov, M., *Commutator length of symplectomorphisms*, Comm. Math. Helv. **79** (2004), 58–104.

[20] Entov, M. and Polterovich, L., *Calabi quasimorphism and quantum homology*, Int. Math. Res. Not. **30** (2003), 1635-1676.

[21] ———, *Quasi-states and symplectic intersections*, Comm. Math. Helv. **81** (2006), 75–99.

[22] Entov, M. and Polterovich, L., *Symplectic quasi-states and semi-simplicity of quantum homology*, in Toric Topology (eds. M.Harada, Y.Karshon, M.Masuda and T.Panov), 47–70. Contemp. Math. 460, AMS, Providence RI, 2008.

[23] ———, *Rigid subsets of symplectic manifolds*, Compositio Math. **145** (2009), 773–826.

[24] ———, *Lie quasi-states*, J. Lie Theory **19** (2009), 613–637.

[25] ———, *C^0-rigidity of Poisson brackets*, in Proceedings of the Joint Summer Research Conference on Symplectic Topology and Measure-Preserving Dynamical Systems (eds. A. Fathi, Y.-G. Oh and C. Viterbo), 25–32. Contemp. Math. 512, AMS, Providence RI, 2010.

[26] Entov, M., Polterovich, L., and Py, P., *On continuity of quasimorphisms for symplectic maps,* With an appendix by Michael Khanevsky, in Perspectives in analysis, geometry, and topology, 169–197. Progr. Math. 296, Birkhauser/Springer, New York, 2012.

[27] Entov, M., Polterovich L., and Rosen, D., *Poisson brackets, quasi-states and symplectic integrators*, Discr. and Cont. Dyn. Systems **28** (2010), 1455–1468.

[28] Entov, M., Polterovich, L., and Zapolsky, F., *Quasi-morphisms and the Poisson bracket*, Pure and Appl. Math. Quarterly **3** (2007), 1037–1055.

[29] ———, *An anti-Gleason phenomenon and simultaneous measurements in classical mechanics*, Found. of Physics **37** (2007), 1306–1316.

[30] Fukaya, K., Oh, Y.-G., Ohta, H., and Ono, K., *Lagrangian intersection Floer theory: anomaly and obstruction Parts I, II.*, AMS, Providence, RI, International Press, Somerville, MA, 2009.

[31] ———, *Spectral invariants with bulk, quasimorphisms and Lagrangian Floer theory*, arXiv:1105.5123, 2011.

[32] Galkin, S., *The conifold point*, preprint, arXiv:1404.7388, 2014.

[33] Gambaudo, J.-M. and Ghys, E., *Commutators and diffeomorphisms of surfaces*, Erg. Th. Dyn. Sys. **24** (2004), 1591–1617.

[34] Givental, A., *The nonlinear Maslov index*, in Geometry of low-dimensional manifolds, 2 (Durham, 1989), 35–43. London Math. Soc. Lecture Note Ser. 151, Cambridge Univ. Press, Cambridge, 1990.

[35] ———, *Nonlinear generalization of the Maslov index*, in Theory of singularities and its applications, 71–103. Adv. Soviet Math. 1, AMS, Providence, RI, 1990.

[36] Gleason, A.M., *Measures on the closed subspaces of a Hilbert space*, J. Math. Mech. **6** (1957), 885–893.

[37] Hastings, M.B., *Making almost commuting matrices commute*, Comm. Math. Phys. **291** (2009), 321–345.

[38] Kawasaki, M., *Superheavy subsets and noncontractible Hamiltonian circle actions*, preprint, 2013.

[39] Khanevsky, M., *Hofer's metric on the space of diameters*, J. Topol. and Analysis **1** (2009), 407–416.

[40] Kislev, A., *Compactly supported Hamiltonian loops with non-zero Calabi invariant*, preprint, arXiv:1310.1555, 2013.

[41] Knudsen, F.F., *Topology and the construction of extreme quasi-measures*, Adv. Math. **120** (1996), 302–321.

[42] Lanzat, S., *Quasi-morphisms and symplectic quasi-states for convex symplectic manifolds*, Int. Math. Res. Not. (2013), 5321–5365.

[43] Le Roux, F., *Six questions, a proposition and two pictures on Hofer distance for Hamiltonian diffeomorphisms on surfaces*, in Symplectic topology and measure preserving dynamical systems, 33–40. Contemp. Math. 512, AMS, Providence, RI, 2010.

[44] Maydanskiy, M. and Mirabelli, B.P., *Semisimplicity of the quantum cohomology for smooth Fano toric varieties associated with facet symmetric polytopes*, Electron. Res. Announc. Math. Sci. **18** (2011), 131–143.

[45] McDuff, D., *Monodromy in Hamiltonian Floer theory*, Comment. Math. Helv. **85** (2010), 95–133.

[46] ———, *Displacing Lagrangian toric fibers via probes*, in Low-dimensional and symplectic topology, 131–160. Proc. Sympos. Pure Math. 82, AMS, Providence, RI, 2011.

[47] McDuff, D. and Salamon, D., *Introduction to symplectic topology*, Second edition. Oxford Univ. Press, New York, 1998.

[48] ———, *J-holomorphic curves and symplectic topology*, Second edition. AMS Colloquium Publ. 52, AMS, Providence, RI, 2012.

[49] Monzner, A., Vichery, N., and Zapolsky, F., *Partial quasimorphisms and quasistates on cotangent bundles, and symplectic homogenization*, J. Mod. Dyn. **6** (2012), 205–249.

[50] Oakley, J. and Usher, M., *On certain Lagrangian submanifolds of $S^2 \times S^2$ and $\mathbb{C}P^n$*, preprint, arXiv:1311.5152, 2013.

[51] Oh, Y.-G., *Symplectic topology as the geometry of action functional I*, J. Diff. Geom. **46** (1997), 499–577.

[52] ———, *Symplectic topology as the geometry of action functional II*, Comm. Analysis Geom. **7** (1999), 1–55.

[53] ———, *Construction of spectral invariants of Hamiltonian paths on general symplectic manifolds*, in The breadth of symplectic and Poisson geometry, 525–570. Birkhäuser, Boston, 2005.

[54] Ostrover, Y., *Calabi quasi-morphisms for some non-monotone symplectic manifolds*, Algebr. Geom. Topol. **6** (2006), 405–434.

[55] Ostrover, Y. and Tyomkin, I., *On the quantum homology algebra of toric Fano manifolds*, Selecta Math. (N.S.) **15** (2009), 121–149.

[56] Polterovich, L., *Quantum unsharpness and symplectic rigidity*, Lett. Math. Phys. **102** (2012), 245–264.

[57] ———, *Symplectic geometry of quantum noise*, preprint, arXiv:1206.3707, 2012. To appear in Comm. Math. Phys.

[58] Polterovich, L. and Rosen., D., *Function theory on symplectic manifolds*, book draft, 2014.

[59] Py, P., *Quasi-morphismes et invariant de Calabi*, Ann. Sci. École Norm. Sup. (4) **39** (2006), 177–195.

[60] ———, *Quasi-morphismes de Calabi et graphe de Reeb sur le tore*, C. R. Math. Acad. Sci. Paris **343** (2006), 323–328.

[61] Schwarz, M., *On the action spectrum for closed symplectically aspherical manifolds*, Pacific J. Math. **193** (2000), 419–461.

[62] Seyfaddini, S., *Descent and C^0-rigidity of spectral invariants on monotone symplectic manifolds*, J. Topol. Analysis **4** (2012), 481–498.

[63] ———, *Unboundedness of the Lagrangian Hofer distance in the Euclidean ball*, preprint, arXiv:1310.1057, 2013.

[64] Shelukhin, E., *Remarks on invariants of Hamiltonian loops*, J. Topol. Analysis **2** (2010), 277–325.

[65] ———, *The action homomorphism, quasimorphisms and moment maps on the space of compatible almost complex structures*, preprint, arXiv:1105.5814, 2011.

[66] Shtern, A.I., *The Kazhdan-Milman problem for semisimple compact Lie groups*, Russian Math. Surveys **62** (2007), 113–174.

[67] Tamarkin, D., *Microlocal condition for non-displaceablility*, preprint, arXiv:0809.1584, 2008.

[68] Usher, M., *Spectral numbers in Floer theories*, Compositio Math. **144** (2008), 1581–1592.

[69] ———, *Duality in filtered Floer-Novikov complexes*, J. Topol. Analysis **2** (2010), 233–258.

[70] ———, *Deformed Hamiltonian Floer theory, capacity estimates and Calabi quasi-morphisms*, Geom. Topol. **15** (2011), 1313–1417.

[71] Viterbo, C., *Symplectic topology as the geometry of generating functions*, Math. Ann. **292** (1992), 685–710.

[72] Von Neumann, J., *Mathematical foundations of quantum mechanics*, Princeton University Press, Princeton, 1955. (Translation of Mathematische Grundlagen der Quantenmechanik. Springer, Berlin, 1932.)

[73] Wu, W., *On an exotic Lagrangian torus in* $\mathbb{C}P^2$, preprint, arXiv:1201.2446.

Department of Mathematics, Technion – Israel Institute of Technology, Haifa 32000, Israel.
E-mail: entov@math.technion.ac.il

Representation stability

Benson Farb

Abstract. Representation stability is a phenomenon whereby the structure of certain sequences X_n of spaces can be seen to stabilize when viewed through the lens of representation theory. In this paper I describe this phenomenon and sketch a framework, the theory of FI-modules, that explains the mechanism behind it.

Mathematics Subject Classification (2010). 11T06, 14F20, 55N99.

Keywords. Configuration space, symmetric group, cohomology, representation, character.

1. Introduction

Sequences V_n of representations of the symmetric group S_n occur naturally in topology, combinatorics, algebraic geometry and elsewhere. Examples include the cohomology of configuration spaces $\mathrm{Conf}_n(M)$, moduli spaces of n-pointed Riemann surfaces and congruence subgroups $\Gamma_n(p)$; spaces of polynomials on rank varieties of $n \times n$ matrices; and n-variable diagonal co-invariant algebras.

Any S_n-representation is a direct sum of irreducible representations. These are parameterized by partitions of n. Following a 1938 paper of Murnaghan, one can pad a partition $\lambda = \sum_{i=1}^{r} \lambda_i$ of any number d to produce a partition $(n - |\lambda|) + \lambda$ of n for all $n \geq |\lambda| + \lambda_1$. The decomposition of V_n into irreducibles thus produces a sequence of multiplicities of partitions λ, recording how often λ appears in V_n.

A few years ago Thomas Church and I discovered that for many important sequences V_n arising in topology including the examples mentioned above, these multiplicities become constant once n is large enough. With Jordan Ellenberg and Rohit Nagpal, we built a theory to explain this stability, converting it to a finite generation property for a single object. We applied this to prove stability in these and many other examples. As a consequence, the character of V_n is given (for all $n \gg 1$) by a single polynomial, called a *character polynomial*, studied by Frobenius but not so widely known today. One of the main points of our work is that the mechanism for this stability comes from a common structure underlying all of these examples.

After giving an overview of this theory, we explain how it applies and connects to an array of counting problems for polynomials over finite fields, and for maximal tori in the finite groups $\mathrm{GL}_n \mathbb{F}_q$. In particular, the stability of such counts reflects, and is reflected in, the representation stability of the cohomology of an associated algebraic variety. We begin with a motivating example.

Proceedings of the International Congress of Mathematicians, Seoul, 2014

2. Configuration spaces and representation theory

Let M be any connected, oriented manifold. For any $n \geq 1$ let $\mathrm{Conf}_n(M)$ be the space of configurations of ordered n-tuples of distinct points in M:

$$\mathrm{Conf}_n(M) := \{(z_1, \ldots, z_n) \in M^n : z_i \neq z_j \text{ if } i \neq j\}.$$

The symmetric group S_n acts freely on $\mathrm{Conf}_n(M)$ by permuting the coordinates:

$$\sigma \cdot (z_1, \ldots, z_n) := (z_{\sigma(1)}, \ldots, z_{\sigma(n)}).$$

This action induces for each $i \geq 0$ an action of S_n on the complex vector space $H^i(\mathrm{Conf}_n(M); \mathbb{C})$, making $H^i(\mathrm{Conf}_n(M); \mathbb{C})$ into an S_n-representation. Here we have chosen $\mathbb{C}$ coefficients for simplicity of exposition.

2.1. Cohomology of configuration spaces. The study of configuration spaces and their cohomology is a classical topic. We concentrate on the following fundamental problem.

Problem 2.1 (Cohomology of configuration spaces). Let M be a connected, oriented manifold. Fix a ring R. Compute $H^*(\mathrm{Conf}_n(M); R)$ as an S_n-representation.

Problem 2.1 was considered in special cases by Brieskorn, F. Cohen, Stanley, Orlik, Lehrer-Solomon and many others; see, e.g. [20] and the references contained therein.

What exactly does "compute as an S_n-representation" mean? Well, by Maschke's Theorem, any S_n-representation over $\mathbb{C}$ is a direct sum of irreducible S_n-representations. In 1900, Alfred Young gave an explicit bijection between the set of (isomorphism classes of) irreducible S_n-representations and the set of partitions $n = n_1 + \cdots + n_r$ of n with $n_1 \geq \cdots \geq n_r > 0$.

Let $\lambda = (a_1, \ldots, a_r)$ be an r-tuple of integers with $a_1 \geq \cdots \geq a_r > 0$ and such that $n - \sum a_i \geq a_1$. We denote by $V(\lambda)$, or $V(\lambda)_n$ when we want to emphasize n, the representation in Young's classification corresponding to the partition $n = (n - \sum_{i=1}^r a_i) + a_1 + \cdots + a_r$. With this terminology we have, for example, that $V(0)$ is the trivial representation, $V(1)$ is the $(n-1)$-dimensional irreducible representation $\{(z_1, \ldots, z_n) \in \mathbb{C}^n : \sum z_i = 0\}$, and $V(1,1,1)$ is the irreducible representation $\bigwedge^3 V(1)$. For each $n \geq 1$ and each $i \geq 0$ we can write

$$H^i(\mathrm{Conf}_n(M); \mathbb{C}) = \bigoplus_{\lambda} d_{i,n}(\lambda) V(\lambda)_n \tag{2.1}$$

for some integers $d_{i,n}(\lambda) \geq 0$. The sum on the right-hand side of (2.1) is taken over all partitions λ of numbers $\leq n$ for which $V(\lambda)_n$ is defined. The coefficient $d_{i,n}(\lambda)$ is called the *multiplicity* of $V(\lambda)$ in $H^i(\mathrm{Conf}_n(M); \mathbb{C})$.

Problem 2.1 over $\mathbb{C}$, restated: Compute the multiplicities $d_{i,n}(\lambda)$.

Why should we care about solving this problem? Here are a few reasons:

1. Even the multiplicity $d_{i,n}(0)$ of the trivial representation $V(0)$ is interesting: it computes the i^{th} Betti number of the space $\mathrm{UConf}_n(M) := \mathrm{Conf}_n(M)/S_n$ of unordered n-tuples of distinct points in M. In other words,

$$\dim_{\mathbb{C}} H^i(\mathrm{UConf}_n(M); \mathbb{C}) = d_{i,n}(0). \tag{2.2}$$

This follows from transfer applied to the finite cover $\mathrm{Conf}_n(M) \to \mathrm{UConf}_n(M)$.

2. More generally, the $d_{i,n}(\lambda)$ for other partitions λ of n compute the Betti numbers of other (un)labelled configuration spaces. For example, for fixed $a, b, c \geq 0$, consider the space $\mathrm{Conf}_n(M)[a, b, c]$ of configurations of n distinct labelled points on M where one colors a of the points blue, b red, and c yellow, and where points of the same color are indistinguishable from each other. Then $H_i(\mathrm{Conf}_n(M)[a, b, c]; \mathbb{C})$ can be determined from $d_{i,n}(\mu)$ for certain $\mu = \mu(a, b, c)$. See [6] for a discussion, and see [32] for an explanation of how these spaces arise naturally in algebraic geometry.
3. The representation theory of S_n provides strong constraints on the possible values of $\dim_{\mathbb{C}} H^i(\mathrm{Conf}_n(M); \mathbb{C})$. As a simple example, if the action of S_n on $H^i(\mathrm{Conf}_n(M); \mathbb{C})$ is *essential* in a specific sense (cf. §2.3) for $n \gg 1$, then one can conclude for purely representation-theoretic reasons that $\lim_{n\to\infty} \dim_{\mathbb{C}} H^i(\mathrm{Conf}_n(M); \mathbb{C}) = \infty$. This happens for example for every $i \geq 1$ when $M = \mathbb{C}$. See §2.3 below.
4. For certain special smooth projective varieties M, the multiplicities $d_{i,n}(\lambda)$ encode and are encoded by delicate information about the combinatorial statistics of the $\mathbb{F}_q$-points of M or related varieties; see §5 below for two specific applications.
5. The decomposition (2.1) can have geometric meaning, and can point the way for us to guess at meaningful topological invariants. We now discuss this in a specific example.

2.2. A case study: the invariants of loops of configurations. Consider the special case where M is the complex plane $\mathbb{C}$. Elements of $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ are homomorphisms $\pi_1(\mathrm{Conf}_n(\mathbb{C})) \to \mathbb{C}$. Computing $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ is thus answering the basic question:

What are the ways of attaching a complex number to each loop of configurations of n points in the plane, in a way that is natural (= additive)?

To construct examples, pick $1 \leq i, j \leq n$ with $i \neq j$. Given any loop $\gamma(t) = (z_1(t), \ldots, z_n(t))$ in $\mathrm{Conf}_n(\mathbb{C})$, we can ignore all points except for $z_i(t)$ and $z_j(t)$ and measure how much $z_j(t)$ winds around $z_i(t)$; namely we let $\alpha_{ij} : [0, 1] \to \mathbb{C}$ be the loop $\alpha_{ij}(t) := z_j(t) - z_i(t)$ and set

$$\omega_{ij}(\gamma) := \frac{1}{2\pi i} \int_{\alpha_{ij}} \frac{dz}{z}$$

It is easy to verify that $\omega_{ij} : \pi_1(\mathrm{Conf}_n(\mathbb{C})) \to \mathbb{C}$ is indeed a homomorphism, that $\omega_{ij} = \omega_{ji}$, and that the set $\{\omega_{ij} : i < j\}$ is linearly independent in $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$; see Figure 2.1.

Linear combinations of the ω_{ij} are in fact the only natural invariants of loops of configurations in $\mathbb{C}$.

$$\omega_{ij}(\alpha) = 1, \text{ but } \omega_{k\ell}(\alpha) = 0, \text{ for all other } k, \ell.$$

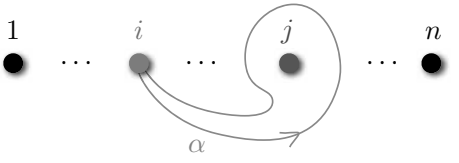


Figure 2.1. The proof that $\{\omega_{ij} : i < j\}$ is linearly independent in $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$.

Theorem 2.2 (Artin(1925), Arnol'd(1968) [1]). *The set* $\{\omega_{ij} : 1 \leq i < j \leq n\}$ *is a basis for* $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ *for any* $n \geq 2$. *Thus* $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C}) \approx \mathbb{C}^{\binom{n}{2}}$.

There is more to say. The S_n action on $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ is determined by its action on the basis via $\sigma \cdot \omega_{ij} = \omega_{\sigma(i)\sigma(j)}$, from which we can deduce that

$$H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C}) = V(0) \oplus V(1) \oplus V(2) \quad \text{for } n \geq 4 \tag{2.3}$$

using only elementary representation theory. We can see from this algebraic picture that the subspace of vectors fixed by all of S_n is 1-dimensional, spanned by the vector

$$\Omega := \sum_{1 \leq i < j \leq n} \omega_{ij} \in H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C}).$$

This implies the following geometric statement: the only natural invariant of loops of configurations of n distinct *unordered* points in $\mathbb{C}$ is total winding number Ω; in particular, $H^1(\mathrm{Conf}_n(\mathbb{C})/S_n; \mathbb{C}) \approx \mathbb{C}$.

Looking again at (2.3) we see a copy of the standard permutation representation $\mathbb{C}^n = V(0) \oplus V(1)$ given by $\sigma \cdot u_i = u_{\sigma(i)}$, with $u_i = \sum_{j \neq i} \omega_{ij}$. This indicates that the u_i should be geometrically meaningful, which indeed they are: u_i gives the total winding number of all points z_j around the point z_i.

I hope that even the simple example of $\mathrm{Conf}_n(\mathbb{C})$ convinces the reader that understanding $H^i(\mathrm{Conf}_n(M); \mathbb{C})$ as an S_n-representation and not just as a naked vector space gives us a much richer geometric picture.

2.3. Homological (in)stability. The above discussion fits in to a broader context. Let X_n be a sequence of spaces or groups. The classical theory of (co)homological stability (over a fixed ring R) in topology produces results of the form: the homology $H_i(X_n; R)$ (resp. $H^i(X_n; R)$) does not depend on n for $n \gg i$. This converts an *a priori* infinite computation to a finite one. Examples of such sequences X_n include symmetric groups S_n (Nakaoka), braid groups B_n (Arnol'd and F. Cohen), the space $\mathrm{UConf}_n(M)$ of n-point subsets of the interior of a compact, connected manifold M with nonempty boundary (McDuff, Segal), special linear groups $\mathrm{SL}_n \mathbb{Z}$ (Borel and Charney), the moduli space $\mathcal{M}_n$ of genus $n \geq 2$ Riemann surfaces (Harer), and automorphism groups of free groups (Hatcher-Vogtmann-Wahl) ; see [15] for a survey.

For many natural sequences X_n homological stability fails in a strong way. We saw above that $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C}) \approx \mathbb{C}^{\binom{n}{2}}$ for all $n \geq 4$. In fact, one can prove for each $i \geq 1$ that

$$\lim_{n \to \infty} \dim_{\mathbb{C}}(H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})) = \infty. \tag{2.4}$$

The underlying mechanism behind this instability is symmetry. Call a representation V of S_n *not essential* if each $\sigma \in S_n$ acts on V with every eigenvalue ± 1. Basic representation theory of S_n implies that any essential representation V of $S_n, n \geq 5$ satisfies $\dim(V) \geq n - 1$. It is not hard to check that the S_n-representation $H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C}))$ is essential, implying the blowup (2.4). One hardly needs to know topology to prove (2.4)! The driving force behind this is the representation theory of S_n.

More generally, whenever we have a sequence of larger and larger groups G_n, and a sequence V_n of "essential" G_n-representations, one expects that $\dim(V_n) \to \infty$. The general slogan of representation stability is: in many situations the *names* of the representations V_n

should stabilize as $n \to \infty$. The question is, how can we formalize this slogan, and how can we use such information? We focus on the case $G_n = S_n$; see §7 for a discussion of other examples, such as $G_n = \mathrm{SL}_n \mathbb{F}_p, \mathrm{GL}_n \mathbb{Z}$ and $\mathrm{Sp}_{2n} \mathbb{Z}$.

3. Representation stability (the S_n case)

With our notation, the description of $H^1(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ given in (2.3) does not depend on n once $n \geq 4$. In 2010 Thomas Church and I guessed that such a phenomenon might be true for cohomology in all degrees. Using an inductive description of $H^*(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ as a sum of induced representations (a weak form of a theorem of Lehrer-Solomon [20]), one can convert this question into a purely representation-theoretic one. After doing this, we asked David Hemmer about the case $i = 2$. He wrote a computer program that produced the following output; we use the notation C_n for $\mathrm{Conf}_n(\mathbb{C})$ to save space.

$$
\begin{aligned}
H^2(C_4; \mathbb{C}) &= V(1)^{\oplus 2} \oplus V(1,1) \oplus V(2) \\
H^2(C_5; \mathbb{C}) &= V(1)^{\oplus 2} \oplus V(1,1)^{\oplus 2} \oplus V(2)^{\oplus 2} \oplus V(2,1) \\
H^2(C_6; \mathbb{C}) &= V(1)^{\oplus 2} \oplus V(1,1)^{\oplus 2} \oplus V(2)^{\oplus 2} \oplus V(2,1)^{\oplus 2} \oplus V(3) \\
H^2(C_7; \mathbb{C}) &= V(1)^{\oplus 2} \oplus V(1,1)^{\oplus 2} \oplus V(2)^{\oplus 2} \oplus V(2,1)^{\oplus 2} \oplus V(3) \oplus V(3,1) \\
H^2(C_8; \mathbb{C}) &= V(1)^{\oplus 2} \oplus V(1,1)^{\oplus 2} \oplus V(2)^{\oplus 2} \oplus V(2,1)^{\oplus 2} \oplus V(3) \oplus V(3,1) \\
H^2(C_9; \mathbb{C}) &= V(1)^{\oplus 2} \oplus V(1,1)^{\oplus 2} \oplus V(2)^{\oplus 2} \oplus V(2,1)^{\oplus 2} \oplus V(3) \oplus V(3,1) \\
H^2(C_{10}; \mathbb{C}) &= V(1)^{\oplus 2} \oplus V(1,1)^{\oplus 2} \oplus V(2)^{\oplus 2} \oplus V(2,1)^{\oplus 2} \oplus V(3) \oplus V(3,1)
\end{aligned}
\tag{3.1}
$$

This was compelling. Indeed, it turns out that this decomposition holds for $H^2(C_n; \mathbb{C})$ for all $n \geq 7$, so the decomposition of $H^2(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ into irreducible S_n-representations *stabilizes*. The most useful way we found to encode this type of behavior was via the notion of a representation stable sequence, which we now explain.

3.1. Representation stability and $H^i(\mathrm{Conf}_n(M); \mathbb{C})$. Let V_n be a sequence of S_n–representations equipped with linear maps $\phi_n \colon V_n \to V_{n+1}$ so that following diagram commutes for each $g \in S_n$:

$$
\begin{array}{ccc}
V_n & \xrightarrow{\phi_n} & V_{n+1} \\
{\scriptstyle g}\downarrow & & \downarrow{\scriptstyle g} \\
V_n & \xrightarrow[\phi_n]{} & V_{n+1}
\end{array}
$$

Here g acts on V_{n+1} by its image under the standard inclusion $S_n \hookrightarrow S_{n+1}$. We call such a sequence of representations *consistent*. We made the following definition in [11].

Definition 3.1 (Representation stability for S_n-representations). A consistent sequence $\{V_n\}$ of S_n–representations is *representation stable* if there exists $N > 0$ so that for all $n \geq N$, each of the following conditions holds:

1. **Injectivity:** The maps $\phi_n \colon V_n \to V_{n+1}$ are injective.
2. **Surjectivity:** The span of the S_{n+1}–orbit of $\phi_n(V_n)$ is all of V_{n+1}.
3. **Multiplicities:** Decompose V_n into irreducible S_n–representations as

$$V_n = \bigoplus_{\lambda} c_n(\lambda) V(\lambda)_n$$

with multiplicities $0 \leq c_n(\lambda) \leq \infty$. Then $c_n(\lambda)$ does not depend on n.

The number N is called the *stable range*. The sequence $V_n := \wedge^* \mathbb{C}^n$ of exterior algebras is an example of a consistent sequence of S^n-representations that is not representation stable.

It is not hard to check that, given Condition 1 for ϕ_n, Condition 2 for ϕ_n is equivalent to the following : ϕ_n is a composition of the inclusion $V_n \hookrightarrow \mathrm{Ind}_{S_n}^{S_{n+1}} V_n$ with a surjective S_{n+1}–module homomorphism $\mathrm{Ind}_{S_n}^{S_{n+1}} V_n \to V_{n+1}$. This point of view leads to the stronger condition of *central stability*, a very useful concept invented by Putman [26] at around the same time, which he applied in his study of the cohomology of congruence subgroups.

There are variations on Definition 3.1. For example one can allow the stable range N to depend on the partition λ. In [11] we define representation stability for other sequences G_n of groups, with a definition analogous to Definition 3.1 with $G_n = S_n$ replaced by

$$G_n = \mathrm{GL}_n \mathbb{Z}, \mathrm{Sp}_{2g} \mathbb{Z}, \mathrm{GL}_n \mathbb{F}_q, \mathrm{Sp}_{2g} \mathbb{F}_q,$$

and hyperoctahedral groups W_n; see §7 below. In each case one needs a coherent naming system for representations of G_n as n varies.

Remark 3.2. We originally stumbled onto representation stability in [12] while making some computations in the homology of the Torelli group $\mathcal{I}_g$. In this situation the homology $H_i(\mathcal{I}_g; \mathbb{C})$ is a representation of the integral symplectic group $\mathrm{Sp}_{2g} \mathbb{Z}$. We found some $\mathrm{Sp}_{2g} \mathbb{Z}$-submodules of $H_i(\mathcal{I}_g; \mathbb{C})$ whose names did not depend on g for $g \gg 1$. Representation stability (for sequences of $\mathrm{Sp}_{2g} \mathbb{Z}$-representations) arose from our attempt to formalize this. See [12]. After [11, 12] appeared, Richard Hain kindly shared with us some of his unpublished notes from the early 1990s, where he also developed a conjectural picture of the homology $H_i(\mathcal{I}_g; \mathbb{C})$ as an $\mathrm{Sp}_{2g} \mathbb{Z}$-representation that is similar to the idea of representation stability for $\mathrm{Sp}_{2g} \mathbb{Z}$-representations presented in [11].

Using in a crucial way a result of Hemmer [16], we proved in [11] the following.

Theorem 3.3 (Representation stability for $\mathbf{Conf}_n(\mathbb{C})$). *For any fixed $i \geq 0$, the sequence $\{H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})\}$ is representation stable with stable range $n \geq 4i$.*

The stable range $n \geq 4i$ given in Theorem 3.3 predicts that $H^2(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ will stabilize once $n = 8$; in truth it stabilizes starting at $n = 7$.

The problem of computing all of the stable multiplicities $d_{i,n}(\lambda)$ in the decomposition of $H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ is thus converted to a problem which is finite and in principle solvable by a computer. However, putting this into practice is a delicate matter, and the actual answers can be quite complicated. For example, for $n \geq 16$:

$$
\begin{aligned}
&H^4(\mathrm{Conf}_n(\mathbb{C});\mathbb{C}) = \\
&\quad V(1)^{\oplus 2} \oplus V(2)^{\oplus 6} \oplus V(1,1)^{\oplus 6} \oplus V(3)^{\oplus 8} \oplus V(1,1,1)^{\oplus 9} \oplus V(2,1)^{\oplus 16} \\
&\oplus V(4)^{\oplus 6} \oplus V(1,1,1,1)^{\oplus 5} \oplus V(5)^{\oplus 2} \oplus V(2,2)^{\oplus 12} \oplus V(3,1)^{\oplus 19} \\
&\oplus V(2,1,1)^{\oplus 17} \oplus V(4,1)^{\oplus 12} \oplus V(2,1,1,1)^{\oplus 7} \oplus V(3,2)^{\oplus 14} \oplus V(2,2,1)^{\oplus 10} \\
&\oplus V(5,1)^{\oplus 3} \oplus V(3,3)^{\oplus 4} \oplus V(3,1,1)^{\oplus 16} \oplus V(2,2,2)^{\oplus 2} \oplus V(4,2)^{\oplus 7} \\
&\oplus V(4,1,1)^{\oplus 8} \oplus V(5,2) \oplus V(2,2,1,1)^{\oplus 2} \oplus V(3,1,1,1)^{\oplus 5} \oplus V(5,1,1)^{\oplus 2} \\
&\oplus V(4,3)^{\oplus 2} \oplus V(3,2,1)^{\oplus 9} \oplus V(4,1,1,1)^{\oplus 2} \oplus V(3,3,1)^{\oplus 2} \oplus V(3,2,2) \\
&\oplus V(4,2,1)^{\oplus 3} \oplus V(3,2,1,1) \oplus V(5,1,1,1) \oplus V(4,3,1)
\end{aligned}
$$

Theorem 3.3 was greatly extended by Church in [6] from $M = \mathbb{C}$ to M any connected, oriented manifold, as follows.

Theorem 3.4 (Representation stability for configuration spaces). *Let M be any connected, oriented manifold with $\dim(M) \geq 2$ and with $\dim_{\mathbb{C}} H^*(M;\mathbb{C}) < \infty$. Fix $i \geq 0$. Then the sequence $\{H^i(\mathrm{Conf}_n(M);\mathbb{C})\}$ is representation stable with stable range $n \geq 2i$ if $\dim(M) \geq 3$ and $n \geq 4i$ if $\dim(M) = 2$.*

This of course leaves open the following.

Problem 3.5 (Computing stable multiplicities). Given a connected, oriented manifold M, compute explicitly the stable multiplicities $d_{i,n}(\lambda)$ for the decomposition of $H^i(\mathrm{Conf}_n(M);\mathbb{C})$ into irreducibles. Give geometric interpretations of these numbers, as in the case of $H^1(\mathrm{Conf}_n(\mathbb{C});\mathbb{C})$ discussed in §2 above.

The problem of computing the $d_{i,n}(\lambda)$ seems to have been solved in very few cases. For example, I do not know the answer even for M a closed surface of genus $g \geq 1$.

The paper [11] gives many other examples of representation stable sequences V_n that arise naturally in mathematics, from the cohomology of Schubert varieties to composition of Schur functors to many of the examples given in §4.3 below.

3.2. An application to classical homological stability. Consider the space $\mathrm{UConf}_n(M) := \mathrm{Conf}_n(M)/S_n$ of unordered n-tuples of distinct points in M. As mentioned above, when M is the interior of a compact manifold with nonempty boundary, classical homological stability for $H_i(\mathrm{UConf}_n(M);\mathbb{Z})$ was proved by McDuff and Segal, generalizing earlier work of Arnol'd and F. Cohen. The reason that the assumption $\partial M \neq \emptyset$ is needed is that in this case one has a map $\psi_n : \mathrm{Conf}_n(M) \to \mathrm{Conf}_{n+1}(M)$ for each $n \geq 0$ given by "injecting a point at infinity" (see Proposition 4.6 of [8] for details). While ψ_n is really just defined up to homotopy, it induces for each $i \geq 0$ a well-defined homomorphism

$$(\psi_n)_* : H_i(\mathrm{UConf}_n(M);\mathbb{Z}) \to H_i(\mathrm{UConf}_{n+1}(M);\mathbb{Z}) \tag{3.2}$$

which McDuff and Segal prove is an isomorphism for $n \geq 2i+2$. This is the typical way one proves classical homological stability for a sequence of spaces X_n, namely one finds maps $X_n \to X_{n+1}$ and proves that they eventually induce isomorphisms on homology.

What about the case when M is closed? In this case there are no natural maps between $\mathrm{UConf}_n(M)$ and $\mathrm{UConf}_{n+1}(M)$. A natural thing to do would be to consider the S_n-cover $\mathrm{Conf}_n(M) \to \mathrm{UConf}_n(M)$, where there are maps (in fact $n+1$ of them) ϕ_n :

$\mathrm{Conf}_{n+1}(M) \to \mathrm{Conf}_n(M)$ given by "forget the point labeled i", where $1 \leq i \leq n+1$. The problem is, as we've seen above, the maps ϕ_n typically do not induce isomorphisms on homology. What to do?

Representation stability allowed Church to analyze this situation. It provided him with a language so that he could prove (Theorem 3.4) that the maps ϕ_n stabilize. The power of this point of view can be seen by applying Church's theorem (Theorem 3.4) to the trivial representation $V(0)$, which gives (for $\dim(M) \geq 3$) that $d_{i,n}(0)$ is constant for $n \geq 2i$. Transfer implies (see (2.2) above) that $\dim_{\mathbb{C}} H_i(\mathrm{UConf}_n(M); \mathbb{C})$ is constant for $n \geq 2i$, giving classical stability without maps between the spaces! Church actually obtains the better stable range of $n > i$ by a more careful analysis.

One might notice that Church only obtains homological stability over $\mathbb{Q}$, while McDuff and Segal's theorem works over $\mathbb{Z}$. One crucial place where $\mathbb{Q}$ is needed is the use of transfer. However, this is not just an artifact of Church's proof, it is a feature of the situation: classical homological stability for $\mathrm{UConf}_n(M)$ with M closed is *false* for general closed manifolds M! For example, $H_1(\mathrm{UConf}_n(S^2); \mathbb{Z}) = \mathbb{Z}/(2n-2)\mathbb{Z}$. After Church's paper appeared, other proofs of homological stability over for $H_i(\mathrm{UConf}_n(M); \mathbb{Q})$ were given by Randal-Williams [27] and then by Bendersky-Miller [2].

By plugging other representations into Theorem 3.4, Church deduces classical homological stability for a number of other colored configuration spaces. The above discussion illustrates how representation stability can be used as a useful method to discover and prove classical homological stability theorems.

3.3. Murnaghan's theorem. The stabilization of names of natural sequences of representations is not new. The notation $V(\lambda)$ that we gave in §2 above goes back at least to the 1938 paper [24] of Murnaghan, where he discovered the following theorem, first proved by Littlewood [21] in 1957.

Theorem 3.6 (Murnaghan's theorem). *For any two partitions λ, μ there is a finite set P of partitions so that for all sufficiently large n:*

$$V(\lambda)_n \otimes V(\mu)_n = \bigoplus_{\nu \in P} d_{\lambda\mu}(\nu) V(\nu)_n \tag{3.3}$$

for some non-negative integers $d_{\lambda\mu}(\nu)$.

The integers $d_{\lambda\mu}(\nu)$ are called *Kronecker coefficients*. In his original paper [24] Murnaghan computes the $d_{\lambda\mu}(\nu)$ explicitly for 58 of the simplest pairs μ, ν. The study of Kronecker coefficients remains an active direction for research. It is central to combinatorial representation theory and geometric complexity theory, among other areas. See, for example, [3] and the references contained therein.

One can deduce from Murnaghan's Theorem that the sequence $V(\lambda)_n \otimes V(\mu)_n$ is multiplicity stable in the sense of Definition 3.1; see [11]. In the following section we will describe a theory where Murnaghan's Theorem pops out as a structural feature of the theory.

4. FI-modules

Representation stability for symmetric groups S_n grew in power and applicability in [8], where Thomas Church, Jordan Ellenberg and I developed a theory of FI-modules.

When looking at the sequence $H^i(\mathrm{Conf}_n(M);\mathbb{C})$, we broke symmetry by only considering the map $\mathrm{Conf}_{n+1}(M) \to \mathrm{Conf}_n(M)$ given by "forget the $(n+1)^{\mathrm{st}}$ point". Of course there are really $n+1$ equally natural maps, given by "forget the j^{th} point" for $1 \leq j \leq n+1$. Taking cohomology switches the direction of arrows, and we have $n+1$ homomorphisms $H^i(\mathrm{Conf}_n(M);\mathbb{C}) \to H^i(\mathrm{Conf}_{n+1}(M);\mathbb{C})$, each one corresponding to an injective map $\{1,\dots,n\} \to \{1,\dots,n+1\}$; namely, the injective map whose image misses j. It is useful to consider all of these maps at once. This is the starting point for the study of FI-modules.

4.1. FI-module basics. An *FI-module* V is a functor from the category FI of finite sets and injections to the category of modules over a fixed Noetherian ring k. Thus to each set $\mathbf{n} := \{1,\dots,n\}$ with n elements the functor V associates a k-module $V_n := V(\mathbf{n})$, and to each injection $\mathbf{m} \to \mathbf{n}$ the functor V associates a linear map $V_m \to V_n$. The set of self-injections $\mathbf{n} \to \mathbf{n}$ is the symmetric group S_n. Thus an FI-module gives a sequence of S_n-representations V_n and linear maps between them, one for each injection of finite sets:

$$\begin{array}{ccccccccc}
\{1\} & \longrightarrow & \{1,2\} & \longrightarrow & \{1,2,3\} & \longrightarrow & \cdots \longrightarrow & \{1,\dots,n\} & \longrightarrow & \cdots \\
\circlearrowleft & & \circlearrowleft & & \circlearrowleft & & & \circlearrowleft & & \\
S_1 & & S_2 & & S_3 & & & S_n & &
\end{array}$$

$$V \downarrow$$

$$\begin{array}{ccccccccc}
V_1 & \longrightarrow & V_2 & \longrightarrow & V_3 & \longrightarrow & \cdots \longrightarrow & V_n & \longrightarrow & \cdots \\
\circlearrowleft & & \circlearrowleft & & \circlearrowleft & & & \circlearrowleft & & \\
S_1 & & S_2 & & S_3 & & & S_n & &
\end{array}$$

Of course each single horizontal arrow really represents many arrows, one for each injection between the corresponding finite sets. Using functors from the category FI to study sequences of objects is not new: FI-spaces were known long ago (under different names) to homotopy theorists.

A crucial observation is that one should think of an FI-module as a module in the classical sense. Many of the familiar notions from the theory of modules, such as submodule and quotient module, carry over to FI-modules in the obvious way: one performs the operations pointwise. So, for example, W is an FI-submodule of V if $W_n \subset V_n$ for each $n \geq 1$. One theme of [8] is that there is conceptual power in the encoding of this large amount of (potentially complicated) data into a single object V.

The property of being an FI-module itself does not guarantee much structure. One of the main insights in [8] was to find a finite generation condition that has strong implications but that one can also prove to hold in many examples.

Definition 4.1 (Finite generation). An FI-module V is *finitely generated* if there is a finite set S of elements in $\coprod_i V_i$ so that no proper sub-FI-module of V contains S.

Example 4.2. Let $k[x_1,\dots,x_n]_{(3)}$ denote the vector space of homogeneous polynomials of degree 3 in n variables over a field k. It is not hard to check that $\{1,\dots,n\} \mapsto k[x_1,\dots,x_n]_{(3)}$ is an FI-module. We claim that this FI-module is finitely generated by the elements $x_1^3, x_1^2x_2$ and $x_1x_2x_3$:

$$\begin{array}{llll} k[x_1]_{(3)} \longrightarrow & k[x_1,x_2]_{(3)} \longrightarrow & k[x_1,x_2,x_3]_{(3)} \longrightarrow & k[x_1,x_2,x_3,x_4]_{(3)} \longrightarrow \cdots \\ \boxed{x_1^3} & x_1^3, x_2^3 & x_1^3, x_2^3, x_3^3 & x_1^3, x_2^3, x_3^3, x_4^3 \\ & \boxed{x_1^2x_2}, x_2^2x_1 & x_1^2x_2, x_1^2x_3, x_2^2x_1 & \vdots \\ & & x_2^2x_3, x_3^2x_1, x_3^2x_2 & \\ & & \boxed{x_1x_2x_3} & \end{array}$$

Here we have written below each $k[x_1,\dots,x_n]_{(3)}$ its basis as a vector space. Finite generation of the FI-module $k[x_1,\dots,x_n]_{(3)}$ is simply the fact that every vector in every $k[x_1,\dots,x_n]_{(3)}$ lies in the k-span of the set of vectors that can be obtained from the three boxed vectors by performing all possible morphisms, i.e. by changing the labels of the x_i. In other words, there are, up to labeling and taking linear combinations, only three homogeneous degree three polynomials in any number $n \geq 3$ of variables: $x_1^3, x_1^2x_2$ and $x_1x_2x_3$. Note that we need $n \geq 3$ to obtain all of the generators. Similarly, $k[x_1,\dots,x_n]_{(87)}$ is finitely generated, but the full generating set appears only for $n \geq 87$.

The connection of FI-modules with representation stability is the following, proved in [8].

Theorem 4.3 (Finite generation vs. representation stable). *Let V be an FI-module over a field k of characteristic 0. Then V is finitely generated if and only if $\{V_n\}$ is a representation stable sequence of S_n-representations with $\dim_k V_n < \infty$ for all n.*

Theorem 4.3 thus converts a somewhat complicated property about a sequence V_n of representations into a single property – finite generation – of a single object V. One example of the power of this viewpoint is the following.

Proof of Murnaghan's Theorem (Theorem 3.6). Since $V(\lambda)$ and $V(\mu)$ are finitely generated FI-modules [8, §2.8], so is $V(\lambda) \otimes V(\mu)$. Theorem 4.3 implies that $V(\lambda)_n \otimes V(\mu)_n$ is representation stable, and so the theorem follows. □

Thus a combinatorial theorem about an infinite list of numbers falls out of a basic structural property of FI-modules.

4.2. Character polynomials. One of the main discoveries of [8] is that character polynomials, studied by Frobenius but not so widely known today, are ubiquitous, and are an incredibly concise way to encode stability phenomena for sequences of S_n-representations.

Fix the ground field $\mathbb{C}$. Recall that the *character* of a representation $\rho : G \to \mathrm{GL}(V)$ of a finite group G over $\mathbb{C}$ is defined to be the function $\chi_V : G \to \mathbb{C}$ given by

$$\chi_V(g) := \mathrm{Trace}(\rho(g)).$$

We view χ_V as an element of the vector space $\mathcal{C}(G)$ of *class functions* on G; that is, those functions that are constant on each conjugacy class in G. A fundamental theorem in the representation theory of finite groups is that any G-representation is determined by its character:

$$\chi_V = \chi_W \text{ in } \mathcal{C}(G) \text{ if and only if } V \approx W \text{ as } G\text{-representations.}$$

For each $i \geq 1$ let $X_i \colon \coprod_n S_n \to \mathbb{N}$ be the class function defined by

$$X_i(\sigma) = \text{number of } i\text{-cycles in the cycle decomposition of } \sigma.$$

A *character polynomial* is any polynomial $P \in \mathbb{Q}[X_1, X_2, \ldots]$. Such a polynomial gives a class function on all the S_n at once. The study of character polynomials goes back to work of Frobenius, Murnaghan, Specht, and Macdonald; see, e.g. [23, Example I.7.14]).

It is easy to see for any fixed $n \geq 1$ that $\mathcal{C}(S_n)$ is spanned by character polynomials, so the character of any representation can be described by such a polynomial. For example, if $\mathbb{C}^n$ is the standard permutation representation of S_n then the character $\chi_{\mathbb{C}^n}(\sigma)$ is the number of fixed points of σ, so $\chi_{\mathbb{C}^n} = X_1$ for any $n \geq 1$. As another example, consider the S_n-representation $\bigwedge^2 \mathbb{C}^n$. Since $\sigma \cdot (e_i \wedge e_j) = \pm e_i \wedge e_j$ according to whether σ contains $(i)(j)$ or (ij), respectively, it follows that

$$\chi_{\bigwedge^2 \mathbb{C}^n} = \binom{X_1}{2} - X_2 = \frac{1}{2}X_1^2 - \frac{1}{2}X_1 - X_2$$

for any $n \geq 1$. These descriptions of characters are uniform in n. On the other hand, if one fixes r then for $n \gg r$ it is incredibly rare for an S_n-representation to be given by a character polynomial $P(X_1, \ldots, X_r)$ depending only on cycles of length at most r. A simple example is the sign representation: for $n \gg r$ one cannot determine the sign of an arbitrary $\sigma \in S_n$ just by looking at cycles in σ of length at most r.

One of the main discoveries in [8] is that finitely-generated FI-modules in characteristic 0 admit such a uniform description.

Theorem 4.4 (Polynomiality of characters). *Let V be a finitely-generated FI-module over a field k of characteristic 0. Then the sequence of characters χ_{V_n} of the S_n-representations V_n is* eventually polynomial*: there exists $N \geq 0$ and a polynomial $P(X_1, \ldots, X_r)$ for some $r > 0$ so that*

$$\chi_{V_n} = P(X_1, \ldots, X_r) \quad \textit{for all } n \geq N$$

In particular $\dim_k(V_n)$ is a polynomial in n for $n \geq N$.

The claim on $\dim_k(V_n)$ is obtained by noting that

$$\dim_k(V_n) = \chi_{V_n}(\mathrm{Id}) = P(n, 0, \ldots, 0).$$

The fact that $\dim_k(V_n)$ is eventually a polynomial was extended to the case $\mathrm{char}(k) > 0$ in [10]. In situations of interest one can often give explicit bounds on r and N. This converts the problem of finding all the characters χ_{V_n} into a concrete finite computation. In some cases one can even get $N = 0$.

We again emphasize that the impact of Theorem 4.4 comes not just from the fact that a single polynomial gives all characters of all V_n with $n \gg 1$ at the same time, but it gives an extremely strong constraint on each individual V_n for $n \gg r$, since $\chi_{V_n} = P(X_1, \ldots, X_r)$ depends only on cycles of length at most r.

4.3. Examples/Applications. Part of the usefulness of finitely generated FI-modules is that they are common. This is illustrated in Table 4.1. We define only a few of these examples here; see [8] for a detailed discussion.

Theorem 4.5 (Finite generation). *Each of the FI-modules (1)-(9) given in Table 4.1 is finitely generated.*

FI-module $V = \{V_n\}$	Description
1. $H^i(\mathrm{Conf}_n(M);\mathbb{Q})$	$\mathrm{Conf}_n(M) =$ configuration space of n distinct ordered points on a connected, oriented manifold M, $\dim(M) > 1$
2. $R_J^{(r)}(n)$	$J = (j_1, \dots j_r)$, $R^{(r)}(n) = \bigoplus_J R_J^{(r)}(n)$= r-diagonal coinvariant algebra on r sets of n variables
3. $H^i(\mathcal{M}_{g,n};\mathbb{Q})$	$\mathcal{M}_{g,n} =$ moduli space of n-pointed genus $g \geq 2$ curves
4. $\mathcal{R}^i(\mathcal{M}_{g,n})$	i^{th} graded piece of the tautological ring of $\mathcal{M}_{g,n}$
5. $\mathcal{O}(X_{P,r}(n))_i$	space of degree i polynomials on the rank variety $X_{P,r}(n)$ of $n \times n$ matrices of P-rank $\leq r$
6. $G(A_n/\mathbb{Q})_i$	degree i part of the Bhargava–Satriano Galois closure of $A_n = \mathbb{Q}[x_1, \dots, x_n]/(x_1, \dots, x_n)^2$
7. $\langle H^1(\mathcal{I}_n;\mathbb{Q})\rangle_{(i)}$	degree i part of the subalgebra of $H^*(\mathcal{I}_n;\mathbb{Q})$ generated by $H^1(\mathcal{I}_n;\mathbb{Q})$, where $\mathcal{I}_n =$ genus n Torelli group
8. $H^i(\mathrm{BDiff}_n(M);\mathbb{Q})$	$\mathrm{BDiff}_n(M) =$ Classifying space of diffeos leaving a given set of n points invariant, for many manifolds M (see [18])
9. $\mathrm{gr}(\Gamma_n)_i$	ith graded piece of associated graded Lie algebra of many groups Γ_n, including $\mathcal{I}_n$, IA_n and the pure braid group P_n

Table 4.1. Some examples of finitely generated FI-modules. Any parameter not equal to n should be considered fixed and nonnegative.

Items 3 and 8 of Theorem 4.5 are due to Jimenez Rolland [17, 18]; the other items are due to Church-Ellenberg-Farb [8].

That each of (1)-(9) in Table 4.1 is an FI-module is not difficult to prove. More substantial is proving finite generation. To do this one of course needs detailed information about the specific example. In some of the cases this involves significant (but known) results; see below.

Except for a few special (e.g. $M = \mathbb{R}^d$) and low-complexity (i.e. small i, d, g, J, etc.) cases, explicit formulas for the characters (or even the dimensions) of the vector spaces (1)-(9) of Table 4.1 do not seem to be known, or even conjectured. Exact computations may be quite difficult. Applying Theorem 4.4 and Theorem 4.5 to these examples gives us an answer, albeit a non-explicit one, in all cases.

Theorem 4.6 (Ubiquity of character polynomials). *For each of the sequences V_n in Table 4.1 there are numbers $N \geq 0, r \geq 1$ and a polynomial $P(X_1, \dots, X_r)$ so that*

$$\chi_{V_n} = P(X_1, \dots, X_r) \quad \textit{for all } n \geq N$$

In particular $\dim(V_n)$ is a polynomial in n for $n \geq N$.

We emphasize that we are claiming eventual *equality* to a polynomial, not just polynomial growth. As a contrasting example, if $\overline{\mathcal{M}}_{g,n}$ is the Deligne-Mumford compactification of the moduli space of n-pointed genus g curves, the dimension of $H^2(\overline{\mathcal{M}}_{g,n};\mathbb{Q})$ grows exponentially with n; in particular the character of $H^2(\overline{\mathcal{M}}_{g,n};\mathbb{Q})$ is not given by a character polynomial. Although $V_n := H^2(\overline{\mathcal{M}}_{g,n};\mathbb{Q})$ is an FI-module, this FI-module is not finitely generated.

As an explicit example of Theorem 4.6, the character of the S_n-representation $H^2(\mathrm{Conf}_n(\mathbb{C});\mathbb{Q})$ is given for all $n \geq 0$ by the character polynomial

$$\chi_{H^2(\mathrm{Conf}_n(\mathbb{R}^2);\mathbb{C})} = 2\binom{X_1}{3} + 3\binom{X_1}{4} + \binom{X_1}{2}X_2 - \binom{X_2}{2} - X_3 - X_4. \quad (4.1)$$

Note that for general finitely-generated FI-modules we only know such information for $n \gg 1$. Recall from (3.1) of §3 how decomposition into irreducibles of $H^2(\mathrm{Conf}_n(\mathbb{R}^2);\mathbb{C})$ was shown to change with n, only to stabilize once $n \geq 7$. I encourage the reader to try to see this via (4.1), which holds for all $n \geq 0$.

Although we can sometimes give explicit upper bounds on their degree, the polynomials produced by Theorem 4.6 are known explicitly in only a few special cases. Thus the following is one of the main open problems in this direction.

Problem 4.7. Compute the polynomials $P(X_1,\ldots,X_r)$ produced by Theorem 4.6.

One difficulty in solving this problem is that, in many examples, the proof of finite generation of the corresponding FI-module uses a Noetherian property (see below), and the proof of this property is not effective.

4.4. The Noetherian property. The following theorem, joint work with Thomas Church, Jordan Ellenberg, and Rohit Nagpal, is central to the theory of FI-modules; it is perhaps the most useful general tool for proving that a given FI-module is finitely generated.

Theorem 4.8 (Noetherian property). *Let V be a finitely-generated FI-module over a Noetherian ring k. Then any sub-FI-module of V is finitely generated.*

Theorem 4.8 was proved in this generality by Church-Ellenberg-Farb-Nagpal [10]. For fields k of characteristic 0 it was proved earlier by Church-Ellenberg-Farb [8, Theorem 2.60] and by Snowden [29, Theorem 2.3], who actually proved a version (in a different language) for modules for many twisted commutative algebras, of which FI-modules are an example. The Noetherian property for FI-modules over fields k of positive characteristic is crucial for the study of the cohomology of congruence subgroups from this point of view; see §6 below. Lück proved a version of Theorem 4.8 for finite categories in [22], but since FI is infinite these do not occur in our context.

One can see how Theorem 4.8 is used in practice via the following.

Theorem 4.9. *Suppose $E_*^{p,q}$ is a first-quadrant spectral sequence of FI-modules over a Noetherian ring k, and that $E_*^{p,q}$ converges to an FI-module $H^{p+q}(X;k)$. If the FI-module $E_2^{p,q}$ is finitely generated for each $p,q \geq 0$, then the FI-module $H^i(X;k)$ is finitely generated for each fixed $i \geq 0$.*

See [6] and [8] for earlier versions of Theorem 4.9, and the paper [18], where Jimenez Rolland gives explicit bounds on the stability degree, etc.

Spectral sequences as in Theorem 4.9 arise in many computations. For example, following Cohen-Taylor [14] and Totaro [31], one computes $H^i(\mathrm{Conf}_n(M);k)$ by using the Leray spectral sequence for the natural inclusion $\mathrm{Conf}_n(M) \to M^n$. As n varies we obtain a sequence of spectral sequences, one for each n. In fact this gives a spectral sequence of FI-modules. Another example is the computation of the homology of congruence subgroups (see §6 below).

The proof of Theorem 4.9 is that, while we have no idea what the differentials might be, or at which page the spectral sequence stabilizes (and this may depend on n), the terms $E_\infty^{p,q}$ are obtained from the $E_2^{p,q}$ terms by repeatedly taking submodules and quotient modules. Since the property of finite generation for an FI-module is preserved by taking submodules (by the Noetherian property Theorem 4.8) and quotients, then if $E_2^{p,q}$ is finitely generated so is $E_j^{p,q}$ for every $j \geq 2$.

4.5. Some remarks on the general theory. There are many other aspects of the general theory of FI-modules that I am not describing due to lack of space. This includes a more quantitative version of the theory, with notions such as stability degree and weight of an FI-module, allowing for explicit estimates on stable ranges and degrees of character polynomials. Co-FI-modules are useful when one has maps going the wrong way. Also useful are FI-spaces, FI-varieties, and FI-hyperplane arrangements (see [9] for the latter); these give FI-modules by applying the (co)homology functor. Church-Putman [13] have developed the theory of FI-groups in order to prove a kind of relative finite generation theorem in group theory; they apply this in [13] to certain subgroups of Torelli groups. Sam and Snowden have given in [28] a more detailed analysis of the algebraic structure of the category of FI-modules in characteristic 0.

5. Combinatorial statistics for varieties over finite fields

In [9] we exposed a close connection between representation stability in cohomology and the stability of various combinatorial statistics for polynomials over finite fields and for maximal tori in $\mathrm{GL}_n(\mathbb{F}_q)$. We now give a brief sketch of how this works.

5.1. The space of polynomials over $\mathbb{F}_q$. Consider the following basic questions: how many square-free (i.e. having no repeated roots), degree n monic polynomials in $\mathbb{F}_q[T]$ are there? How many linear factors does one expect such a polynomial to have? factors of degree d? What is the variance of this expectation?

If one fixes q and allows n to increase, something interesting happens. A good example of what I'd like to describe is the expected *quadratic excess* of a polynomial in $\mathbb{F}_q[T]$; that is, the expected difference of the number of reducible quadratic factors and the number of irreducible quadratic factors. This number can be computed by adding up the quadratic excess of each degree n, monic square-free polynomial in $\mathbb{F}_q[T]$ and then dividing by the total number $q^n - q^{n-1}$ of such polynomials. Here are some values for small n:

	total:	expectation:
$n=3:$	q^2-q	$\frac{1}{q}$
$n=4:$	q^3-3q^2+2q	$\frac{1}{q}-\frac{2}{q^2}$
$n=5:$	$q^4-4q^3+5q^2-2q$	$\frac{1}{q}-\frac{3}{q^2}+\frac{2}{q^3}$
$n=6:$	$q^5-4q^4+7q^3-7q^2+3q$	$\frac{1}{q}-\frac{3}{q^2}+\frac{4}{q^3}-\frac{3}{q^4}$
$n=7:$	$q^6-4q^5+7q^4-8q^3+8q^2-4q$	$\frac{1}{q}-\frac{3}{q^2}+\frac{4}{q^3}-\frac{4}{q^4}+\frac{4}{q^5}$
$n=8:$	$q^7-4q^6+7q^5-8q^4+9q^3-10q^2+4q$	$\frac{1}{q}-\frac{3}{q^2}+\frac{4}{q^3}-\frac{4}{q^4}+\frac{5}{q^5}-\frac{5}{q^6}$

Notice that in each column in the counts above, the coefficient changes as n increases, until n is sufficiently large, and then this coefficient stabilizes. For example the third column gives coefficients $0, 2, 5, 7, 7, 7, \ldots$ Theorem 5.2 below implies that these formulas must converge term-by-term to a limit. A somewhat involved computation (see below) allowed us in [9] to compute this limit as:

$$q^{n-1}-4q^{n-2}+7q^{n-3}-8q^{n-4}+\cdots \qquad \text{and} \qquad \frac{1}{q}-\frac{3}{q^2}+\frac{4}{q^3}-\frac{4}{q^4}+\cdots \quad (5.1)$$

This numerical stabilization is a reflection of something deeper. To explain, consider the space Z_n of all monic, square-free, degree n polynomials with coefficients in the finite field $\mathbb{F}_q$. Recall that the *discriminant* $\Delta_n \in \mathbb{Z}[x_1, \ldots, x_{n-1}]$ is a polynomial with the property that an arbitrary monic polynomial $f(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z_1 + a_0 \in \mathbb{C}[z]$ is square-free if and only if $\Delta_n(a_0, \ldots a_{n-1}) \neq 0$. Thus Z_n is a complex algebraic variety. For example

$$Z_2 = \{z^2 + bz + c \in \mathbb{C}[z] : b^2 - 4c \neq 0\}$$

and

$$Z_3 = \{z^3 + bz^2 + cz + d \in \mathbb{C}[z] : b^2c^2 - 4c^3 - 4b^3d - 27d^2 + 18bcd \neq 0\}.$$

The complex variety Z_n is also an algebraic variety over the finite field $\mathbb{F}_q$ for any prime power q. The set of $\mathbb{F}_q$-points $Z_n(\mathbb{F}_q)$ is exactly the set of monic, square-free, degree n polynomials in $\mathbb{F}_q[T]$. From this point of view we should think of the original complex algebraic variety as the complex points $Z_n(\mathbb{C})$. There is a remarkable relationship between $Z_n(\mathbb{C})$ and $Z_n(\mathbb{F}_q)$, given by the Grothendieck–Lefschetz fixed point theorem in étale cohomology, which we now explain.

5.2. The Grothendieck-Lefschetz formula. It is a fundamental observation of Weil that for an algebraic variety Z defined over $\mathbb{F}_q$, one can realize $Z(\mathbb{F}_q)$ as the fixed points of a dynamical system, as follows. Denote by $\overline{\mathbb{F}}_q$ the algebraic closure of $\mathbb{F}_q$. The *geometric Frobenius* morphism $\mathrm{Frob}_q\colon Z(\overline{\mathbb{F}}_q) \to Z(\overline{\mathbb{F}}_q)$ acts (in an affine chart) on the coordinates of Z by $x \mapsto x^q$. Fermat's Little Theorem implies that

$$Z(\mathbb{F}_q) = \mathrm{Fix}[(\mathrm{Frob}_q : Z(\overline{\mathbb{F}}_q) \to Z(\overline{\mathbb{F}}_q)].$$

In the case of the varieties Z_n that we are considering, the Grothendieck-Lefschetz fixed point theorem takes the form:

$$|Z_n(\mathbb{F}_q)| = \sum_{f \in \mathrm{Fix}(\mathrm{Frob}_q)} 1 = \sum_i (-1)^i q^{n-i} \dim_{\mathbb{C}} H^i(Z_n(\mathbb{C}); \mathbb{C}) = q^n - q^{n-1} \tag{5.2}$$

where the last equality comes from the theorem of Arnol'd that $H^i(Z_n(\mathbb{C}); \mathbb{C}) = 0$ unless $i = 0, 1$, in which case it is $\mathbb{C}$.

We want to compute more subtle counts than just $|Z_n(\mathbb{F}_q)|$. To this end, we can *weight* the points of $Z_n(\mathbb{F}_q) = \mathrm{Fix}(\mathrm{Frob}_q)$, as follows. For each $f \in \mathrm{Fix}(\mathrm{Frob}_q) = Z_n(\mathbb{F}_q)$ the map Frob_q permutes the set

$$\mathrm{Roots}(f) := \{y \in \overline{\mathbb{F}}_q : f(y) = 0\}$$

giving a conjugacy class σ_f in S_n. Thus $X_i(\sigma_f)$ is well defined. Let $d_i(f)$ denote the number of irreducible (over $\mathbb{F}_q$) degree i factors of f. A crucial observation is that for any $i \geq 1$:

$$X_i(\sigma_f) = d_i(f). \tag{5.3}$$

Any $P \in \mathbb{C}[x_1, \ldots, x_r]$ (here $r \geq 1$ is arbitrary) determines a character polynomial $P(X_1, \ldots, X_r)$ (cf. §4.2 above). The polynomial P gives a way to weight points $f \in Z_n(\mathbb{F}_q)$ via

$$P(f) := P(d_1(f), \ldots, d_r(f)) = P(X_1(\sigma_f), \ldots, X_r(\sigma_f)).$$

So for example $P(f) = d_1(f)$ counts the number of linear factors of f (i.e. the number of roots of f that lie in $\mathbb{F}_q$), and $P(f) = d_2(f) - d_1(f)^2$ counts the quadratic excess of f. In general we call such a P a *polynomial statistic*. The expected value of $P(f)$ for $f \in Z_n(\mathbb{F}_q)$ is then given by $[\sum_{f \in Z_n(\mathbb{F}_q)} P(f)]/(q^n - q^{n-1})$.

$H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ enters the picture. Computing this expectation for a given P is where $H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ comes in. We can identify $Z_n(\mathbb{C})$ with the space $\mathrm{UConf}_n(\mathbb{C}) = \mathrm{Conf}_n(\mathbb{C})/\ S_n$ of unordered n-tuples of distinct points in $\mathbb{C}$ via the bijection that sends $f \in Z_n(\mathbb{C})$ to its set of roots. We thus have a covering $\mathrm{Conf}_n(\mathbb{C}) \to Z_n(\mathbb{C})$ of algebraic varieties, with deck group S_n.

Now, it's something of a long story, and there are a number of technical details to worry about, but the theory of étale cohomology and the twisted Grothendieck-Lefschetz formula, together with work of Lehrer [19], who proved that this machinery can be applied in this case, can be used to give the following theorem of [9]. Let $\langle \phi, \psi \rangle_{S_n} := \sum_{\sigma \in S_n} \phi(\sigma)\overline{\psi(\sigma)}$ be the standard inner product on the space of $\mathbb{C}$-valued functions on S_n.

Theorem 5.1 (Twisted Grothendieck–Lefschetz for Z_n). *For each prime power q, each positive integer n, and each character polynomial P, we have*

$$\sum_{f \,\in\, Z_n(\mathbb{F}_q)} P(f) = \sum_{i=0}^{n} (-1)^i q^{n-i} \left\langle P, \chi_{H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})} \right\rangle_{S_n}. \tag{5.4}$$

For example, when $P = 1$ the inner product $\langle P, H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C}) \rangle$ is the multiplicity of the trivial S_n-representation in $H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$, which by transfer is the dimension of $H^i(Z_n(\mathbb{C}); \mathbb{C})$, giving the formula (5.2) above. Theorem 5.1 tells us that we can compute various weighted point counts on $Z_n(\mathbb{F}_q)$ if we understand the cohomology of the S_n-cover $\mathrm{Conf}_n(\mathbb{C})$ of $Z_n(\mathbb{C})$ as an S_n-representation.

5.3. Representation stability and Grothendieck-Lefschetz. Now we can bring representation stability into the picture. According to Theorem 4.6, for each $i \geq 0$ the character of $H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})$ is given by a character polynomial for $n \gg 1$ (actually in this case it holds for all $n \geq 1$). The inner product of two character polynomials is, for $n \gg 1$, constant. Keeping track of stable ranges, and defining $\deg P$ as usual but with $\deg x_k = k$, we deduce in [10] the following.

Theorem 5.2 (Stability of polynomial statistics). *For any polynomial $P \in \mathbb{Q}[x_1, x_2, \ldots]$, the limit*

$$\big\langle P, H^i(\mathrm{Conf}_\bullet(\mathbb{C}); \mathbb{C})\big\rangle := \lim_{n\to\infty} \big\langle P, \chi_{H^i(\mathrm{Conf}_n(\mathbb{C}); \mathbb{C})}\big\rangle_{S_n}$$

exists; in fact, this sequence is constant for $n \geq 2i + \deg P$. Furthermore, for each prime power q:

$$\lim_{n\to\infty} q^{-n} \sum_{f \in Z_n(\mathbb{F}_q)} P(f) = \sum_{i=0}^{\infty} (-1)^i \big\langle P, \chi_{H^i(\mathrm{Conf}_\bullet(\mathbb{C}); \mathbb{C})}\big\rangle q^{-i} \tag{5.5}$$

In particular, both the limit on the left and the series on the right in (5.5) *converge, and they converge to the same limit.*

Plugging $P = \binom{X_1}{2} - X_2$ into Theorem 5.2 gives the stable formula (5.1) for quadratic excess of a square-free degree n polynomial in $\mathbb{F}_q[T]$. The limiting values of other polynomial statistics P are computed in [9]; some of these are given in Table 5.1 below. One can actually apply Equation (5.5) of Theorem 5.2 in reverse, using number theory to compute the left-hand side in order to determine the right-hand side, as we do in §4.3 of [9].

The above method is applied in [9] to a different counting problem. Consider the complex algebraic variety of ordered n-frames in $\mathbb{C}^n$:

$$Z_n(\mathbb{C}) = \big\{(L_1, \ldots, L_n) \,\big|\, L_i \text{ a line in } \mathbb{C}^n,\ \ L_1, \ldots, L_n \text{ linearly independent}\big\}.$$

The group S_n acts on $Z_n(\mathbb{C})$ via $\sigma \cdot L_i = L_{\sigma(i)}$. The quotient $Z_n(\mathbb{C})/S_n$ is also an algebraic variety, and its $\mathbb{F}_q$-points parametrize the set of maximal tori in the finite group $\mathrm{GL}_n \mathbb{F}_q$. In analogy with the case of square-free polynomials, each $P \in \mathbb{C}[X_1, \ldots, X_r]$ counts maximal tori in $\mathrm{GL}_n \mathbb{F}_q$ with different weights. Since $H^i(Z_n(\mathbb{C}); \mathbb{C})$ is known to be representation stable (essentially by a theorem of Kraskiewicz-Weyman, Lustig, and Stanley - see §7.1 of [11]), we can apply an analogue of Theorem 5.2 in this context to compute this weighted point count.

Table 5.1 lists some examples of specific asymptotics that are computed in [9] using this method. The formulas in each column are obtained from Theorem 5.2 (and its analogue for $Z_n(\mathbb{C})$) with $P = 1$, $P = X_1$, $P = \binom{X_1}{2} - X_2^2$, the character χ_{sign} of the sign representation, and the characteristic function $\chi_{n\mathrm{cyc}}$ of the n-cycle, respectively. Note that the latter two are not character polynomials.

The formulas for square-free polynomials in Table 5.1 can be proved by direct means, for example using analytic number theory (e.g. weighted L-functions). In contrast, the formulas for maximal tori in $\mathrm{GL}_n \mathbb{F}_q$ may be known but are not so easy to prove. For example, the formula for the number of maximal tori in $\mathrm{GL}_n \mathbb{F}_q$ is a well-known theorem of Steinberg; proofs using the Grothendieck–Lefschetz formula have been given by Lehrer and Srinivasan (see e.g. [30]). Regardless, a central message of [9] is that representation stability provides a single underlying mechanism for all such formulas.

P	Counting theorem for squarefree polys in $\mathbb{F}_q[T]$	Counting theorem for maximal tori in $\mathrm{GL}_n\,\mathbb{F}_q$
1	# of degree n squarefree polynomials $= q^n - q^{n-1}$	# of maximal tori in $\mathrm{GL}_n\,\mathbb{F}_q$ (both split and non-split) $= q^{n^2-n}$
x_1	expected # of linear factors $= 1 - \frac{1}{q} + \frac{1}{q^2} - \frac{1}{q^3} + \cdots \pm \frac{1}{q^{n-2}}$	expected # of eigenvectors in $\mathbb{F}_q^n$ $= 1 + \frac{1}{q} + \frac{1}{q^2} + \cdots + \frac{1}{q^{n-1}}$
$\binom{x_1}{2} - x_2^2$	expected excess of *reducible* vs. *irreducible* quadratic factors $\to \frac{1}{q} - \frac{3}{q^2} + \frac{4}{q^3} - \frac{4}{q^4} + \frac{5}{q^5} - \frac{7}{q^6} + \frac{8}{q^7} - \frac{8}{q^8} + \cdots$ as $n \to \infty$	expected excess of *reducible* vs. *irreducible* dim-2 subtori $\to \frac{1}{q} + \frac{1}{q^2} + \frac{2}{q^3} + \frac{2}{q^4} + \frac{3}{q^5} + \frac{3}{q^6} + \frac{4}{q^7} + \frac{4}{q^8} + \cdots$ as $n \to \infty$
χ_{sign}	discriminant of random squarefree polynomial is equidistributed in $\mathbb{F}_q^\times$ between residues and nonresidues	# of irreducible factors is more likely to be $\equiv n \bmod 2$ than not, with bias $\sqrt{\#\text{ of tori}}$
$\chi_{n\text{cyc}}$	Prime Number Theorem for $\mathbb{F}_q[T]$: # of irreducible polynomials $= \sum_{d\mid n} \frac{\mu(n/d)}{n} q^d \sim \frac{q^n}{n}$	# of irreducible maximal tori $= \frac{q^{\binom{n}{2}}}{n}(q-1)(q^2-1)\cdots(q^{n-1}-1) \sim c \cdot \frac{q^{n^2-n}}{n}$

Table 5.1. Some asymptotics from [9], computed using Theorem 5.2 and its $Z_n(\mathbb{C})$ analogue.

Remark 5.3 (Stable range vs. rate of convergence). The dictionary between representation stability and stability of point-counts goes one level deeper. One can of course ask how quickly the formulas in Table 5.1 converge. As discussed in [9], the speed of convergence of any such formula depends on the stable range of the corresponding representation stability problem. For example, let L denote the limit of each side of Equation (5.5). The fact that $\langle \chi_P, H^i(\mathrm{Conf}_n(\mathbb{C}))\rangle_{S_n}$ is stable with stable range $n \geq 2i + \deg P$ can be used to deduce that

$$q^{-n} \sum_{f(T)\in Z_n(\mathbb{F}_q)} P(f) = L + O(q^{(\deg P - n)/2}) = L + O(q^{-n/2}).$$

We thus have a *power-saving bound* on the error term. See [9] for more details.

6. FI-modules in characteristic p

The Noetherian property for FI-modules was extended from fields of characteristic 0 to arbitrary Noetherian rings by Church-Ellenberg-Farb-Nagpal [10]. The proof is significantly more difficult in this case, and new ideas were needed. Indeed, [10] brought in more categorical and homological methods into the theory, for example with a homological reformulation of finite generation, and a certain shift functor that plays a crucial role. This line of ideas has culminated in the recent theory of FI-homology of Church-Ellenberg [7], which is an

exciting and powerful new tool.

One reason that we care about characteristic $p > 0$ is that in some examples this case contains most of the information. As an example, let K be a number field with ring of integers $\mathcal{O}_K$, and let $\mathfrak{p} \subset \mathcal{O}_K$ be any proper ideal. Define the *congruence subgroup* $\Gamma_n(\mathfrak{p}) \subset \mathrm{GL}_n(\mathcal{O}_k)$ to be

$$\Gamma_n(\mathfrak{p}) := \text{kernel}[\mathrm{GL}_n(\mathcal{O}_k) \to \mathrm{GL}_n(\mathcal{O}_k/\mathfrak{p})].$$

As shown by Charney [5], when one considers coefficients localized at $\mathfrak{p}$ then $\Gamma_n(\mathfrak{p})$ and $\mathrm{GL}_n(\mathcal{O}_K)$ have the same homology. Thus the interesting new information about $\Gamma_n(\mathfrak{p})$ comes via coefficients k with $\mathrm{char}(k) = p > 0$. While $H_i(\Gamma_n(\mathfrak{p}); k)$ is most naturally a representation of $\mathrm{SL}_n(\mathcal{O}_k/\mathfrak{p})$, one can restrict this action to a copy of S_n, and show that $H_i(\Gamma_n(\mathfrak{p}); k)$ is an FI-module. The Noetherian condition is crucial for proving that this FI-module is finitely generated, since the proof uses a spectral sequence argument (see below).

The proof of Theorem 4.4, that for a finitely-generated FI-module V the character χ_{V_n} is a character polynomial for $n \gg 1$, works only over a field k with $\mathrm{char}(k) = 0$. However, for fields k with $\mathrm{char}(k) > 0$, we were still able to prove [10, Theorem B] that there is a polynomial $P \in \mathbb{Q}[T]$ so that $\dim_k(V_n) = P(n)$ for all $n \gg 1$. Following the approach of Putman [26], we were able to apply this to $H_i(\Gamma_n(\mathfrak{p}); k)$, giving the following theorem, first proved by Putman [26] for fields of large characteristic.

Theorem 6.1 (**mod p Betti numbers of congruence subgroups**). *Let K be a number field, $\mathcal{O}_K$ its ring of integers, and $\mathfrak{p} \subsetneq \mathcal{O}_K$ any proper ideal. For any $i \geq 0$ and any field k, there exists a polynomial $P(T) = P_{\mathfrak{p},i,k}(T) \in \mathbb{Q}[T]$ so that for all sufficiently large n,*

$$\dim_k H_i(\Gamma_n(\mathfrak{p}); k) = P(n).$$

The exact numbers $\dim_k H_i(\Gamma_n(\mathfrak{p}); k)$ for $i > 1$ are known in very few cases, even for the simplest case $K = \mathbb{Q}, \mathfrak{p} = (p), k = \mathbb{F}_p$. Frank Calegari [4] has recently determined the rate of growth of the mod p Betti numbers of the level p^d congruence subgroup of $\mathrm{SL}_n(\mathcal{O}_K)$. He proves for example in [4, Lemma 3.5] that for $p \geq 5, d \geq 1$:

$$\dim_{\mathbb{F}_p} H_i(\Gamma_n(p^d); \mathbb{F}_p) = \binom{n^2-1}{i} + O(n^{2i-4}).$$

Calegari's result tells us the leading term of the polynomial guaranteed by Theorem 6.1. It should be noted that Calegari's proof uses (Putman's version of) Theorem 6.1.

Problem 6.2 ([10]). Compute the polynomials $P_{\mathfrak{p},i,k} \in \mathbb{Q}[T]$ given by Theorem 6.1. Do the Brauer characters of $H_i(\Gamma_n(\mathfrak{p}); k)$, or indeed of an arbitrary finitely-generated FI-module over a finite field k with $\mathrm{char}(k) > 0$, exhibit polynomial behavior in n for $n \gg 1$?

The more categorical setup in [10] allowed us to find an inductive description for any finitely generated FI-module.

Theorem 6.3 (**Inductive description of f.g. FI-modules**). *Let V be a finitely-generated FI-module over a Noetherian ring R. Then there exists some $N \geq 0$ such that for all $n \in \mathbb{N}$, there is an isomorphism of S_n-representations:*

$$V_n \approx \varinjlim V(S) \tag{6.1}$$

where the direct limit is taken over the poset of subsets $S \subset \{1, \ldots, n\}$ with $|S| \leq N$.

The condition (6.1) in Theorem 6.3 can be viewed as a reformulation of Putman's central stability condition [26, §1].

Since we proved in [10] that $H_m(\Gamma_n(\mathfrak{p}); k)$ is a finitely generated FI-module, Theorem 6.3 thus gives the following inductive presentation of $H_m(\Gamma_n(\mathfrak{p}); \mathbb{Z})$. Let $\Gamma^{(i)}_{n-1}(\mathfrak{p})$ with $1 \leq i \leq n$ denote the n standard subgroups of $\Gamma_n(\mathfrak{p})$ isomorphic to $\Gamma_{n-1}(\mathfrak{p})$. Let $\Gamma^{(i,j)}_{n-2}(\mathfrak{p}) := \Gamma^{(i)}_{n-1}(\mathfrak{p}) \cap \Gamma^{(j)}_{n-1}(\mathfrak{p})$. As the notation suggests, each $\Gamma^{(i,j)}_{n-2}(\mathfrak{p})$ is isomorphic to $\Gamma_{n-2}(\mathfrak{p})$. As with the Mayer-Vietoris sequence, the difference of the two inclusions gives a map

$$H_m(\Gamma^{(i,j)}_{n-2}(\mathfrak{p})) \to H_m(\Gamma^{(i)}_{n-1}(\mathfrak{p})) \oplus H_m(\Gamma^{(j)}_{n-1}(\mathfrak{p}))$$

whose image vanishes in $H_m(\Gamma_n(\mathfrak{p}))$. A version of the following theorem for coefficients in a sufficiently large finite field was first proved by Putman [26].

Theorem 6.4 (**A presentation for** $H_m(\Gamma_n(\mathfrak{p}); \mathbb{Z})$)**.** *Let K be a number field, let $\mathcal{O}_K$ be its ring of integers, and let $\mathfrak{p}$ be a proper ideal in $\mathcal{O}_K$. Fix $m \geq 0$. Then for all sufficiently large n,*

$$H_m(\Gamma_n(\mathfrak{p}); \mathbb{Z}) \simeq \frac{\bigoplus_{i=1}^{n} H_m(\Gamma^{(i)}_{n-1}(\mathfrak{p}); \mathbb{Z})}{\operatorname{im} \bigoplus_{i<j} H_m(\Gamma^{(i,j)}_{n-2}(\mathfrak{p}); \mathbb{Z})}.$$

We think of Theorem 6.4 as giving a presentation for $H_m(\Gamma_n(\mathfrak{p}); \mathbb{Z})$, with copies of $H_m(\Gamma_{n-1}(\mathfrak{p}); \mathbb{Z})$ as generators and copies of $H_m(\Gamma_{n-2}(\mathfrak{p}); \mathbb{Z})$ as relations. Theorem 6.3 is applied in [10] to give a similar description for $H_m(\mathrm{Conf}_n(M); \mathbb{Z})$ and for graded pieces of diagonal coinvariant algebras. Nagpal [25] has recently extended this point of view considerably, and has applied it to prove that the groups $H_m(\mathrm{UConf}_n(M); \mathbb{F}_p)$ are periodic in n.

7. Representation stability for other sequences of representations

In this paper we focused our attention on sequences V_n of S_n-representations. This is just one of the examples from [11], where we introduced and studied representation stability (and variations) for other families G_n of groups whose representation theory has a consistent naming system. Examples include $G_n = \mathrm{GL}_n \mathbb{Q}$, $\mathrm{Sp}_{2g} \mathbb{Q}$ and the hyperoctahedral groups. We also explored the case of modular representations of algebraic groups over finite fields, where instead of stability we found representation periodicity. The reader is referred to [11] for precise definitions and many examples.

I would like to illustrate here how these kinds of examples arise. For brevity let's stick to the calculation of group homology. Here is the general setup. Let Γ be a group with normal subgroup N and quotient $A := \Gamma/N$. The conjugation action of Γ on N induces a Γ–action on $H_i(N; R)$ for any coefficient ring R. This action factors through an A–action on $H_i(N, R)$, making $H_i(N, R)$ into an A–module.

The structure of $H_i(N, R)$ as an A–module encodes fine information. For example, the transfer isomorphism shows that when A is finite and $R = \mathbb{Q}$, the space $H_i(\Gamma; \mathbb{Q})$ appears precisely as the subspace of A–fixed vectors in $H_i(N; \mathbb{Q})$. But there are typically many other summands, and knowing the representation theory of A (over R) gives us a language with which to access these.

There are many natural examples of families Γ_n of this type, with normal subgroups N_n and quotients A_n. Table 7 summarizes some examples that fit into this framework.

kernel N_n	group Γ_n	acts on	quotient A_n	$H_1(N_n; R)$ for big n
pure braid group P_n	braid group B_n	$\{1, \dots, n\}$	S_n	$\mathrm{Sym}^2 V_n / V_n$
Torelli group $\mathcal{I}_n$	mapping class group Mod_n	$H_1(\Sigma_n, \mathbb{Z})$	$\mathrm{Sp}_{2n}\,\mathbb{Z}$	$\bigwedge^3 V_n / V_n$
$\mathrm{IAut}(F_n)$	$\mathrm{Aut}(F_n)$	$H_1(F_n, \mathbb{Z})$	$\mathrm{GL}_n\,\mathbb{Z}$	$V_n^* \otimes \bigwedge^2 V_n$
congruence subgroup $\Gamma_n(p)$	$\mathrm{SL}_n\,\mathbb{Z}$	$\mathbb{F}_p^n$	$\mathrm{SL}_n\,\mathbb{F}_p$	$\mathfrak{sl}_n\,\mathbb{F}_p$
level p subgroup $\mathrm{Mod}_n(p)$	Mod_n	$H_1(\Sigma_n; \mathbb{F}_p)$	$\mathrm{Sp}_{2n}\,\mathbb{F}_p$	$\bigwedge^3 V_n / V_n \oplus \mathfrak{sp}_{2n}\,\mathbb{F}_p$

Table 7.1. Some natural sequences of representations.

In each case the group N_n arises as the kernel of a natural Γ_n-action. Each example is explained in detail in [11]. Here $R = \mathbb{Q}$ in the first three examples, $R = \mathbb{F}_p$ in the fourth and fifth, and V_n stands in each case for the "standard representation" of A_n. In the last example p is an odd prime.

In each of the examples in Table 7, the groups Γ are known to satisfy classical homological stability. In contrast, the rightmost column of Table 7 shows that none of the groups N satisfies homological stability, even in dimension 1. In fact, except for the example of P_n, very little is known about the A_n–module $H_i(N_n, R)$ for $i > 1$, and indeed it is not clear if there is a nice closed form description of these homology groups. However, the appearance of some kind of stability can already be seen in the rightmost column, as the names of the irreducible composition factors of these A_n–modules are constant for large enough n; this is discussed in detail in [11].

A crucial common property of the examples in Table 7 is that each of the sequences A_n has an inherent stability in the naming of its irreducible algebraic representations over R. For example, an irreducible algebraic representation of $\mathrm{SL}_n\,\mathbb{Q}$ is determined by its highest weight vector, and these vectors can be described uniformly without reference to n. For example, for $\mathrm{SL}_n\,\mathbb{Q}$ the irreducible representation $V(L_1 + L_2 + L_3)$ with highest weight $L_1 + L_2 + L_3$ is isomorphic to $\bigwedge^3 V$ regardless of n, where V is the standard representation of SL_n.

In [11] we defined a notion of representation stability for each of the sequences of groups A_n given in Table 7. We gave some examples, gave some conjectures using this language, and worked out some of the basic theory. The powerful FI-module point of view has only been developed in the special case of $A_n = S_n$. This is completely missing in general.

Problem 7.1 (FI-theory for other sequences of groups). For each of the sequences $A_n = \mathrm{Sp}_{2n}\,\mathbb{Z}, \mathrm{GL}_n\,\mathbb{Z}, \mathrm{SL}_n\,\mathbb{F}_p, \mathrm{Sp}_{2n}\,\mathbb{F}_p$, work out a theory of FI_{A}-modules, where:

(1) Finite generation (perhaps with an additional condition) is equivalent to representation stability for A_n-represntations, as defined in [11].

(2) The theory gives uniform descriptions (uniform in n) of the characters of the examples in Table 7.

(3) FI_{A}-modules satisfy a Noetherian property.

In [34] J. Wilson extended the theory of FI-modules from the case of S_n to the other two sequences of classical Weyl groups (of type B/C and D); this includes the hyperoctahedral groups. New phenomena occur here. For example, character polynomials must be given with two distinct sets $\{X_i\}, \{Y_i\}$ of variables. Wilson applies this theory to a number of examples, including the cohomology of the pure string motion groups (see also [33]), the cohomology of various hyperplane arrangements, and diagonal co-invariant algebras for Weyl groups.

Acknowledgements. The author gratefully acknowledges support from the National Science Foundation. He would also like to thank his collaborators Thomas Church and Jordan Ellenberg, without whom the work discussed herein would not exist. Thanks to them and also to Dan Margalit for making extensive comments on an earlier version of this paper.

References

[1] V.I. Arnol'd, *The cohomology ring of the colored braid group*, Mathematical Notes **5**, no. 2 (1969), 138–140.

[2] M. Bendersky and J. Miller, *Localization and homological stability of configuration spaces*, Quarterly Jour. of Math., to appear, arXiv:1212.3596.

[3] C. Bowman, M. De Visscher, and R. Orellana, *The partition algebra and the Kronecker coefficients*, preprint, February 2013, arXiv:1210.5579v6.

[4] F. Calegari, *The stable homology of congruence subgroups*, preprint, November 2013, arXiv:1311.5190.

[5] R. Charney, *On the problem of homology stability for congruence subgroups*, Comm. in Alg., 12:17-18, 2081–2123 (1984).

[6] T. Church, *Homological stability for configuration spaces of manifolds*, Invent. Math. **188**, (2012) 2, 465–504.

[7] T. Church and J. S. Ellenberg, *Homological properties of FI-modules and stability*, in preparation.

[8] T. Church, J. S. Ellenberg, and B. Farb, *FI-modules: a new approach to stability for S_n-representations*, preprint, June 2012, arXiv:1204.4533.

[9] ———, *Representation stability in cohomology and asymptotics for families of varieties over finite fields*, to appear in Algebraic Topology: Applications and New Directions, AMS Contemp. Math. Series.

[10] T. Church, J. S. Ellenberg, B. Farb, and R. Nagpal, *FI-modules over Noetherian rings*, Geometry & Topology, to appear.

[11] T. Church and B. Farb, *Representation theory and homological stability*, Advances in Math., Vol. **245** (2013), pp. 250–314.

[12] T. Church and B. Farb, *Parameterized Abel–Jacobi maps and abelian cycles in the Torelli group*, Journal of Topology, **5** (2012), no. 1, 15–38.

[13] T. Church and A. Putman, *Generating the Johnson filtration*, preprint, November 2013, arXiv:1311.7150.

[14] F. R. Cohen and L.R. Taylor, *Computations of Gel'fand-Fuks cohomology, the cohomology of function spaces, and the cohomology of configuration spaces, in " Geometric applications of homotopy theory (Proc. Conf., Evanston, Ill., 1977)" I*, pp. 106–143, Springer Lect. in Math., Vol. **657**, 1978.

[15] R. Cohen, *Stability phenomena in the topology of moduli spaces, in "Surveys in differential geometry, Vol. XIV, Geometry of Riemann surfaces and their moduli spaces"*, pp. 23–56, Surv. Differ. Geom., **14**, Int. Press, 2009.

[16] D. Hemmer, *Stable decompositions for some symmetric group characters arising in braid group cohomology*, J. Combin. Theory Ser. A **118** (2011), 1136–1139.

[17] R. Jimenez Rolland, *Representation stability for the cohomology of the moduli space* $\mathcal{M}_g^n$, Algebr. Geom. Topol. **11** (2011), no. 5, 3011–3041.

[18] ———, *On the cohomology of pure mapping class groups as FI-modules*, J. Homotopy Relat. Struct., 2013.

[19] G.I. Lehrer, *The ℓ-adic Cohomology of Hyperplane Complements*, Bull. Lond. Math. Soc. 24 (1992) **1**, 76–82.

[20] G. Lehrer and L. Solomon, *On the action of the symmetric group on the cohomology of the complement of its reflecting hyperplanes*, J. Algebra 104 (1986), no. 2, 410–424.

[21] D.E. Littlewood, *Products and plethysms of characters with orthogonal, symplectic and symmetric groups*, Canad. J. Math. **10** (1958), 17–32.

[22] W. Lück, *Transformation Groups and Algebraic K-Theory*, Lecture Notes in Math., Vol. **1408**, Springer-Verlag, Berlin, 1989.

[23] I.G. Macdonald, *Symmetric functions and Hall polynomials*, second ed., Oxford Math. Mon., Clarendon Press, Oxford, 1995.

[24] F.D. Murnaghan, *The analysis of the Kronecker product of irreducible representations of the symmetric group*, American Jour. of Math., Vol. **60**, No. 3 (July 1938), pp. 761–784.

[25] R. Nagpal, *FI-Modules: Cohomology of modular S_n-representations*, in preparation.

[26] A. Putman, *Stability in the homology of congruence subgroups*, arXiv1201.48764, revised August 2012.

[27] O. Randal-Williams, *Homological stability for unordered configuration spaces*, Quarterly Jour. of Math., Vol. **64**, No. 1, pp. 303-326 (2013).

[28] S. Sam and A. Snowden, GL-*equivariant modules over polynomial rings in infinitely many variables*, arXiv1206.22332, revised October 2013.

[29] A. Snowden, *Syzygies of Segre embeddings and Δ-modules*, Duke Math J., **162** (2013) 2, 225–277.

[30] B. Srinivasan, *Representations of Finite Chevalley Groups*, Lecture Notes in Math. Vol. **764**, Springer-Verlag, 1979.

[31] B. Totaro, *Configuration spaces of algebraic varieties*, Topology 35 (1996), no. 4, 1057–1067.

[32] R. Vakil and M.M. Wood, *Discriminants in the Grothendieck ring*, arXiv:1208.3166.

[33] J.C.H. Wilson, *Representation stability for the cohomology of the pure string motion groups*, Alg. and Geom. Top. (12) (2012) 909–93.

[34] ———, *FI_W-modules and stability criteria for representations of classical Weyl groups*, preprint, Sept. 2013, arXiv:1309.3817.

Department of Mathematics, University of Chicago, 5734 S. University Ave. , Chicago, IL 60637.
E-mail: farb@math.uchicago.edu

Moduli spaces of manifolds

Søren Galatius

Abstract. This article surveys some recent advances in the topology of moduli spaces, with an emphasis on moduli spaces of manifolds.

Mathematics Subject Classification (2010). 57R90, 57R15, 57R56, 55P47.

Keywords. Manifolds, moduli spaces, classifying spaces, infinite loop spaces, surgery theory, diffeomorphism groups.

1. Introduction: Cohomology of $B\mathrm{Diff}(W)$

In this paper I shall survey some recent developments in the topological approach to moduli spaces of manifolds. The moduli spaces in question are classifying spaces $B\mathrm{Diff}(W)$, where $\mathrm{Diff}(W)$ denotes the topological group of diffeomorphisms of a smooth compact manifold W, restricting to the identity near ∂W, and variations thereof. These spaces show up in numerous contexts, but at the most basic level they are classifying spaces for smooth fiber bundles whose fibers are diffeomorphic to W: For a smooth manifold B there is a natural bijection between the set of smooth fiber bundles $\pi : E \to B$ whose fibers are diffeomorphic to W, and the set of homotopy classes of maps $X \to B\mathrm{Diff}(W)$. From this point of view, it is especially interesting to understand the *cohomology* of $B\mathrm{Diff}(W)$, as it is the ring of characteristic classes of such fiber bundles.

A motivating previous result concerns the case where W is an oriented connected 2-manifold, where the theorems of Harer ([25]) and of Madsen and Weiss ([35]) give a complete description of $H^*(B\mathrm{Diff}(W))$ in a range of degrees growing with the genus of W. These results are recalled in Section 1.2 below.

In Section 1.3 we shall state precise analogues of these theorems for $B\mathrm{Diff}(W)$ when W is an (even-dimensional) manifold of higher dimension, leading to a similar understanding of $H^*(B\mathrm{Diff}(W))$, in a range of degrees depending on a suitable analogue of genus. Some calculational aspects, and two interesting examples are discussed in Section 1.4.

1.1. Classification of $2n$-manifolds by tangential n-type. In the classical approach (cf. e.g. [9, 56]) to the classification of (high dimensional) manifolds, the basic homotopy theoretic invariant attached to a manifold M is its homotopy type, and one attempts only to classify manifolds inside one homotopy type at a time: If a space X is fixed, the *structure set* $\mathscr{S}(X)$ parametrizes pairs consisting of a smooth manifold and a homotopy equivalence $M \to X$. In this section we briefly recall another approach to classification, due to Kreck ([32]).

Proceedings of the International Congress of Mathematicians, Seoul, 2014

As the basic homotopy theoretic invariant of a smooth $2n$-manifold W, we shall instead consider the *tangential n-type* defined below (a minor variation of the normal n-type from [32]). Let W be a smooth compact $2n$-dimensional manifold. Writing γ_{2n} for the canonical vector bundle over $BO(2n)$, the space of vector bundle maps $TW \to \gamma_{2n}$ is contractible. Picking an element in this contractible space gives an underlying map $W \to BO(2n)$, and we may form the Moore–Postnikov factorization

$$W \xrightarrow{\ell} X_W \xrightarrow{\theta_W} BO(2n),$$

uniquely characterized up to weak homotopy equivalence by the facts that ℓ is an n-connected cofibration and θ_W is an n-coconnected Serre fibration. (Being n-connected means that the induced map in homotopy groups is surjective for π_n and bijective for $\pi_{<n}$; being n-coconnected means that the induced map in homotopy groups is injective in π_n and bijective in $\pi_{>n}$.)

We shall consider the fiber homotopy equivalence class of the fibration $\theta_W : X_W \to BO(2n)$ a primary homotopy theoretic invariant of W. If W has non-empty boundary $\partial W = P$, we shall consider the factorization $P \to X_W \to BO(2n)$, up to weak equivalence over $BO(2n)$ and under P. (Here, over and under are used in the usual sense: the map is required to commute with the maps to $BO(2n)$ and with the maps from P.)

Definition 1.1. Let $\theta : X \to BO(2n)$ be an n-coconnected Serre fibration. Let $\mathscr{Y}(\theta)$ be the set of diffeomorphism classes of closed manifolds W for which there exists a weak equivalence $X_W \to X$ over $BO(2n)$.

More generally, given a closed $(2n-1)$-manifold P and a bundle map $\ell_P : \varepsilon^1 \oplus TP \to \theta^*\gamma_{2n}$, let $\mathscr{Y}(\theta, P, \ell_P)$ be the set of diffeomorphism classes of compact manifolds W with $\partial W = P$, up to diffeomorphism relative to P, for which there exists a weak equivalence $X_W \to X$ over $BO(2n)$ and under P.

For some purposes it might be cleaner to define a set $\mathscr{Y}'(\theta)$ classifying pairs of a closed manifold W together with a choice of weak equivalence $X_W \to X$ over $BO(2n)$, up to homotopy through such maps. These sets contain essentially the same information, since $\mathscr{Y}(\theta)$ is the set of orbits of the monoid $\mathrm{Aut}(\theta)$ of weak equivalences $X \to X$ over $BO(2n)$ acting on $\mathscr{Y}'(\theta)$ in the obvious way. These sets admit self-maps induced by $[W] \mapsto [W\#(S^n \times S^n)]$, which we shall denote by s, for "stabilization". (Here we are assuming X and hence W is connected, so the diffeomorphism class of $W\#(S^n \times S^n)$ is well defined.) A remarkable theorem of Kreck ([32]) determines the direct limit of sets $\mathscr{Y}'(\theta) \xrightarrow{s} \mathscr{Y}'(\theta) \xrightarrow{s} \dots$ by constructing a bijection to a certain bordism group. Let us recall his result.

The composition $X \xrightarrow{\theta} BO(2n) \subset BO \xrightarrow{-1} BO$ defines a bordism theory $\Omega_*^{-\theta}$ in the usual way, cf. e.g. [51]: representatives are closed manifolds with a lift of their stable normal bundle to X, and these are considered up to bordism with the same structure. By definition, an element in $\mathscr{Y}'(\theta)$ represents an element of $\Omega_{2n}^{-\theta}$, and [32, Theorem C] (when translated from normal n-types to the tangential n-types used here) implies that this induces an bijection

$$\mathrm{colim}\left(\mathscr{Y}'(\theta) \xrightarrow{s'} \mathscr{Y}'(\theta) \xrightarrow{s'} \dots\right) \to \Omega_{2n}^{-\theta}.$$

Thus, a connected closed manifold W with tangential n-type θ is uniquely classified by a single invariant in the orbit set $(\Omega_{2n}^{-\theta})/\mathrm{Aut}(\theta)$, up to *stable diffeomorphism*, i.e. the equivalence

relation generated by $W \sim W'$ if $W \# k(S^n \times S^n)$ is diffeomorphic to $W' \# k(S^n \times S^n)$ for some k.

1.2. Classification of smooth (surface) bundles. The classification problem for manifolds has a parametrized version: for a given manifold B, describe the set of smooth fiber bundles $\pi : E \to B$, say with a fixed diffeomorphism type of the fibers, up to fiberwise diffeomorphism. If the fiber is a closed manifold W, this set is in natural bijection with the set $[B, B\mathrm{Diff}(W)]$ of homotopy classes of maps. If $P = \partial W \neq \emptyset$, the set $[B, B\mathrm{Diff}(W)]$ is in bijection with smooth fiber bundles $\pi : E \to B$ with an identification $\partial E = B \times P$, such that all fibers are diffeomorphic to W relative to P. (Recall that we write $\mathrm{Diff}(W)$ for the diffeomorphisms of W restricting to the identity near ∂W.) Thus the classification theory for smooth fiber bundles amounts to the understanding of the homotopy type of $B\mathrm{Diff}(W)$.

The association of the factorization $W \to X_W \to BO(2n)$ to the manifold W is, when carefully constructed, continuous and functorial in W and gives rise to a continuous map

$$B\mathrm{Diff}(W) \to B\mathrm{Aut}(\theta_W, \partial W), \tag{1.1}$$

where $\mathrm{Aut}(\theta_W, \partial W)$ is the topological monoid of self-homotopy equivalences $\phi : X \to X$ with $\phi \circ \ell|_{\partial W} = \ell|_{\partial W}$ and $\theta \circ \phi = \phi$.

Definition 1.2. For a compact manifold W, let $\theta = \theta_W : X_W \to BO(2n)$ be its tangential n-type, and write

$$B\mathrm{Diff}^{\theta}(W) = \mathrm{hofib}\big(B\mathrm{Diff}(W) \to B\mathrm{Aut}(\theta, \partial W)$$

for the homotopy fiber of the map (1.1).

(Despite the notation, this space is not always the classifying space of a group, and may have several path components.) By construction, the space $B\mathrm{Diff}^{\theta}(W)$ comes with an action of the monoid $\mathrm{Aut}(\theta, \partial W)$ and the ordinary $B\mathrm{Diff}(W)$ can be reconstructed as the homotopy orbit space of that action.

Example 1.3. If $n = 1$ and W is a connected, closed, orientable 2-manifold, the tangential 1-type of W is that of $\theta : BSO(2) \to BO(2)$. In this case, $B\mathrm{Diff}(W)$ is the classifying space of the group of all diffeomorphisms of W, not necessarily preserving orientation. The monoid $\mathrm{Aut}(\theta)$ is homotopy equivalent to the discrete group $\mathbb{Z}^{\times} \cong \mathbb{Z}/2$, and the map $B\mathrm{Diff}(W) \to B\mathrm{Aut}(\theta)$ is induces by the action of $\mathrm{Diff}(W)$ on $H_2(W) \cong \mathbb{Z}$. Therefore, $B\mathrm{Diff}^{\theta}(W)$ is homotopy equivalent to the classifying space of the group of orientation preserving diffeomorphisms.

If $n = 1$ and W is a connected orientable 2-manifold with *non-empty* boundary, the tangential 1-type is still $BSO(2) \to BO(2)$, but now $B\mathrm{Aut}(\theta, \partial W)$ is contractible and hence $B\mathrm{Diff}^{\theta}(W, \partial W) \simeq B\mathrm{Diff}(W, \partial W)$, corresponding to the fact that diffeomorphisms fixing the boundary are automatically orientation preserving.

By the classification theorem for connected oriented compact 2-manifolds, the diffeomorphism type of W is given by two numbers, viz. the genus and the number of boundary components. If $W = W_{g,n}$ has genus g and n boundary components, the moduli space $B\mathrm{Diff}^{\theta}(W_{g,n})$ is known to have several interesting geometric models (at least when excluding the exceptional surfaces of non-negative Euler characteristic), most notably the moduli space $\mathscr{M}_{g,n}$ of isomorphism classes of genus g Riemann surfaces, with n marked points and

a non-zero tangent vector at each point. This space has a wealth of geometric structure, including all the structure which follows from being the complex points of a variety defined over $\mathbb{Z}$ (a mixed Hodge structure on its rational cohomology, and a Galois action on its profinite completion), as well as metric and dynamic aspects related to the Teichmüller and Weil-Petersson metrics. This shall barely be touched upon here, except to note that it would be interesting to better understand the connections between these more geometric aspects of moduli spaces and the homotopy theoretic aspects covered in the present paper. In particular it would be interesting to understand how much, if any, of that geometric structure admits an analogue for moduli spaces of higher dimensional manifolds.

Let us recall two important theorems about the cohomology of this space. A version of the following result was first proved by Harer.

Theorem 1.4. *For any inclusion $j : W \subset W'$ of connected oriented compact 2-dimensional manifolds, the map*

$$H_k(B\mathrm{Diff}^\theta(W)) \to H_k(B\mathrm{Diff}^\theta(W'))$$

induced by extending a diffeomorphism of W by the identity map of $W' \setminus jW$, is an isomorphism for $k \leq (g(W) - 1)/1.5$, where $g(W)$ is the genus of W.

Harer's original paper [25] proved a weaker version of this result where the linear term $g/1.5$ in the range is replaced by $g/3$. Later results [31] improved this to $g/2$ and allowed closed surfaces. The slope $g/1.5$ above is due to Boldsen ([8]), inspired by unpublished ideas of Harer ([27]), and is known to be optimal ([41]).

Determining the cohomology of $B\mathrm{Diff}^\theta(W)$ in the stable range is then equivalent to determining the inverse limit

$$H^*(B\mathrm{Diff}^\theta(W_\infty)) = \varprojlim H^*(B\mathrm{Diff}^\theta(W_{g,1})),$$

where $W_{g,1}$ denotes a genus g surface with one boundary component. The following result is due to Madsen and Weiss.

Theorem 1.5 ([35]). *Let $BSO(2)^{-\theta}$ denote the Thom spectrum of the map $BSO(2) \xrightarrow{\theta} BO(2) \subset BO \xrightarrow{-1} BO$, graded such that the Thom class is in degree -2. There is a map*

$$B\mathrm{Diff}^\theta(W_\infty) \to \Omega^\infty(BSO(2)^{-\theta})$$

which is a homology equivalence (i.e. it induces an isomorphism in singular homology with integral coefficients) after restricting to a map between path connected spaces.

Analogues of the theorems of Harer and of Madsen–Weiss have been established for unoriented surfaces ([47, 53]) and for surfaces with certain tangential structures (cf. e.g. [4, 11, 26], or the comprehensive treatment in [47]). It would be interesting to understand how these results are related to the finite-generation criteria of A. Kupers and J. Miller ([33]). For manifolds of dimension higher than 2, A. Hatcher and N. Wahl ([28]) have proved homological stability for *mapping class groups* of many 3-manifolds.

The theorems of Harer and Madsen–Weiss combine to a formula for the homology of $B\mathrm{Diff}^\theta(W)$ for any orientable 2-manifold W, in the range where homological stability applies. The rational cohomology of a path component of $\Omega^\infty(BSO(2)^{-\theta})$ is easy to calculate, implying that in the stable range, the ring $H^*(B\mathrm{Diff}^\theta(W); \mathbb{Q})$ is polynomial on a sequence of classes κ_i of degree $2i$, $i \geq 1$.

1.3. Bundles of higher dimensional manifolds. Next, I shall discuss analogues for higher-dimensional manifolds of the results of Harer and Madsen–Weiss. To begin with, the following turns out to be useful generalizations of *genus*.

Definition 1.6. Write $W_1 = S^n \times S^n$ and $W_{1,1} = W_1 \setminus \mathrm{int}(D^{2n})$ for some choice of embedding $D^{2n} \to W_1$. The *genus* of a connected compact smooth $2n$-manifold W (with or without boundary) is the maximal number $g(W)$ for which there exist $g(W)$ disjoint embeddings $W_{1,1} \to W$. The *stable genus* is the maximal number $\overline{g}(W)$ such that for some k there exist $k + \overline{g}(W)$ disjoint embeddings $W_{1,1} \to W \# kW_1$.

If $j : W \hookrightarrow W'$ is an embedding of compact connected $2n$-manifolds, then $g(W') \geq g(W)$ and $\overline{g}(W') \geq \overline{g}(W)$. Such an embedding also induces a map of tangential n-types $\theta_W \to \theta_{W'}$, i.e. a map $X_W \to X_{W'}$ over $BO(2n)$. The fibration (1.1) can be made functorial with respect to embeddings and j induces a diagram

$$\begin{array}{ccccc} B\mathrm{Diff}^{\theta_W}(W) & \longrightarrow & B\mathrm{Diff}(W) & \longrightarrow & B\mathrm{Aut}(\theta_W, \partial W) \\ \downarrow & & \downarrow & & \downarrow \\ B\mathrm{Diff}^{\theta_{W'}}(W') & \longrightarrow & B\mathrm{Diff}(W') & \longrightarrow & B\mathrm{Aut}(\theta_{W'}, \partial W') \end{array}$$

whose middle vertical arrow is induced by extending diffeomorphisms of W by the identity on $W' \setminus j(W)$. We shall be especially concerned with embeddings j which induce a weak equivalence of tangential n-types, i.e. where the map $X_W \to X_{W'}$ is a weak equivalence. In this case we can identify θ_W and $\theta_{W'}$ and denote them both by θ. The following generalization of Harer's theorem is proved in joint work with O. Randal-Williams.

Theorem 1.7 ([18, 19]). *Let $j : W \hookrightarrow W'$ be an embedding of compact, connected manifolds of dimension $2n$, inducing an equivalence of tangential n-types. If W is simply connected and $n > 2$, then the map*

$$B\mathrm{Diff}^{\theta}(W) \to B\mathrm{Diff}^{\theta}(W')$$

induces an isomorphism in $H_k(-; \mathbb{Z})$ for $k \leq (\overline{g}(W) - 3)/2$, when restricted to a map between path connected spaces.

If we write $W_g = g(S^n \times S^n)$ for the connected sum of g copies of $S^n \times S^n$, then for any connected manifold W with non-empty boundary we get embeddings $W \hookrightarrow W \# W_1 \hookrightarrow W \# W_2 \hookrightarrow \dots$, by first using a collar of the boundary to embed M into its own interior, then performing the connected sum outside the image of that self-embedding. Each embedding induces an equivalence of tangential n-types, and we will write $B\mathrm{Diff}^{\theta}(W \# W_\infty)$ for the homotopy direct limit of the resulting sequence

$$B\mathrm{Diff}^{\theta}(W \# W_\infty) = \operatorname*{hocolim}_{g \to \infty} B\mathrm{Diff}^{\theta}(W \# W_g). \tag{1.2}$$

The limiting case $g \to \infty$ of Theorem 1.7 holds with weaker assumptions. Let us note that if $j : W \to W'$ is an embedding which maps a chosen disk $D^{2n-1} \approx D \subset \partial W$ into $\partial W'$, then j induces maps $B\mathrm{Diff}(W \# W_g) \to B\mathrm{Diff}(W' \# W_g)$ compatible with the maps in the direct limit (1.2), and similarly for $B\mathrm{Diff}^{\theta}$, provided the connected sum is performed near the chosen disk. Therefore j induces a map $B\mathrm{Diff}^{\theta}(W \# W_\infty) \to B\mathrm{Diff}^{\theta}(W' \# W_\infty)$. The following infinite-genus version of Theorem 1.7 is proved in joint work with O. Randal-Williams.

Addendum 1.8 ([19]). *Let $j : W \hookrightarrow W'$ be an embedding of compact, connected manifolds of dimension $2n > 0$, inducing an equivalence of tangential n-types. If j takes a disk in ∂W to a disk in $\partial W'$ as above, the induced map*

$$B\mathrm{Diff}^{\theta}(W \# W_{\infty}) \to B\mathrm{Diff}^{\theta}(W' \# W_{\infty})$$

is a homology equivalence, when restricting to a map between path connected spaces.

The following generalization of Madsen and Weiss' theorem to higher dimensional manifolds determines the homology in the stable range. It is proved in joint work with O. Randal-Williams.

Theorem 1.9 ([19, 20]). *Let W be any connected manifold of dimension $2n > 0$ with non-empty boundary. There is a map*

$$B\mathrm{Diff}^{\theta}(W \# W_{\infty}) \to \Omega^{\infty} X^{-\theta}$$

which, when restricted to a map between path connected spaces, induces an isomorphism in integral homology.

The map to $\Omega^{\infty} X^{-\theta}$ is defined using the Pontryagin–Thom construction. It exists also for closed manifolds, and can be made functorial under gluing, cf. Section 2 below. When W is simply connected and $n > 2$, the two theorems can be combined to deduce that the map

$$B\mathrm{Diff}^{\theta}(W) \to \Omega^{\infty} X^{-\theta}$$

induces an isomorphism in $H_k(-; \mathbb{Z})$ when $k \leq (\overline{g}(W) - 3)/2$, when restricted to a map between path connected spaces.

If constructed carefully, this map is also equivariant with respect to the action of $\mathrm{Aut}(\theta, \partial W)$ on both sides. The induced map of homotopy orbit spaces (= Borel constructions)

$$B\mathrm{Diff}(W) \to (\Omega^{\infty} X^{-\theta})_{h\mathrm{Aut}(\theta, \partial W)}, \tag{1.3}$$

then also induces an isomorphism in homology of path components in the stable range.

If n is large and W is very complicated, the monoid $\mathrm{Aut}(\theta, \partial W)$ may well be quite complicated too. However, it is always an $(n-1)$-type (i.e. the homotopy groups $\pi_{\geq n}$ all vanish, with all basepoints) and is an $(n-c-2)$-type if the inclusion $\partial W \to W$ is c-connected, since the fibers of θ are all $(n-1)$-types. In many examples of interest, the homotopy type of $B\mathrm{Aut}(\theta, \partial W)$ is easy to understand, and in particular it is contractible if $(W, \partial W)$ is $(n-1)$-connected.

1.4. Explicit calculations. For a given $\theta : X \to BO(2n)$, the integral homology of the infinite loop spaces $\Omega^{\infty} X^{-\theta}$ appearing in Theorem 1.9 is almost certainly very complicated in any interesting example. The rational cohomology is usually easy; in this section we spell out two interesting examples.

The integral cohomology is probably best approached one prime at a time, where several calculations have been carried out in the case $2n = 2$, cf. [16, 17, 46]. It would be useful to get a more concrete hold of the cohomology classes resulting from those calculations (perhaps in the style of [23]). It also seems interesting to investigate the real or complex K-theory, as well as more exotic cohomology theories.

1.4.1. Miller–Morita–Mumford classes. The calculations of the cohomology rings $H^*(B\mathrm{Diff}^\theta(W);\mathbb{Q})$ and $H^*(B\mathrm{Diff}(W);\mathbb{Q})$ in the stable range implied by Theorems 1.7 and 1.9 involve the characteristic classes known as MMM classes, after Miller, Morita and Mumford, who defined these classes in [38, 39, 42] in the case of oriented surfaces. Their definition readily generalizes to higher dimensions.

A smooth fiber bundle $\pi : E \to B$ has a fiberwise tangent bundle $T_\pi E$ which is a vector bundle on E. If the fibers of E are compatibly oriented d-manifolds and $p \in H^{k+2n}(BSO(2n))$, we obtain a class $p(T_\pi E) \in H^{k+2n}(E)$. If the fibers of π are closed, in addition to being oriented, we have a fiber integration map $\pi_! : H^{k+2n}(E) \to H^k(B)$, and we define the MMM class associated to $p \in H^{k+2n}(BSO(2n))$ as

$$\kappa_p(\pi) = \pi_!(T_\pi E) \in H^{*-2n}(B).$$

There are universal classes $\kappa_p \in H^*(B\mathrm{Diff}^+(W))$ for any closed oriented manifold W, where $\mathrm{Diff}^+(W)$ denotes the group of orientation preserving diffeomorphisms.

These characteristic classes and their variants appear in applications of Theorems 1.7 and 1.9, as well as multiple other contexts (cf. [12, 13, 24, 30] and others).

1.4.2. Connected sums of products of spheres. An interesting special case concerns the manifold $W_g^{2n} = g(S^n \times S^n)$, which could perhaps be regarded as a higher-dimensional analogue of the genus g surfaces. This manifold is $(n-1)$-connected and $W_g \to BO(2n)$ induces the trivial map in π_n, so in this case the Moore–Postnikov factorization $X \to BO(2n)$ is the n-connected cover of $BO(2n)$, denoted $\theta^n : BO(2n)\langle n\rangle \to BO(2n)$ in [20] (the notation $BO(2n)\langle n+1\rangle$ is sometimes used for the same thing). The inclusion $D^{2n} \hookrightarrow W_g$ is $(n-1)$-connected, so $B\mathrm{Aut}(\theta, D^{2n})$ is contractible and Theorem 1.9 asserts that the map

$$B\mathrm{Diff}(W_g^{2n}, D^{2n}) \to \Omega^\infty(BO(2n)\langle n\rangle^{-\gamma})$$

induces an isomorphism in homology (onto a path component of the target) in the range $* \le (g-3)/2$.

The rational cohomology of $\Omega^\infty(BO(2n)\langle n\rangle^{-\gamma})$ corresponds to certain MMM-classes in the cohomology of $B\mathrm{Diff}(W_g^{2n}, D^{2n})$ and $B\mathrm{Diff}^\theta(W_g^{2n})$, namely those corresponding to the subring $H^*(BO(2n)\langle n\rangle;\mathbb{Q}) \subset H^*(BSO(2n);\mathbb{Q})$. To obtain $B\mathrm{Diff}^+(W_g)$ from $B\mathrm{Diff}^\theta(W_g)$ we take homotopy orbit space by the monoid of fiber homotopy equivalences of the fibration $BO(2n)\langle n\rangle \to BSO(2n)$, whose classifying space is weakly equivalent to $BO[0,n]$. This leads to the following consequence of Theorems 1.7 and 1.9.

Theorem 1.10 ([18–20]). *Let $\mathcal{B} \subset H^*(BSO(2n))$ be the set of monomials in the classes $\{e, p_i | \frac{n+1}{4} \le i \le n-1\}$ of total degree more than $2n$ and let $\mathcal{B}' = \mathcal{B} \cup \{ep_i | 1 \le i \le \frac{n}{4}\}$. Then the natural maps*

$$\mathbb{Q}[\kappa_c | c \in \mathcal{B}] \to H^*(B\mathrm{Diff}^\theta(W_g);\mathbb{Q}) \to H^*(B\mathrm{Diff}(W_g, D^{2n});\mathbb{Q})$$

and

$$\mathbb{Q}[\kappa_c | c \in \mathcal{B}'] \to H^*(B\mathrm{Diff}^+(W_g);\mathbb{Q})$$

are all isomorphisms in degrees $* \le (g-3)/2$.

1.4.3. Complete intersections. Theorems 1.7 and 1.9 are most useful when applied to manifolds for which the genus is easily estimated. As in [32], an interesting class of such examples arises from the Lefschetz hyperplane theorem. I will briefly outline the simplest non-trivial example.

Let $V_d \subset \mathbb{C}P^4$ denote the variety defined by a homogeneous polynomial of degree d (a section of the line bundle $\mathcal{O}(d)$). For generically chosen polynomial, V_d is a smooth manifold whose diffeomorphism type is independent of the polynomial. By [55], the genus of V_d is half the third Betti number, and an easy Chern class calculation then proves that $g(V_d) = (d^4 - 5d^3 + 10d^2 - 10d + 4)/2$, allowing for a calculation of cohomology in a range of degree growing as $d^4/4$.

The embedding $V_d \subset \mathbb{C}P^4$ classifies a natural complex line bundle over V_d which we shall denote L_d, and $c_1(L_d)$ is a generator of $H_2(V_d;\mathbb{Z}) \cong \mathbb{Z}$. We consider the group $\mathrm{Diff}(V_d, L_d)$ consisting of pairs of a diffeomorphism $\phi : V_d \to V_d$ and a (specified) isomorphism $L_d \to \phi^* L_d$. The classifying space $B\mathrm{Diff}(V_d, L_d)$ classifies smooth fiber bundles $\pi : E \to B$ with fibers diffeomorphic to V_d, together with a complex line bundle $L \to E$ such that the class $x = c_1(L) \in H_2(E;\mathbb{Z})$ restricts to a generator of H^2 of each fiber. For such bundles, the relevant characteristic classes are defined as follows. For each monomial $p = p_1^i p_2^j e^k \in H^*(BSO(6))$ and each $n \in \mathbb{Z}_{\geq 0}$, we have the class $x^n p(T_\pi E) \in H^{2n+4i+8j+6k}(E)$ and may define the corresponding MMM class

$$\kappa_{x^n p}(\pi) = \pi_!(x^n p(T_\pi E)) \in H^{2n+4i+8j+6k-6}(B)$$

with a universal class $\kappa_{x^n p} \in H^*(B\mathrm{Diff}(V_d, L_d))$. The space $B\mathrm{Diff}(V_d, L_d)$ is rationally equivalent to the space $B\mathrm{Diff}^\theta(V_d)$ of Theorems 1.7 and 1.9, so we may use those theorems to calculate the rational cohomology.

Theorem 1.11 ([19]). *Let $\mathcal{B} \subset H^*(BSO(6))$ denote the set of monomials in p_1, p_2 and e. The natural map*

$$\mathbb{Q}[\kappa_{x^n p} | p \in \mathcal{B}, 2n + |p| > 6] \to H^*(B\mathrm{Diff}(V_d, L_d); \mathbb{Q})$$

is an isomorphism in degrees $* \leq (d^4 - 5d^3 + 10d^2 - 10d - 2)/4$.

The forgetful map $B\mathrm{Diff}(V_d, L_d) \to B\mathrm{Diff}^+(V_d)$ can also be analyzed in cohomology. Up to homotopy, it is a principal $K(\mathbb{Z}, 2)$-bundle, and the restriction of the action map to $\mathbb{C}P^1 \subset \mathbb{C}P^\infty = K(\mathbb{Z}, 2)$ gives rise to a homomorphism

$$D : H^*(B\mathrm{Diff}(V_d, L_d); \mathbb{Q}) \to H^{*-2}(B\mathrm{Diff}(V_d, L_d); \mathbb{Q}), \tag{1.4}$$

encoding the differential in the Serre spectral sequence for the fibration associated to the action of $K(\mathbb{Z}, 2)$. That spectral sequence can then be used to deduce the following result from Theorem 1.11.

Corollary 1.12. *In the stable range, the homomorphism*

$$H^*(B\mathrm{Diff}^+(V_d); \mathbb{Q}) \to H^*(B\mathrm{Diff}(V_d, L_d); \mathbb{Q})$$

is injective and its image is precisely the kernel of the homomorphism D in (1.4). The homomorphism D is a derivation with respect to cup product, and in the stable range it is explicitly determined by the formula $D(\kappa_{x^n p}) = n\kappa_{x^{n-1} p}$.

Three of the MMM classes, namely κ_{x^3}, κ_{xp_1} and κ_e, have degree 0 and are really characteristic *numbers* of the fiber V_d, depending only on the number d. These appear in the formula for three of the classes $D(\kappa_{x^n p})$ in Theorem 1.11, leading to a numerically somewhat complicated formula for the subring $H^*(B\mathrm{Diff}^+(V_d);\mathbb{Q}) \subset \mathbb{Q}[\kappa_{x^n p}]$ in the stable range.

The space $B\mathrm{Diff}(V_d, L_d)$ admits a map from an algebraic counterpart, defined as follows. Let $\Gamma(\mathcal{O}(d))$ be the algebraic sections of the line bundle $\mathcal{O}(d)$ on $\mathbb{P}^4$. The open set $U \subset \mathbb{P}(\Gamma(\mathcal{O}(d)))$ consisting of homogeneous polynomials defining a smooth variety $V_d \subset \mathbb{P}^4$ admits an action of the group $\mathrm{Aut}(\mathbb{P}^4) = \mathrm{PGL}(5)$. The quotient stack $U/\!\!/\mathrm{Aut}(\mathbb{P}^4)$ is a moduli space of complete intersections, whose complex points map in a natural way to $B\mathrm{Diff}(V_d, L_d)$. It would be interesting to understand more about that map and its variations.

2. Moduli spaces and bordism categories

In this section we wish to explain the role played by bordism categories in the proof of Theorem 1.9.

Theorem 1.9 concerns a very particular *gluing* of manifolds, namely the iterated gluing $S^n \times S^n$ to W near the boundary of W, but it has long been realized (cf. [22, 35, 52]) that more general gluing constructions are important. A convenient encoding of such gluing is through *bordism categories*, whose objects are closed $(2n-1)$-dimensional manifolds, morphisms are $2n$-dimensional bordisms, and composition defined by gluing of bordisms.

2.1. Higher and lower categories. Categories of bordisms have appeared in many other contexts, and in many variations. The cobordism categories of [51] are mainly used to keep track of transitivity of the bordism relation. The category theoretical aspects play an important role in Atiyah's and Segal's definition of topological quantum field theory ([1, 50]), defined as functors from a category of bordisms to the category of vector spaces, subject to certain axioms. More recently, the most mathematically popular field theories have been the *extended* field theories, defined as functors out of a higher category of bordisms, bordisms between bordisms, etc. Much of the recent interest has been inspired by the classification theorem for field theories announced by Lurie ([34]). Giving the right definitions has proved delicate, and popular approaches, e.g. the Θ_n-spaces of [49] or the iterated complete Segal spaces of [2], have only recently been proved equivalent ([3, 7]).

For the purpose of calculating cohomology of $B\mathrm{Diff}(W)$ or $B\mathrm{Diff}^\theta(W)$, it is not desirable to involve any higher categories of bordisms, and in this document we shall work entirely with simplicial categories whose objects are closed manifolds. On the contrary, the full bordism category of all compact bordisms between closed manifolds is already too large. For example, in the oriented 2-dimensional case Theorem 1.9 concerns *connected* surfaces of high genus, and it is not desirable to have bordisms which have many path components of low genus. In fact, an important step (going back to [52]) in the proof is to reduce to a subcategory where morphisms are highly connected relative to one end.

2.2. A bordism category. We define a simplicial version of a category of bordisms, useful for proving Theorems 1.7 and 1.9, after explaining some conventions.

Remark 2.1 (Conventions on simplicial sets). In the following we shall often consider sim-

plicial sets whose p-simplices are certain *families* of objects, parametrized by the extended simplices $\Delta^p_e = \{t \in \mathbb{R}^{p+1} | \sum t_i = 1\}$. The families are certain maps $\pi : E \to \Delta^p_e$ and the simplicial structure is given by pull-back along face inclusion $\Delta^{p-1}_e \to \Delta^p_e$. There are well known ways to deal with the inherent set-theoretical problems involved in this type of definition, cf. e.g. [45, §1.1] or [35, §2.1], as follows. Choose once and for all a set Ω and insist that E is *equal* to a subset of $\Delta^p_e \times \Omega$ and that $\pi : E \to \Delta^p_e$ is equal to the composition $E \subset (\Delta^p_e \times \Omega) \to \Delta^p_e$, where the second map is the projection. When the cardinality of Ω is sufficiently large, the homotopy type of the resulting simplicial set will not depend on Ω. This requirement shall be imposed without further mention in all the following definitions, whenever a map $\pi : E \to \Delta^p_e$ or $\theta : X \to \Delta^p_e$ appears. When a pair of maps $(\pi, f) : E \to \Delta^p_e \times \mathbb{R}$ appears, we shall instead assume that $E \subset (\Delta^p \times \mathbb{R}) \times \Omega$ and that (π, f) is equal to the projection.

Remark 2.2 (Conventions on boundaries). We shall consider various moduli spaces of manifolds with boundary. However, we shall generally eschew actual boundries, replacing them instead by germs of open manifolds containing them. For example, instead of a compact bordism W with incoming boundary P_0 and outgoing boundary P_1, we will consider a triple (t, E, f), where $t \geq 0$, E is a smooth manifold without boundary, $f : E \to \mathbb{R}$ is a smooth map which has 0 and t as regular values and such that $W = f^{-1}([0, t]) \subset E$ is compact. Such triples will be considered up to the equivalence relation generated by identifying (t, E, f) with $(t, E', f|'_E)$ whenever $E' \subset E$ is an open subset containing W. Instead of taking the actual incoming boundary $P = f^{-1}(0) \subset W$, we would consider (E, f) up to the coarser equivalence relation generated by being equal near P.

This convention has various technical advantages. For example, we will only ever need to discuss tangential structures on manifolds of the same dimension. More importantly, the germs give a well defined gluing of manifolds: a diffeomorphism $\partial W \cong \partial W'$ does not quite induce a well defined smooth structure on the topological manifold $W \cup_\partial W'$.

We shall need a simplicial category of bordisms equipped with tangential structures. Since the entire definition is rather long, we first define a category of bordisms without structure.

Definition 2.3. Let $\mathscr{C}'$ be the simplicial category where $N_q\mathscr{C}'_p$ is the set of equivalence classes of triples (t, π, f), where $t : \Delta^p_e \to (\mathbb{R}_{\geq 0})^q$ is a smooth function, $\pi : E \to \Delta^p_e$ is a submersion from a smooth manifold E of dimension $p + 2n$, and $f : E \to \mathbb{R}$ is a smooth function. Writing $a_i = \sum_{j=1}^i t_j$, this data is subject to the requirement that the functions $(\pi, f - a_i \circ \pi) : E \to \Delta^p_e \times \mathbb{R}$ are transverse to the submanifold $\Delta^p_e \times \{0\}$ for all $i = 0, \dots, q$. Writing $E_{ij} \subset E$ for the submanifold $(f - a_i \circ \pi)^{-1}(\mathbb{R}_{\geq 0}) \cap (f - a_j \circ \pi)^{-1}(\mathbb{R}_{\leq 0})$, we require that the restriction $\pi|_{E_{0q}} : E_{0q} \to \Delta^p_e$ is a proper map. Finally, the data is subject to the equivalence relation generated by replacing (π, f) by their restriction to an open subset $E' \subset E$ containing E_{0q}.

The face maps in the p-direction are defined in the obvious way, pulling all the data back along a face map $\Delta^{p-1}_e \to \Delta^p_e$. The face map $d_i : N_q\mathscr{C}'_p \to N_{q-1}\mathscr{C}'_p$ replaces $(t_1, \dots, t_q)$ by $(t_1, \dots, t_i + t_{i+1}, \dots, t_q)$ if $0 < i < q$. If $i = q$ it forgets t_q, and if $i = 0$ it forgets t_1 and additionally replaces $(\pi, f) : E \to \Delta^p_e \times \mathbb{R}$ by its pullback along the map $(x, s) \mapsto (x, s - t_1 \circ \pi(x))$.

With appropriate set-theoretic conventions, cf. Remark 2.1, the simplicial space $[q] \mapsto N_q\mathscr{C}'$ is the nerve of a simplicial category $\mathscr{C}'$. Informally, the morphisms are (Δ^p_e-parameter

families of) compact $2n$-dimensional bordisms E_{01}, and composition is defined by gluing; the germ of a thickening $E \supset E_{01}$ could be thought of as a technical detail.

Definition 2.4. Let $\mathscr{C}$ be the simplicial category where $N_q\mathscr{C}_p$ is the set of (t, π, f, x) where $(t, \pi, f) \in N_q\mathscr{C}'_p$ is as before, and $x = (c, \theta, \ell)$ consists of the following bundle data. Writing $T_\pi E$ for the fiberwise tangent bundle of π, $c : E \to BO(2n)$ is a map covered by a vector bundle map $\bar{c} : T_\pi E \to \gamma_{2n}$. θ is an n-coconnected Serre fibration $\theta : X \to \Delta^p_e \times BO(2n)$ such that X is weakly equivalent to a CW complex with finite n-skeleton. Finally, $\ell : E \to X$ is a map with $\theta \circ \ell = (\pi, c)$.

If we fix an n-coconnected Serre fibration $\theta_0 : X_0 \to BO(2n)$, the simplicial subcategory of $\mathscr{C}$ where $\theta : X \to \Delta^p_e \times BO(2n)$ is equal to $\mathrm{id} \times \theta_0 : \Delta^p_e \times X_0 \to \Delta^p_e \times BO(2n)$ shall be denoted $\mathscr{C}_{\theta_0}$.

Remark 2.5. The simplicial category $\mathscr{C}_\theta$ is a simplicial model for the topological category denoted $\mathscr{C}_\theta$ in [22]. Allowing θ to vary as in the above definition of $\mathscr{C}$ models the functoriality of $\theta \mapsto \mathscr{C}_\theta$. In fact, we can let $\mathscr{H}_p$ be the category whose objects are the n-coconnected Serre fibrations $\theta : X \to \Delta^p_e \times BO(2n)$ such that X is weakly equivalent to a CW complex with finite n-skeleton, and morphisms $\mathscr{H}_p(\theta, \theta')$ the set of maps $\phi : X \to X'$ with $\theta' \circ \phi = \theta$. This defines a simplicial category $\mathscr{H}$ and there is an obvious forgetful functor $\mathscr{C} \to \mathscr{H}$ which could perhaps be viewed as modelling $\mathscr{C}$ as a homotopy colimit of a functor $\theta \mapsto \mathscr{C}_\theta$ from $\mathscr{H}$ to the self-enriched category of small simplicial categories.

Let us write $B\mathscr{C}_\theta$ for the classifying space of the simplicial category $\mathscr{C}_\theta$, i.e. the topological space obtained by realizing in both directions. The following result, obtained in joint work with I. Madsen, U. Tillmann, and M. Weiss, determines the homotopy type of this classifying space.

Theorem 2.6 ([22]). *Let $\theta : X \to BO(d)$ and write $X^{-\theta}$ for the Thom space of the inverse of the bundle classified by θ (graded so that the Thom class is in degree $-d$). There is a weak equivalence*

$$B\mathscr{C}_\theta \simeq \Omega^{\infty-1} X^{-\theta}. \tag{2.1}$$

Using this result, the connection between $B\mathrm{Diff}^\theta(W)$ and $\Omega^\infty X^{-\theta}$ expressed in Theorems 1.7 and 1.9 goes via the based loop space $\Omega B\mathscr{C}_\theta$, or more precisely a space of paths in $B\mathscr{C}_\theta$ with fixed endpoints.

2.3. The moduli spaces. Let us explain how the bordism categories $\mathscr{C}$ and $\mathscr{C}_\theta$ are related to the moduli spaces $B\mathrm{Diff}^\theta(W)$ from Section 1.

The unique zero-simplex of $N_0\mathscr{C}_\theta$ with empty underlying manifold shall be denoted simply by $\emptyset$. If $P \in N_0\mathscr{C}_\theta$ is some other zero-simplex, the morphism space $\mathscr{C}_\theta(P, \emptyset)$ is quite closely related to the spaces $B\mathrm{Diff}^\theta(W)$. Indeed, the space $\mathscr{C}_\theta(P, \emptyset)$ is the "moduli space of null bordisms" of P. Ignoring the distinction between closed $(2n-1)$-dimensional manifolds and germs of $2n$-dimensional thickenings, $\mathscr{C}(P, \emptyset)$ is the space of compact $2n$-manifolds W with $\partial W = P$, equipped with a lift of the classifying map $TW : W \to BO(2n)$ over the fixed $\theta_0 : X_0 \to BO(2n)$, extending a given lift over P. Let us define the following subspaces.

Definition 2.7. For a zero-simplex $P \in N_0\mathscr{C}_\theta$, define

$$\mathscr{N}(P) \subset \mathscr{C}(P, \emptyset)$$

as the subspace defined by the requirement that, in the notation of Definition 2.4, the map $\ell : E \to X$ is n-connected.

The space $\mathscr{N}(P)$ is then the moduli space of null bordisms W of P, equipped with lifts to X_0 of a classifying map for their tangent bundle, extending a specified lift over $P = \partial W$, subject to the requirement that $W \to X$ be n-connected. The bundle condition is then nothing but an identification, up to homotopy equivalence, of the factorization $W \to X \to BO(2n)$ with the Moore–Postnikov factorization of $W \to BO(2n)$. From these considerations a weak equivalence

$$\mathscr{N}(P) \simeq \coprod_W B\mathrm{Diff}^\theta(W), \tag{2.2}$$

may be deduced, where the disjoint union is over representatives W for the set $\mathscr{Y}(P)$ from Definition 1.1.

Unfortunately $\mathscr{N}(-)$ is not a subfunctor of the representable functor $\mathscr{C}(-, \emptyset)$, because pre-composing with a morphism does not always preserve the connectivity conditions imposed in $\mathscr{N}$. Motivated by this observation, we consider the following subcategories.

Definition 2.8. Define subcategories $\mathscr{C}^i_\theta, \mathscr{C}^o_\theta \subset \mathscr{C}_\theta$, with the same objects as $\mathscr{C}_\theta$, in the following way. A morphism $f \in \mathscr{C}_\theta(P_0, P_1)$ is in $\mathscr{C}^o_\theta(P_0, P_1)$ provided the underlying bordism satisfies that the inclusion of the outgoing boundary be $(n-1)$-connected (in the notation of Definition 2.4, the inclusion $(f - \pi \circ t)^{-1}(0) \hookrightarrow E$ should be $(n-1)$-connected). To be in $\mathscr{C}^i_\theta$ we instead require that the inclusion of the incoming boundary be $(n-1)$-connected.

In the special case $n = 1$ the categories $\mathscr{C}^i_\theta$ and $\mathscr{C}^o_\theta$ are sometimes called *positive boundary* subcategories. In this case the condition is simply that every path component of a bordism is required to have non-empty incoming boundary, respectively non-empty outgoing boundary.

2.4. Group completion and stable homology. The association $P \mapsto \mathscr{N}(P) \subset \mathscr{C}_\theta(P, \emptyset)$ defines a functor from $\mathscr{C}^o_\theta$ to simplicial sets. Using the weak equivalences (2.1) and (2.2), we can now explain the map appearing in Theorem 1.9, which arises from a special case of a very general principle. If C is a simplicial category and $F : C \to \mathrm{sSet}$ is a simplicially enriched functor, there is a simplicial category $C \wr F$ and a functor $C \wr F \to C$. The fiber of the induced map $B(C \wr F) \to BC$ over a point $c \in N_0 C \subset C$ is $F(c)$, and the inclusion into the homotopy fiber gives a map $F(c) \to \mathrm{hofib}_c(B(C \wr F) \to BC)$. This homotopy fiber is weakly equivalent to ΩBC provided $B(C \wr F)$ contractible, which is the case if for example F is representable, or even just homotopy ind-representable (i.e. a filtered homotopy colimit of representable functors). Given any inverse system $k = (P_0 \xleftarrow{K_1} P_1 \xleftarrow{K_2} \dots)$ in $\mathscr{C}_\theta$ we therefore obtain a map

$$\mathscr{C}_\theta(k, \emptyset) \to \Omega B \mathscr{C}_\theta,$$

where $\mathscr{C}_\theta(k, -)$ denotes the ind-representable functor $Q \mapsto \mathrm{hocolim}\, \mathscr{C}_\theta(P_i, Q)$. If the morphisms in the inverse system are in the subcategory $\mathscr{C}^i_\theta$ we likewise obtain a map

$$\mathscr{C}^i_\theta(k, \emptyset) \to \Omega B \mathscr{C}^i_\theta. \tag{2.3}$$

We may restrict to the subspace $\mathscr{N}(k) = \mathrm{hocolim}\, \mathscr{N}(P_i) \subset \mathscr{C}_\theta(k, \emptyset)$ to obtain a map

$$\mathscr{N}(k) \to \Omega B \mathscr{C}_\theta.$$

By Theorem 2.6, the right hand side is a model for $\Omega^\infty X^{-\theta}$, and by the equivalence (2.2), the left hand side is the disjoint union of direct limits of spaces of the form $B\mathrm{Diff}^\theta(W)$. If we choose an appropriate inverse system k, these direct limits become $B\mathrm{Diff}^\theta(W\#W_\infty)$. More precisely, we may for each object $P \in \mathscr{C}_\theta$ choose an endomorphism $t_P \in \mathscr{C}_\theta(P,P)$ whose underlying bordism is diffeomorphic to $([0,1]\times P)\#(S^n\times S^n)$ and form the direct system $t_P^{-1} = (P \xleftarrow{t_P} P \xleftarrow{t_P} \dots)$. Then taking direct limit of the weak equivalence (2.2) gives a weak equivalence

$$\begin{aligned}\mathscr{N}(t_P^{-1}) &\simeq \mathrm{hocolim}\left(\coprod_W B\mathrm{Diff}^\theta(W) \xrightarrow{s} \coprod_W B\mathrm{Diff}^\theta(W) \xrightarrow{s} \dots\right)\\ &\simeq \mathbb{Z}\times\coprod_W B\mathrm{Diff}^\theta(W\#W_\infty),\end{aligned}$$

where s again denotes the map which forms connected sum with $S^n\times S^n$, the first disjoint union is over $[W]\in\mathscr{Y}_\theta$, and the second is over the quotient of $\mathscr{Y}_\theta$ by the equivalence relation generated by $[W]\sim[W\#(S^n\times S^n)]$.

With this notation, the following result is a strengthening of Theorem 1.9. The implied bijection on π_0 is essentially equivalent to [32, Theorem C], as discussed in Section 1.1 above.

Theorem 2.9. *For connected X and any non-empty object $P\in\mathscr{C}_\theta$, the map $\mathscr{N}(t_P^{-1})\to\Omega B\mathscr{C}_\theta$ induces an isomorphism in integral homology.*

Notes on the proof. Most of the proof is in [20], but the result claimed here is slightly stronger. We explain the two main ingredients in the proof, full details will appear in [19].

The first ingredient is the infinite-genus homological stability theorem stated as Addendum 1.8 above. As stated there, nothing is said about π_0, but a slightly modified version of the statement, also proved in [19], asserts that an appropriate disjoint union of the maps in Addendum 1.8 is a homology equivalence (not just after restricting to a map between path connected spaces). It implies that for certain inverse systems $k = (P_0 \xleftarrow{K_0} P_1 \xleftarrow{K_1} \dots)$, the ind-representable functor $Q\mapsto\mathscr{C}_\theta^i(k,Q)$ sends all morphisms in $\mathscr{C}_\theta^i$ to homology equivalences. By a general result of [36], used in a similar way as in [52], this implies that the map $\mathscr{C}_\theta^i(k,\emptyset)\to\Omega B\mathscr{C}_\theta^i$ from (2.3) is a homology equivalence as well.

The requirement on the direct system $k = (P_0 \xleftarrow{K_0} P_1 \xleftarrow{K_1} \dots)$ is that it must be a "universal θ-end", in the language of [20], which will be the case if the morphisms K_i are in both $\mathscr{C}_\theta^o$ and $\mathscr{C}_\theta^i$ and the map $K_i\to X$ underlying the tangential structure is n-connected for all i. That condition also implies that $\mathscr{N}(k)\simeq\mathscr{C}^i(k,\emptyset)$. The homological stability results of [19] is then used again to produce homology equivalences $\mathscr{N}(t_P^{-1})\to\mathscr{N}(k)$ for any non-empty object P. This combines to a homology equivalence $\mathscr{N}(t_P^{-1})\to\Omega B\mathscr{C}_\theta^i$ for any non-empty object P.

The second ingredient is a theorem of [20] (and in the case $n=1$, also [21] and [22]) asserting that the inclusion $B\mathscr{C}_\theta^i\to B\mathscr{C}_\theta$ is a weak equivalence. □

3. Other approaches to $B\mathrm{Diff}(W)$

A well established approach to understanding $B\mathrm{Diff}(W)$ goes via the *block* diffeomorphism group $\widetilde{\mathrm{Diff}}(W)$ and the corresponding moduli space $B\widetilde{\mathrm{Diff}}(W)$. It seems quite different

from the theory presented here, and a deeper understanding of their relationship seems worthwhile.

For a fixed dimension d, there is a simplicial set Man^d whose p-simplices are the maps $\pi : E \to \Delta^p_e$ where E is a smooth $(d+p)$-manifold without boundary and π is a smooth map which is proper (i.e. the inverse image of a compact set is compact) and a submersion. Its homotopy type is the disjoint union of $B\mathrm{Diff}(W)$ over all closed smooth d-manifolds W, one in each diffeomorphism class.

There is a larger space $\widetilde{\mathrm{Man}}^d$ defined in the same way, except that the condition on π is weakened to requiring only that for all morphisms $\theta : [q] \to [p]$ in Δ, the map $\pi : E \to \Delta^p_e$ is transverse to $\Delta^q_e \to \Delta^p_e$, and the resulting square

$$\begin{array}{ccc} \theta^* E & \longrightarrow & E \\ \downarrow & & \downarrow \pi \\ \Delta^q & \xrightarrow{\theta} & \Delta^p \end{array}$$

is homotopy cartesian (i.e. the top horizontal map is a homotopy equivalence). Thus the 0-simplices are the same for the two spaces, essentially the set of all closed d-manifolds, whereas the 1-simplices of $\widetilde{\mathrm{Man}}^d$ are essentially the h-cobordisms, etc.

The path component of Man^d containing a 0-simplex W is a model for the space $B\mathrm{Diff}(W)$, and the path component of $\widetilde{\mathrm{Man}}^d$ containing W is a model for $B\widetilde{\mathrm{Diff}}(W)$. F. Quinn's thesis [45] established a homotopy fiber sequence

$$G(W)/\widetilde{\mathrm{Diff}}(W) \to \mathrm{Map}(W, G/O) \to \Omega^{\infty+d} L(\pi_1(W)),$$

where $G(W)$ denotes the monoid of self-homotopy equivalences of W, $G(W)/\widetilde{\mathrm{Diff}}(W)$ denotes the homotopy fiber of a map $B\widetilde{\mathrm{Diff}}(W) \to BG(W)$, G/O denotes the homotopy fiber of the stable J-homomorphism $BO \to BG$, and $L(\pi_1 W)$ is the quadratic L-theory spectrum. (More precisely, $G(W)/\widetilde{\mathrm{Diff}}(W)$ is one of the path components of the fiber, which in general need not be path connected.) The latter is defined purely algebraically and is well understood, at least when W is simply connected. The space $\mathrm{Map}(W, G/O)$ is a purely homotopy theoretic object, although in general quite a difficult one to understand explicitly. This in principle pins down the homotopy type of the space $B\widetilde{\mathrm{Diff}}(W)$, as the homotopy orbit space of $G(W)$ acting on $G(W)/\widetilde{\mathrm{Diff}}(W)$ whose homotopy type is in turn described by Quinn's fibration sequence. In practice, the homotopy types of G/O and $G(W)$ are quite complicated themselves, but for instance in rational homotopy one can sometimes say a great deal about $B\widetilde{\mathrm{Diff}}(W)$ this way, cf. e.g. [5].

To obtain information about the more geometric object $B\mathrm{Diff}(W)$ by this method, the second step is to understand $\widetilde{\mathrm{Diff}}(W)/\mathrm{Diff}(W)$, the homotopy fiber of $B\mathrm{Diff}(W) \to B\widetilde{\mathrm{Diff}}(W)$. This is done via algebraic K-theory, in the form of Waldhausen's functor $A(W)$. M. Weiss and B. Williams ([57]) constructed a highly connected map

$$\widetilde{\mathrm{Diff}}(W)/\mathrm{Diff}(W) \to \Omega^\infty(\Sigma^{-1}\mathrm{Wh}^{\mathrm{Diff}}(W)_{hC_2}), \tag{3.1}$$

where $\mathrm{Wh}^{\mathrm{Diff}}(W)$ is the Whitehead spectrum of W, in turn described by the splitting $A(W) \simeq Q(W_+) \times \mathrm{Wh}^{\mathrm{Diff}}(W)$ (cf. [54] and Rognes' article in these proceedings), and

the subscript hC_2 denotes the homotopy orbit spectrum. The connectivity of the map (3.1) depends on the *concordance stable range* which is known [29] to be $\gtrsim \dim(M)/3$.

This two-step approach to $B\mathrm{Diff}(W)$ has an analogue for topological manifolds, describing $B\mathrm{Top}(W)$, where $\mathrm{Top}(W)$ denotes the topological group of homeomorphisms of W. There are block version $B\widetilde{\mathrm{Top}}(W)$ and $G(W)/\widetilde{\mathrm{Top}}(W)$, whose homotopy type is understood by an analogue of the fibration sequence (3.1); the homotopy fiber $\widetilde{\mathrm{Top}}(W)/\mathrm{Top}(W)$ of a map $B\mathrm{Top}(W) \to B\widetilde{\mathrm{Top}}(W)$ is again described in a concordance stable range via algebraic K-theory, cf. [54] and [57]. The two-step approach via block homeomorphisms can also be combined into one, cf. [58].

In comparison with the approach to $B\mathrm{Diff}(W)$ outlined in Section 1.3, the following are some of the main conceptual differences.

- Theorems 1.7 and 1.9 are fundamentally homological, and at best determines the homotopy type of $B\mathrm{Diff}(W)^+$ (or $B\mathrm{Diff}^\theta(W)^+$), where the plus denotes Quillen's plus construction. In particular, it would be very difficult to deduce much about the diffeomorphism group itself from this formula. In contrast, the approach via $B\widetilde{\mathrm{Diff}}(W)$ is homotopical, and the homotopy theoretic approximation to $B\mathrm{Diff}(W)$ it provides may be looped to an approximation of $\mathrm{Diff}(W)$.
- Both approaches describe $B\mathrm{Diff}(W)$ only below a certain stable range. In theorems 1.7 and 1.9 the range depends on the *genus* of W, whereas in the approach via $B\widetilde{\mathrm{Diff}}(W)$ the stable range depends on the *dimension* of W.
- The homotopy theoretic calculations involved are somewhat different. For example, the approach via $B\widetilde{\mathrm{Diff}}(W)$ involves the monoid $G(W)$ of all homotopy equivalences of W, whereas the approach via Theorem 1.9 instead involves the monoid $\mathrm{Aut}(\theta)$.

4. Unstable cohomology

Let me briefly discuss what is known about about the cohomology of moduli spaces of manifold outside the stable range, focusing on the case $W = W_g = g(S^n \times S^n)$. In the notation of Section 1.4, we consider the ring homomorphisms

$$\mathbb{Q}[\kappa_c | c \in \mathcal{B}] \to H^*(B\mathrm{Diff}(W_g, D^{2n}); \mathbb{Q}) \tag{4.1}$$

and

$$\mathbb{Q}[\kappa_c | c \in \mathcal{B}'] \to H^*(B\mathrm{Diff}^+(W_g); \mathbb{Q}) \tag{4.2}$$

which are isomorphisms in the stable range, known to be at least $* \leq (g-1)/2$ for $n \geq 3$ and $* \leq (g-1)/1.5$ for $n = 1$. However, the homomorphisms (4.1) and (4.2) are still defined outside this stable range and one may ask for an estimate of how non-trivial they are. We shall call the image of either map the *tautological subring* of cohomology, following established terminology in the case $n = 1$. The kernel is the ideal consisting of the relations satisfied universally by these classes.

Even for $n = 1$, the cokernel is still largely mysterious: various constructions are known to produce non-zero elements in the cokernel, but no pattern (even conjecturally) is known. By contrast, much effort by many people has gone into understanding the kernel of these maps when $n = 1$, leading to a complete understanding of the tautological subring for $g \leq 23$, as well as a conjectural description for all g, cf. e.g. [14, 15, 44].

For larger n, I. Grigoriev has devised a method for producing relations among the tautological classes for finite g, at least when $n = \dim(W_g)/2$ is odd. His method generalizes [40] and [48], and puts a lower bound on the size of the kernels of (4.1) and (4.2).

Theorem 4.1 ([24]). *For $n = (\dim(W_g))/2$ odd, the images of the ring maps* (4.1) *and* (4.2) *are finitely generated $\mathbb{Q}$-algebras. The kernel is non-trivial in degree $6g + 6$ if $n \equiv 1 \mod 4$ and in degree $2g + 2$ if $n \equiv 3 \mod 4$.*

The non-triviality of the kernel in particular shows that if the bound $* \leq (g - 3)/2$ in Theorem 1.7 can be improved to $* \leq ag + b$, then a is at most 2, at least if $n \equiv 3 \mod 4$. Hence the optimal such a is somewhere in the interval $[.5, 2]$.

Recent joint work between Grigoriev, Randal-Williams and myself has addressed the question of whether the tautological ring might in fact be finite dimensional as a *vector space* over $\mathbb{Q}$. We prove the following result.

Theorem 4.2. *Let $n = \dim(W_g)/2$ be an odd number and let $g \geq 2$. Then the image of the ring map* (4.1) *is a finite dimensional $\mathbb{Q}$-vector space and the image of the ring map* (4.2) *has Krull dimension precisely $(n - 1)$.*

In particular, the space $B\mathrm{Diff}^+(W_g)$ is not rationally equivalent to a finite CW complex except when $n = 1$. It seems unlikely that $B\mathrm{Diff}(W_g, D^{2n})$ could be homotopy equivalent to a finite dimensional CW complex for any $n > 1$.

5. Further moduli spaces

The theory presented in this survey is most well developed for smooth manifolds of even dimension. I now wish to briefly discuss moduli spaces of other objects, which seem sufficiently similar to $B\mathrm{Diff}(W)$ that one might expect that some or all of the theory will have an analogue for those moduli spaces.

5.1. Embedded manifolds. The group $\mathrm{Diff}(W)$ acts freely on the space of embeddings of W into any other manifold. The orbit space $\mathrm{Emb}(W, M)/\mathrm{Diff}(W)$ is the space of submanifolds of M diffeomorphic to W. When $M = \mathbb{R}^\infty$, the quotient space $\mathrm{Emb}(W, \mathbb{R}^\infty)/\mathrm{Diff}(W)$ is a convenient model for $B\mathrm{Diff}(W)$. It is also interesting to study $\mathrm{Emb}(W, M)/\mathrm{Diff}(W)$ for $M = \mathbb{R}^k$ for finite k, or for even more general manifolds M. In the case where W is an oriented 2-manifold, F. Cantero and O. Randal-Williams ([10]) have proved a homological stability result very similar to Harer's, provided $\dim M \geq 5$, and described the stable homology in terms of a space of compactly supported sections of a certain fibration over M.

5.2. Discretized diffeomorphism groups. The moduli spaces from Section 1.4.2 are classifying spaces of $\mathrm{Diff}(W_g^{2n}, D^{2n})$ considered as a topological group in the C^∞ topology. However, it makes perfect sense to consider also the underlying group with the discrete topology. Thus we write $\mathrm{Diff}^\delta(W_g, D)$ for the discrete group of diffeomorphisms of W_g which restrict to the identity on a neighborhood of D. S. Nariman has proved an analogue of Theorem 1.7 for these: There is a natural group homomorphism $\mathrm{Diff}^\delta(W_g, D) \to \mathrm{Diff}^\delta(W_{g+1}, D)$, induced by forming connected sum with $S^n \times S^n$ and extending by the identity diffeomorphism. Up to conjugation, this homomorphism is independent of choices, and [43] shows that the induced map in group homology is an isomorphism in a stable range.

The stable homology is expressed in terms of classifying spaces for foliations.

5.3. Odd dimensions. As proved by Meyer ([37]), any manifold which fibres over an odd dimensional manifold must have vanishing signature. Another proof was given by Ebert ([12]), who also pointed out that this implies that any analogue of Theorem 1.9 for odd dimensional manifolds must differ significantly from the even dimensional case.

Consider for example the manifolds $W_g^{2n+1} = g(S^n \times S^{2n+1})$ and choose some embedded disk $D^{2n+1} \subset W_g^{2n+1}$. Writing $\mathrm{Diff}(W_g^{2n+1}, D^{2n+1})$ for the topological group of diffeomorphisms fixing the disk pointwise, the Pontryagin–Thom construction can be used to define a map

$$B\mathrm{Diff}(W_g^{2n+1}, D^{2n+1}) \to \Omega^\infty(BO(2n+1)\langle n\rangle^{-\theta^*\gamma})$$

where $\theta : BO(2n+1)\langle n\rangle \to BO(2n+1)$ denotes the n-connected cover. It seems tempting to conjecture that the induced map in homology be an isomorphism in some stable range. As pointed out in [12] this cannot be true, even rationally: there are non-trivial classes in the rational cohomology of the codomain (viz. the MMM classes associated to the $\mathcal{L}$-classes in Hirzebruch's signature formula) which pull back to the zero class in $B\mathrm{Diff}(W_g^{2n+1}, D^{2n+1})$, for all large g.

One would expect that the limiting homology is described by some infinite loop space, but it is an open question which one it is. For $n = 1$, A. Hatcher has announced a proof that it is $Q(BSO(4)_+)$.

5.4. Other categories of manifolds. It is an interesting question whether the methods discussed in this survey may be applied in the setting of topological manifolds, or piecewise linear manifolds.

In a related direction, I. Madsen and A. Berglund ([6], [5]) have studied the case where $\mathrm{Diff}(W)$ is replaced by the monoid $G(W)$ consisting of self-homotopy equivalences of W (restricting to the identity near ∂W). Making extensive use of rational homotopy theory, they prove homological stability for $G(W_{g,1})$ and find a formula for the rational stable cohomology which intriguingly involves the *unstable* rational cohomology of automorphism groups of free groups.

Acknowledgements. S. Galatius was partially supported by NSF grant DMS-1105058. It is a pleasure to thank my long-term collaborator Oscar Randal-Williams, with whom many of the results presented here are obtained. I also wish to acknowledge the lasting impact of the work by Ib Madsen, Ulrike Tillmann and Michael Weiss, which has greatly influenced the work presented here.

References

[1] M. Atiyah, *Topological quantum field theories*, Inst. Hautes Études Sci. Publ. Math., **68**:175–186 (1989), 1988.

[2] C. Barwick, *(infinity, n)-Cat as a closed model category*, 2005, Thesis (Ph.D.)–University of Pennsylvania.

[3] C. Barwick and C. Schommer-Pries, *On the Unicity of the Homotopy Theory of Higher*

Categories, Preprint, arXiv:1112.0040.

[4] T. Bauer, *An infinite loop space structure on the nerve of spin bordism categories*, Q J Math., **55**(2):117–133, 2004.

[5] A. Berglund and I. Madsen, *Rational homotopy theory of automorphisms of highly connected manifolds*, Preprint, arXiv:1401.4096.

[6] ———, *Homological stability of diffeomorphism groups*, Pure Appl. Math. Q., **9**(1):1–48, 2013.

[7] J. E. Bergner and C. Rezk, *Comparison of models for (∞, n)-categories*, I. Geom. Topol., **17**(4):2163–2202, 2013.

[8] S. Boldsen, *Improved homological stability for the mapping class group with integral or twisted coefficients*, Mathematische Zeitschrift, **270**:297–329, 2012.

[9] W. Browder, *Surgery on simply-connected manifolds*, Springer-Verlag, New York-Heidelberg, 1972, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 65.

[10] F. Cantero and O. Randal-Williams, *Homological stability for spaces of surfaces*, Preprint, arXiv:1304.3006.

[11] R. Cohen and I. Madsen, *Surfaces in a background space and the homology of mapping class group*, Proc. Symp. Pure Math., **80**(1):43–76, 2009.

[12] J. Ebert, *A vanishing theorem for characteristic classes of odd-dimensional manifold bundles*, J. Reine Angew. Math. **684**:1–29, 2013.

[13] ———, *Algebraic independence of generalized MMM-classes*, Algebr. Geom. Topol., **11**(1):69–105, 2011.

[14] C. Faber, *A conjectural description of the tautological ring of the moduli space of curves*, In Moduli of curves and abelian varieties, Aspects Math. **E33**, pp. 109–129, Vieweg, Braunschweig, 1999.

[15] ———, *Tautological algebras of moduli spaces of curves*, In Moduli spaces of Riemann surfaces, volume 20 of IAS/Park City Math. Ser., pp. 197–219, Amer. Math. Soc., Providence, RI, 2013.

[16] S. Galatius, *Mod p homology of the stable mapping class group*, Topology, **43**(5):1105–1132, 2004.

[17] ———, *Mod 2 homology of the stable spin mapping class group*, Math. Ann., **334**(2):439–455, 2006.

[18] S. Galatius and O. Randal-Williams, *Homological stability for moduli spaces of high dimensional manifolds. I*, Preprint, arXiv:1403.2334.

[19] ———, *Homological stability for moduli spaces of high dimensional manifolds. II*, In preparation.

[20] ———, *Stable moduli spaces of high dimensional manifolds*, Acta Math., **212**(2): 257–377, 2014.

[21] ———, *Monoids of moduli spaces of manifolds*, Geom. Topol., **14**(3):1243–1302, 2010.

[22] S. Galatius, U. Tillmann, I. Madsen, and M. Weiss, *The homotopy type of the cobordism category*, Acta Math., **202**(2):195–239, 2009.

[23] C. Giusti, P. Salvatore, and D. Sinha, *The mod-2 cohomology rings of symmetric groups*, J. Topol., **5**(1):169–198, 2012.

[24] I. Grigoriev, *Relations among characteristic classes of manifold bundles*, Preprint, arXiv:1310.6804.

[25] J. L. Harer, *Stability of the homology of the mapping class groups of orientable surfaces*, Ann. of Math.(2), **121**(2):215–249, 1985.

[26] ———, *Stability of the homology of the moduli spaces of Riemann surfaces with spin structure*, Math. Ann., **287**(2):323–334, 1990.

[27] ———, *Improved stability for the homology of the mapping class groups of surfaces*, Duke University preprint, 1993.

[28] A. Hatcher and N. Wahl, *Stabilization for mapping class groups of 3-manifolds*, Duke Math. J., **155**(2):205–269, 2010.

[29] K. Igusa, *The stability theorem for smooth pseudoisotopies*, K-Theory, **2**(1-2):vi+355, 1988.

[30] ———, *Higher Franz-Reidemeister torsion*, volume 31 of *AMS/IP Studies in Advanced Mathematics*, American Mathematical Society, Providence, RI, 2002.

[31] N. V. Ivanov, *On the homology stability for Teichmüller modular groups: closed surfaces and twisted coefficients*, In Mapping class groups and moduli spaces of Riemann surfaces (Göttingen, 1991/Seattle, WA, 1991), volume 150 of Contemp. Math., pp. 149–194, Amer. Math. Soc., Providence, RI, 1993.

[32] M. Kreck, *Surgery and duality*, Ann. of Math. (2), **149**(3):707–754, 1999.

[33] A. Kupers and J. Miller, *E_n-cell attachments and a local-to-global principle for homological stability*, In preparation.

[34] J. Lurie, *On the classification of topological field theories*, In Current developments in mathematics, 2008, pp. 129–280, Int. Press, Somerville, MA, 2009.

[35] I. Madsen and M. Weiss, *The stable moduli space of Riemann surfaces: Mumford's conjecture*, Ann. of Math. (2), **165**(3):843–941, 2007.

[36] D. McDuff and G. Segal, *Homology fibrations and the "group-completion" theorem*, Invent. Math., **31**(3):279–284, 1975–76.

[37] W. Meyer, *Die Signatur von lokalen Koeffizientensystemen und Faserbündeln*, Bonn. Math. Schr., (53):viii+59, 1972.

[38] E. Y. Miller, *The homology of the mapping class group*, J. Differential Geom., **24**(1):1–14, 1986.

[39] S. Morita, *On the homology groups of the mapping class groups of orientable surfaces with twisted coefficients*, Proc. Japan Acad. Ser. A Math. Sci., **62**(4):148–151, 1986.

[40] ———, *Families of Jacobian manifolds and characteristic classes of surface bundles*, I. Ann. Inst. Fourier (Grenoble), **39**(3):777–810, 1989.

[41] ———, *Generators for the tautological algebra of the moduli space of curves*, Topology, **42**(4):787–819, 2003.

[42] D. Mumford, *Towards an enumerative geometry of the moduli space of curves*, In Arithmetic and geometry, Vol. II, volume 36 of *Progr. Math.*, pp. 271–328, Birkhäuser Boston, Boston, MA, 1983.

[43] S. Nariman, *Homological stability of diffeomorphism groups made discrete*, In preparation.

[44] R. Pandharipande and A. Pixton, *Relations in the tautological ring of the moduli space of curves*, Preprint, arXiv:1301.4561.

[45] F. S. Quinn, III, *A geometric formulation of surgery*, ProQuest LLC, Ann Arbor, MI, 1970, Thesis (Ph.D.)–Princeton University.

[46] O. Randal-Williams, *The homology of the stable nonorientable mapping class group*, Algebr. Geom. Topol., **8**(3):1811–1832, 2008.

[47] ———, *Resolutions of moduli spaces and homological stability*, arXiv:0909.4278, 2009.

[48] ———, *Relations among tautological classes revisited*, Adv. Math., **231**(3-4):1773–1785, 2012.

[49] C. Rezk, *A Cartesian presentation of weak n-categories*, Geom. Topol., **14**(1):521–571, 2010.

[50] G. Segal, *The definition of conformal field theory*, In Topology, geometry and quantum field theory, volume 308 of London Math. Soc. Lecture Note Ser., pp. 421–577, Cambridge Univ. Press, Cambridge, 2004.

[51] R. E. Stong, *Notes on cobordism theory*, Mathematical notes, Princeton University Press, Princeton, N.J., University of Tokyo Press, Tokyo, 1968.

[52] U. Tillmann, *On the homotopy of the stable mapping class group*, Invent. Math., **130**(2):257–275, 1997.

[53] N. Wahl, *Homological stability for the mapping class groups of non-orientable surfaces*, Invent. Math., **171**(2):389–424, 2008.

[54] F. Waldhausen, B. Jahren, and J. Rognes, *Spaces of PL manifolds and categories of simple maps*, volume 186 of Annals of Mathematics Studies, Princeton University Press, Princeton, N. J., 2013.

[55] C. T. C. Wall, *Classification problems in differential topology*, V. On certain 6-manifolds, Invent. Math. **1** (1966), 355–374, corrigendum, ibid, 2:306, 1966.

[56] ———, *Surgery on compact manifolds*, Academic Press, London, 1970, London Mathematical Society Monographs, No. 1.

[57] M. Weiss and B. Williams, *Automorphisms of manifolds and algebraic K-theory*, I. K-Theory, **1**(6):575–626, 1988.

[58] ———, *Automorphisms of Manifolds and Algebraic K-Theory: Part III*, Preprint, arXiv:1308.4047.

Department of Mathematics, Stanford University, Stanford CA, 94305
E-mail: galatius@stanford.edu

On the non-existence of elements of Kervaire invariant one

Michael A. Hill, Michael J. Hopkins, and Douglas C. Ravenel

Abstract. We sketch a proof of our solution to the Kervaire invariant one problem, showing that there are Kervaire invariant one manifolds only in dimensions 2, 6, 14, 30, 62, and possibly 126. This resolves a long-standing problem in algebraic and differential topology.

Mathematics Subject Classification (2010). Primary 55Q45; Secondary 57R60.

Keywords. Kervaire invariant, algebraic topology, equivariant homotopy, bordism, slice filtration.

1. Introduction

At the 1936 ICM in Oslo, Norway, Lefschetz presented on behalf of Pontryagin the following result:

Pontryagin's Theorem. *The second stable homotopy group of spheres is trivial.*

His approached this computation via geometry, establishing a perfect dictionary between geometric information, in this case the bordism classes of framed manifolds, and algebraic topology, here the homotopy groups of spheres. He deduced his theorem from the classification of surfaces, using his newly developed technique of framed surgery to reduce the genus of a framed surface until he had a framed sphere. Unfortunately, Pontryagin made a subtle mistake: it is not always possible to reduce the genus of a framed manifold by framed surgery. In fact, there is a quadratic obstruction to performing framed surgery, and this obstruction is non-trivial on, say, the 2-torus with the framing induced by the Lie multiplication.

Pontryagin's techniques have remained influential in algebraic topology, and his approach to bordism via surgery is still a very powerful tool. The obstruction to surgery is our primary object of study, the Kervaire invariant.

The early 1960s saw the topology community rocked by two related theorems. Milnor showed that there are smooth manifolds homeomorphic to the 7 sphere but not diffeomorphic to it [34]. Kervaire then showed that there are topological (really piecewise-linear), framed 10-manifolds that do not admit any smooth structure [24]. Kervaire did this by defining a quadratic form μ_f associated to a framing f on M:

$$\mu_f \colon H_5(M; \mathbb{Z}/2) \to \mathbb{Z}/2$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

which refines the intersection form $- \cap -$ in the sense that

$$\mu(x+y) = \mu(x) + \mu(y) + x \cap y.$$

The distinguishing feature of quadratic forms with values in $\mathbb{Z}/2$ is their Arf invariant, and Kervaire showed that for smooth manifolds M, the Arf invariant of μ_f vanishes for every f. He then produced a topological manifold and framing for which μ_f had nontrivial Arf invariant, proving the result. Kervaire's quadratic form is exactly the obstruction encountered by Pontryagin, just in dimension 10.

These results set the stage for Kervaire-Milnor's treatment of smooth structures on spheres [25]. They first observed that the set of smooth structures on the n-sphere is finite if $n > 4$, and that the diffeomorphism classes of n-spheres forms a group Θ_n under connected sum. Kervaire and Milnor observed that any exotic smooth structure on an n-sphere still produces an n-sphere, and hence is a framed manifold. They used this to produce a map

$$\psi_n \colon \Theta_n \to \pi_n S^0 / Im(J),$$

where $Im(J)$ is the subgroup of $\pi_n S^0$ consisting of all possible framings on the standard n-sphere. This uses Pontryagin's dictionary, as $\pi_n S^0$ is the same as the bordism classes of framed n-manifolds.

The kernel of ψ_n is, by definition, the collection of exotic spheres which bound frameable manifolds, and Kervaire-Milnor call this group bp_n. This group is cyclic, and

$$|bp_n| = \begin{cases} 1 & n \equiv 0 \mod 2 \\ 1 \text{ or } 2 & n \equiv 1 \mod 4 \\ 2^{2k-2}(2^{2k-1}-1)num\left(\frac{4B_k}{k}\right) & n = 4k-1 > 3, \end{cases}$$

where B_k is the k^{th} Bernoulli number. If n is not congruent to 2 modulo 4, then Kervaire and Milnor showed that ψ_n is surjective. When n is congruent to 2 modulo 4, there is a longer exact sequence

$$0 \to bp_n \to \Theta_n \xrightarrow{\Psi_n} \pi_n^s / Im(J) \xrightarrow{\Phi_n} \mathbb{Z}/2 \to \Theta_{n-1}^{bp} \to 0.$$

The map Φ_n is *exactly* the Kervaire invariant, so this allows us to describe our problem.

Problem. *For which n is there a manifold of Kervaire invariant 1?*

Equivalently, for which n are there smooth manifolds *not* framed bordant to an exotic sphere?

Browder greatly reduced the problem, tethering it to the fate of elements in the Adams spectral sequence. To describe this, we need a small amount of background.

Adams introduced his eponymous spectral sequence to solve the Hopf invariant one problem (and thereby to also determine which vector spaces $\mathbb{R}^n$ admit a division algebra structure) [1]. The spectral sequence starts by considering mod 2 cohomology together with its ring of natural endomorphisms $\mathcal{A}$, the Steenrod algebra. For Y of finite type, the Adams spectral sequence has the form

$$E_2^{s,t} = \operatorname{Ext}_{\mathcal{A}}^{s,t}\left(H^*(Y), \mathbb{Z}/2\right) \Rightarrow \pi_{t-s}(Y) \otimes \mathbb{Z}_2.$$

The most interesting cases for us is $Y = S^0$, where we are computing the derived $\mathcal{A}$-module endomorphisms of $\mathbb{Z}/2$. This is a bi-graded commutative algebra, the values of which for small s have been completely determined.

Adem showed that

$$\mathrm{Ext}_{\mathcal{A}}^{1,t}(\mathbb{Z}/2, \mathbb{Z}/2) = \begin{cases} \mathbb{Z}/2 \cdot h_j & t = 2^j \\ 0 & \text{otherwise.} \end{cases}$$

These elements have geometric content: the survival of h_j in the Adams spectral sequence is equivalent to the existence of an element of Hopf invariant one.

Adams showed that these elements generate Ext for small values of s. In particular

$$\mathrm{Ext}_{\mathcal{A}}^{2,?} = \{h_i h_j | |i - j| \neq 1\}.$$

Moreover, he produced a differential

$$d_2(h_j) = h_0 h_{j-1}^2$$

for $j \geq 4$.

Browder's reformulation of the Kervaire invariant one problem also uses the Adams spectral sequence.

Theorem (Browder [10]). *There can only be elements of Kervaire invariant one in dimensions of the form $2^k - 2$. Any element of Kervaire invariant one is detected in the Adams spectral sequence by h_{k-1}^2.*

Since h_1, h_2, and h_3 are themselves permanent cycles (corresponding to the framings on S^1, S^3, and S^7 coming from viewing them as the units in a division ring), so their squares are. These manifolds were known to be Kervaire invariant one manifolds, the first being the obstruction into which Kervaire ran.

Barratt and Mahowald showed the next case.

Theorem. *There is a 30 manifold of Kervaire invariant one.*

Jones produced the manifold. He began with a genus 5 surface Σ. Arranging 4 of the holes at the points of a square and placing the fifth at the center shows that there is a free action of the dihedral group of order 8, D_8 on this surface. Jones then produced a framing on

$$\Sigma \times_{D_8} (S^7)^4,$$

where D_8 acts on $(S^7)^4$ via a coinducing up the antipodal action.

Barratt-Jones-Mahowald then showed the next case.

Theorem (Barratt-Jones-Mahowald [8]). *The class h_5^2 survives the Adams spectral sequence.*

To date, there is no explicit construction of this manifold.

Our theorem resolves all but one remaining case, showing that h_j^2 is not a permanent cycle for $j \geq 7$ [20].

Main Theorem. *There are manifolds of Kervaire invariant one only in dimensions 2, 6, 14, 30, 62, and maybe 126.*

The proof follows from four smaller theorems concerning a C_8-equivariant spectrum Ω and its basic properties. This spectrum is by construction periodic in the same way that K-theory is periodic, and we use this natural self-symmetry to simplify the problem.

At this point, the appearance of equivariance might be surprising. The problem, as stated, is one of manifolds with no equivariance. Our method has an important antecedent, Ravenel's proof for primes larger than 3 of the analogous result [38].

1.1. Ravenel's proof for very odd primes and the Adams-Novikov Spectral Sequence. The first step in our argument is to lift the computation to a more universal example. At this point, we can no longer avoid spectra. A more detailed description will be given below when we discuss equivariant homotopy; we use them here mainly for expository flavor.

The Adams spectral sequence is really a recipe. Given any (ring) spectrum R, we can form an Adams tower which, if R has nice homological properties (like the R-homology of R itself is flat as an R_*-module), then we can identify the E_2 term of the spectral sequence. If R is moreover commutative, then we can describe this as a kind of descent spectral sequence approximating a spectrum X by R-module spectra via the Amitsur complex. As a bit of jargon, when used without any modifiers, the phrase "Adams spectral sequence" refers to the one described above, built out of the mod 2 Eilenberg-MacLane spectrum.

While $H\mathbb{F}_2$ is a beautiful commutative ring spectrum, it has the disadvantage that any algebra R over $H\mathbb{F}_2$ is again Eilenberg-MacLane. This means that the Adams spectral sequence based on an $H\mathbb{F}_2$-algebra R is *isomorphic* to the classical Adams spectral sequence. In some sense, that makes this spectral sequence terminal. There are no Adams spectral sequences which receive a map from the classical one that are not isomorphic to it.

Novikov observed that if instead we use the bordism theory of (almost) complex manifolds, MU, then we have an Adams spectral sequence, here called the Adams-Novikov spectral sequence, which maps to the classical Adams spectral sequence. The homotopy groups of MU were computed by Milnor to be a polynomial ring with one generator in every positive, even degree [35], and the MU-homology of MU is also polynomial:

$$MU_*MU \cong MU_*[b_1, \dots].$$

Quillen showed that this is closely related to formal group laws, as MU_* is the Lazard ring classifying formal group laws and MU_*MU classifies the universal isomorphism [37]. The descent formulation of the Adams spectral sequence then identifies the E_2-term of the Adams-Novikov spectral sequence with a descent spectral sequence on the moduli stack of formal groups. This algebro-geometric perspective has been deeply influential in algebraic topology for the last 50 years.

We need very little of the algebraic geometry here. However, the myriad spectra which receive ring maps from MU provide a host of other Adams spectral sequences in which we can detect elements. In particular, we will produce a theory (Ω) which sees the Kervaire classes at 2. We must first lift these elements.

Since we have a map of spectral sequences

$$ANSS \to ASS,$$

any element in the Adams-Novikov spectral sequence which detects the Kervaire classes must have filtration at most 2. The zero line of the Adams-Novikov E_2 term is simple: it is a copy of $\mathbb{Z}$ in dimension zero corresponding to the MU-Hurewicz image. Thus we cannot

see the Kervaire classes here. The one line, at all primes, is the image of J, $Im(J)$, which showed up in the Kervaire-Milnor result. This was computed up to a factor of 2 by Adams [2], and the ambiguity was resolved by Quillen and Mahowald [36], [29]. In particular, the Kervaire classes are not detected in $Im(J)$, and hence not on the one line. Thus if the Kervaire classes exist, then they are detected on the two line of the Adams-Novikov spectral sequence. Here work of Miller-Ravenel-Wilson, described in Ravenel's ICM talk, allows us to identify exactly what we see [33, 39]. For $p = 2$, the reader is also directed to work of Shimomura [41].

Theorem (Miller-Ravenel-Wilson [33]).

(1) *At odd primes, the* 2*-line is generated by elements* $\beta_{i/j,k}$.

(2) *At two, the* 2*-line is generated by elements* $\beta_{i/j,k}$ *and* $\alpha_1\alpha_j$*, where* α_i *are the generators on the* 1*-line.*

(3) *The elements* $\beta_{i/j,k}$ *and* α_i *are all readily and algebraically described using short exact sequences of* Ext *groups.*

Since we are mapping the two line of the Adams-Novikov spectral sequence to the two line of the Adams spectral sequence, this can be directly computed.

Theorem. *If* x *is an element on the Adams-Novikov* 2 *line maps to* h_j^2*, then*

$$x = \beta_{2^j/2^j,1} + (\textit{sum of monomials not involving} \beta_{2^j/2^j,1}).$$

In other words, any lift of h_j^2 to the Adams-Novikov spectral sequence involves $\beta_{2^j/2^j,1}$.

Ravenel considered the analogous problem for primes greater than 3 (sometimes called "very odd primes") [38].

Theorem (Ravenel [38]). *For primes* $p > 3$ *and for any* $j > 0$*, the classes*

$$\beta_{p^j/p^j,1}$$

are not permanent cycles in the Adams-Novikov spectral sequence.

Ravenel's strategy was to chose an appropriately simplified cohomology theory into which the classes $\beta_{p^j/p^j,1}$ map non-trivially and in which the images of these classes support differentials. The modern language of the higher real K-theories of Hopkins and Miller make this more transparent.

Theorem (Hopkins-Miller [40]). *There is a* C_p*-equivariant spectrum* E *such that*

(1) $\beta_{p^i/p^i,1}$ *maps to a non-trivial element in the Adams-Novikov spectral sequence for* E^{hC_p}.

(2) *The Adams-Novikov* E_2*-term is given by the group cohomology* $H^*(C_p; \pi_* E)$.

(3) *The image of* $\beta_{p^i/p^i,1}$ *supports a differential.*

We copy this argument at $p = 2$. Here, we need to use a larger group than C_2 to distinguish between the Kervaire elements; we use C_8.

1.2. Outline of the proof of the main theorem. We first need that out of Ω, we can build a spectrum which detects the Kervaire classes. In this case, we can mirror a refinement of an odd-primary argument of Ravenel and look in the homotopy fixed points of Ω.

Detection Theorem. *If there is a manifold of Kervaire invariant one, then the corresponding homotopy class is non-zero in the homotopy groups of homotopy fixed points of* Ω*:* Ω^{hC_8}.

This reduces the problem to studying the homotopy groups of the homotopy fixed points Ω^{hC_8}. This is still a tall order, so we approach an *a priori* harder computation and determine the actual fixed points.

Gap Theorem. *The group* $\pi_{-2}\Omega^{C_8}$ *is zero.*

The periodicity of Ω now also comes into play, as does the connection between the fixed and homotopy fixed points.

Periodicity Theorem. *The homotopy groups of* Ω^{hC_8} *are* 256*-periodic:*

$$\pi_{k+256}\Omega^{hC_8} = \pi_k\Omega^{hC_8}.$$

Homotopy Fixed Points Theorem. *The natural map*

$$\Omega^{C_8} \to \Omega^{hC_8}$$

is a weak equivalence, and hence the homotopy groups of the fixed and homotopy fixed points are the same.

The proof of the detection theorem is essentially identical to Ravenel's argument for primes bigger than 3, and so we will not sketch a proof here. We will provide a sketch of the remaining theorems below.

2. Equivariant homotopy, real K-theory, and real bordism

2.1. Equivariant homotopy. Equivariant homotopy theory is homotopy theory for spaces with an action of a fixed group G. For our purposes, G will be a finite group. Group actions on spaces are always assumed to be continuous. Excellent introductions are found in the books of tom Dieck [42, 43].

2.1.1. G-Spaces. There is a category of G-spaces. In $\mathcal{T}op^G$, the objects are G-spaces and the morphisms are equivariant maps, maps which commute with the G-action. This category is enriched in topological spaces, in that there is a space of maps between any two objects. It is also tensored and cotensored over space.

For any subgroup $H \subset G$, we have an obvious forgetful functor

$$i_H^* \colon \mathcal{T}op^G \to \mathcal{T}op^H,$$

and this functor has both adjoints. The left adjoint, induction, is given by the balanced product:

$$X \mapsto G \times_H X.$$

We should think of this as taking the disjoint union (the coproduct in $\mathcal{T}op^G$) of G/H copies of X and letting the group act by permuting the coordinates. The right adjoint, coinduction, is given by the mapping space

$$X \mapsto \mathrm{Map}_H(G, X),$$

where $\mathrm{Map}_H(-,-)$ is the space of H-equivariant maps. Here H and G both act on G on different sides. We think of this as the product of G/H copies of X, letting the group act by permuting the coordinates.

The values of induction and coinduction on the image of the restriction have other descriptions. We have natural homeomorphisms

$$G \times_H i_H^* X \cong G/H \times X \text{ and } \mathrm{Map}_H(G, i_H^* X) \cong \mathrm{Map}(G/H, X),$$

where G acts diagonally on $G/H \times X$ and where the final Map is the G-space of all continuous maps G/H to X, with the conjugation action.

Any G-space can be approximated by a G-CW complex. The definition is essentially identical to that of an ordinary CW-complex: we built our space by iterative forming mapping cones out of a distinguished collection of spaces, spaces of the form $T \times S^n$, where T is a discrete G-set and $n \geq 0$.

As is standard, in both spaces and spectra, we will let $[X, Y]^G$ denote the set of G-homotopy classes of equivariant maps from X to Y. When there is no ambiguity, we will suppress the G from the notation.

2.1.2. *G*-Spectra. We need the stable version of this. We actually work in the rigid, point-set category of orthogonal G-spectra, but for expository purposes, we describe Lewis-May-Steinberger spectra [27]. Adams gives a beautiful treatment of the classical approach to stabilizing the category of finite G-CW complexes [4], and this gives intuition for us. Adams also gives a nice general introduction to spectra and their connection to cohomology [3].

Classically, a spectrum is a sequence of spaces $\ldots, X_0, X_1, \ldots$ together with structure maps

$$\Sigma X_i \to X_{i+1}$$

whose adjoints are equivalences. Brown Representability [11] provides a dictionary linking cohomology theories and spectra via

$$E^n(X) \cong [X, E_n].$$

In the equivariant context, we grade on a larger set: the set of (isomorphism classes of) finite dimensional representations of G. These form a category, which we will denote by $\mathcal{S}p^G$. There is an obvious forgetful functor

$$i_H^* \colon \mathcal{S}p^G \to \mathcal{S}p^H,$$

and this again has both adjoints: $G_+ \wedge_H (-)$ and $\mathrm{Map}_H(G_+, -)$.

The key benefit of grading on all finite dimensional representations is the Wirthmüller isomorphism:

Theorem 2.1 (Wirthmüller [47]). *If X is a G-spectrum, then we have a natural equivalence*

$$G_+ \wedge_H X \simeq \mathrm{Map}_H(G_+, X).$$

This means that maps into

$$G/H_+ \cong G_+ \wedge_H S^0 \simeq \mathrm{Map}_H(G_+, S^0)$$

is nontrivial. In fact, the Wirthmüller isomorphism makes finite G-sets stably self-dual. In many ways, this is the defining feature of the "genuine" equivariant stable homotopy, as it endows G-spectra and their homotopy groups with much richer structure. The homotopy classes of maps between two G-spectra has a richer structure than classically: they are Mackey functors in the sense of Dress [12]. For expository reasons, we use a slight refinement of Dress's original definition [42].

Definition 2.2. Let $\mathcal{A}$ denote the Burnside category. The objects are finite G-sets, and the morphisms are given by the group completion of the monoid isomorphism classes of finite G-sets which map to $S \times T$ together with disjoint union. The composition is given by pullback.

This category is obviously isomorphic to its dual, and disjoint union is both the product and coproduct. The canonical isomorphisms $T \to T \times *$ and $T \to * \times T$ give distinguished elements in $\mathcal{A}(T, *)$ and $\mathcal{A}(*, T)$. We call the former the "transfer" and the latter the "restriction". The action of the automorphisms of T is the "Weyl action".

Definition 2.3. A Mackey functor is an additive functor

$$\underline{M} \colon \mathcal{A} \to \mathcal{A}b.$$

Example 2.4. The most important Mackey functors we will consider are "fixed point" Mackey functors. If M is a $\mathbb{Z}[G]$-module, then the assignment

$$T \mapsto \mathrm{Map}_G(T, M)$$

defines a Mackey functor, the value at G/H of which is just M^H. The restrictions are the natural inclusions, while the transfers are "summing over cosets". We will denote this $\underline{M}$. If $M = \mathbb{Z}$ with the trivial action, then we get the "constant Mackey functor $\underline{\mathbb{Z}}$". Here, the value at G/H is $\mathbb{Z}$ for all subgroups H, the Weyl action is trivial, the restriction maps are the identity, and the transfer from H to K is the index.

Theorem 2.5 (Tom Dieck). *The assignment*

$$T \mapsto \Sigma^\infty T_+$$

extends to a fully-faithful embedding of $\mathcal{A}$ into the homotopy category of G-spectra.

Corollary 2.6. *If $n \in \mathbb{Z}$, then we have a Mackey functor*

$$\underline{\pi}_n(X)(T) := [T_+ \wedge S^n, X]^G.$$

When evaluating on a G-set of the form G/H, we will often denote this by

$$\pi_n^H(X) := \underline{\pi}_n(X)(G/H).$$

On the G-set G/H, this gives us the n^{th} homotopy group of the H-fixed points of X:

$$\underline{\pi}_n(X)(G/H) \cong \pi_n(X^H).$$

However, we have a huge warning here: the fixed points of a G-spectrum are not what one would immediately think! For example, tom Dieck showed that the fixed points of the sphere spectrum S^0 split as a wedge of classifying spaces of Weyl groups of conjugacy classes of subgroups. This is the topological underpinning of the theorem cited above.

2.1.3. Bredon homology. Associated to any Mackey functor $\underline{M}$ is an Eilenberg-MacLane spectrum $H\underline{M}$. This is the homotopy type defined by the property

$$\underline{\pi}_n H\underline{M} = \begin{cases} \underline{M} & n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Just as in the classical case, the Eilenberg-MacLane spectra represent ordinary homology. In this case, the homology theory was described by Bredon [9]. We will use a slight refinement: our version of Bredon homology is actually Mackey functor valued and uses a chain complex of Mackey functors. Bredon's original treatment handled a slightly more general type of coefficients for spaces.

By definition, we have

$$\underline{H}_k(X; \underline{M}) \cong \underline{\pi}_k(X \wedge H\underline{M}). \tag{2.1}$$

In particular, we know immediately the homology of spheres, and more generally, of induced spheres:

$$\underline{H}_k(G/H_+ \wedge S^n; \underline{M}) \cong \begin{cases} \underline{M}_{G/H} & k = n \\ 0 & \text{otherwise,} \end{cases}$$

where $\underline{M}_{G/H}(T) = \underline{M}(G/H \times T)$. Now if X is a finite G-CW complex, then as above, X has a filtration $X^{[n]}$ such that each $X^{[n]}$ is a homotopy cofiber

$$T_{n+} \wedge S^{n-1} \xrightarrow{e_n} X^{[n-1]} \xrightarrow{\partial_n} X^{[n]}.$$

Our cellular chain complex arises from the obvious Mayer-Vietoris sequence for this, using the dimension axiom. As an aside, we assume here that T_n is finite (otherwise, we take the colimit). The boundary map in the complex comes from

$$T_{n+} \wedge S^{n-1} \to X^{[n-1]} \to T_{(n-1)+} \wedge S^{n-1},$$

This is an element in

$$[T_{n+} \wedge S^{n-1}, T_{(n-1)+} \wedge S^{n-1}] \cong [\Sigma^\infty T_{n+}, \Sigma^\infty T_{(n-1)+}] \cong \mathcal{A}(T_n, T_{n-1}),$$

and by functoriality, we know exactly how to evaluate a Mackey functor on this! This is the cellular differential. We summarize this into a proposition.

Proposition 2.7. *The Bredon cellular chain complex for a G-CW complex X is the chain complex $C_*^{cell}(X; \underline{M})$ with*

$$C_n^{cell}(X; \underline{M}) = \underline{M}_{T_n},$$

and boundary induced by the composite

$$\underline{M}\big((T_n \to T_{(n-1)}) \times -\big).$$

The Bredon cochain complex is the Bredon chain complex on the Spanier-Whitehead dual of X.

We now restrict attention to $G = C_{2^n}$, $\underline{M} = \underline{\mathbb{Z}}$, and X a representation or dual representation sphere. Proposition 2.7 says we need only find convenient G-CW models for representation spheres.

For this, we need to list the irreducible representations.

Definition 2.8.

(1) Let λ_n denote the 1-dimensional complex representation of C_{2^n} arising from the inclusion of the $2^{n\text{th}}$ roots of unity into $\mathbb{C}^\times$.

(2) Let $\lambda_n(k)$ denote the k^{th} complex tensor power of λ_n.

(3) Let σ_n denote the sign representation of C_{2^n}

(4) Let ρ_n denote the regular representation of C_{2^n}, and let $\bar{\rho}_n$ denote the quotient of the regular representation by the trivial subrepresentation.

We implicitly work 2-locally, and in this context, it is not difficult to see that if the 2-adic valuations of k and ℓ are equal, then

$$S^{\lambda_n(k)} \simeq S^{\lambda_n(\ell)}.$$

This lets us restrict attention to those k which are powers of 2. These representations are also more easily described as the composite

$$C_{2^n} \to C_{2^n}/C_{2^k} \cong C_{2^{n-k}} \to \mathbb{C}^\times.$$

In other words, we pull the representation λ_{n-k} back along the canonical quotient. The cell-structure then also pulls back in the obvious way (recognizing that $C_{2^{n-k}}/H \cong C_{2^n}/H'$ for the subgroup H' that projects to H). Similarly, the sign representation comes from

$$C_{2^n} \to C_2 \to \mathbb{R}^\times.$$

We therefore need only describe the cell structure for S^{λ_n} and S^{σ_1}, as all others follow from this.

For the former, we have a single S^0 for our zero cells: the point at infinity and the origin. We have a copy of $C_{2^n+} \wedge e^1$ for the one cells: the rays from the origin to the point at infinity and passing through the various roots of unity. We then have a single 2-cell: $C_{2^n+} \wedge e^2$ filling in the wedges between the rays. In particular, the boundary of our two cell comes in as the difference $1 - \gamma$ of 1 and a fixed generator γ. This gives the following picture:

$$S^{\lambda_n} = S^0 \cup C_{2^n+} \wedge e^1 \cup_{1-\gamma} C_{2^n+} \wedge e^2. \tag{2.2}$$

Again, the map labeled $1 - \gamma$ describes the cellular attaching map. The unlabeled attaching map is the canonical projection $C_{2^n+} \to S^0$.

The sign representation is even easier:

$$S^{\sigma_n} = S^0 \cup C_{2+} \wedge e^1.$$

We can view this as the unit sphere in the complex numbers together with complex conjugation.

Again, for $\lambda_n(2^k)$, we replace all instances of C_{2^n} with C_{2^n}/C_{2^k} and similarly for σ_n, we replace C_2 with $C_{2^n}/C_{2^{n-1}}$.

This is actually sufficient to build any other representation sphere. First note that if $\ell \geq k$, then

$$C_{2^n}/C_{2^k+} \wedge S^{\lambda_n(2^\ell)} \cong C_{2^n+} \wedge_{C_{2^k}} S^{\downarrow \lambda_n(2^\ell)} \cong C_{2^n+} \wedge_{C_{2^k}} S^2 \cong C_{2^n}/C_{2^k+} \wedge S^2. \tag{2.3}$$

In other words, from the point of view of cells induced up from C_{2^k} for $k \leq \ell$, the $S^{\lambda_n(2^\ell)}$ sphere is trivial. An easy induction argument then shows the following.

Proposition 2.9. *If V is a representation of C_{2^n} such that $V^{C_{2^k}} = V$, then for any $j \leq k$, there is a cell structure on $S^{V+\lambda_n(2^j)}$ of the form*

$$S^V \cup C_{2^n}/C_{2^j+} \wedge e^{\dim V+1} \cup C_{2^n}/C_{2^j+} \wedge e^{\dim V+2}.$$

If V is orientable (arising from a map $C_{2^n} \to SO(V)$), then the top cellular boundary map is again $1-\gamma$.

Proof. Smash S^V with the cell structure given by Equation 2.2. The isomorphisms given by Equation 2.3 then prove the first part of the proposition. The second part follows from observing that any oriented representation of C_{2^n} arises from a representation of S^1, and the map $1-\gamma$ is the map in the cell structure for S^1 as a C_{2^n}-space. □

We close with two computations. Choose once and for all an orientation of S^2. This gives an orientation for all other even spheres.

Proposition 2.10. *If V is orientable, then*

$$\underline{H}_{\dim V}(S^V; \underline{\mathbb{Z}}) \cong \underline{\mathbb{Z}}.$$

Proof. This follows immediately from Proposition 2.9 once we understand the base case of $V = \lambda_n(2^k)$. Here our Bredon cellular chain complex is

$$\underline{\mathbb{Z}} \leftarrow \underline{\mathbb{Z}}_{C_{2^n}/C_{2^k}} \xleftarrow{1-\gamma} \underline{\mathbb{Z}}_{C_{2^n}/C_{2^k}}.$$

An elementary-but-illuminating computation shows that the kernel of $1-\gamma$ is exactly $\underline{\mathbb{Z}}$. □

Remark 2.11. The bottom boundary map is non-trivial. It is the covariant map associated to the canonical quotient $C_{2^n}/C_{2^k} \to *$, and that is the transfer. Thus the differential is a Mackey functor lift of the transfer. The dual version, giving the Bredon cohomology is therefore a Mackey functor lift of the restriction map. This is the key ingredient in Theorem 2.16 below.

Definition 2.12. Let u_V be the generator of $\underline{H}_2(S^V; \underline{\mathbb{Z}})(G/G)$ which restricts to the fixed orientation of $S^{\dim V}$.

For the rest of the homology groups, we need Euler classes in the homotopy groups of spheres.

Definition 2.13. Let V be a representation of a finite group G. If V is such that $V^G = \{0\}$, then let

$$a_V \colon S^0 \to S^V$$

denote the inclusion of the origin and the point at infinity into S^V. This is an element in

$$\underline{\pi}_0 S^V(G/G).$$

The Hurewicz map provides for us elements in $H_0(S^V)$, and we will also denote these by a_V. Then we have the following relations connecting the orientation classes u_V and the Euler classes a_W.

Proposition 2.14. *For $G = C_{2^n}$, we have*

(1) $a_{V+W} = a_V a_W$ *and if* V *and* W *are oriented, then* $u_{V+W} = u_V u_W$.

(2) *If* $H \subset G$ *has* $V^H \neq \{0\}$, *then* $|G/H| a_V = 0$.

(3) *If* $k \leq \ell$, *then*

$$a_{\lambda_n(2^\ell)} u_{\lambda_n(2^k)} = 2^{\ell-k} a_{\lambda_n(2^k)} u_{\lambda_n(2^\ell)}.$$

Only the relation involving swapping the subscripts on a and u is any work, and this is a direct computation [21].

With this, we can completely determine a portion of the homology of representation spheres. Since $H\underline{\mathbb{Z}}$ is a ring spectrum, we have a natural map

$$(S^V \wedge H\underline{\mathbb{Z}}) \wedge (S^W \wedge H\underline{\mathbb{Z}}) \to S^{V+W} \wedge H\underline{\mathbb{Z}}.$$

This means that the collection of all groups

$$\underline{\pi}_*(S^\star \wedge H\underline{\mathbb{Z}})(G/G),$$

where $* \in \mathbb{Z}$ and where $\star$ runs over all isomorphism classes of representations, forms a ring. Proposition 2.14 completely determines this ring, and this in turn gives the homology of any representation sphere.

Theorem 2.15. *The ring* $\underline{\pi}_*(S^\star \wedge H\underline{\mathbb{Z}})(G/G)$ *is isomorphic to*

$$\mathbb{Z}[a_\sigma, a_{\lambda_n(2^k)}, u_{\lambda_n(2^j)} | 0 \leq k \leq n-2, 0 \leq j \leq n-1]/(rels)$$

where $(rels)$ *is the ideal*

$$(rels) = (2a_\sigma, 2^{n-k} a_{\lambda_n(2^k)}, a_{\lambda_n(2^\ell)} u_{\lambda_n(2^k)} = 2^{\ell-k} a_{\lambda_n(2^k)} u_{\lambda_n(2^\ell)}).$$

For the Gap Theorem, we need a small number of Bredon cohomology groups. In particular, we need

$$\underline{H}^2(S^{k\rho_n}; \underline{\mathbb{Z}})(G/G) = \underline{H}_{-2}(S^{-k\rho_n}; \underline{\mathbb{Z}})(G/G).$$

Theorem 2.16. *For all* $n \neq 0$ *and for all* $k \in \mathbb{Z}$, *the group*

$$\underline{H}_{-2}(S^{k\rho_n}; \mathbb{Z})(G/G)$$

is zero.

Proof. If $k \geq 0$, then $S^{k\rho_n}$ is a space, and hence it has no negative homology. Similarly, if $k \leq -3$, then $S^{k\rho_n}$ is (-2)-coconnected, and again, we have no

$$\underline{\pi}_{-2}(S^{k\rho_n} \wedge H\underline{\mathbb{Z}}) = \underline{H}{-2}(S^{k\rho_n}; \mathbb{Z}).$$

We need only consider the cell structure for S^{ρ_n} and $S^{2\rho_n}$. If $n = 1$, then the argument is left as an exercise. Otherwise, we have the following cell structures from Proposition 2.9, using the fact that the trivial representation is stabilized by everything and the index 2 subgroup stabilizes only 1 and σ_n:

$$S^{\rho_n} : S^1 \cup C_{2^n}/C_{2^{n-1}+} \wedge e^2 \cup C_{2^n}/C_{2^{n-2}} \wedge e^3 \cup \dots$$

and

$$S^{2\rho_2} : S^2 \cup C_{2^n}/C_{2^{n-1}+} \wedge e^3 \cup_{1-\gamma} C_{2^n}/C_{2^{n-1}+} \wedge e^4 \dots.$$

Spanier-Whitehead duality swaps all the signs and the arrows. Thus for the first, the relevant cochain complex is

$$\underline{\mathbb{Z}} \to \mathbb{Z}_{C_{2^n}/C_{2^{n-1}}} \to \mathbb{Z}_{C_{2^n}/C_{2^{n-2}}} \cdots$$

and

$$\underline{\mathbb{Z}} \to \mathbb{Z}_{C_{2^n}/C_{2^{n-1}}} \xrightarrow{1-\gamma} \mathbb{Z}_{C_{2^n}/C_{2^{n-1}}} \cdots.$$

The unlabeled map is the restriction map, by Remark 2.11, which is an isomorphism here! This lets us determine the homology in the second position in the first sequence and that in the first position in the second, getting zero. □

2.2. Real theories.

2.2.1. Real K-theory. Just as the heart of the Kervaire invariant one problem is about bordism classes of manifolds, the heart of our proof is a bordism theory: the bordism theory of Real manifolds. Real (always with a capital "R" to distinguish them from unReal ones) theories arose from work of Atiyah on K-theory [7]. Atiyah defined a $C_2 = Gal(\mathbb{C}/\mathbb{R})$-equivariant version of K-theory, Real K-theory $K_{\mathbb{R}}$ which records Galois descent for vector bundles.

If a space X has a trivial action, then there is a canonical real bundle associated to any Real bundle on X, namely the fixed points of the Real bundle. This identifies the fixed points of $K_{\mathbb{R}}$ with ordinary real K-theory:

$$K_{\mathbb{R}}^{C_2} = KO.$$

Similarly, if X has a free C_2-action, then the Real structure provides an identification of the fibers over a point x and those over its translates under the group action. This identifies the underlying spectrum of $K_{\mathbb{R}}$ with complex K-theory:

$$i_e^* K_{\mathbb{R}} = KU.$$

Atiyah also determined the relationship between the fixed and homotopy fixed points of $K_{\mathbb{R}}$. Using the nilpotence of η, he showed that the fixed and homotopy fixed points coincide:

$$KO = K_{\mathbb{R}}^{C_2} = K_{\mathbb{R}}^{hC_2}.$$

Real K-theory provides a model for the entire argument that follows. We have a Gap Theorem ($\pi_{-2}KO = 0$ by work of Bott), Atiyah proved the Homotopy Fixed Points theorem (and gave a model for how to prove such things in general), and he also showed how one can deduce the 8-fold periodicity of KO (giving another proof of Bott periodicity). We will produce the same sort of argument, but for a more complicated equivariant spectrum.

2.2.2. Real bordism. Building on Atiyah's work, Landweber and Araki considered a Real version of bordism, $MU_{\mathbb{R}}$ [6, 26]. This is the bordism theory of C_2-manifolds with a Real structure on their stable normal bundle.

First Araki and then Hu-Kriz extensively studied the fixed points of the spectrum $MU_{\mathbb{R}}$, determining the homotopy groups and showing that the fixed and homotopy fixed points agree [5, 23]. Similarly, essentially by construction, the underlying spectrum is the ordinary complex bordism spectrum MU. One of the most important features is that the homotopy groups indexed by regular representations are especially nice:

$$\pi_{*\rho_2} MU_{\mathbb{R}} \cong \pi_{2*} MU,$$

and by a celebrated theorem of Quillen, this is the Lazard ring classifying formal group laws [37].

This is related to a Real version of formal groups and formal group laws, and just as in the classical case, the spectrum $MU_{\mathbb{R}}$ classifies Real oriented spectra. This allows us to use classical techniques and approaches. We use this to help determine the homotopy groups of related spectra, but a key starting point is the determination of the homotopy type of the smash square of $MU_{\mathbb{R}}$:

$$MU_{\mathbb{R}} \wedge MU_{\mathbb{R}} \simeq MU_{\mathbb{R}}[\bar{b}_1, \bar{b}_2, \dots],$$

where $|\bar{b}_i| = i\rho_2$. This in particular applies to all higher smash powers of $MU_{\mathbb{R}}$, letting us completely determine the homotopy groups in multiples of ρ_2:

$$\pi_{*\rho_2}\big(MU_{\mathbb{R}}^{\wedge k}\big) \cong \big(\pi_{2*}MU\big)^{\otimes k}. \tag{2.4}$$

While $MU_{\mathbb{R}}$ has a very satisfying story, it is not sufficient to detect the Kervaire classes. For this, we need to study a larger group. Experience with the techniques similar to Ravenel's results for $p > 3$ show that C_8 is the smallest group which detects the Kervaire classes appropriately. For this, we need to produce a new ring spectrum which is a C_8 analogue of $MU_{\mathbb{R}}$. Additive induction is insufficient: the C_8-equivariant homotopy theory of that is essentially the same as the C_2-equivariant theory of $MU_{\mathbb{R}}$ by the Wirthmüller isomorphism. Instead, we use the norm functor provides a way to do this.

2.3. The norm. Equivariant spectra are a bisymmetric monoidal category. The wedge provides the sum in the category, while the smash product provides a multiplication. The norm functor is a multiplicative kind of induction. It is the rigid, spectral version of the Evens' transfer in group cohomology [14], which was imported into stable homotopy by Greenlees-May, [16].

Theorem 2.17. *There is a strong symmetric monoidal functor*

$$N_H^G \colon \mathcal{S}p^H \to \mathcal{S}p^G$$

which commutes with certain colimits and for which

$$N_H^G S^V \cong S^{\mathrm{Ind}_H^G V}$$

for every representation V of H.

The way we should think of $N_H^G(X)$ is "smashing together G/H copies of X", and letting the group act by permuting the factors as well. This describes a functor from spectra with an H-action to spectra with a G-action, and showing that it is homotopically meaningful is one of the most difficult parts of the proof.

2.3.1. Relation to commutative rings. Since the smash product is the coproduct on commutative ring spectra, the norm is the corresponding "induction" functor.

Theorem 2.18. *The norm lifts to the left adjoint to the forgetful functor i_H^*:*

$$N_H^G \colon \mathcal{C}omm^H \rightleftarrows \mathcal{C}omm^G \colon i_H^*.$$

This means that commutative ring maps out of the a norm are easy to compute, as they are determined by the underlying H-equivariant maps. Since the norm is the left adjoint to the forgetful funtor, for any G-eqivaraint commutative ring R, we have a natural map of G-equivariant commutative rings

$$N_H^G i_H^* R \to R.$$

In particular, this allows us to define several kinds of internal norms on the cohomology theory given by R.

Definition 2.19. If $x \in \left(i_H^* R\right)^0(X)$, then let $N_H^G(x) \in R^0(N_H^G X)$ also denote the composite

$$N_H^G X \xrightarrow{N_H^G x} N_H^G i_H^* R \to R,$$

where the last map is the counit to the adjunction.

Since the sphere is a commutative ring spectrum, we have norms of the Euler classes, and since $H\underline{\mathbb{Z}}$ is a commutative ring spectrum as well, we have norms of orientation classes.

Proposition 2.20. *If V is a representation of H of dimension d (orientable for the second line), we have*

$$N_H^G a_V = a_{\mathrm{Ind}_H^G V}$$
$$u_{d\,\mathrm{Ind}_H^G 1} N_H^G u_V = u_{\mathrm{Ind}_H^G V}.$$

With the machinery aside, we can move towards defining Ω.

Definition 2.21. Let $MU^{((G))}$ be the commutative G-equivariant ring spectrum

$$MU^{((G))} = N_{C_2}^G MU_{\mathbb{R}}.$$

Our spectrum Ω is a localization of this, the exact form of which is forced by the Detection, Periodicity, and Homotopy Fixed Points Theorems. We will return to this as we delve into their proofs.

3. Slice filtration and the Gap Theorem

The main computational tool we use is the "slice filtration". This generalizes groundbreaking work of Dugger [13], where he describes the C_2-equivariant version. The filtration and name are motivated by the motivic slice filtration of Voevodsky, especially as applied by Hopkins-Morel in their study of the motivic bordism spectrum MGL [44–46].

The idea behind the slice filtration is to mirror the construction of the Postnikov tower. Classically, we form the Postnikov tower by considering the localization functors $P^n(-)$ whose acyclics are exactly the n-connected spectra or spaces. The category of n-connected spectra is a localizing category in the sense of Farjoun [15]. We build a different collection of localizing subcategories, using non-trivial representation spheres as our basic ingredients [18]. This gives us a tower of equivariant spectra and an associated spectral sequence. For simplicity, we only describe what is seen for $MU^{((G))}$.

3.1. The slice filtration of norms of $MU_{\mathbb{R}}$. For the spectra $MU^{((G))}$, for G a cyclic 2-group, the slice tower is beautifully simple. To describe it, we need to use Equation 2.4 to produce enough homotopy elements.

Theorem 3.1. *Let $G = C_{2^n}$ be a cyclic 2-group, generated by an element γ. Then there are elements $\bar{r}_i \in \pi_{i\rho_2} MU^{((G))}$ such that*

$$\{\bar{r}_1, \dots, \gamma^{2^{n-1}-1}\bar{r}_1, \bar{r}_2, \dots\}$$

is a set of polynomial generators for

$$\pi_{*\rho_1} MU^{((G))} \cong \left(\pi_{*\rho_1} MU_{\mathbb{R}}\right)^{\otimes 2^{n-1}}.$$

If we need to consider multiple groups G, then we will add them as superscripts: $\bar{r}_i^G$.

This is a refinement of Equation 2.4. The restriction to C_2 of $MU^{((G))}$ is the $|G|/2$-fold smash power of $MU_{\mathbb{R}}$. This classifies a $|G|/2$-tuple of Real formal group laws, and the group G acts by permuting these formal groups. The elements $\bar{r}_i$ are then essentially the coefficients of the universal isomorphism between a chosen first formal group and γ on it.

Definition 3.2. Let $\mathcal{P}$ be the set of monic monomials in $\pi_{*\rho_1} MU^{((G))}$ with the generators $\bar{r}_j$.

If $\bar{p} \in \mathcal{P}$, let $H_{\bar{p}}$ denote the stabilizer of the reduction modulo 2 of $\bar{p}$.

If $\bar{p} \in \mathcal{P}$ is in $\pi_{k\rho_1} MU^{((G))}$, then let

$$|p| = 2k \text{ and } |\bar{p}| = \tfrac{2k}{|H_{\bar{p}}|}\rho_{H_{\bar{p}}}.$$

Then the heart of our identification of the slice tower is the following proposition (which follows formally from properties of the norm).

Proposition 3.3. *The spectrum*

$$A = \bigvee_{\bar{p}\in\mathcal{P}} S^{|\bar{p}|} = \bigvee_{\bar{p}\in\mathcal{P}/G} G_+ \wedge_{H_{\bar{p}}} S^{|\bar{p}|}$$

is an associative ring spectrum and the maps

$$\bar{r}_i \colon S^{i\rho_1} \to i_{C_2}^* MU^{((G))}$$

give a G-equivariant associative algebra map $A \to MU^{((G))}$.

This has an obvious monomial filtration by degree, so let

$$I_k = \bigvee_{p\in\mathcal{P}_n/G, |p|\geq 2k} G_+ \wedge_{H_{\bar{p}}} S^{|\bar{p}|}.$$

Then each I_k is an A-A-bimodule, and I_k/I_{k+1} is a wedge of induced regular representations spheres.

This filtration of A by bimodules induces a filtration on any other A-module, so in particular, on $MU^{((G))}$. This is one of our main theorems.

Theorem 3.4. *We have an equivalence*

$$MU^{((G))} \wedge_A S^0 = H\underline{\mathbb{Z}}$$

and the filtration of $MU^{((G))}$ by the ideals I_k has associated graded

$$A \wedge H\underline{\mathbb{Z}} = \bigvee_{p \in \mathcal{P}_n/G} G_+ \wedge_{H_{\bar{p}}} S^{|\bar{p}|} \wedge H\underline{\mathbb{Z}}.$$

This is the slice filtration of $MU^{((G))}$.

The proof of the second part follows from the first immediately. The first is quite difficult and lengthy, and we will not provide a proof here. There are several approaches that can be taken, all of which are underlain by simple geometry. In fact, this theorem can be proved by computing appropriate characteristic numbers for certain C_{2^n}-equivariant manifolds. Determining these manifolds is non-trivial.

Corollary 3.5 (Slice Spectral Sequence). *For any virtual representation V of G, there is a spectral sequence*

$$E_2^{s,t} = \bigoplus_{p \in \mathcal{P}_n/G} \pi_{t-s}^G\big(G_+ \wedge_{H_{\bar{p}}} S^{|\bar{p}|} \wedge S^V \wedge H\underline{\mathbb{Z}}\big) \Rightarrow \pi_{t-s-V} MU^{((G))}.$$

The differentials are Adams type differentials.

The t in $E_2^{s,t}$ records the degree of p and is suppressed in this formulation.

Thus to understand the slice E_2-term, we need only determine

$$\pi_*^G\big(G_+ \wedge_H S^{k\rho_H} \wedge S^V \wedge H\underline{\mathbb{Z}}\big) = H_*^H\big(S^{k\rho_H} \wedge S^V; \underline{\mathbb{Z}}\big),$$

but this was computed in Theorem 2.15.

3.2. The Gap Theorem. We can now prove the Gap Theorem. We give a slightly stronger form, as this subsumes the case of interest.

Gap Theorem. *If $G = C_{2^n}$, and $\bar{D} \in \pi_{m\rho_G}^G MU^{((G))}$ then the group*

$$\pi_{-2}^G\big(\bar{D}^{-1} MU^{((G))}\big)$$

is zero.

Proof. Consider the slice spectral sequence computing

$$\pi_*^G \Sigma^{mk\rho_G} MU^{((G))} = \pi_{*-mk\rho_G}^G MU^{((G))}.$$

By Corollary 3.5, the portion of the E_2 term contributing to π_{-2} is

$$\bigoplus_{p \in \mathcal{P}_n/G} H_2^{H_{\bar{p}}}\big(S^{|\bar{p}|} \wedge S^{\downarrow mk\rho_G}; \underline{\mathbb{Z}}\big).$$

Since the restriction of $mk\rho_G$ to $H(p)$ is of the $mk[G : H(p)]\rho_{H(p)}$, we see by Theorem 2.16 that all of these groups are zero. Thus

$$\pi_{-2} \Sigma^{mk\rho_G} MU^{((G))} = 0.$$

Since inverting an element in dimension $m\rho_G$ is given by a directed colimit of spectra of the form $\Sigma^{mk\rho_G} MU^{((G))}$ and since homotopy groups commute with directed colimits, we conclude that $\pi_{-2} \bar{D}^{-1} MU^{((G))} = 0$. □

4. Geometric fixed points and the homotopy fixed points and periodicity theorems

The Periodicity and Homotopy Fixed Points Theorems are two sides of the same coin, and we will see that the proofs are closely intertwined. Both rely on an argument like Atiyah's: we show that certain classes vanish, which given the equality between fixed and homotopy fixed points, and the reason these classes vanish is because of certain differentials. The vanishing of these classes causes other classes to be permanent cycles, and this gives the Periodicity Theorem. Both rely on the geometric fixed points.

4.1. Geometric fixed points and the norm. When discussing $MU_{\mathbb{R}}$, at no point did we discuss the fixed points of the Thom spaces making up the spectrum. In fact, in contrast to the case of $K_{\mathbb{R}}$, the fixed points of $MU_{\mathbb{R}}$ seem to not have geometric content. The connection back to the geometry is provided by the "geometric fixed points". This is a second kind of fixed points that is actually the one most people would guess when confronted with fixed points.

Theorem. *There is a functor*

$$\Phi^G \colon \mathcal{S}p^G \to \mathcal{S}p$$

that commutes with most homotopy colimits and which has the following properties:

(1) *If X has only cells induced up from proper subgroups, then $\Phi^G(X)$ is contractible.*
(2) *If $X = \Sigma^\infty Y$, then $\Phi^G(X) \simeq \Sigma^\infty Y^G$.*
(3) *For G-CW spectra X and Y, $\Phi^G(X \wedge Y) \simeq \Phi^G(X) \wedge \Phi^G(Y)$.*

Thus the geometric fixed points functor has all of the properties that we would expect a fixed point functor to have, thinking only of what we learn from spaces.

Since it commutes with smash products and essentially interacts nicely with suspensions, it is no surprise that it plays nicely with the norm functor.

Theorem 4.1. *The diagonal map induces a weak equivalence*

$$\Phi^H(X) \simeq \Phi^G N_H^G X.$$

Finally, generalizing the connection to suspension spectra, the geometric fixed points functor has the expected, geometric content for Thom spectra.

Theorem 4.2. *The geometric fixed points of $MU_{\mathbb{R}}$ is MO.*

To describe the geometric fixed points, we introduce the isotropy separation sequence. Let $\mathcal{P}$ denote the family of proper subgroups of G. If $G = C_{2^n}$, then let $E\mathcal{P}$ be the space

$$E\mathcal{P} \simeq \lim_{\to} S(n\sigma),$$

where $S(V)$ is the unit sphere in an orthogonal representation V. This has the property that

$$(E\mathcal{P}_+)^H \simeq \begin{cases} S^0 & H \subsetneq G \\ * & \text{otherwise.} \end{cases}$$

We have an obvious map $E\mathcal{P}_+ \to S^0$. Let $\tilde{E}\mathcal{P}$ denote the homotopy cofiber. In our case, this can be taken to be the infinite sign sphere:

$$\tilde{E}\mathcal{P} \simeq \lim_{\to} S^{n\sigma}.$$

The maps in the direct system are all multiplication by a_σ, so $\tilde{E}\mathcal{P}$ amounts to inverting a_σ.

Definition 4.3. Given a G-CW spectrum X, define the geometric fixed points by

$$\Phi^G(X) = \left(\tilde{E}\mathcal{P} \wedge X\right)^G.$$

The isotropy separation sequence is

$$E\mathcal{P}_+ \wedge X \to X \to \tilde{E}\mathcal{P} \wedge X.$$

The isotropy separation sequence has a purely algebraic description. Since $\tilde{E}\mathcal{P}$ is the infinite sign sphere, the map

$$X \to \tilde{E}\mathcal{P} \wedge X$$

is the map

$$X \to X[a_\sigma^{-1}].$$

This map also inverts all of the other Euler classes a_V; there are classes in the $RO(G)$-graded homotopy groups of S^0 of the form a_σ^k/a_V.

The geometric fixed points are the final tool we need. In particular, the nilpotence of any of the Euler classes results in a contractible geometric fixed point spectrum. We will use our understanding that the geometric fixed points of $MU^{((G))}$ is MO to produce differentials in the slice spectral sequence hitting multiples of powers of a_σ. Inverting the multiples will then produce contractible geometric fixed points.

4.2. A differential in the slice spectral sequence. First, a quick simplification from the observation that smashing with $\tilde{E}\mathcal{P}$ is the nullification of induced cells. For notation, let

$$\bar{n}_i = N_{C_2}^{C_{2^n}} \bar{r}_i,$$

let $\mathcal{N}$ denote the set of monic monomials in

$$\mathbb{Z}[\bar{n}_1, \bar{n}_2, \dots],$$

and let

$$|\bar{n}_i| = i\rho_{2^n}.$$

Theorem 4.4. *The inclusion*

$$\bigvee_{n \in \mathcal{N}} S^{|n|} \to A$$

induces an equivalence after smashing with $\tilde{E}\mathcal{P}$.

Moreover, since smashing with $\tilde{E}\mathcal{P}$ inverts all Euler classes, the regular representation spheres which appear are also simplified.

Proposition 4.5. *The inclusion*

$$a_{i\bar{\rho}}\colon S^i \to S^{i\rho}$$

induces an equivalences after smashing with $\tilde{E}\mathcal{P}$.

In light of this, let f_i denote the composite

$$S^i \xrightarrow{a_{i\bar{\rho}}} S^{i\rho} \xrightarrow{\bar{n}_i} MU^{((G))}.$$

This is detected in the slice spectral sequence by the eponymous element of filtration $i(|G|-1)$.

We want to compute the homotopy groups of the localization $a_\sigma^{-1}MU^{((G))}$. We know, since this gives the homotopy groups of the geometric fixed points, the answer: these are just the homotopy groups of MO! They are packaged in a non-trivial way, and we see a single pattern of differentials which achieves this.

The computational content of Theorem 4.4 is given in the following theorem about the $RO(G)$-graded slice spectral sequence. In particular, it says that only a very narrow collection of elements can contribute to the spectral sequence computing the homotopy of MO.

Theorem 4.6. *The inclusion*

$$\mathbb{Z}[u_{2\sigma}, f_1, f_2, \dots, a_\sigma, \dots]/2a_\sigma, \cdots \to E_2^{*,\star}$$

induces an isomorphism upon inverting a_σ. (Here the dots after a_σ stands for all of the other Euler classes, and the ones after the relation $2a_\sigma$ stand for the truncation of the other Euler classes.)

The classes a_V are all in the Hurewicz image, so they are necessarily permanent cycles. The classes f_i are explicit homotopy classes in $\pi_i^G MU^{((G))}$. Thus the only class that is not a permanent cycle in the algebra listed in Theorem 4.6 is $u_{2\sigma}$. Since the homotopy groups of MO are polynomial on classes in ever dimension not of the form 2^k-1, we have a single pattern of differentials which achieves this.

Theorem 4.7. *There are differentials*

$$d_{2^{k+1}+(2^{k+1}-1)(2^n-1)}(u_{2\sigma}^{2^k}) = a_\sigma^{2^{k+1}} f_{2^{k+1}-1}.$$

Remark 4.8. Theorem 4.6 also shows that the differential on $u_{2\sigma}^{2^k}$ is also the last possible, as the target is on the vanishing line for the slice spectral sequence.

This gives us an immediate corollary.

Corollary 4.9. *In the slice spectral sequence for $\bar{n}_{2^k-1}^{-1}MU^{((G))}$, the element $u_{2\sigma}^{2^k}$ is a permanent cycle.*

Corollary 4.10. *The geometric fixed points of $\bar{n}_{2^k-1}^{-1}MU^{((G))}$ are contractible.*

This is the essential step in both the Periodicity and Homotopy Fixed Points theorem. It is also the key step in proving Atiyah's result for Real K-theory! We include a picture of the slice spectral sequence for $K_{\mathbb{R}}$ as Figure 4.1. This was exactly the impetus for Dugger, and our Figure is essentially the same as his [13].

The element $\bar{v}_1$ is our $\bar{r}_1$ for C_2. The differential on $\bar{v}_1^2 u_{2\sigma}$ is the differential from Theorem 4.7:

$$\bar{v}_1^2 u_{2\sigma} \mapsto \bar{v}_1^3 a_\sigma^3.$$

We see that there is a horizontal vanishing line in the spectral sequence at filtration 3. The differentials on $u_{2\sigma}^2$ and higher are d_7 or longer, and hence these are permanent cycles.

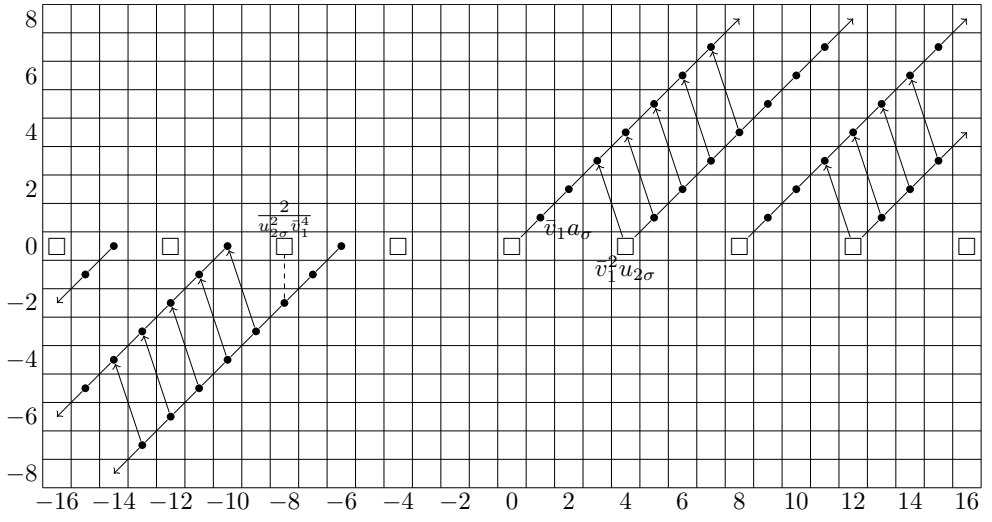


Figure 4.1. The slice spectral sequence for $K_{\mathbb{R}}$.

4.3. Periodicity and homotopy fixed points theorems. We first need a general result linking contractibility of geometric fixed points to the equivalence of fixed and homotopy fixed points.

Lemma 4.11. *Let R be a G-equivariant ring spectrum. If for all non-trivial H, the spectrum $\Phi^H(R)$ is contractible, then*

$$R \to F(EG_+, R)$$

is an equivalence.

Lemma 4.12. *Let M be an $MU^{((G))}$-module. If u_V is a permanent cycle in the slice spectral sequence for M, then u_V induces an equivalence*

$$\Sigma^{\dim V} M^{hG} \simeq (\Sigma^{\dim V} M)^{hG} \to (\Sigma^V M)^{hG}.$$

Proof. Since u_V survives the slice spectral sequence, it gives rise to an equivariant map

$$\Sigma^{\dim V} \bar{D}^{-1} MU^{((G))} \to \Sigma^V \bar{D}^{-1} MU^{((G))}.$$

Any map detected by u_V in the slice spectral sequence has the property that the underlying map is a weak equivalence. However, an equivariant map underlain by a weak equivalence induces a weak equivalence on homotopy fixed points, giving the result. □

If a class $\bar{D}$ we invert is a norm, then the resulting spectrum $\bar{D}^{-1}MU^{((G))}$ is again a commutative ring spectrum [19]. This means we can take the internal norms, which, on classes of the form u_V, were determined by Proposition 2.20.

We now have all of the ingredients. Corollary 4.10 provides a way to show that various geometric fixed point spectra are contractible, while Corollary 4.9 shows that classes $u_{2\sigma}^{2^k}$ are permanent cycles. This plays into the Periodicity Theorem, as the classes u_V are all orientation classes.

Definition 4.13. Let

$$\bar{D} = N_{C_2}^{C_8} \bar{r}_{1,3} N_{C_2}^{C_8} \bar{r}_{3,2} N_{C_2}^{C_8} \bar{r}_{15,1} \colon S^{19\rho_8} \to MU^{((C_8))}.$$

Let

$$\Omega = \bar{D}^{-1} MU^{((C_8))}.$$

Periodicity Theorem. *The homotopy groups of Ω^{hC_8} are 256-periodic.*

Homotopy Fixed Points Theorem. *The natural map*

$$\Omega^{C_8} \to \Omega^{hC_8}$$

is a weak equivalence.

We prove both simultaneously.

Proof. Since for any homotopy class x and for any groups $H \subset K \subset G$, the element $N_H^K(x)$ divides $i_K^* N_H^G(x)$, we know that inverting $\bar{D}$ inverts $N_{C_2}^{C_4} \bar{r}_{3,2}$ and $\bar{r}_{15,1}$. Thus $i_{C_4}^* \Omega$ is a module over $N_{C_2}^{C_4} \bar{r}_{3,2}^{-1} MU^{((C_4))}$ and similarly for $i_{C_2}^* \Omega$. Since by Corollary 4.10 each of these rings has contractible geometric fixed points, we conclude that for all nontrivial H, $\Phi^H \Omega$ is contractible. Lemma 4.11 then gives the desired equivalence.

By Corollary 4.9, we then conclude that $u_{2\sigma_8}^2$, $u_{2\sigma_4}^4$, and $u_{2\sigma_2}^{16}$ are permanent cycles. Combining the parts of Proposition 2.20 then shows that $u_{32\rho_8}$ is a permanent cycle in the slice spectral sequence for Ω.

Inverting $\bar{D}$ automatically makes the homotopy groups of Ω ρ_8-periodic. Combining this with Lemma 4.12 then shows that the homotopy groups of Ω^{hC_8} are 256-periodic. □

Acknowledgements. The authors were supported by DARPA Grant FA9550-07-1-0555; NSF Grants DMS-0905160, 1307896, 0906285, and 1207774; and the Sloan Foundation. We also thank MSRI for its hospitality while the author was in residence, Spring 2014, with support provided by NSF Grant 0932078-000

References

[1] J. F. Adams, *On the non-existence of elements of Hopf invariant one*, Ann. of Math. (2) **72** (1960), 20–104.

[2] ———, *On the groups $J(X)$. IV*, Topology **5** (1966) 21–71.

[3] ———, *Stable homotopy and generalised homology*, University of Chicago Press, Chicago, 1974.

[4] ———, *Prerequisites (on equivariant stable homotopy) for Carlsson's lecture*, Algebraic topology, Aarhus 1982 (Aarhus, 1982), Lecture Notes in Math., vol. 1051, Springer, Berlin, 1984, pp. 483–532.

[5] Shôrô Araki, *Coefficients of MR-theory*, Available online at www.math.rochester.edu/u/faculty/doug/.

[6] ———, *Orientations in τ-cohomology theories*, Japan. J. Math. (N.S.), **5**(2):403–430, 1979.

[7] M. F. Atiyah, *K-theory and reality*, Quart. J. Math. Oxford Ser. (2), **17**:367–386, 1966.

[8] M. G. Barratt, J. D. S. Jones, and M. E. Mahowald, *Relations amongst Toda brackets and the Kervaire invariant in dimension 62*, Journal of the London Mathematical Society, **30** (1985), 533–550.

[9] G. E. Bredon, *Equivariant cohomology theories*, Lecture Notes in Mathematics, No. 34. Springer-Verlag, Berlin-New York (1967).

[10] W. Browder, *The Kervaire invariant of framed manifolds and its generalization*, Annals of Mathematics, **90** (1969), 157–186.

[11] E. H. Brown, Jr., *Cohomology theories* Ann. of Math. (2) **75** (1962), 467–484.

[12] A. W. M. Dress, *Contributions to the theory of induced representations, In Algebraic K-theory, II: "Classical" algebraic K-theory and connections with arithmetic (Proc. Conf., Battelle Memorial Inst., Seattle, Wash., 1972)*, pp. 183–240, Lecture Notes in Math., Vol. **342**. Springer, Berlin, 1973.

[13] D. Dugger, *An Atiyah-Hirzebruch spectral sequence for KR-theory*, K-Theory, **35**(3-4):213–256 (2006), 2005.

[14] L. Evens, *A generalization of the transfer map in the cohomology of groups*, Trans. Amer. Math. Soc., **108**:54–65, 1963.

[15] E. Dror Farjoun, *Cellular spaces, null spaces and homotopy localization*, volume 1622 of Lecture Notes in Mathematics, Springer-Verlag, New York, 1996.

[16] J. P. C. Greenlees and J. P. May, *Equivariant stable homotopy theory*, In Handbook of algebraic topology, pp. 277–323, North-Holland, Amsterdam, 1995.

[17] ———, *Localization and completion theorems for MU-module spectra*, Ann. of Math. (2), **146**(3):509–544, 1997.

[18] M. A. Hill, *The equivariant slice filtration: a primer* Homology Homotopy Appl., **14** (2012), no 2, 143–166.

[19] M. A. Hill and M. J. Hopkins, *Equivariant Multiplicative Closure*, arXiv:1303.4479[math.AT], March 2013.

[20] M. A. Hill, M. J. Hopkins, and D. C. Ravenel, *On the non-existence of elements of Kervaire invariant one*, Available on the arxiv, 2009.

[21] ———, *The Slice Spectral*, Sequence for $RO(C_{p^n})$-graded Suspensions of $H\mathbf{Z}$ I, In preparation.

[22] M. J. Hopkins and F. Morel, *On the zero slice of MGL and S^0*, In preparation.

[23] P. Hu and I. Kriz, *Real-oriented homotopy theory and an analogue of the Adams-Novikov spectral sequence*, Topology, **40**(2):317–399, 2001.

[24] M. A. Kervaire, A manifold which does not admit any differentiable structure, *Comment. Math. Helv.* **34** (1960), 257–270.

[25] M. A. Kervaire and J. W. Milnor, *Groups of homotopy spheres. I*, Ann. of Math. (2) **77** (1963), 504–537.

[26] P. S. Landweber, *Conjugations on complex manifolds and equivariant homotopy of* MU, Bull. Amer. Math. Soc., **74**:271–274, 1968.

[27] L. G. Lewis, J. P. May, and M. Steinberger, *Equivariant Stable Homotopy Theory*, volume 1213 of Lecture Notes in Mathematics, Springer-Verlag, New York, 1986.

[28] L. G. Lewis, Jr, *The* $R\mathrm{O}(G)$*-graded equivariant ordinary cohomology of complex projective spaces with linear* $\mathbf{Z}/p$ *actions*, In Algebraic topology and transformation groups (Göttingen, 1987), volume 1361 of Lecture Notes in Math., pp. 53–122, Springer, Berlin, 1988.

[29] M. Mahowald, *On the order of image of* J, Topology, **6** (1967), 371–378.

[30] M. A. Mandell and J. P. May, *Equivariant orthogonal spectra and* S*-modules*, Mem. Amer. Math. Soc., **159**(755):x+108, 2002.

[31] M. A. Mandell, J. P. May, S. Schwede, and B. Shipley, *Model categories of diagram spectra*, Proc. London Math. Soc. (3), **82**(2):441–512, 2001.

[32] J. P. May, *Equivariant homotopy and cohomology theory*, volume 91 of *CBMS Regional Conference Series in Mathematics*, Published for the Conference Board of the Mathematical Sciences, Washington, D.C., 1996, With contributions by M. Cole, G. Comezaña, S. Costenoble, A. D. Elmendorf, J. P. C. Greenlees, L. G. Lewis, Jr., R. J. Piacenza, G. Triantafillou, and S. Waner.

[33] H. R. Miller, D. C. Ravenel, and W. S. Wilson, *Periodic phenomena in the Adams–Novikov spectral sequence*, Annals of Mathematics, **106** (1977), 469–516.

[34] J. W. Milnor, *On manifolds homeomorphic to the 7-sphere*, Ann. of Math. (2), **64** (1956), 399–405.

[35] ———, *On the cobordism ring* Ω^* *and a complex analogue*, Part I, American Journal of Mathematics, **82** (1960), 505–521.

[36] D. G. Quillen, *The Adams conjecture*, Topology, **10** (1971), 67–80.

[37] ———, *On the formal group laws of unoriented and complex cobordism theory*, Bulletin of the American Mathematical Society, **75** (1969), 1293–1298.

[38] D. C. Ravenel, *The nonexistence of odd primary Arf invariant elements in stable homotopy theory*, Math. Proc. Cambridge Phil. Soc., **83** (1978), 429–443.

[39] ———, *Complex cobordism and its applications to homotopy theory*, Proceedings of the International Congress of Mathematicians (Helsinki, 1978), 491–496.

[40] C. Rezk, *Notes on the Hopkins-Miller theorem*, In Homotopy theory via algebraic geometry and group representations (Evanston, IL, 1997), volume 220 of Contemp. Math., pp. 313–366, Amer. Math. Soc., Providence, RI, 1998.

[41] K. Shimomura, *Novikov's* Ext^2 *at the prime* 2, *Hiroshima Math. J.*, **11** (1981), no. 3, 499–513.

[42] T. tom Dieck, *Transformation groups and representation theory*, Lecture Notes in Mathematics, vol. 766, Springer, Berlin, 1979.

[43] ———, *Transformation groups*, de Gruyter Studies in Mathematics, vol. 8, Walter de Gruyter & Co., Berlin, 1987.

[44] V. Voevodsky, *Open problems in the motivic stable homotopy theory. I, Motives, polylogarithms and Hodge theory*, Part I (Irvine, CA, 1998), Int. Press Lect. Ser., vol. 3, Int. Press, Somerville, MA, 2002, pp. 3–34.

[45] ———, *A possible new approach to the motivic spectral sequence for algebraic* K*-theory*, Recent progress in homotopy theory (Baltimore, MD, 2000), Contemp. Math., vol. 293, Amer. Math. Soc., Providence, RI, 2002, pp. 371–379.

[46] ———, *On the zero slice of the sphere spectrum*, Tr. Mat. Inst. Steklova, **246** (2004), no. Algebr. Geom. Metody, Svyazi i Prilozh., 106–115.

[47] K. Wirthmüller, *Equivariant homology and duality*, Manuscripta Math., **11** (1974), 373–390.

Department of Mathematics, University of Virginia, Charlottesville, VA 22904
E-mail: mikehill@virginia.edu

Department of Mathematics, Harvard University, Cambridge, MA 02138
E-mail: mjh@math.harvard.edu

Department of Mathematics, University of Rochester, Rochester, NY
E-mail: douglas.ravenel@rochester.edu

Heegaard splittings of 3-manifolds

Tao Li

Abstract. Heegaard splitting is one of the most basic and useful topological structures of 3-manifolds. In the past few years, much progress has been made on Heegaard splittings and several long-standing questions have been answered. In this paper, we review some recent progress in studying Heegaard splittings and discuss related open problems.

Mathematics Subject Classification (2010). Primary 57N10, 57M50; Secondary 57M25.

Keywords. Heegaard splitting, 3-manifold.

1. Introduction

A (3-dimensional) *handlebody* is a compact orientable 3–manifold with boundary that is homeomorphic to a closed regular neighborhood of a graph in $\mathbb{R}^3$. A *Heegaard splitting* of a closed orientable 3–manifold M is a decomposition of M into two handlebodies along an embedded surface called *Heegaard surface*. A Heegaard splitting can also be naturally associated to a Morse function $f\colon M \to \mathbb{R}$. The idea of Heegaard splitting was introduced by Poul Heegaard in his 1898 thesis [15] in which Heegaard constructed a counterexample to an early version of Poincaré's duality theorem. Heegaard splitting has been extensively studied by many mathematicians, such as Haken and Waldhausen in the 1960s and 70s, and it was a major tool in their efforts to prove the Poincaré Conjecture. In the 1980s, Casson and Gordon introduced an extremely powerful new concept called strongly irreducible Heegaard splitting, and this totally changed the study of Heegaard splittings. In particular, it leads to the resolution of several long-standing conjectures, e.g. [24, 25, 28, 30].

The definition of Heegaard splitting can be extended to 3-manifolds with boundary by replacing handlebody with compression body. A compression body can be viewed as the manifold obtained from a handlebody by removing a regular neighborhood of a portion of its core graph. A handlebody can be viewed as a special compression body. It follows from a theorem of Bing and Moise [2, 35] that every orientable 3–manifold has a Heegaard splitting.

Three–manifolds have some very nice topological and geometric structures. For example, a 3–manifold has a canonical prime decomposition along essential 2–spheres; see [16, 18, 21, 34]. A 3–manifold that contains no essential 2–sphere is said to be irreducible. An irreducible 3–manifold can be canonically decomposed into simpler pieces along essential tori, which is called a JSJ decomposition; see [19, 20]. Heegaard splittings also have close connections with these decompositions. Note that Thurston's geometrization, proved by Perelman, says that each closed 3–manifold with trivial prime and JSJ decompositions has one of eight geometries [42–44, 58].

Proceedings of the International Congress of Mathematicians, Seoul, 2014

Definition 1.1. Let $M = V \cup_S W$ be a Heegaard splitting of a 3–manifold M, where V and W are compression bodies and S is the Heegaard surface. The Heegaard splitting is *reducible* if there is a nontrivial simple loop in S that bounds disks in both V and W, otherwise the splitting is said to be *irreducible*. The Heegaard splitting is *weakly reducible* if there are disks $D_V \subset V$ and $D_W \subset W$ such that ∂D_V and ∂D_W are a pair of disjoint nontrivial loops in S. The Heegaard splitting is *strongly irreducible* if it is not weakly reducible.

Irreducible and strongly irreducible Heegaard splittings give tremendous information about the 3-manifolds. For example, by Haken's lemma [5, 10], a Heegaard splitting of a reducible 3–manifold is always reducible. If the 3–manifold contains an incompressible torus, then the distance (see section 3) of any Heegaard splitting is at most 2.

In their seminal work [5], Casson and Gordon showed that if a 3–manifold admits an irreducible but weakly reducible Heegaard splitting, then the manifold must be a Haken 3–manifold. The theorem of Casson and Gordon implies that an irreducible non-Haken 3–manifold always has a strongly irreducible Heegaard splitting. For Haken manifolds, Scharlemann and Thompson [49] gave a natural construction called untelescoping of a Heegaard splitting, which is a decomposition of the Haken 3–manifold into several "blocks" along incompressible surfaces, such that there is a strongly irreducible Heegaard splitting in each block and the original Heegaard splitting can be recovered by amalgamating these strongly irreducible Heegaard splittings along the incompressible surfaces (see [51] for a more detailed description).

Strongly irreducible Heegaard splittings have many remarkable properties. In the next few sections, we will discuss some of these properties and recent progress in studying Heegaard splittings.

2. Normal surfaces and branched surfaces

A Heegaard splitting $M = U \cup_S V$ can be viewed as a sweepout $H : S \times (I, \partial I) \to (M, \Sigma_U \cup \Sigma_V)$, where $I = [0, 1]$, Σ_U and Σ_V are the cores of the compression bodies U and V respectively, and $H|_{S\times(0,1)}$ is an embedding. If one considers the intersection of the sweepout of a strongly irreducible Heegaard splitting with another object (e.g. a surface), then being strongly irreducible puts some great restrains on the intersection patterns. This leads to many remarkable properties of strongly irreducible Heegaard surfaces. For example, by studying how the sweepout meets the 2–skeleton of a triangulation of the 3–manifold, Rubinstein [48] (also see [56]) discovered that a strongly irreducible Heegaard surface can be isotoped into a nice position with respect to any given triangulation. A surface in such a position is called an almost normal surface, which is a slight generalization of the classical normal surface theory studied by Kneser and Haken [9, 21]. A surface is normal with respect to a triangulation if its intersection with every tetrahedron is a collection of normal disks (i.e. triangles or quadrilaterals that meet each edge in at most one point and meet each face of the tetrahedron in at most one edge). A surface is almost normal if its intersection with the tetrahedra consists of normal disks and exactly one almost normal piece, where an almost normal piece in a tetrahedron is either an octagon or an annulus obtained by connecting two normal disks by an unknotted tube; see [56] for details.

Theorem 2.1 (Rubinstein [48], Stocking [56]). *Given a triangulation of a compact orientable 3–manifold M, any strongly irreducible Heegaard surface is isotopic to a normal or*

an almost normal surface with respect to the triangulation.

The idea of sweepout and thin position used in Theorem 2.1 is very powerful in solving some decision problems in 3–manifold topology. In particular, Rubinstein and Thompson proved:

Theorem 2.2 (Rubinstein [48], Thompson [57]). *There is an algorithm to decide whether a closed 3–manifold is S^3.*

There is a natural connection between normal surfaces and branched surfaces. A branched surface can be viewed as a 2-dimensional analogue of a train track; see [23, 24, 28] for a discussion. A useful property of normal and almost normal surfaces is that these surfaces have only a finite number of different types of pieces in each tetrahedron. By identifying the normal disks of the same type, we obtain a branched surface. This implies that there is a finite collection of branched surfaces such that, after isotopy, every strongly irreducible Heegaard surface is carried by a branched surface in this collection. Using branched surfaces and measured laminations, we proved the so-called generalized Waldhausen Conjecture [24].

Theorem 2.3 ([24]). *A closed, orientable, irreducible and atoroidal 3–manifold has only finitely many Heegaard splittings in each genus, up to isotopy.*

A modified proof of Theorem 2.3 gives an algorithm to determine the Heegaard genus of a 3–manifold [28].

Theorem 2.4 ([28]). *There is an algorithm to determine the Heegaard genus of a closed orientable 3–manifold.*

A famous example of Casson and Gordon [6] shows that a 3–manifold can have an infinite family of strongly irreducible Heegaard splittings with different genera. This is constructed by "spinning" a Heegaard surface around an incompressible surface, and in particular, the 3–manifold is Haken. In [25], we used branched surfaces and laminations to study Heegaard surfaces and showed that this phenomenon happens only if there is an incompressible surface in the 3–manifold (i.e. the manifold is Haken); see Theorem 2.5. The proof is a limiting argument: Suppose a 3–manifold M contains an infinite family of strongly irreducible Heegaard splittings. Since all strongly irreducible Heegaard surfaces are isotopic to normal or almost normal surfaces, there is a branched surface carrying infinitely many strongly irreducible Heegaard surfaces. Thus, a subsequence of these Heegaard surfaces has a limit in the projective measured lamination space. By exploring some properties of strongly irreducible Heegaard splittings, we were able to show in [25] that this limit measured lamination is an essential lamination, which implies that the branched surface carries an incompressible surface and hence the manifold is Haken.

Theorem 2.5 ([25]). *A closed orientable non-Haken 3–manifold has only finitely many irreducible Heegaard splittings, up to isotopy.*

In the construction of Casson and Gordon [6], the sequence of strongly irreducible Heegaard surfaces $\{S_n\}$ can be expressed as $S_n = S + nF$, where S is a Heegaard surface, F is a closed incompressible surface and the sum is similar to the Haken sum for normal surfaces. Theorem 2.5 says that if a 3–manifold contains infinitely many strongly irreducible Heegaard surfaces, then the manifold must contain an incompressible surface. However, the limiting argument in the proof does not give a satisfactory picture. The following is an interesting open question. This question was also studied in [39].

Question 2.6. *Suppose a closed orientable 3–manifold M contains infinitely many pairwise nonisotopic strongly irreducible Heegaard surfaces. Does M always contain a sequence of strongly irreducible Heegaard surfaces $\{S_n\}$ that can be expressed as $S_n = S + nF$, where S is a Heegaard surface and F is an incompressible surface?*

In a certain sense, normal surface can be viewed as a combinatorial analogue of minimal surface. This is particularly true for incompressible surfaces and strongly irreducible Heegaard surfaces. For example, using the idea of sweepout, Pitts and Rubinstein proved the following theorem. This theorem provides a useful tool to study Heegaard surfaces using geometry.

Theorem 2.7 (Pitts-Rubinstein). *Let S be a strongly irreducible Heegaard surface in a closed 3–manifold. Then there is a minimal surface F such that S is either isotopic to F or to the surface obtained from the boundary of a regular neighborhood of F by attaching a 1–handle.*

3. Curve complex and Hempel distance

Let S be a closed orientable surface of genus at least 2. The curve complex of S, introduced by Harvey [13], is the complex whose vertices are the isotopy classes of essential simple closed curves in S, and $k+1$ vertices determine a k–simplex if they are represented by pairwise disjoint curves. We denote the curve complex of S by $\mathcal{C}(S)$. The curve complex of a torus is defined similarly, and $\mathcal{C}(T^2)$ is the same as the Farey graph. For any two vertices in $\mathcal{C}(S)$, the distance $d(x,y)$ is the minimal number of 1–simplices in a simplicial path jointing x to y.

Curve complex has been an important ingredient in the solutions of several important questions in low-dimensional geometry and topology, e.g. [30, 32, 33]. In [17], Hempel used the curve complex to define a certain complexity for Heegaard splittings: Given a Heegaard splitting $M = H_1 \cup_S H_2$. Let $\mathcal{H}_i$ $(i = 1, 2)$ be the set of vertices in $\mathcal{C}(S)$ represented by curves bounding disks in H_i. The distance $d(S)$ is defined to be the distance between $\mathcal{H}_1$ and $\mathcal{H}_2$ in the curve complex $\mathcal{C}(S)$. Clearly, the Heegaard splitting is reducible if and only if $d(S) = 0$, and if $g(S) \geq 2$, the Heegaard splitting is strongly irreducible if and only if $d(S) \geq 2$.

Hartshorn first observed that the Hempel distance of a Heegaard splitting is related to the genus of an incompressible surface [12]. Later, Scharlemann and Tomova showed that the same is true for strongly irreducible surfaces and in a much more general setting [50] (see [26] for another proof).

Theorem 3.1 (Hartshorn [12]). *Let $M = H_1 \cup_S H_2$ be a Heegaard splitting of a compact 3–manifold M. Suppose M contains a closed orientable incompressible surface P. Then $d(S) \leq 2g(P)$.*

Theorem 3.2 (Scharlemann-Tomova [50]). *Let $M = H_1 \cup_S H_2$ be a Heegaard splitting of a compact 3–manifold M. Suppose M contains a closed strongly irreducible surface P. Then $d(S) \leq 2g(P)$.*

Theorem 3.1 and Theorem 3.2 imply that if the Hempel distance of a Heegaard splitting is large, then the 3–manifold M contains no other small-genus Heegaard surface, and in

particular, this Heegaard splitting is the unique minimal-genus Heegaard splitting of the 3–manifold.

In [29], Hempel distance is generalized to measure the complexity of an amalgamation of any two compact 3-manifolds; see also [27]. Let F be a closed orientable surface embedded in a closed orientable 3–manifold M, and suppose F divides M into two submanifolds U_1 and U_2 with $\partial U_1 = \partial U_2 = F$. We may view $M = U_1 \cup_F U_2$ as an amalgamation of U_1 and U_2. Let $\mathcal{U}_i$ $(i = 1, 2)$ be the set of vertices in $\mathcal{C}(F)$ represented by boundary curves of properly embedded essential surfaces in U_i with maximal Euler characteristic. The amalgamation distance can be defined as $d(M) = d(\mathcal{U}_1, \mathcal{U}_2)$ in the curve complex $\mathcal{C}(F)$. For example, if F is compressible in U_i, then $\mathcal{U}_i$ consists of vertices represented by curves bounding disks in U_i. So the amalgamation distance is the same as the Hempel distance if F is a Heegaard surface. In [29], we proved the following theorem.

Theorem 3.3 ([29]). *Let $M = U_1 \cup_F U_2$ as above. If the amalgamation distance $d(M)$ is sufficiently large, then all small-genus Heegaard surfaces are standard.*

Theorem 3.3 is a fundamental tool used in answering the rank versus genus conjecture for closed hyperbolic 3-manifolds, which we will discuss in the next section.

4. Rank and Heegaard genus

Given a Heegaard splitting $M = U \cup_S V$ of a compact orientable 3-manifold M, if U is a handlebody, then the Heegaard splitting gives a natural presentation of the fundamental group of M: the core graph of U give a set of generators and a set of compressing disks in the compression body V gives a set of relators.

The Heegaard genus of a 3–manifold M, denoted by $g(M)$, is defined as the minimal genus of all the Heegaard surfaces of M, and we define the *rank* of M to be the minimal number of elements needed to generate $\pi_1(M)$. Suppose M is either closed or has connected boundary. Then every Heegaard splitting of M contains a handlebody, hence giving a presentation of $\pi_1(M)$. Thus we have $r(M) \leq g(M)$.

In the 1960s, Waldhausen asked whether $r(M) = g(M)$ for all M; see [11, 59]. This was called the generalized Poincaré Conjecture in [11], as the case $r(M) = 0$ is the Poincaré conjecture. In fact, many 3-manifolds satisfy this conjecture; e.g. if M is a small Seifert fibered space, then $r(M) = g(M)$. However, this conjecture is not true in general. In [4], Boileau and Zieschang found a Seifert fiber space with $r(M) = 2$ and $g(M) = 3$. Later, many other Seifert fiber spaces and graph manifolds have been found with similar properties; see e.g. [52]. A crucial ingredient in all these examples is that the fundamental group of a Seifert fiber space has a nontrivial center and, for a certain class of Seifert fiber spaces, one can use this property to find a smaller generating set of $\pi_1(M)$ than the one prescribed by a Heegaard splitting. However, these examples are very special. The fundamental group of a closed hyperbolic 3-manifold does not contain such commuting elements, so the modernized version of this old conjecture is whether $r(M) = g(M)$ holds for hyperbolic 3-manifolds; see [53, Conjecture 1.1]. This conjecture is sometimes called the Rank versus Genus Conjecture or the Rank Conjecture, as $r(M)$ can be viewed as the algebraic rank, and $g(M)$ can be regarded as the geometric rank of M. This conjecture is also related to the Fixed Price Conjecture in topological dynamics [1].

Some progress on this conjecture was made in the past few years. In [55], Souto proved

$r(M) = g(M)$ for any fiber bundle whose monodromy is a high power of a pseudo-Anosov map. In [41], Namazi and Souto showed that rank equals genus if the gluing map of a Heegaard splitting is a high power of a generic pseudo-Anosov map. This means that, in some sense, a sufficiently complicated hyperbolic 3-manifold satisfies this conjecture. On the other hand, many simple hyperbolic 3-manifolds also satisfy the conjecture; e.g., if $g(M) = 2$ then $\pi_1(M)$ cannot be cyclic and hence $r(M) = g(M) = 2$. In [30], we gave a negative answer to this conjecture.

Theorem 4.1 ([30]). *There is a closed orientable hyperbolic 3-manifold with rank of its fundamental group smaller than its Heegaard genus. Moreover, the discrepancy between its rank and Heegaard genus can be arbitrarily large.*

A basic idea of proving Theorem 4.1 is an observation based on the curve complex and Theorem 3.3: if there is a 3–manifold M such that ∂M is connected and $r(M) < g(M)$, then one can construct a closed 3–manifold $\hat{M}$ with $r(\hat{M}) < g(\hat{M})$ by capping off ∂M using a handlebody, via a complicated gluing map. This means that it suffices to construct such a 3–manifold M with connected boundary. The construction in [30] is an annulus sum of three pieces according to how generators of their fundamental groups are conjugate to each other.

Although rank and Heegaard genus are among the most basic invariants of 3-manifolds, they are surprisingly not well-studied. Theorem 4.1 says that they are not the same for hyperbolic 3-manifolds, but their relation is still not clear.

Question 4.2. *Is there a function $f(x)$, such that $g(M) \leq f(r(M))$ for any closed hyperbolic 3–manifold M.*

Another closely related but more difficult question is whether the function in Question 4.2 is linear.

Question 4.3. *Is there a number $\delta > 0$, such that $\frac{r(M)}{g(M)} > \delta$ for every closed hyperbolic 3–manifold M?*

Note that Biringer and Souto [3] made significant progress toward solving Question 4.2. They showed that, given any $\epsilon > 0$, there is a function f, such that for any closed hyperbolic 3–manifold M with injectivity radius bigger than ϵ, $g(M) \leq f(r(M))$.

The examples constructed in [30] are Haken manifolds. In a certain sense, non-Haken manifolds are more rigid than Haken manifolds; see e.g. [24, 25]. So it is interesting to know whether rank equals genus for non-Haken manifolds.

Question 4.4. *Is there a non-Haken 3-manifold M with $r(M) < g(M)$?*

Knot group has been an important topic in low-dimensional topology. Ever since 1960s (see [11]), people have been trying to answer the question of rank versus genus for knot exteriors, but this question remains open:

Question 4.5. *Is there a knot k in S^3 such that $r(S^3 - N(k)) < g(S^3 - N(k))$? How about a prime knot k?*

It is conceivable that one can use the methods in [30] to produce a composite knot k in S^3 whose exterior has rank smaller than Heegaard genus. But it is much harder to find a prime-knot example.

The examples in [30] are very complicated. It would be interesting to know what is the simplest hyperbolic 3–manifold with rank smaller than genus. For Seifert fibered spaces, the simplest such examples were discovered first, and the more complicated examples are, in a sense, built on these simple examples; see e.g. [4, 52]. It would be interesting to know whether there is a simpler hyperbolic example, or a simple local structure, such that the examples in [30] can be viewed as an extension of this simpler example or structure.

Question 4.6. *Among all hyperbolic 3-manifolds M with $r(M) < g(M)$, what is the minimal value for $r(M)$?*

Another interesting question related to Question 4.6 is whether the minimal value for $r(M)$ is 2 for hyperbolic 3-manifolds.

Question 4.7. *Let M be a hyperbolic 3-manifold with $r(M) = 2$. Is $g(M) = 2$?*

A fundamental tool in the construction of [30] is Theorem 3.3, which gives a nice formula to compute the Heegaard genus of an amalgamated 3–manifold. It is conceivable that a similar formula also holds for rank.

Question 4.8. *Is there an analogue of Theorem 3.3 for the rank of fundamental group?*

The following question is more specific.

Question 4.9. *Let M_1 and M_2 be compact 3-manifolds with connected boundary and $\partial M_1 \cong \partial M_2$. Let $\phi\colon \partial M_1 \to \partial M_2$ be a homeomorphism and let M be the closed manifold obtained by gluing M_1 to M_2 via ϕ. If ϕ is sufficiently complicated, then is it true that $r(M) = r(M_1) + r(M_2) - g(\partial M_i)$?*

Question 4.9 is true if both M_1 and M_2 are handlebodies and ϕ is a high power of a pseudo-Anosov map [41]. However, the question is unknown if only one of the two manifolds is a handlebody, and it is not even known in the case of Dehn filling, i.e., M_2 is a solid torus.

Theorem 4.1 says that the discrepancy $g(M) - r(M)$ can be arbitrarily large for hyperbolic 3-manifolds. However, the 3-manifolds constructed in [30] have the same ratio $\frac{r(M)}{g(M)}$. Thus the following is a natural question and is a slight variation of Question 4.3.

Question 4.10. *How small can the ratio $\frac{r(M)}{g(M)}$ be? Can the infimum of the ratio $\frac{r(M)}{g(M)}$ be zero for 3-manifolds?*

A map between two topological spaces is one of the most fundamental objects in topology. Degree-one maps are a particularly important class of maps. For 3-manifolds, such maps have a close relation with Thurston's geometrization of 3-manifolds (see [14, 54, 60]). It has been known for a long time that maps of nonzero degree between surfaces are standard [7]. However, many important questions remain open for maps between 3-manifolds. One of the most basic questions about degree-one maps between two 3-manifolds is the relation between their Heegaard genera.

Conjecture 4.11. *Let M and N be closed orientable 3-manifolds and suppose there is a degree-one map $f\colon M \to N$. Then $g(M) \geq g(N)$.*

Conjecture 4.11 is an old and difficult question in 3–manifold topology. It implies the Poincaré Conjecture: If a closed 3–manifold X is homotopy equivalent to S^3, since a homotopy equivalence is a degree-one map, Conjecture 4.11 implies that $g(X) = g(S^3) = 0$ and X must be S^3.

If $f\colon M \to N$ is a degree-one map, then $f_*\colon \pi_1(M) \to \pi_1(N)$ is a surjection and hence $r(M) \geq r(N)$. Thus, a counterexample to Conjecture 4.11 also gives a 3–manifold N with $g(N) > r(N)$. This suggests that Conjecture 4.11 is closely related to the rank versus genus question. In particular, one approach to constructing a counterexample to Conjecture 4.11 is to start with the examples in Theorem 4.1.

5. Dehn surgery on knots

One of the most useful constructions in low-dimensional topology is Dehn surgery. Let K be a nontrivial knot in the 3–sphere S^3 and let $E(K) = S^3 - N(K)$, where $N(K)$ is an open tubular neighborhood of K. Let $g(E(K))$ be the Heegaard genus of the knot exterior $E(K)$. Let $K(s)$ ($s \in \mathbb{Q} \cup \{\infty\}$) be the closed 3–manifold obtained by performing a Dehn surgery on K along slope s. An important question in 3–manifold topology is to determine the Heegaard genus of $K(s)$.

Theorem 5.1 (Gordon-Luecke [8]). *Let K be a nontrivial knot in S^3. If $g(K(s)) = 0$ (i.e. $K(s) = S^3$), then the Dehn surgery must be a trivial surgery, i.e. $s = \infty$.*

Theorem 5.1 concludes that there is no other way to obtain S^3 by performing Dehn surgery on a nontrivial knot in S^3. However, Dehn surgery on many knots can produce lens spaces. Berge observed that if a knot K lies on a genus-2 Heegaard surface of S^3 and K is primitive in each of the two handlebodies, then the genus-2 Heegaard splitting becomes a genus-1 splitting after a Dehn surgery, and hence the Dehn surgery yields a lens space. Such knots are called doubly primitive knots or Berge knots. A difficult question is whether the converse is true:

Conjecture 5.2 (Berge Conjecture). *Let K be a nontrivial knot in S^3. If $K(s)$ is a lens space (i.e. $g(K(s)) = 1$), then K must be a doubly primitive knot.*

When discussing Heegaard genus of knot exterior, sometimes it is more convenient to use another invariant called *tunnel number*. For any knot K is S^3, there is always a collection of disjoint and embedded arcs $\tau_1, \dots, \tau_t$ in S^3, such that $\tau_i \cap K = \partial\tau_i$ for each i, and $H = S^3 - N((\cup_{i=1}^t \tau_i) \cup K)$ is a handlebody. This means that ∂H is a Heegaard surface of $E(K)$. These arcs $\tau_1, \dots, \tau_t$ are called unknotting tunnels of K, and we say that they form a tunnel system for K. The tunnel number of K, denoted by $t(K)$, is the minimal number of arcs in a tunnel system for K. Clearly $g(E(K)) = t(K) + 1$.

By considering $E(K)$ as a submanifold of $K(s)$, it is easy to see that any Heegaard surface of $E(K)$ is a Heegaard surface of $K(s)$. Thus $g(K(s)) \leq g(E(K))$. Moreover, it is known that if the surgery slope s is "complicated", then a minimal-genus Heegaard surface in $E(K)$ remains a minimal-genus Heegaard surface of $K(s)$, in particular $g(K(s)) = g(E(K))$. Moriah and Rubinstein [38] first proved this for hyperbolic knots; Rieck and Sedgwick [45–47] gave a topological proof that works for all knots. This has also been generalized to handlebody surgery; see [29].

Theorem 5.3 (Moriah-Rubinstein [38]; Rieck-Sedgwick [45–47]). *Let K and $K(s)$ be as above. Then there is a finite set of slopes $\mathcal{N}$ and a finite set of lines of slopes $\mathcal{H}$, so that if $s \notin \mathcal{N} \cup \mathcal{H}$, then $g(K(s)) = g(E(K))$.*

In fact, they proved something stronger. They showed that, except for finitely many slopes, $g(K(s)) \geq g(E(K)) - 1$. Note that the theorems of Moriah-Rubinstein and Rieck-Sedgwick are for all 3-manifolds with torus boundary. If we only consider knots in S^3, it seems that this inequality always holds except for the trivial surgery:

Conjecture 5.4. *Let K be a nontrivial knot in S^3. Then $g(K(s)) \geq g(E(K)) - 1$ unless the Dehn surgery is a trivial surgery, i.e. $s = \infty$.*

There is an easy picture to see how Heegaard genus may drop after Dehn surgery. Given a minimal-genus Heegaard splitting $E(K) = H \cup_S W$ of the knot exterior $E(K)$, where H is a handlebody and W is a compression body. The Heegaard splitting is said to be s-primitive if there is a compressing disk D in H and an annulus A properly embedded in W, such that

1. one boundary circle of A lies in the Heegaard surface S and the other boundary circle is a circle of slope s in the boundary torus $\partial E(K)$, and
2. $\partial D \cap \partial A$ is a single point in S.

If the Heegaard splitting is s-primitive, then the annulus A extends to a disk in $K(s)$, which means that the corresponding Heegaard splitting in $K(s)$ is stabilized. Since the splitting $E(K) = H \cup_S W$ is a minimal-genus splitting, this means that $g(K(s)) < g(E(K))$. However, the converse of this phenomenon is not true: it is possible that a minimal-genus Heegaard splitting is not s-primitive in $E(K)$ but becomes stabilized in $K(s)$. Nonetheless, it is conceivable that, in this case, one can modify the Heegaard splitting into a new s-primitive Heegaard splitting of the same genus. The following conjecture says that if the Heegaard genus drops after an integer surgery, then the picture described above always occurs.

Conjecture 5.5. *Let K be a nontrivial knot in S^3. Suppose $g(K(s)) < g(E(K))$ for some integer s. Then $E(K)$ admits a minimal-genus and s-primitive Heegaard splitting.*

If s is an integer and $E(K)$ has an s-primitive genus-2 Heegaard splitting, then the knot K is doubly primitive. Hence Conjecture 5.5 implies the Berge Conjecture for knots with Heegaard genus 2 (i.e. tunnel number one knots).

If a Heegaard splitting of $E(K)$ is s-primitive and s is the meridian, then the splitting is called meridionally primitive or μ-primitive. A knot K is said to be meridionally primitive or μ-primitive, if $E(K)$ has a minimal-genus Heegaard splitting that is meridionally primitive.

Although we are focused on knots in S^3, the concept of s-primitive Heegaard splittings can be extended to all knot manifolds. Let K' be the core curve of the solid torus in the Dehn filling. We may view K' as a knot in $K(s)$. If $E(K)$ has a minimal-genus and s-primitive Heegaard splitting, then the knot K' is meridionally primitive in $K(s)$. Thus one way to understand Conjecture 5.5 is to study meridionally primitive knots in a 3–manifold. This approach seems particularly useful in attacking the Berge Conjecture, as lens spaces are among the best understood 3-manifolds.

Meridionally primitive knots have some very interesting properties. For example, if a knot K in S^3 is meridionally primitive, then for any other knot K', $g(E(K\#K')) <$

$g(E(K)) + g(E(K'))$; in other words, $t((K\#K') \leq t(K) + t(K')$; see [31] for an explanation. In [40], Morimoto proved a converse of this fact for small knots (a knot is small if its exterior contains no nonperipheral incompressible surface). He showed that if both K and K' are small knots, then $t((K\#K') \leq t(K) + t(K')$ only if one of K and K' is meridionally primitive. Morimoto and Moriah conjectured that the converse is always true [36, 40]. Kobayashi and Rieck [22] first gave a counterexample of which the two factors are both composite knots. Since composite knots are special, the conjecture was modified for prime knots; see [37, Conjecture 7.14]). Recently, this conjecture was also shown to be false [31]. In fact, we proved a much more general theorem:

Theorem 5.6 ([31]). *For any integer $n \geq 3$, there is a prime knot K such that*

1 *K is not meridionally primitive, and*

2 *for every m-bridge knot K' with $m \leq n$, the tunnel numbers satisfy $t(K\#K') \leq t(K)$.*

In light of Theorem 5.6, the following question was raised in [31].

Conjecture 5.7. *For any two knots K and K' in S^3, $t(K\#K') \geq t(K)$.*

Note that there is a degree-one map from $E(K\#K')$ to $E(K)$. Thus a counterexample to Conjecture 5.7 also gives a counterexample to Conjecture 4.11.

Acknowledgements. The author is partially supported by NSF grant DMS–1305613.

References

[1] Miklós Abért and Nikolay Nikolov, *Rank gradient, cost of groups and the rank versus Heegaard genus problem*, arXiv:math/0701361.

[2] R. H. Bing, *An alternative proof that 3-manifolds can be triangulated*, Ann. of Math., **69** (1959), 37–65.

[3] Ian Biringer and Juan Souto, Manuscript in preparation.

[4] M. Boileau and H. Zieschang, *Heegaard genus of closed orientable Seifert 3-manifolds*, Invent. Math., **76** (1984), 455–468.

[5] Andrew Casson and Cameron Gordon, *Reducing Heegaard splittings*, Topology and its Applications, **27** (1987), 275–283.

[6] ———, unpublished.

[7] A. Edmonds, *Deformation of maps to branched covering in dimension 2*, Ann. of Math., **110** (1979), 113–125.

[8] Cameron Gordon and John Luecke, *Knots are determined by their complements*, J. Amer. Math. Soc., **2** (1989), 371–415.

[9] Wolfgang Haken, *Theorie der Normalflächen: Ein Isotopickriterium für der Kreisknoten*, Acta Math., **105** (1961), 245–375.

[10] ———, *Some results on surfaces in 3-manifolds*, Studies in Modern Topology, Math. Assoc. Amer., distributed by Pretice-Hall, (1968) 34–98.

[11] ———, *Various aspects of the 3-dimensional Poincaré problem*, Topology of manifolds, Markham-Chicago (1970), 140–152.

[12] Kevin Hartshorn, *Heegaard splittings of Haken manifolds have bounded distance*, Pacific J. Math., **204** (2002), 61–75.

[13] W. J. Harvey, *Boundary structure of the modular group*, Ann. of Math. Stud., **97**, Princeton (1981), 245-251.

[14] C. Hayat-Legrand, S. Wang, and H. Zieschang, *Any 3–manifold 1-dominates only finitely many 3-manifolds supporting S^3 geometry*, Proc. Amer. Math. Soc., **130** (2002), 3117–3123.

[15] Poul Heegaard, *Forstudier til en topologisk Teori for de algebraiske Fladers Sammenhang*, thesis (1898).

[16] John Hempel, *3-manifolds*, Ann. of Math. Studies 86, Princeton University Press, Princeton, New Jersey, 1976.

[17] ———, *3-manifolds as viewed from the curve complex*, Topology, **40** (2001), 631–657.

[18] William Jaco, *Lectures on Three-Manifold Topology*, CBMS Regional Conference Series in Mathematics, **43** (1977).

[19] William Jaco and Peter Shalen, *Seifert fibered spaces in 3-manifolds*, Mem. Amer. Math. Soc., **21** (1979), no. 220.

[20] Klaus Johannson, *Homotopy equivalences of 3-manifolds with boundary*, Lecture Notes in Mathematics, **761**, Springer, (1979).

[21] H. Kneser, *Die kleinste Bedeckungszahl innerhalb einer Klasse von Flächenabbildungen*, Math., Annalen **103** (1930), 347–358.

[22] Tsuyoshi Kobayashi and Yo'av Rieck, *Knot exteriors with additive Heegaard genus and Morimoto's conjecture,* Algebr. Geom. Topol., **8** (2008), 953–969.

[23] Tao Li, *Laminar branched surfaces in 3-manifolds*. Geometry and Topology, Vol. **6** (2002), 153–194.

[24] ———, *Heegaard surfaces and measured laminations, I: the Waldhausen conjecture*, Invent. Math., **167** (2007), 135–177.

[25] ———, *Heegaard surfaces and measured laminations, II: non-Haken 3-manifolds*, J. Amer. Math. Soc., **19** (2006), 625–657.

[26] ———, *Saddle tangencies and the distance of Heegaard splittings*, Algebraic & Geometric Topology, **7** (2007), 1119–1134.

[27] ———, *On the Heegaard splittings of amalgamated 3-manifolds*, Geometry & Topology Monographs, **12** (2007), 157–190.

[28] ———, *An algorithm to determine the Heegaard genus of an atoroidal 3–manifold*, Geometry & Topology, **15** (2011), 1029–1106.

[29] ———, *Heegaard surfaces and the distance of amalgamation*, Geometry & Topology, **14** (2010), 1871–1919.

[30] ———, *Rank and genus of 3-manifolds*, J. Amer. Math. Soc., **26** (2013), 777–829.

[31] Tao Li and Ruifeng Qiu, *On the degeneration of tunnel numbers under connected sum*, arXiv:1310.5054.

[32] Howard Masur and Yair Minsky, *Geometry of the complex of curves. I. Hyperbolicity*, Invent. Math., **138** (1999), 103–149.

[33] ———, *Geometry of the complex of curves. II. Hierarchical structure*, Geom. Funct. Anal., **10** (2000), 902–974.

[34] J. Milnor, *A unique factorization theorem for 3-manifolds*, Amer. J. Math., **84** (1960), 1–7.

[35] E. Moise, *Affine structures in 3-manifolds V: the triangulation theorem and Hauptvermutung*, Ann. of Math., **55** (1952), 96-114.

[36] Yoav Moriah, *Connected sums of knots and weakly reducible Heegaard splittings*, Topology Appl., **141** (2004), 1–20.

[37] ———, *Heegaard splittings of knot exteriors*, Geometry & Topology Monographs, **12** (2007), 191–232.

[38] Yoav Moriah and Hyam Rubinstein, *Heegaard structures of negatively curved 3-manifolds*, Comm. Anal. Geom., **5** (1997), 375–412.

[39] Yoav Moriah, Saul Schleimer, and Eric Sedgwick, *Heegaard splittings of the form $H + nK$*, Comm. Anal. Geom., **14** (2006), 215–247.

[40] Kanji Morimoto, *On the super additivity of tunnel number of knots,* Math. Ann., **317** (2000), 489-508.

[41] Hossein Namazi and Juan Souto, *Heegaard splittings and pseudo-Anosov maps*, Geom. Funct. Anal., **19** (2009), 1195–1228.

[42] Grisha Perelman, *The entropy formula for the Ricci flow and its geometric applications*, arXiv:math.DG/0211159.

[43] ———, *Ricci flow with surgery on three-manifolds*, arXiv:math.DG/0303109.

[44] ———, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*, arXiv:math.DG/0307245.

[45] Yo'av Rieck, *Heegaard structure of manifolds in the Dehn filling space*, Topology, **39** (2000), 619–641.

[46] Yo'av Rieck and Eric Sedgwick, *Finiteness results for Heegaard surfaces in surgered manifolds*, Comm. Anal. Geom., **9** (2001), 351–367.

[47] ———, *Persistence of Heegaard structures under Dehn filling*, Topology, Appl., **109** (2001), 41–53.

[48] Hyam Rubinstein, *Polyhedral minimal surfaces, Heegaard splittings and decision problems for 3-dimensional manifolds*, Proc. Georgia Topology Conference, Amer. Math. Coc./Intl. Press, 1993.

[49] Martin Scharlemann and Abigail Thompson, *Thin position for 3–manifold*, Comptemp. Math., **164** (1992), 231–238.

[50] Martin Scharlemann and Maggy Tomova, *Alternate Heegaard genus bounds distance*, Geometry and Topology, **10** (2006), 593–617.

[51] Jenifer Schultens, *The classification of Heegaard splittings for (compact orientable surface)*$\times S^1$, Proc. London Math. Soc., **67** (1993), 425–448.

[52] Jennifer Schultens and Richard Weidmann, *On the geometric and the algebraic rank of graph manifolds*, Pacific J. Math., **231** (2007), 481–510.

[53] Peter Shalen, *Hyperbolic volume, Heegaard genus and ranks of groups*, Geom. Topol. Monogr., **12** (2007), 335–349.

[54] T. Soma, *Non-zero degree maps onto hyperbolic 3-manifolds*, J. Diff. Geom., **49** (1998), 517–546.

[55] Juan Souto, *The rank of the fundamental group of certain hyperbolic 3-manifolds fibering over the circle*, Geom. Topol. Monogr., **14** (2008), 505–518.

[56] Michelle Stocking, *Almost normal surfaces in 3-manifolds*, Trans. Amer. Math. Soc., **352** (2000), 171–207.

[57] Abigail Thompson, *Thin position and the recognition problem for* S^3, Math. Res. Lett., **1** (1994), 613–630.

[58] William Thurston, *Three-dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. Amer. Math. Soc., **6** (1982), no. 3, 357–381.

[59] Friedhelm Waldhausen, *Some problems on 3-manifolds*, Proc. Symposia in Pure Math., **32** (1978), 313–322.

[60] Shicheng Wang and Qing Zhou, *Any 3–manifold 1-dominates at most finitely many geometric 3- manifolds*, Math. Ann., **332** (2002), 525–535.

Department of Mathematics, Boston College, Chestnut Hill, MA 02467 USA
E-mail: taoli@bc.edu

Algebraic K-theory of strict ring spectra

John Rognes

Abstract. We view strict ring spectra as generalized rings. The study of their algebraic K-theory is motivated by its applications to the automorphism groups of compact manifolds. Partial calculations of algebraic K-theory for the sphere spectrum are available at regular primes, but we seek more conceptual answers in terms of localization and descent properties. Calculations for ring spectra related to topological K-theory suggest the existence of a motivic cohomology theory for strictly commutative ring spectra, and we present evidence for arithmetic duality in this theory. To tie motivic cohomology to Galois cohomology we wish to spectrally realize ramified extensions, which is only possible after mild forms of localization. One such mild localization is provided by the theory of logarithmic ring spectra, and we outline recent developments in this area.

Mathematics Subject Classification (2010). Primary 19D10, 55P43; Secondary 19F27, 57R50.

Keywords. Arithmetic duality, automorphisms of manifolds, brave new rings, étale descent, logarithmic ring spectrum, logarithmic topological André–Quillen homology, logarithmic topological Hochschild homology, motivic truncation, replete bar construction, sphere spectrum, tame ramification, topological K-theory.

1. Strict ring spectra

First, let R be an abelian group. Ordinary singular cohomology with coefficients in R is a contravariant homotopy functor that associates to each based space X a graded cohomology group $\tilde{H}^*(X;R)$. It is stable, in the sense that there is a natural isomorphism $\tilde{H}^*(X;R) \cong \tilde{H}^{*+1}(\Sigma X;R)$, and this implies that it extends from the category of based spaces to the category of spectra. The latter is a category of space-like objects, where the suspension is invertible up to homotopy equivalence, and which has all colimits and limits. The extended cohomology functor becomes representable, meaning that there is a spectrum HR, called the Eilenberg–Mac Lane spectrum of R, and a natural isomorphism $\tilde{H}^*(X;R) \cong [X, HR]_{-*}$, where $[X, HR]_{-n}$ is the group of homotopy classes of morphisms $X \to \Sigma^n HR$.

Next, let R be a ring. Then the cohomology theory is multiplicative, meaning that there is a bilinear cup product $\tilde{H}^*(X;R) \times \tilde{H}^*(X;R) \to \tilde{H}^*(X;R)$. This is also representable in the category of spectra, by a morphism $\mu\colon HR \wedge HR \to HR$, where $\wedge$ denotes the smash product of spectra. With the modern models for the category of spectra [26, 40, 48] we may arrange that μ is strictly unital and associative, so that HR is a *strict ring spectrum*. Equivalent terms are A_∞ ring spectrum, S-algebra, symmetric ring spectrum and orthogonal ring spectrum.

If R is commutative, then the cup product is graded commutative, which at the representing level means that $\mu\tau \simeq \mu$, where $\tau\colon HR \wedge HR \to HR \wedge HR$ denotes the twist iso-

Proceedings of International Congress of Mathematicians, 2014, Seoul

morphism. In fact, we may arrange that μ is strictly commutative, in the sense that $\mu\tau = \mu$ as morphisms of spectra, so that HR is a *strictly commutative ring spectrum*. Equivalent phrases are E_∞ ring spectrum, commutative S-algebra, commutative symmetric ring spectrum and commutative orthogonal ring spectrum. This leads to a compatible sequence of Σ_k-equivariant morphisms $E\Sigma_{k+} \wedge HR^{\wedge k} \to HR$ for $k \geq 0$. At the represented level these morphisms give rise to power operations in cohomology, including Steenrod's operations Sq^i for $R = \mathbb{F}_2$ and $\beta^\epsilon P^i$ for $R = \mathbb{F}_p$.

One now realizes that the Eilenberg–Mac Lane ring spectra HR exist as special cases within a much wider class of ring spectra. Each spectrum B represents a generalized cohomology theory $X \mapsto \tilde{B}^*(X) = [X, B]_{-*}$ and a generalized homology theory $X \mapsto \tilde{B}_*(X) = \pi_*(B \wedge X)$. Examples of early interest include the spectrum KU that represents complex topological K-theory, $KU^*(X) = [X_+, KU]_{-*}$, and the spectrum MU that represents complex bordism, $MU_*(X) = \pi_*(MU \wedge X_+)$. A fundamental example is given by the sphere spectrum S, which is the image of the based space S^0 under the stabilization functor from spaces to spectra. It represents stable cohomotopy $\pi^*_S(X) = \tilde{S}^*(X)$ and stable homotopy $\pi^S_*(X) = \tilde{S}_*(X)$. Each of these three examples, KU, MU and S, is naturally a strictly commutative ring spectrum, representing a multiplicative cohomology theory with power operations, etc. Furthermore, there are interesting multiplicative morphisms connecting these ring spectra to the Eilenberg–Mac Lane ring spectra previously considered, as in the diagram

$$\begin{array}{ccccccc} & & & & KU & \longrightarrow & H\mathbb{Q} \\ & & & & \uparrow & & \uparrow \\ S & \longrightarrow & MU & \longrightarrow & ku & \longrightarrow & H\mathbb{Z}, \end{array}$$

where $ku = KU[0, \infty)$ denotes the connective cover of KU.

By placing the class of traditional rings inside the wider realm of all strict ring spectra, a new world of possibilities opens up. Following Waldhausen [47, p. xiii] we may refer to strict ring spectra as "brave new rings". If we think in algebro-geometric terms, where commutative rings appear as the rings of functions on pieces of geometric objects, then strictly commutative ring spectra are the functions on affine pieces of brave new geometries, more general than those realized by ordinary schemes.

How vast is this generalization? In the case of connective ring spectra B, i.e., those with $\pi_i(B) = 0$ for i negative, there is a natural ring spectrum morphism $B \to H\pi_0(B)$ that induces an isomorphism on π_0. This behaves for many purposes like a topologically nilpotent extension, and in geometric terms, B can be viewed as the ring spectrum of functions on an infinitesimal thickening of $\operatorname{Spec} \pi_0(B)$.

This infinitesimal thickening can be quite effectively controlled in terms of diagrams of Eilenberg–Mac Lane spectra associated with simplicial rings. The Hurewicz map $B \cong S \wedge B \to H\mathbb{Z} \wedge B$ is 1-connected, and there is an equivalence $H\mathbb{Z} \wedge B \simeq HR_\bullet$ for some simplicial ring $R_\bullet$. The square

$$\begin{array}{ccc} S \wedge S \wedge B & \longrightarrow & H\mathbb{Z} \wedge S \wedge B \\ \downarrow & & \downarrow \\ S \wedge H\mathbb{Z} \wedge B & \longrightarrow & H\mathbb{Z} \wedge H\mathbb{Z} \wedge B \end{array}$$

induces a 2-connected map from $B \cong S \wedge S \wedge B$ to the homotopy pullback. More generally, for each $n \geq 1$ there is an n-dimensional cubical diagram that induces an n-connected map

from the initial vertex B to the homotopy limit of the remainder of the cube, and the terms in that remainder have the form $HR_\bullet$ for varying simplicial rings $R_\bullet$. Dundas [24] used a clever strengthening of this statement to prove that relative algebraic K-theory is p-adically equivalent to relative topological cyclic homology for morphisms $A \to B$ of connective strict ring spectra, under the assumption that $\pi_0(A) \to \pi_0(B)$ is a surjection with nilpotent kernel. He achieved this by reducing to the analogous statement for homomorphisms $R_\bullet \to T_\bullet$ of simplicial rings, which had been established earlier by McCarthy [49]. This confirmed a conjecture of Goodwillie [33], motivated by a similar result for rational K-theory and (negative) cyclic homology [32].

How about the case of non-connective ring spectra? Those that arise as homotopy fixed points $B^{hG} = F(EG_+, B)^G$ for a group action may be viewed as the functions on an orbit stack for the induced G-action on the geometry associated to B. Those that arise as smashing Bousfield localizations $L_E B$, with respect to a homology theory E_*, may be viewed as open subspaces in a finer topology than the one derived from the Zariski topology on $\operatorname{Spec} \pi_0(B)$ [63, §9.3]. In the general case the connection to classical geometry is less clear.

2. Automorphisms of manifolds

Why should we be interested in brave new rings and their ring-theoretic invariants, like algebraic K-theory [26, Ch. VI], [55], other than for the sake of generalization? One good justification comes from the tight connection between the geometric topology of high-dimensional manifolds and the algebraic K-theory of strict ring spectra. This connection is given by the higher simple-homotopy theory initiated by Hatcher [36], which was fully developed by Waldhausen in the context of his algebraic K-theory of spaces [82] and the stable parametrized h-cobordism theorem [83]. On the geometric side, this theory concerns the fundamental problem of finding a parametrized classification of high-dimensional compact manifolds, up to homeomorphism, piecewise-linear homeomorphism or diffeomorphism, as appropriate for the respective geometric category. The set of path components of the resulting moduli space corresponds to the set of isomorphism classes of such manifolds, and each individual path component is a classifying space for the automorphism group $\mathrm{Aut}(X)$ of a manifold X in the respective isomorphism class.

This parametrized classification is finer than the one provided by the Browder–Novikov–Sullivan–Wall surgery theory [21, 84], which classifies manifolds up to h-cobordism (or s-cobordism), and whose associated moduli space has path components that classify the block automorphism groups of manifolds, rather than their actual automorphism groups. The difference between these two classifications is controlled by the space $H(X)$ of h-cobordisms $(W; X, Y)$ with a given manifold X at one end. Here W is a compact manifold with $\partial W = X \cup Y$, and the inclusions $X \to W$ and $Y \to W$ are homotopy equivalences. More precisely, there is one h-cobordism space $H^{\mathrm{Cat}}(X)$ for each of the three flavors of manifolds mentioned, namely $\mathrm{Cat} = \mathrm{Top}$, PL or Diff.

The original h-cobordism theorem enumerates the isomorphism classes of h-cobordisms with X at one end, i.e., the set $\pi_0 H(X)$ of path components of the h-cobordism space of X, in terms of an algebraic K-group of the integral group ring $\mathbb{Z}[\pi]$, where $\pi = \pi_1(X)$ is the fundamental group of X. One defines the Whitehead group as the quotient $\mathrm{Wh}_1(\pi) = K_1(\mathbb{Z}[\pi])/(\pm\pi)$, and associates a Whitehead torsion class $\tau(W; X, Y) \in \mathrm{Wh}_1(\pi)$ to each h-cobordism on X.

Theorem 2.1 (Smale [73], Barden, Mazur, Stallings). *Let X be a compact, connected n-manifold with $n \geq 5$ and $\pi = \pi_1(X)$. The Whitehead torsion defines a bijection*

$$\pi_0 H(X) \cong \mathrm{Wh}_1(\pi)\,.$$

These constructions involve using Morse functions or triangulations to choose a relative CW complex structure on the pair (W, X), or equivalently, a π-equivariant relative CW complex structure on the pair of universal covers $(\tilde{W}, \tilde{X})$, and to study the associated cellular chain complex $C_*(\tilde{W}, \tilde{X})$ of free $\mathbb{Z}[\pi]$-modules. This works fine as long as one is only concerned with a classification of h-cobordisms up to isomorphism, but for the parametrized problem, i.e., the study of the full homotopy type of $H(X)$, the passage from a CW complex structure to the associated cellular chain complex loses too much information. One should remember the actual attaching maps from the boundaries of cells to the preceding skeleta, not just their degrees. In a stable range this amounts to working with maps from spheres to spheres and coefficients in the sphere spectrum S, rather than with degrees and coefficients in the integers $\mathbb{Z}$. Likewise, the passage to the π-equivariant universal cover $\tilde{X} \to X$ should be replaced to a passage to a G-equivariant principal fibration $P \to X$, where P is contractible and $G \simeq \Omega X$ is a topological group that is homotopy equivalent to the loop space of $X \simeq BG$. To sum up, the parametrized analog of the Whitehead torsion must take values in a Whitehead space that is built from the algebraic K-theory of the spherical group ring $S[G] = S \wedge G_+$, a strict ring spectrum, rather than that of its discrete reduction, the integral group ring $\pi_0(S[G]) \cong \mathbb{Z}[\pi]$.

Waldhausen's algebraic K-theory of spaces, traditionally denoted $A(X)$, was first introduced without reference to strict ring spectra [80], but can be rewritten as the algebraic K-theory $A(X) = K(S[G])$ of the strict ring spectrum $S[G]$, cf. [26, Ch. VI] and [81]. This point of view is convenient for the comparison of algebraic K-theory with other ring-theoretic invariants.

In the case of differentiable manifolds, the Whitehead space $\mathrm{Wh}^{\mathrm{Diff}}(X)$ is defined to sit in a split homotopy fiber sequence of infinite loop spaces

$$\Omega^\infty(S \wedge X_+) \xrightarrow{\iota} K(S[G]) \longrightarrow \mathrm{Wh}^{\mathrm{Diff}}(X)\,.$$

In the topological case there is a homotopy fiber sequence of infinite loop spaces

$$\Omega^\infty(K(S) \wedge X_+) \xrightarrow{\alpha} K(S[G]) \longrightarrow \mathrm{Wh}^{\mathrm{Top}}(X)\,,$$

where α is known as the assembly map. The piecewise-linear Whitehead space is the same as the topological one. The stable parametrized h-cobordism theorem reads as follows.

Theorem 2.2 (Waldhausen–Jahren–Rognes [83, Thm. 0.1]). *Let X be a compact* Cat *manifold, for* $\mathrm{Cat} = \mathrm{Top}$, PL *or* Diff. *There is a natural homotopy equivalence*

$$\mathcal{H}^{\mathrm{Cat}}(X) \simeq \Omega\,\mathrm{Wh}^{\mathrm{Cat}}(X)\,,$$

where $\mathcal{H}^{\mathrm{Cat}}(X) = \mathrm{colim}_k\, H^{\mathrm{Cat}}(X \times I^k)$ is the stable Cat *h-cobordism space of X.*

When combined with connectivity results about the dimensional stabilization map

$$H(X) = H^{\mathrm{Cat}}(X) \to \mathcal{H}^{\mathrm{Cat}}(X)\,,$$

and here the main result is Igusa's stability theorem for smooth pseudoisotopies [42], knowledge of $K(S)$ and $K(S[G])$ gives good general results on the h-cobordism space $H(X)$ and the automorphism group $\mathrm{Aut}(X)$ of a high-dimensional manifold X.

Example 2.3. When G is trivial, so that $S[G] = S$ and X is contractible, the π_0-isomorphism and rational equivalence $S \to H\mathbb{Z}$ induces a rational equivalence $K(S) \to K(\mathbb{Z})$. Here $\pi_* K(\mathbb{Z}) \otimes \mathbb{Q}$ was computed by Borel [20], so

$$\pi_i \operatorname{Wh}^{\mathrm{Diff}}(*) \otimes \mathbb{Q} \cong \begin{cases} \mathbb{Q} & \text{for } i = 4k+1 \neq 1, \\ 0 & \text{otherwise.} \end{cases}$$

For $X = D^n$, Farrell–Hsiang [27] used this to show that

$$\pi_i \operatorname{Diff}(D^n) \otimes \mathbb{Q} \cong \begin{cases} \mathbb{Q} & \text{for } i = 4k-1,\ n \text{ odd}, \\ 0 & \text{otherwise,} \end{cases}$$

for i up to approximately $n/3$, where $\operatorname{Diff}(D^n)$ denotes the group of self-diffeomorphisms of D^n that fix the boundary. For instance, $\pi_3 \operatorname{Diff}(D^{13})$ is rationally nontrivial. By contrast, the group $\operatorname{Top}(D^n)$ of self-homeomorphisms of D^n that fix the boundary is contractible. Similar results follow for n-manifolds that are roughly $n/3$-connected.

The case of spherical space forms, when G is finite with periodic cohomology, has been studied by Hsiang–Jahren [41]. For closed, non-positively curved manifolds X, Farrell–Jones [28] showed that $\operatorname{Wh}^{\mathrm{Diff}}(X)$ can be assembled from copies of $\operatorname{Wh}^{\mathrm{Diff}}(*)$ and $\operatorname{Wh}^{\mathrm{Diff}}(S^1)$, indexed by the points and the closed geodesics in X, respectively. These correspond to the special cases G trivial and G infinite cyclic, respectively, so $K(S)$ and $K(S[\mathbb{Z}])$ are of fundamental importance for the parametrized classification of this large class of Riemannian manifolds. In this paper we shall focus on the case of $K(S)$, but see Hesselholt's paper [37] for the case of $K(S[\mathbb{Z}])$, and see Weiss–Williams [86] for a detailed survey about automorphisms of manifolds and algebraic K-theory.

Remark 2.4. More recent papers of Madsen–Weiss [46], Berglund–Madsen [12] and Galatius –Randal-Williams [29] give precise results about automorphism groups of manifolds of a fixed even dimension $n = 2d \neq 4$, at the expense of first forming a connected sum with many copies of $S^d \times S^d$. The latter results are apparently not closely related to the algebraic K-theory of strict ring spectra.

3. Algebraic K-theory of the sphere spectrum

We can strengthen the rational results about $A(X) = K(S[G])$, $\operatorname{Wh}^{\mathrm{Diff}}(X)$ and $\operatorname{Diff}(X)$ to integral results, or more precisely, to p-adic integral results for each prime p. From here on it will be convenient to think of algebraic K-theory as a spectrum-valued functor, and likewise for the Whitehead theories, so that there are homotopy cofiber sequences of spectra

$$S \wedge X_+ \xrightarrow{\iota} K(S[G]) \longrightarrow \operatorname{Wh}^{\mathrm{Diff}}(X)$$
$$K(S) \wedge X_+ \xrightarrow{\alpha} K(S[G]) \longrightarrow \operatorname{Wh}^{\mathrm{Top}}(X)\,,$$

and the first one is naturally split.

A key tool for this study is the cyclotomic trace map $\operatorname{trc}\colon K(B) \to TC(B;p)$ from algebraic K-theory to the topological cyclic homology of Bökstedt–Hsiang–Madsen [16]. The latter invariant of the strict ring spectrum B can sometimes be calculated by analyzing the S^1-equivariant homotopy type of the topological Hochschild homology spectrum $THH(B)$. Its power is illustrated by the following previously mentioned theorem.

Theorem 3.1 (Dundas [24]). *Let B be a connective strict ring spectrum. The square*

$$\begin{array}{ccc} K(B) & \longrightarrow & K(\pi_0(B)) \\ {\scriptstyle \mathrm{trc}}\downarrow & & \downarrow{\scriptstyle \mathrm{trc}} \\ TC(B;p) & \longrightarrow & TC(\pi_0(B);p) \end{array}$$

becomes homotopy Cartesian upon p-completion.

In the basic case $B = S$, when $K(S) \simeq S \vee \mathrm{Wh}^{\mathrm{Diff}}(*)$ determines $\mathrm{Diff}(D^n)$ for large n, this square takes the form below. Three of the four corners are quite well understood, but for widely different reasons.

$$\begin{array}{ccc} K(S) & \longrightarrow & K(\mathbb{Z}) \\ {\scriptstyle \mathrm{trc}}\downarrow & & \downarrow{\scriptstyle \mathrm{trc}} \\ TC(S;p) & \longrightarrow & TC(\mathbb{Z};p)\,. \end{array}$$

These reasons were tied together by the author for $p = 2$ in [61], and for p an odd regular prime in [62], to compute the mod p cohomology

$$H^*(K(S);\mathbb{F}_p) \cong \mathbb{F}_p \oplus H^*(\mathrm{Wh}^{\mathrm{Diff}}(*);\mathbb{F}_p)$$

as a module over the Steenrod algebra $\mathscr{A}$ of stable mod p cohomology operations. This sufficed to determine the E_2-term of the Adams spectral sequence

$$E_2^{s,t} = \mathrm{Ext}_{\mathscr{A}}^{s,t}(H^*(K(S);\mathbb{F}_p),\mathbb{F}_p) \Longrightarrow \pi_{t-s}K(S)_p$$

in a large range of degrees, and to determine the homotopy groups of $K(S)_p$ and $\mathrm{Wh}^{\mathrm{Diff}}(*)_p$ in a smaller range of degrees.

The structure of the algebraic K-theory of the integers, $K(\mathbb{Z})$, was predicted by the Lichtenbaum–Quillen conjectures [55], which were confirmed for $p = 2$ by Voevodsky [78] with contributions by Rognes–Weibel [67], and for p odd by Voevodsky [79] with contributions by Rost and Weibel. For $p = 2$ or p a regular odd prime, this led to a p-complete description of the spectrum $K(\mathbb{Z})$ in terms of topological K-theory spectra, which in turn led to an explicit description of the spectrum cohomology $H^*(K(\mathbb{Z});\mathbb{F}_p)$ as an $\mathscr{A}$-module.

The topological cyclic homology of the integers, $TC(\mathbb{Z};p)$, was computed for odd primes p by Bökstedt–Madsen [17, 18] and for $p = 2$ by the author [57–60], in papers that start with knowledge of the mod p homotopy of the S^1-spectrum $THH(\mathbb{Z})$ and inductively determine the mod p homotopy of the C_{p^n}-fixed points $THH(\mathbb{Z})^{C_{p^n}}$ for $n \geq 1$. It is then possible to recognize the p-completed spectrum level structure by comparisons with known models, using [56] for p odd, and to obtain the $\mathscr{A}$-module $H^*(TC(\mathbb{Z};p);\mathbb{F}_p)$ from this.

The topological cyclic homology of the sphere spectrum, $TC(S;p)$, was determined in the original paper [16]. There is an equivalence of spectra $TC(S;p) \simeq S \vee \Sigma\mathbb{C}P^\infty_{-1}$ after p-completion, where $\Sigma\mathbb{C}P^\infty_{-1}$ is the homotopy fiber of the dimension-shifting S^1-transfer map $t\colon \Sigma\mathbb{C}P^\infty_+ \to S$. The mod p cohomology $H^*(\Sigma\mathbb{C}P^\infty_{-1};\mathbb{F}_p)$ is well known as an $\mathscr{A}$-module. For $p = 2$ it is cyclic with

$$H^*(\Sigma\mathbb{C}P^\infty_{-1};\mathbb{F}_2) \cong \Sigma^{-1}\mathscr{A}/C\,,$$

where the ideal $C \subset \mathscr{A}$ is generated by the admissible Sq^I where $I = (i_1, \dots, i_n)$ with $n \geq 2$ or $I = (i)$ with i odd. The determination of the homotopy groups of $TC(S; p)$ is of comparable difficulty to the computation of the homotopy groups of S, due to our extensive knowledge about the attaching maps in the usual CW spectrum structure on $\Sigma\mathbb{C}P^\infty_{-1}$, cf. Mosher [52].

The linearization map $TC(S; p) \to TC(\mathbb{Z}; p)$ is only partially understood [45], but for p regular the cyclotomic trace map $K(\mathbb{Z}) \to TC(\mathbb{Z}; p)$ can be controlled by an appeal to global Tate–Poitou duality [76, Thm. 3.1], see [62, Prop. 3.1]. This leads to the following conclusion for $p = 2$. See [62, Thm. 5.4] for the result at odd regular primes.

Theorem 3.2 ([61, Thm. 4.5]). *The mod* 2 *cohomology of the spectrum* $\mathrm{Wh}^{\mathrm{Diff}}(*)$ *is given by the unique non-trivial extension of* $\mathscr{A}$*-modules*

$$0 \to \Sigma^{-2}C/\mathscr{A}(Sq^1, Sq^3) \longrightarrow H^*(\mathrm{Wh}^{\mathrm{Diff}}(*); \mathbb{F}_2) \longrightarrow \Sigma^3\mathscr{A}/\mathscr{A}(Sq^1, Sq^2) \to 0\,.$$

Using the Adams spectral sequence and related methods, the author obtained the following explicit calculations. Less complete information, in a larger range of degrees, is provided in the cited references. Previously, Bökstedt–Waldhausen [19, Thm. 1.3] had computed $\pi_i \mathrm{Wh}^{\mathrm{Diff}}(*)$ for $i \leq 3$.

Theorem 3.3 ([61, Thm. 5.8], [62, Thm. 4.7]). *The homotopy groups of* $\mathrm{Wh}^{\mathrm{Diff}}(*)$ *in degrees* $i \leq 18$ *are as follows, modulo* p*-power torsion for irregular primes* p*.*

i	0	1	2	3	4	5	6	7	8	9
$\pi_i \mathrm{Wh}^{\mathrm{Diff}}(*)$	0	0	0	$\mathbb{Z}/2$	0	$\mathbb{Z}$	0	$\mathbb{Z}/2$	0	$\mathbb{Z} \oplus \mathbb{Z}/2$

i	10	11	12	13	14
$\pi_i \mathrm{Wh}^{\mathrm{Diff}}(*)$	$\mathbb{Z}/8 \oplus (\mathbb{Z}/2)^2$	$\mathbb{Z}/6$	$\mathbb{Z}/4$	$\mathbb{Z}$	$\mathbb{Z}/36 \oplus \mathbb{Z}/3$

i	15	16	17	18
$\pi_i \mathrm{Wh}^{\mathrm{Diff}}(*)$	$(\mathbb{Z}/2)^2$	$\mathbb{Z}/24 \oplus \mathbb{Z}/2$	$\mathbb{Z} \oplus (\mathbb{Z}/2)^2$	$\mathbb{Z}/480 \oplus (\mathbb{Z}/2)^3$

Example 3.4. For $X = D^n$ with n sufficiently large, it follows that $\pi_{4p-4}\mathrm{Diff}(D^n)$ or $\pi_{4p-4}\mathrm{Diff}(D^{n+1})$ contains an element of order p, for each regular $p \geq 5$, and that $\pi_9\mathrm{Diff}(D^n)$ or $\pi_9\mathrm{Diff}(D^{n+1})$ contains an element of order 3, see [62, Thm. 6.4]. To get more precise results one needs to investigate the canonical involution on $\mathrm{Wh}^{\mathrm{Diff}}(*)$ and apply Weiss–Williams [85, Thm. A].

Remark 3.5. It would be interesting to extend these results to irregular primes. Dwyer–Mitchell [25] described the spectrum $K(\mathbb{Z})_p$ in terms of the p-primary Iwasawa module of the rationals. It should be possible to turn this into a description of the $\mathscr{A}$-module $H^*(K(\mathbb{Z}); \mathbb{F}_p)$. Next one must control the cyclotomic trace map $K(\mathbb{Z}) \to TC(\mathbb{Z}; p)$, or the closely related completion map $K(\mathbb{Z}) \to K(\mathbb{Z}_p)$, whose behavior is governed by special values of p-adic L-functions, cf. Soulé [74, Thm. 3].

4. Algebraic K-theory of topological K-theory

The calculations reviewed in the previous section extracted detailed information about $\pi_*K(S) \cong \pi_*(S) \oplus \pi_* \mathrm{Wh}^{\mathrm{Diff}}(*)$ from our knowledge of $\pi_*(\mathbb{C}P^\infty_{-1})$. However, this understanding was not

presented to us in as conceptual a way as the understanding we have of $K(\mathbb{Z})$, say in terms of Quillen's localization sequence

$$K(\mathbb{F}_p) \longrightarrow K(\mathbb{Z}) \longrightarrow K(\mathbb{Z}[1/p])$$

and the étale descent property

$$\pi_i K(\mathbb{Z}[1/p])_p \xrightarrow{\cong} K_i^{\text{ét}}(\mathbb{Z}[1/p]; \mathbb{Z}_p)$$

for $i > 0$, cf. [54, §5] and [55, §9]. It would be desirable to have a similarly conceptual understanding of $K(S)_p$ in terms of a comparison with $K(B)_p$ for suitably local strict ring spectra B, a descent property describing $K(B)_p$ as a homotopy limit of $K(C)_p$ for appropriate extensions $B \to C$, and a simple description of $K(\Omega)_p$ for a sufficiently large such extension $B \to \Omega$.

To explore this problem, we first simplify the number theory involved by working with the p-adic integers $\mathbb{Z}_p$ in place of the rational integers $\mathbb{Z}$, and then seek a conceptual understanding of $K(B)_p$ for some of the strictly commutative ring spectra B that are closest to $H\mathbb{Z}_p$, namely the p-complete connective complex K-theory spectrum ku_p and its Adams summand ℓ_p. Here $\pi_*(ku_p) = \mathbb{Z}_p[u]$ and $\pi_*(\ell_p) = \mathbb{Z}_p[v_1]$, with $|u| = 2$ and $|v_1| = 2p-2$. Let KU_p and L_p denote the associated periodic spectra, with $\pi_*(KU_p) = \mathbb{Z}_p[u^{\pm 1}]$ and $\pi_*(L_p) = \mathbb{Z}_p[v_1^{\pm 1}]$. There are multiplicative morphisms

$$\begin{array}{ccccccc} & & L_p & \xrightarrow{\phi} & KU_p & & \\ & & \uparrow & & \uparrow & & \\ S_p & \longrightarrow & \ell_p & \xrightarrow{\phi} & ku_p & \longrightarrow & H\mathbb{Z}_p \end{array} \tag{4.1}$$

of strictly commutative ring spectra, where $\phi_*(v_1) = u^{p-1}$. The group $\Delta \cong \mathbb{F}_p^\times$ of p-adic roots of unity acts by Adams operations on KU_p, and $\phi\colon L_p \to KU_p$ is a Δ-Galois extension in the sense of [63, p. 3].

Definition 4.1. Let $V(1) = S \cup_p e^1 \cup_{\alpha_1} e^{2p-1} \cup_p e^{2p}$ be the type 2 Smith–Toda complex, defined as the mapping cone of the Adams self-map $v_1\colon \Sigma^{2p-2} S/p \to S/p$ of the mod p Moore spectrum $S/p = S \cup_p e^1$. It is a ring spectrum up to homotopy for $p \geq 5$, which we now assume. We write $V(1)_*B = \pi_*(V(1) \wedge B)$ for the "mod p and v_1 homotopy" of any spectrum B. It is naturally a module over the polynomial ring $\mathbb{F}_p[v_2]$, where $v_2 \in \pi_{2p^2-2} V(1)$.

The mod p and v_1 homotopy of the topological cyclic homology of the connective Adams summand ℓ was computed by Ausoni and the author, by starting with knowledge of $V(1)_*THH(\ell)$ from [51] and inductively determining the mod p and v_1 homotopy of the fixed points $THH(\ell)^{C_{p^n}}$ for $n \geq 1$. The calculations were later extended to the full connective complex K-theory spectrum ku by Ausoni. To avoid introducing too much notation, we only describe the most striking features of the answers, referring to the original papers for more precise statements.

Theorem 4.2 (Ausoni–Rognes [5, Thm. 0.3, Thm. 0.4]). *$V(1)_*TC(\ell; p)$ is a finitely generated free $\mathbb{F}_p[v_2]$-module on $4(p+1)$ generators, which are located in degrees $-1 \leq * \leq 2p^2 + 2p - 2$. There is an exact sequence of $\mathbb{F}_p[v_2]$-modules*

$$0 \to \Sigma^{2p-3}\mathbb{F}_p \longrightarrow V(1)_*K(\ell_p) \xrightarrow{\text{trc}} V(1)_*TC(\ell; p) \longrightarrow \Sigma^{-1}\mathbb{F}_p \to 0\,.$$

Theorem 4.3 (Ausoni [4, Thm. 7.9, Thm. 8.1]). *$V(1)_*TC(ku;p)$ is a finitely generated free $\mathbb{F}_p[v_2]$-module on $4(p-1)(p+1)$ generators, which are located in degrees $-1 \le * \le 2p^2+2p-2$. There is an exact sequence of $\mathbb{F}_p[v_2]$-modules*

$$0 \to \Sigma^{2p-3}\mathbb{F}_p \longrightarrow V(1)_*K(ku_p) \xrightarrow{\mathrm{trc}} V(1)_*TC(ku;p) \longrightarrow \Sigma^{-1}\mathbb{F}_p \to 0\,,$$

and the natural map $K(\ell_p) \to K(ku_p)^{h\Delta}$ is a p-adic equivalence.

Blumberg–Mandell [14] constructed homotopy cofiber sequences

$$K(\mathbb{Z}_p) \longrightarrow K(\ell_p) \longrightarrow K(L_p) \tag{4.2}$$

and

$$K(\mathbb{Z}_p) \to K(ku_p) \to K(KU_p)\,,$$

which lead to calculations of $V(1)_*K(L_p)$ and $V(1)_*K(KU_p)$, cf. [4, Thm. 8.3]. The natural map $K(L_p) \to K(KU_p)^{h\Delta}$ is also a p-adic equivalence, which confirms the étale descent property for algebraic K-theory in this particular case.

Remark 4.4. The examples discussed above are the case $n = 1$ of a series of approximations to S associated with the Lubin–Tate spectra E_n, with coefficient rings $\pi_*E_n = \mathbb{W}\mathbb{F}_{p^n}[[u_1,\dots,u_{n-1}]][u^{\pm 1}]$, which are known to be strictly commutative ring spectra by the Goerss–Hopkins–Miller obstruction theory [31]. There are multiplicative morphisms

$$\begin{array}{ccccccc}
\hat{L}_nS & \longrightarrow & \hat{L}_nE(n) & \xrightarrow{\phi} & E_n & & \\
\uparrow & & \uparrow & & \uparrow & & \\
L_nS & \longrightarrow & BP\langle n\rangle & \longrightarrow & e_n & \longrightarrow & H\pi_0(e_n)\,,
\end{array}$$

where L_n and $\hat{L}_n$ denote Bousfield localization with respect to the Johnson–Wilson spectrum $E(n)$ and the Morava K-theory spectrum $K(n)$, respectively, and $BP\langle n\rangle$ is the truncated Brown–Peterson spectrum. The n-th extended Morava stabilizer group $\mathbb{G}_n$ acts on E_n, and $\hat{L}_nE(n) \to E_n$ is an H-Galois extension for $H \cong \mathbb{F}_{p^n}^\times \rtimes \mathbb{Z}/n$. We write e_n for the connective cover of E_n.

The algebraic K-theory computations above provide evidence for the chromatic redshift conjecture, see [5, p. 7] and [6], predicting that the algebraic K-theory $K(B)$ of a purely v_n-periodic strictly commutative ring spectrum B, such as E_n, is purely v_{n+1}-periodic in sufficiently high degrees.

5. Motivic truncation and arithmetic duality

The proven Lichtenbaum–Quillen conjectures subsume a spectral sequence

$$E^2_{s,t} = H^{-s}_{\text{ét}}(R;\mathbb{Z}_p(t/2)) \Longrightarrow \pi_{s+t}K(R)_p\,,$$

which converges for reasonable R and $s+t$ sufficiently large. Here $H^*_{\text{ét}}$ denotes étale cohomology, R is a commutative $\mathbb{Z}[1/p]$-algebra, and $\mathbb{Z}_p(t/2) = \pi_t(KU_p)$ is $\mathbb{Z}_p(m)$ when

$t = 2m$ is even, and 0 otherwise. For instance, we may take $R = \mathscr{O}_F[1/p]$ to be the ring of p-integers in a number field F, or R may be a p-adic field, i.e., a finite extension of $\mathbb{Q}_p$.

The proven Beilinson–Lichtenbaum conjectures, cf. [75] and [30], provide a more precise convergence statement. For each field F containing $1/p$ there is a spectral sequence

$$E^2_{s,t} = H^{-s}_{\mathrm{mot}}(F; \mathbb{Z}(t/2)) \Longrightarrow \pi_{s+t}K(F)\,,$$

converging in all degrees, and similarly with mod p coefficients. Here H^*_{mot} denotes motivic cohomology, which satisfies

$$H^r_{\mathrm{mot}}(F; \mathbb{Z}/p(m)) \cong \begin{cases} H^r_{\text{ét}}(F; \mathbb{Z}/p(m)) & \text{for } 0 \le r \le m, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

In terms of Bloch's higher Chow groups [13], the vanishing of these groups for $r > m$ expresses the fact that there are no codimension r subvarieties of affine m-space over $\operatorname{Spec} F$. Conversely,

$$H^r_{\text{ét}}(F; \mathbb{Z}/p(*)) \cong v_1^{-1} H^r_{\mathrm{mot}}(F; \mathbb{Z}/p(*)) \tag{5.2}$$

with $v_1 \in H^0_{\mathrm{mot}}(F; \mathbb{Z}/p(p-1))$. We refer to the aspects (5.1) and (5.2) of the Beilinson–Lichtenbaum conjectures as the *motivic truncation property* for the field F.

The following prediction expresses a similar conceptual description of $K(B)$ for some strictly commutative ring spectra, and should in particular apply for $B = \ell_p$, L_p, ku_p and KU_p.

Conjecture 5.1. *For purely v_1-periodic strictly commutative ring spectra B there is a spectral sequence*

$$E^2_{s,t} = H^{-s}_{\mathrm{mot}}(B; \mathbb{F}_{p^2}(t/2)) \Longrightarrow V(1)_{s+t}K(B)\,,$$

converging for $s + t$ sufficiently large.

Here H^*_{mot} denotes a currently undefined form of motivic cohomology for strictly commutative ring spectra. The coefficient $\mathbb{F}_{p^2}(t/2)$ may be interpreted as $V(1)_t E_2$, where E_2 is the Lubin–Tate ring spectrum [31] with $\pi_* E_2 = \mathbb{W}\mathbb{F}_{p^2}[[u_1]][u^{\pm 1}]$.

More generally one might consider purely v_n-periodic ring spectra B, replace $V(1)$ by any type $n+1$ finite spectrum $F(n+1)$, see [39], and replace $V(1)_t E_2$ and $V(1)_{s+t}K(B)$ by $F(n+1)_t E_{n+1}$ and $F(n+1)_{s+t}K(B)$, respectively.

Example 5.2. Based on the detailed calculations behind Theorem 4.2, it is fairly evident that the E^2-term for the spectral sequence conjectured to converge to $V(1)_*K(\ell_p)$ will be concentrated in the four columns $-3 \le s \le 0$, and that the free $\mathbb{F}_p[v_2]$-module generators are located in the groups $H^r_{\mathrm{mot}}(\ell_p; \mathbb{F}_{p^2}(m))$ where $0 \le r \le 3$ and $r \le m < r + p^2 + p - 1$. This presumes that the spectral sequence collapses at the E^2-term, for $p \ge 5$. In addition, there is a sporadic copy of $\mathbb{F}_p$ in $V(1)_{2p-3}K(\ell_p)$.

The class $v_2 \in V(1)_{2p^2-2}K(\ell_p)$ is represented in bidegree $(s,t) = (0, 2p^2-2)$, corresponding to $(r,m) = (0, p^2-1)$. The presence of $\mathbb{F}_p[v_2]$-module generators in the range $r + p^2 - 1 \le m$ shows that $H^r_{\mathrm{mot}}(\ell_p; \mathbb{F}_{p^2}(*))$ is *not* isomorphic to $v_2^{-1}H^r_{\mathrm{mot}}(\ell_p; \mathbb{F}_{p^2}(*))$ in several bigradings (r,m) with $r \le m < r + p$. In other words, the motivic truncation property fails for ℓ_p. However, this is to be expected, since ℓ_p has the residue ring spectrum $H\mathbb{Z}_p$ and should not behave as a field.

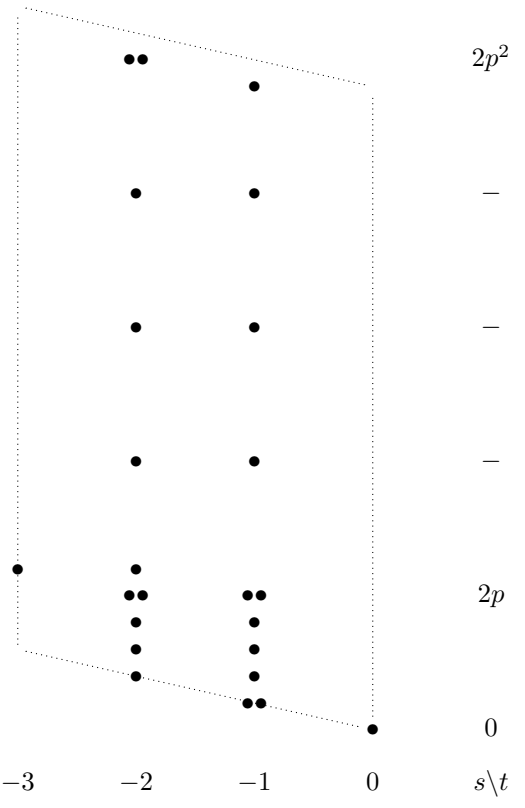


Figure 5.1. $\mathbb{F}_p[v_2]$-generators of $E^2_{s,t} = H^{-s}_{\mathrm{mot}}(L_p; \mathbb{F}_{p^2}(t/2)) \Longrightarrow V(1)_{s+t}K(L_p)$

Example 5.3. Turning instead to $V(1)_*K(L_p)$, as determined from $V(1)_*K(\mathbb{Z}_p)$ and $V(1)_*K(\ell_p)$ by the homotopy cofiber sequence (4.2), free $\mathbb{F}_p[v_2]$-module generators for the E^2-term in Conjecture 5.1 would be concentrated in the groups $H^r_{\mathrm{mot}}(L_p; \mathbb{F}_{p^2}(m))$ with $0 \le r \le 3$ and $r \le m < r + p^2 - 1$. The (s,t)-bidegrees of these $4p+4$ generators are displayed for $p = 5$ in Figure 5.1, lying in a fundamental domain in the shape of a parallelogram, of width 3 and height $2p^2 - 2$. In addition, there are sporadic copies of $\mathbb{F}_p$ in $V(1)_{2p-3}K(L_p)$ and $V(1)_{2p-2}K(L_p)$.

In this case, the motivic truncation property for L_p is perfectly satisfied, in the sense that

$$H^r_{\mathrm{mot}}(L_p; \mathbb{F}_{p^2}(m)) \cong \begin{cases} H^r_{\text{ét}}(L_p; \mathbb{F}_{p^2}(m)) & \text{for } 0 \le r \le m, \\ 0 & \text{otherwise,} \end{cases}$$

where, by definition,

$$H^r_{\text{ét}}(L_p; \mathbb{F}_{p^2}(*)) = v_2^{-1} H^r_{\mathrm{mot}}(L_p; \mathbb{F}_{p^2}(*))$$

is free over the Laurent polynomial ring $\mathbb{F}_p[v_2^{\pm 1}]$ on the same generators as in Figure 5.1.

The (additive) Euler characteristic

$$\chi(L_p; \mathbb{F}_{p^2}(m)) = \sum_{r=0}^{3} (-1)^r \dim_{\mathbb{F}_p} H^r_{\mathrm{mot}}(L_p; \mathbb{F}_{p^2}(m))$$

is zero for each m, cf. [76, Thm. 2.2]. To the eyes of algebraic K-theory and the hypothetical motivic cohomology, the strictly commutative ring spectrum L_p behaves much like a brave new field. We discuss the role of its (non-commutative) residue ring L/p in the next section.

The étale cohomology of a p-adic field F satisfies local Tate–Poitou duality [76, Thm. 2.1]. In the case of mod p coefficients, this is a perfect pairing

$$H^r_{\text{ét}}(F; \mathbb{Z}/p(m)) \otimes H^{2-r}_{\text{ét}}(F; \mathbb{Z}/p(1-m)) \xrightarrow{\cup} H^2_{\text{ét}}(F; \mathbb{Z}/p(1)) \cong \mathbb{Z}/p$$

for each r and m. For general p-power torsion coefficients there is a perfect pairing taking values in the larger group $H^2_{\text{ét}}(F; \mathbb{Z}/p^\infty(1)) \cong \mathbb{Z}/p^\infty$, cf. [72, p. 130]. The multiplicative structure on $V(1)_*K(L_p)$ is compatible with an algebra structure on $H^*_{\mathrm{mot}}(L_p; \mathbb{F}_{p^2}(*))$ such that the resulting multiplicative structure on $H^*_{\text{ét}}(L_p; \mathbb{F}_{p^2}(*))$ also satisfies *arithmetic duality*. This can be seen as a rotational symmetry about $(s,t) = (-3/2, p+1)$ in the variant of Figure 5.1 where v_2 has been inverted.

Conjecture 5.4. *For finite extensions B of L_p there is a perfect pairing*

$$H^r_{\text{ét}}(B; \mathbb{F}_{p^2}(m)) \otimes H^{3-r}_{\text{ét}}(B; \mathbb{F}_{p^2}(p+1-m)) \xrightarrow{\cup} H^3_{\text{ét}}(B; \mathbb{F}_{p^2}(p+1)) \cong \mathbb{Z}/p$$

for each r and m.

Remark 5.5. The dependence of the twist in $\mathbb{F}_{p^2}(p+1)$ on the prime p may be an artifact of the passage to mod p and v_1 coefficients. Let $E_2/(p^\infty, u_1^\infty)$ be the E_2-module spectrum defined by the homotopy cofiber sequences $E_2 \to p^{-1}E_2 \to E_2/p^\infty$ and $E_2/p^\infty \to u_1^{-1}E_2/p^\infty \to E_2/(p^\infty, u_1^\infty)$. Its homotopy groups $\pi_*E_2/(p^\infty, u_1^\infty) = \mathbb{W}\mathbb{F}_{p^2}[[u_1]]/(p^\infty, u_1^\infty)[u^{\pm 1}]$ are Pontryagin dual to those of E_2. Then

$$\mathbb{F}_{p^2}(p+1) = V(1)_{2p+2}E_2 \cong V(1)_{2p+3}(E_2/p^\infty) \cong V(1)_{2p+4}E_2/(p^\infty, u_1^\infty)$$

and

$$\begin{aligned} V(1)_{2p+4}E_2/(p^\infty, u_1^\infty) &\subset (S/p)_5E_2/(p^\infty, u_1^\infty) \\ &\subset \pi_4E_2/(p^\infty, u_1^\infty) = \mathbb{W}\mathbb{F}_{p^2}[[u_1]]/(p^\infty, u_1^\infty)(2)\,. \end{aligned}$$

The conjectured arithmetic duality for mod p and v_1 coefficients may be a special case of a duality for p- and u_1-power torsion coefficients, taking values in

$$H^3_{\text{ét}}(B; \mathbb{W}\mathbb{F}_{p^2}[[u_1]]/(p^\infty, u_1^\infty)(2))\,.$$

It would be desirable to find a canonical identification of this group, like the Hasse invariant in the classical case of p-adic fields and Kato's work [43] on the Galois cohomology of higher-dimensional local fields.

6. Fraction fields and ramified extensions

The étale cohomology of a field is, by construction, the same as its Galois cohomology, i.e., the continuous group cohomology of its absolute Galois group. There is no such direct description of $H^r_{\text{ét}}(L_p; \mathbb{F}_{p^2}(m))$, since according to Baker–Richter [8] the maximal connected pro-Galois extension of L_p is the composite

$$L_p \xrightarrow{\phi} KU_p \longrightarrow KU_p^{\mathrm{nr}}\,,$$

where $\pi_*(KU_p^{\mathrm{nr}}) = \mathbb{W}\bar{\mathbb{F}}_p[u^{\pm 1}]$. The unramified extensions of $\pi_0(KU_p) = \mathbb{Z}_p$ are spectrally realized, using the methods of Schwänzl–Vogt–Waldhausen [71, Thm. 3] or Goerss–Hopkins–Miller [31], but the associated Galois group only has p-cohomological dimension 1, whereas L_p would have p-cohomological dimension 3. Likewise, the maximal connected pro-Galois extension of E_n is E_n^{nr}, with $\pi_*(E_n^{\mathrm{nr}}) = \mathbb{W}\bar{\mathbb{F}}[[u_1, \dots, u_{n-1}]][u^{\pm 1}]$, of p-cohomological dimension 1 over E_n and $\hat{L}_n E(n)$.

To allow for ramification at p, one might simply invert that prime. However, the resulting strictly commutative ring spectrum $p^{-1}L_p$, with $\pi_*(p^{-1}L_p) = \mathbb{Q}_p[v_1^{\pm 1}]$, is an algebra over $p^{-1}S_p = H\mathbb{Q}_p$, so $V(1)_*K(p^{-1}L_p)$ is an algebra over $V(1)_*K(\mathbb{Q}_p)$, where v_2 acts trivially. Hence $H^r_{\text{ét}}(p^{-1}L_p; \mathbb{F}_{p^2}(*))$ would be zero.

A milder form of localization may be appropriate. By Waldhausen's localization theorem [82], the homotopy fiber of $K(L_p) \to K(p^{-1}L_p)$ is given by the algebraic K-theory of the category with cofibrations of finite cell L_p-modules with p-power torsion homotopy, equipped with the usual weak equivalences. We might instead step back to a category with cofibrations of coherent L/p^ν-modules (i.e., having degreewise finite homotopy groups, see Barwick–Lawson [9]), for some natural number ν, and suppose that these have the same algebraic K-theory as the category with cofibrations of finite cell L/p-modules. Here $L/p = K(1)$ is the first Morava K-theory, which by Angeltveit [2] is a strict ring spectrum, but not strictly commutative. By Davis–Lawson [22, Cor. 6.4] the tower $\{L/p^\nu\}_\nu$ is as commutative as possible in the category of pro-spectra:

$$L/p \longleftarrow \{L/p^\nu\}_\nu \longleftarrow L_p \longrightarrow p^{-1}L_p\,.$$

Definition 6.1. Let $K(\!\!f L_p)$ be defined by the homotopy cofiber sequence

$$K(L/p) \xrightarrow{i_*} K(L_p) \longrightarrow K(\!\!f L_p)\,,$$

where i_* is the transfer map associated to $i\colon L_p \to L/p$.

We think of $K(\!\!f L_p)$ as the algebraic K-theory of a hypothetical *fraction field* of L_p, intermediate between L_p and $p^{-1}L_p$, and similar to the 2-dimensional local field $\mathbb{Q}_p(\!(u)\!)$. Its mod p and v_1 homotopy groups can be calculated using the following result, in combination with the homotopy cofiber sequence $K(\mathbb{F}_p) \to K(\ell/p) \to K(L/p)$.

Theorem 6.2 (Ausoni–Rognes [7, Thm. 7.6, Thm. 7.7]). *$V(1)_*TC(\ell/p; p)$ is a finitely generated free $\mathbb{F}_p[v_2]$-module on $2p^2 - 2p + 8$ generators, which are located in degrees $-1 \le * \le 2p^2 + 2p - 2$. There is an exact sequence of $\mathbb{F}_p[v_2]$-modules*

$$0 \to V(1)_*K(\ell/p) \xrightarrow{\mathrm{trc}} V(1)_*TC(\ell/p; p) \longrightarrow \Sigma^{-1}\mathbb{F}_p \oplus \Sigma^{2p-2}\mathbb{F}_p \to 0\,.$$

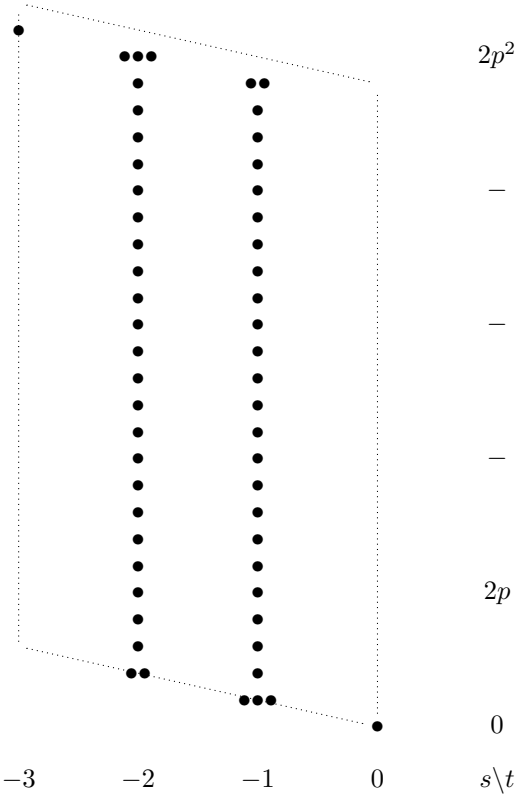


Figure 6.1. $\mathbb{F}_p[v_2]$-generators of $E^2_{s,t} = H^{-s}_{\mathrm{mot}}(\mathit{ff}\, L_p; \mathbb{F}_{p^2}(t/2)) \Longrightarrow V(1)_{s+t}K(\mathit{ff}\, L_p)$

Example 6.3. The expected E^2-term for a spectral sequence

$$E^2_{s,t} = H^{-s}_{\mathrm{mot}}(\mathit{ff}\, L_p; \mathbb{F}_{p^2}(t/2)) \Longrightarrow V(1)_{s+t}K(\mathit{ff}\, L_p)$$

is displayed for $p = 5$ in Figure 6.1. In addition there are four sporadic copies of $\mathbb{F}_p$, in degrees $2p-3$, $2p-2$, $2p-2$ and $2p-1$. The motivic truncation properties for $\mathit{ff}\, L_p$, analogous to (5.1) and (5.2), are clearly visible, and conjecturally there is now a perfect arithmetic duality pairing

$$H^r_{\text{ét}}(\mathit{ff}\, L_p; \mathbb{F}_{p^2}(m)) \otimes H^{3-r}_{\text{ét}}(\mathit{ff}\, L_p; \mathbb{F}_{p^2}(2-m)) \xrightarrow{\cup} H^3_{\text{ét}}(\mathit{ff}\, L_p; \mathbb{F}_{p^2}(2)) \cong \mathbb{Z}/p\,.$$

After such localization of L_p away from L/p, it may be possible to construct enough Galois extensions of $\mathit{ff}\, L_p$ to realize its v_2-localized motivic cohomology as continuous group cohomology

$$H^r_{\text{ét}}(\mathit{ff}\, L_p; \mathbb{F}_{p^2}(m)) \cong H^r_{\mathrm{gp}}(G_{\mathit{ff}\, L_p}; \mathbb{F}_{p^2}(m))\,,$$

for an absolute Galois group $G_{\mathit{ff}\, L_p}$ of p-cohomological dimension 3, corresponding to some maximal extension $\mathit{ff}\, L_p \to \Omega_1$. If each Galois extension of $\mathbb{Q}_p$ can be lifted to an extension of $\mathit{ff}\, L_p$, we get a short exact sequence

$$1 \to I_{v_1} \longrightarrow G_{\mathit{ff}\, L_p} \longrightarrow G_{\mathbb{Q}_p} \to 1\,,$$

with I_{v_1} the inertia group over (v_1). Here $G_{\mathbb{Q}_p}$ has p-cohomological dimension 2, and I_{v_1} will have p-cohomological dimension 1.

In the less structured setting of ring spectra up to homotopy it is possible to construct totally ramified extensions of KU_p, complementary to the unramified extension KU_p^{nr}, without inverting p. Torii [77, Thm. 2.5] shows that for each $r \geq 1$ the homotopy cofiber

$$F(B\mathbb{Z}/p_+^{r-1}, KU_p) \xrightarrow{\tau_r^\#} F(B\mathbb{Z}/p_+^r, KU_p) \longrightarrow KU_p[\zeta_{p^r}]\,,$$

of the map of function spectra induced by the stable transfer map $\tau_r\colon B\mathbb{Z}/p_+^r \to B\mathbb{Z}/p_+^{r-1}$, is a ring spectrum up to homotopy with $\pi_* KU_p[\zeta_{p^r}] \cong \mathbb{Z}_p[\zeta_{p^r}][u^{\pm 1}]$, where ζ_{p^r} denotes a primitive p^r-th root of unity. (He has similar realization results in the $K(n)$-local category.) However, it does not make sense to talk about the algebraic K-theory of a ring spectrum up to homotopy, so these constructions are only helpful if they can be made strict.

It is not possible to realize $KU_p[\zeta_{p^r}]$ as a strictly commutative ring spectrum. Angeltveit [1, Rem. 5.18] uses the identity $\psi^p(x) = x^p + p\theta(x)$ among power operations in π_0 of a $K(1)$-local strictly commutative ring spectrum to show that if $-p$ admits a k-th root in such a ring, with $k \geq 2$, then p is invertible in that ring. For $r = 1$ we have $\mathbb{Z}_p[\zeta_p] = \mathbb{Z}_p[\xi]$ where $\xi^{p-1} = -p$, so for p odd this proves that adjoining ζ_p to KU_p in a strictly commutative context will also invert p.

It is, however, possible to adjoin ζ_p to π_0 of the connective cover ku_p, in the category of strictly commutative ring spectra, without fully inverting p. Instead, one must make some positive power of the Bott element $u \in \pi_2(ku_p)$ singly divisible by p. If one thereafter inverts u, it follows that p has also become invertible. To achieve this, we modify Torii's construction for $r = 1$ by replacing the transfer map with a norm map. This leads to the G-Tate construction

$$B^{tG} = t_G(B)^G = [\widetilde{EG} \wedge F(EG_+, i_*B)]^G$$

for a spectrum B with G-action, cf. Greenlees–May [35, p. 3]. This construction preserves strictly commutative ring structures, see McClure [50, Thm. 1].

Example 6.4. Let $KU'_p[\xi] = (ku_p)^{t\mathbb{Z}/p}$ denote the $\mathbb{Z}/p$-Tate construction on the spectrum ku_p with trivial $\mathbb{Z}/p$-action, and let $ku'_p[\xi] = KU'_p[\xi][0,\infty)$ be its connective cover. Additively, these are generalized Eilenberg–Mac Lane spectra, cf. Davis–Mahowald [23] and [35, Thm. 13.5]. Multiplicatively, $\pi_*(KU'_p[\xi]) \cong \mathbb{Z}_p[\xi][v^{\pm 1}]$ where $p + \xi^{p-1} = 0$ and $|v| = 2$. Furthermore,

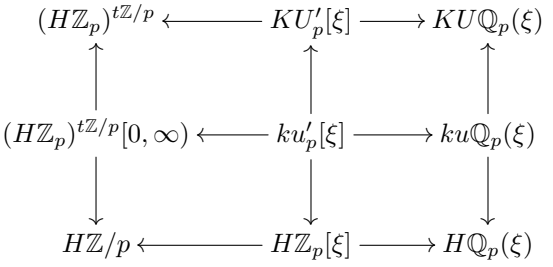
$$\pi_*(ku'_p[\xi]) \cong \mathbb{Z}_p[\xi][v]\,,$$

and a morphism $ku_p \to ku'_p[\xi]$ of strictly commutative ring spectra induces the ring homomorphism $\mathbb{Z}_p[u] \to \mathbb{Z}_p[\xi][v]$ that maps u to $\xi \cdot v$. There is no multiplicative morphism $KU_p \to KU'_p[\xi]$, but

$$KU_p \wedge_{ku_p} ku'_p[\xi] = u^{-1} ku'_p[\xi] \simeq KU\mathbb{Q}_p(\xi) = KU\mathbb{Q}_p(\zeta_p)$$

is a totally ramified extension of $KU\mathbb{Q}_p = p^{-1}KU_p$. We get a diagram of strictly commu-

tative ring spectra



with horizontal maps reducing modulo or inverting ξ, and vertical maps reducing modulo or inverting v. Here $\pi_*((H\mathbb{Z}_p)^{t\mathbb{Z}/p}) \cong \hat{H}^{-*}(\mathbb{Z}/p;\mathbb{Z}_p) = \mathbb{Z}/p[v^{\pm 1}]$. We view $ku'_p[\xi]$ as an integral model for a 2-dimensional local field close to $KU\mathbb{Q}_p(\xi)$, but note that $ku'_p[\xi]$ is not finite as a ku_p-module.

Example 6.5. Let $(\mathbb{Z}/p)^\times$ act on the group $\mathbb{Z}/p$ by multiplication, hence also on the $\mathbb{Z}/p$-Tate construction $KU'_p[\xi] = (ku_p)^{t\mathbb{Z}/p}$. Let

$$KU'_p = (KU'_p[\xi])^{h(\mathbb{Z}/p)^\times}$$

be the homotopy fixed points, and let $ku'_p = KU'_p[0,\infty)$ be its connective cover. These are strictly commutative ring spectra, with $\pi_*(KU'_p) \cong \mathbb{Z}_p[u, w^{\pm 1}]/(pw + u^{p-1})$ where $|w| = 2p-2$, and $\pi_*(ku'_p) \cong \mathbb{Z}_p[u,w]/(pw+u^{p-1})$. Multiplicative morphisms $ku_p \to ku'_p \to ku'_p[\xi]$ induce the inclusions $\mathbb{Z}_p[u] \to \mathbb{Z}_p[u,w]/(pw+u^{p-1}) \to \mathbb{Z}_p[\xi][v]$ where w maps to v^{p-1}.

The morphism $KU'_p \to KU'_p[\xi]$ is a $(\mathbb{Z}/p)^\times$-Galois extension in the sense of [63]. The morphism $ku'_p \to ku'_p[\xi]$ becomes $(\mathbb{Z}/p)^\times$-Galois after inverting p or w. It remains to be determined whether $V(1)_*K(ku'_p)$ remains purely v_2-periodic, i.e., whether the multiplicative approximation $ku_p \to ku'_p$ counters the additive splitting of ku'_p as a sum of suspended Eilenberg–Mac Lane spectra.

Example 6.6. More generally, for $r \geq 1$ let $G = \mathbb{Z}/p^r$, let $\mathscr{P}$ be the family of proper subgroups of G, and let $KU_G[0,\infty)$ be the "brutal" truncation of G-equivariant periodic K-theory, cf. [34, p. 129]. Define

$$KU'_p[\zeta_{p^r}] = (KU_G[0,\infty))^{t\mathscr{P}} = [\widetilde{E\mathscr{P}} \wedge F(E\mathscr{P}_+, KU_G[0,\infty))]^G$$

to be the $\mathscr{P}$-Tate construction, as in Greenlees–May [35, §17], and let $ku'_p[\zeta_{p^r}]$ be its connective cover. Then $\pi_*(KU'_p[\zeta_{p^r}]) \cong \mathbb{Z}_p[\zeta_{p^r}][v^{\pm 1}]$ and

$$\pi_*(ku'_p[\zeta_{p^r}]) \cong \mathbb{Z}_p[\zeta_{p^r}][v]\,.$$

The map $(KU_G[0,\infty))^{t\mathscr{P}} \to (KU_G)^{t\mathscr{P}}$ induces the inclusion $\mathbb{Z}_p[\zeta_{p^r}] \subset \mathbb{Q}_p[\zeta_{p^r}]$ in each even degree. To prove this, one can compute Amitsur–Dress homology for the family $\mathscr{P}$, use the generalized Tate spectral sequence from [35, §22], and then compare with the calculation in [35, §19] of the periodic case. The cyclotomic extension $\mathbb{Z}_p[\zeta_{p^r}]$ arises as $(R(G)/J'\mathscr{P})^\wedge_{J\mathscr{P}}$, where $R(G)$ is the representation ring, $J\mathscr{P}$ is the kernel of the restriction map $R(G) \to R(H)$, and $J'\mathscr{P}$ is the image of the induction map $R(H) \to R(G)$, where H is the index p subgroup in G.

7. Logarithmic ring spectra

The heuristics from the last two sections suggest that we should attempt to construct ramified finite extensions $B \to C$ of strictly commutative ring spectra B like ℓ_p, ku_p and e_n. The Goerss–Hopkins–Miller obstruction theory [31] for strictly commutative B-algebra structures on such spectra C has vanishing obstruction groups in the case of unramified extensions, but appears to be less useful in the case of ramification over (p), due to the presence of nontrivial (topological) André–Quillen cohomology groups [11].

The same heuristics also suggest that we should approach the extension problem by passing to mildly local versions of B, intermediate between B and $p^{-1}B$. In arithmetic algebraic geometry, one such intermediary is provided by logarithmic geometry, cf. Kato [44]. An affine pre-log scheme $(\operatorname{Spec} R, M)$ is a scheme $\operatorname{Spec} R$, a commutative monoid M, and a homomorphism $\alpha\colon M \to (R, \cdot)$ to the underlying multiplicative monoid of R. More precisely, M and α live étale locally on $\operatorname{Spec} R$. In this wider context, there is a factorization

$$\operatorname{Spec} R[M^{-1}] \longrightarrow (\operatorname{Spec} R, M) \longrightarrow \operatorname{Spec} R$$

of the natural inclusion, and the right-hand map is often a well-behaved proper replacement for the composite open immersion. Logarithmic structures on valuation rings in p-adic fields were successfully used by Hesselholt–Madsen [38] to analyze the topological cyclic homology and algebraic K-theory of these classical rings.

A theory of logarithmic structures on strictly commutative ring spectra was started by the author in [64], and developed further in joint work with Sagave and Schlichtkrull. To present it, we take the category $\mathcal{C}\mathrm{Sp}^\Sigma$ of commutative symmetric ring spectra [40], with the positive stable model structure [48, §14], as our model for strictly commutative ring spectra.

By the graded underlying space of a symmetric spectrum A we mean a diagram

$$\Omega^{\mathcal{J}}(A)\colon (\mathbf{n}_1, \mathbf{n}_2) \longmapsto \Omega^{n_2} A_{n_1}$$

of spaces, where $(\mathbf{n}_1, \mathbf{n}_2)$ ranges over the objects in a category $\mathcal{J}$. We call such a diagram a $\mathcal{J}$-space. Following Sagave, the natural category $\mathcal{J}$ to consider turns out to be isomorphic to Quillen's construction $\Sigma^{-1}\Sigma$, where Σ is the permutative groupoid of finite sets and bijections. Its nerve $B\mathcal{J} \cong B(\Sigma^{-1}\Sigma)$ is homotopy equivalent to $QS^0 = \Omega^\infty S$. For any $\mathcal{J}$-space X, the homotopy colimit $X_{h\mathcal{J}} = \operatorname{hocolim}_{\mathcal{J}} X$ is augmented over $B\mathcal{J}$, so we say that X is QS^0-graded. A map $X \to Y$ of $\mathcal{J}$-spaces is called a $\mathcal{J}$-equivalence if the induced map $X_{h\mathcal{J}} \to Y_{h\mathcal{J}}$ is a weak equivalence.

If A is a commutative symmetric ring spectrum, then $\Omega^{\mathcal{J}}(A)$ is a commutative monoid with respect to a convolution product in the category of $\mathcal{J}$-spaces. The category $\mathcal{CS}^{\mathcal{J}}$ of commutative $\mathcal{J}$-space monoids has a positive projective model structure [70, §4], with the $\mathcal{J}$-equivalences as the weak equivalences, and is Quillen equivalent to a category of E_∞ spaces over $B\mathcal{J}$. The functor $\Omega^{\mathcal{J}}\colon \mathcal{C}\mathrm{Sp}^\Sigma \to \mathcal{CS}^{\mathcal{J}}$ admits a left adjoint $M \mapsto S^{\mathcal{J}}[M]$, and $(S^{\mathcal{J}}[-], \Omega^{\mathcal{J}})$ is a Quillen adjunction.

There is a commutative submonoid of graded homotopy units $\iota\colon GL_1^{\mathcal{J}}(A) \subset \Omega^{\mathcal{J}}(A)$. A *pre-log ring spectrum* (A, M, α) is a commutative symmetric ring spectrum A with a pre-log structure (M, α), i.e., a commutative $\mathcal{J}$-space monoid M and a map $\alpha\colon M \to \Omega^{\mathcal{J}}(A)$ in $\mathcal{CS}^{\mathcal{J}}$. If the pullback $\alpha^{-1}(GL_1^{\mathcal{J}}(A)) \to GL_1^{\mathcal{J}}(A)$ is a $\mathcal{J}$-equivalence we call (M, α) a log structure and (A, M, α) a *log ring spectrum*. We often omit α from the notation.

In order to classify extensions $(A, M) \to (B, N)$ of pre-log ring spectra, one is led to study infinitesimal deformations and derivations in this category. Derivations are corepre-

sented by a logarithmic version $TAQ(A, M)$ of topological André–Quillen homology, defined by a pushout

$$\begin{array}{ccc} A \wedge_{S^{\mathcal{J}}[M]} TAQ(S^{\mathcal{J}}[M]) & \xrightarrow{\psi} & A \wedge \gamma(M) \\ \downarrow{\scriptstyle \alpha} & & \downarrow \\ TAQ(A) & \longrightarrow & TAQ(A, M) \end{array}$$

of A-module spectra, cf. [64, Def. 11.19] and [69, Def. 5.20]. Here $TAQ(A)$ is the ordinary topological André–Quillen homology, as defined by Basterra [10], and $\gamma(M)$ is the connective spectrum associated to the E_∞ space $M_{h\mathcal{J}}$. A morphism $(A, M) \to (B, N)$ is formally log étale if $B \wedge_A TAQ(A, M) \to TAQ(B, N)$ is an equivalence.

Let $j\colon e \to E$ be a fibration of commutative symmetric ring spectra. The direct image of the trivial log structure on E is the log structure $(j_*GL_1^{\mathcal{J}}(E), \alpha)$ on e given by the pullback

$$\begin{array}{ccc} j_*GL_1^{\mathcal{J}}(E) & \longrightarrow & GL_1^{\mathcal{J}}(E) \\ \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \iota} \\ \Omega^{\mathcal{J}}(e) & \xrightarrow{\Omega^{\mathcal{J}}(j)} & \Omega^{\mathcal{J}}(E) \end{array}$$

in $\mathcal{CS}^{\mathcal{J}}$. Applying this natural construction to the vertical maps in (4.1), we get the following example. Note that $\ell_p \to ku_p$ is not étale, while $L_p \to KU_p$ is Δ-Galois, hence étale.

Theorem 7.1 (Sagave [69, Thm. 6.1]). *The morphism*

$$\phi\colon (\ell_p, j_*GL_1^{\mathcal{J}}(L_p)) \longrightarrow (ku_p, j_*GL_1^{\mathcal{J}}(KU_p))$$

is log étale.

In order to approximate algebraic K-theory, one is likewise led to study logarithmic topological Hochschild homology and logarithmic topological cyclic homology. The former is defined by a pushout

$$\begin{array}{ccc} S^{\mathcal{J}}[B^{\mathrm{cy}}(M)] & \xrightarrow{\rho} & S^{\mathcal{J}}[B^{\mathrm{rep}}(M)] \\ \downarrow{\scriptstyle \alpha} & & \downarrow \\ THH(A) & \longrightarrow & THH(A, M) \end{array}$$

in $\mathcal{C}\mathrm{Sp}^\Sigma$, cf. [65, §4]. Here $THH(A)$ is the ordinary topological Hochschild homology of A, given by the cyclic bar construction of A in $\mathcal{C}\mathrm{Sp}^\Sigma$. The cyclic bar construction $B^{\mathrm{cy}}(M)$ is formed in $\mathcal{CS}^{\mathcal{J}}$, and is naturally augmented over M. The *replete bar construction* $B^{\mathrm{rep}}(M)$ can be viewed as a fibrant replacement of $B^{\mathrm{cy}}(M)$ over M in a group completion model structure on $\mathcal{CS}^{\mathcal{J}}$, cf. [68, Thm. 1.6], but also has a more direct description as the (homotopy) pullback in the right hand square below:

$$\begin{array}{ccccc} B^{\mathrm{cy}}(M) & \xrightarrow{\rho} & B^{\mathrm{rep}}(M) & \longrightarrow & B^{\mathrm{cy}}(M^{\mathrm{gp}}) \\ \downarrow{\scriptstyle \epsilon} & & \downarrow & & \downarrow{\scriptstyle \epsilon} \\ M & \xrightarrow{=} & M & \xrightarrow{\eta} & M^{\mathrm{gp}}. \end{array}$$

Here $\eta\colon M \to M^{\mathrm{gp}}$ is a group completion in $\mathcal{CS}^{\mathcal{J}}$, which means that $(M^{\mathrm{gp}})_{h\mathcal{J}}$ is a group completion of the E_∞ space $M_{h\mathcal{J}}$. The role of repletion in homotopy theory is similar to that of working within the subcategory of fine and saturated logarithmic structures in the discrete setting [44, §2].

A morphism $(A, M) \to (B, N)$ is formally log thh-étale if $B \wedge_A THH(A, M) \to THH(B, N)$ is an equivalence. The following theorem strengthens the previous result.

Theorem 7.2 (Rognes–Sagave–Schlichtkrull [66, Thm. 1.5]). *The morphism*

$$\phi\colon (\ell_p, j_*GL_1^{\mathcal{J}}(L_p)) \longrightarrow (ku_p, j_*GL_1^{\mathcal{J}}(KU_p))$$

is log thh-étale.

Remark 7.3. These results harmonize with the classical correspondence between tamely ramified extensions and log étale extensions. By Noether's theorem [53], tame ramification corresponds locally to the existence of a normal basis. This conforms with the observation that ku_p is a retract of a finite cell $\ell_p[\Delta]$-module, so that $\ell_p \to ku_p$ is tamely ramified. By contrast, ku_2 is not a retract of a finite cell $ko_2[C_2]$-module, e.g. because $(ku_2)^{tC_2}$ is nontrivial, so $ko_2 \to ku_2$ is wildly ramified.

We say that a commutative symmetric ring spectrum E is d-periodic if d is the minimal positive integer such that $\pi_*(E)$ contains a unit in degree d.

Theorem 7.4 (Rognes–Sagave–Schlichtkrull [65, Thm. 1.5]). *Let E in $\mathcal{C}\mathrm{Sp}^{\Sigma}$ be d-periodic, with connective cover $j\colon e \to E$. There is a natural homotopy cofiber sequence*

$$THH(e) \xrightarrow{\rho} THH(e, j_*GL_1^{\mathcal{J}}(E)) \xrightarrow{\partial} \Sigma THH(e[0, d))\,,$$

where $e[0, d)$ is the $(d-1)$-th Postnikov section of e.

These results allow us to realize the strategy outlined in [3, §10] to compute the $V(1)$-homotopy of $THH(ku_p)$ by way of

$$THH(\ell_p)\,, THH(\ell_p, j_*GL_1^{\mathcal{J}}(L_p)) \text{ and } THH(ku_p, j_*GL_1^{\mathcal{J}}(KU_p)).$$

The details are given in [66, §7, §8].

When $e[0, d) = H\pi_0(e)$ with $\pi_0(e)$ regular, Blumberg–Mandell [15, Thm. 4.2.1] have constructed a map of horizontal homotopy cofiber sequences

$$\begin{array}{ccccc} K(\pi_0(e)) & \xrightarrow{i_*} & K(e) & \xrightarrow{j^*} & K(E) \\ \downarrow & & \downarrow & & \downarrow \\ THH(\pi_0(e)) & \xrightarrow{i_*} & THH(e) & \xrightarrow{j^*} & WTHH^{\Gamma}(e|E) \end{array}$$

where the vertical arrows are trace maps.

Conjecture 7.5. *There is an equivalence of cyclotomic spectra*

$$THH(e, j_*GL_1^{\mathcal{J}}(E)) \simeq WTHH^{\Gamma}(e|E)\,,$$

compatible with the maps from $THH(e)$ and to $\Sigma THH(\pi_0(e))$.

The author hopes that a logarithmic analog of the Goerss–Hopkins–Miller obstruction theory [31] can be developed to classify log extensions of log ring spectra, and that in the case of log étale extensions the obstruction groups will vanish in such a way as to enable the construction of interesting examples. The underlying strictly commutative ring spectra should then provide novel examples of tamely ramified extensions, and realize a larger part of motivic cohomology as a case of Galois cohomology.

References

[1] Vigleik Angeltveit, *Topological Hochschild homology and cohomology of A_∞ ring spectra*, Geom. Topol., **12** (2008), no. 2, 987–1032.

[2] ———, *Uniqueness of Morava K-theory*, Compos. Math., **147** (2011), no. 2, 633–648.

[3] Christian Ausoni, *Topological Hochschild homology of connective complex K-theory*, Amer. J. Math., **127** (2005), no. 6, 1261–1313.

[4] ———, *On the algebraic K-theory of the complex K-theory spectrum*, Invent. Math., **180** (2010), no. 3, 611–668.

[5] Christian Ausoni and John Rognes, *Algebraic K-theory of topological K-theory*, Acta Math., **188** (2002), no. 1, 1–39.

[6] ———, *The chromatic red-shift in algebraic K-theory*, Monographie de L'Enseignement Mathématique, **40** (2008), 13–15.

[7] ———, *Algebraic K-theory of the first Morava K-theory*, J. Eur. Math. Soc. (JEMS), **14** (2012), no. 4, 1041–1079.

[8] Andrew Baker and Birgit Richter, *Galois extensions of Lubin–Tate spectra*, Homology, Homotopy Appl., **10** (2008), no. 3, 27–43.

[9] Clark Barwick and Tyler Lawson, *Regularity of structured ring spectra and localization in K-theory* (2014), available at http://arxiv.org/abs/1402.6038.

[10] M. Basterra, *André-Quillen cohomology of commutative S-algebras*, J. Pure Appl. Algebra, **144** (1999), no. 2, 111–143.

[11] Maria Basterra and Birgit Richter, *(Co-)homology theories for commutative (S-)algebras*, Structured ring spectra, London Math. Soc. Lecture Note Ser., vol. 315, Cambridge Univ. Press, Cambridge, 2004, pp. 115–131.

[12] A. Berglund and I. Madsen, *Rational homotopy theory of automorphisms of highly connected manifolds* (2014), available at http://arxiv.org/abs/1401.4096.

[13] Spencer Bloch, *Algebraic cycles and higher K-theory*, Adv. in Math. **61** (1986), no. 3, 267–304.

[14] Andrew J. Blumberg and Michael A. Mandell, *The localization sequence for the algebraic K-theory of topological K-theory*, Acta Math., **200** (2008), no. 2, 155–179.

[15] ———, *Localization for $THH(ku)$ and the topological Hochschild and cyclic homology of Waldhausen categories* (2014), available at http://arxiv.org/abs/1111.4003v3.

[16] M. Bökstedt, W. C. Hsiang, and I. Madsen, *The cyclotomic trace and algebraic K-theory of spaces*, Invent. Math., **111** (1993), no. 3, 465–539.

[17] M. Bökstedt and I. Madsen, *Topological cyclic homology of the integers*, Astérisque, **226** (1994), 7–8, 57–143. K-theory (Strasbourg, 1992).

[18] ———, *Algebraic K-theory of local number fields: the unramified case*, Prospects in topology (Princeton, NJ, 1994), Ann. of Math. Stud., vol. 138, Princeton Univ. Press, Princeton, NJ, 1995, pp. 28–57.

[19] Marcel Bökstedt and Friedhelm Waldhausen, *The map $BSG \to A(*) \to QS^0$*, Algebraic topology and algebraic K-theory (Princeton, N.J., 1983), Ann. of Math. Stud., vol. 113, Princeton Univ. Press, Princeton, NJ, 1987, pp. 418–431.

[20] Armand Borel, *Stable real cohomology of arithmetic groups*, Ann. Sci. École Norm. Sup. (4), **7** (1974), 235–272 (1975).

[21] William Browder, *Surgery on simply-connected manifolds*, Springer-Verlag, New York, 1972. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 65.

[22] Daniel G. Davis and Tyler Lawson, *Commutative ring objects in pro-categories and generalized Moore spectra* (2013), available at http://arxiv.org/abs/1208.4519v3.

[23] Donald M. Davis and Mark Mahowald, *The spectrum $(P \wedge bo)_{-\infty}$*, Math. Proc. Cambridge Philos. Soc., **96** (1984), no. 1, 85–93.

[24] Bjørn Ian Dundas, *Relative K-theory and topological cyclic homology*, Acta Math., **179** (1997), no. 2, 223–242.

[25] W. G. Dwyer and S. A. Mitchell, *On the K-theory spectrum of a ring of algebraic integers*, K-Theory, **14** (1998), no. 3, 201–263.

[26] A. D. Elmendorf, I. Kriz, M. A. Mandell, and J. P. May, *Rings, modules, and algebras in stable homotopy theory*, Mathematical Surveys and Monographs, vol. 47, American Mathematical Society, Providence, RI, 1997. With an appendix by M. Cole.

[27] F. T. Farrell and W. C. Hsiang, *On the rational homotopy groups of the diffeomorphism groups of discs, spheres and aspherical manifolds*, Algebraic and geometric topology (Proc. Sympos. Pure Math., Stanford Univ., Stanford, Calif., 1976), Part 1, Proc. Sympos. Pure Math., XXXII, Amer. Math. Soc., Providence, R.I., 1978, pp. 325–337.

[28] F. T. Farrell and L. E. Jones, *Stable pseudoisotopy spaces of compact non-positively curved manifolds*, J. Differential Geom., **34** (1991), no. 3, 769–834.

[29] S. Galatius and O. Randal-Williams, *Stable moduli spaces of high dimensional manifolds* (2012), available at http://arxiv.org/abs/1201.3527v2.

[30] Thomas Geisser and Marc Levine, *The Bloch–Kato conjecture and a theorem of Suslin–Voevodsky*, J. Reine Angew. Math., **530** (2001), 55–103.

[31] P. G. Goerss and M. J. Hopkins, *Moduli spaces of commutative ring spectra*, Structured ring spectra, London Math. Soc. Lecture Note Ser., vol. 315, Cambridge Univ. Press, Cambridge, 2004, pp. 151–200.

[32] Thomas G. Goodwillie, *Relative algebraic K-theory and cyclic homology*, Ann. of Math., (2) **124** (1986), no. 2, 347–402.

[33] ———, *The differential calculus of homotopy functors*, Proceedings of the International Congress of Mathematicians, Vol. II (Kyoto, 1990), Math. Soc., Japan, Tokyo, 1991, pp. 621–630.

[34] J. P. C. Greenlees, *Equivariant connective K-theory for compact Lie groups*, J. Pure Appl. Algebra, **187** (2004), no. 1-3, 129–152.

[35] J. P. C. Greenlees and J. P. May, *Generalized Tate cohomology*, Mem. Amer. Math. Soc., **113** (1995), no. 543, viii+178.

[36] A. E. Hatcher, *Higher simple homotopy theory*, Ann. of Math. (2), **102** (1975), no. 1, 101–137.

[37] Lars Hesselholt, *On the Whitehead spectrum of the circle*, Algebraic topology, Abel Symp., vol. 4, Springer, Berlin, 2009, pp. 131–184.

[38] Lars Hesselholt and Ib Madsen, *On the K-theory of local fields*, Ann. of Math. (2), **158** (2003), no. 1, 1–113.

[39] Michael J. Hopkins and Jeffrey H. Smith, *Nilpotence and stable homotopy theory. II*, Ann. of Math. (2), **148** (1998), no. 1, 1–49.

[40] Mark Hovey, Brooke Shipley, and Jeff Smith, *Symmetric spectra*, J. Amer. Math. Soc., **13** (2000), no. 1, 149–208.

[41] W. C. Hsiang and B. Jahren, *A note on the homotopy groups of the diffeomorphism groups of spherical space forms*, Algebraic K-theory, Part II (Oberwolfach, 1980), Lecture Notes in Math., vol. 967, Springer, Berlin, 1982, pp. 132–145.

[42] Kiyoshi Igusa, *The stability theorem for smooth pseudoisotopies*, K-Theory, **2** (1988), no. 1-2, vi+355.

[43] Kazuya Kato, *Galois cohomology of complete discrete valuation fields*, Algebraic K-theory, Part II (Oberwolfach, 1980), Lecture Notes in Math., vol. 967, Springer, Berlin, 1982, pp. 215–238.

[44] ———, *Logarithmic structures of Fontaine–Illusie*, Algebraic analysis, geometry, and number theory (Baltimore, MD, 1988), Johns Hopkins Univ. Press, Baltimore, MD, 1989, pp. 191–224.

[45] John R. Klein and John Rognes, *The fiber of the linearization map* $A(*) \to K(\mathbf{Z})$, Topology, **36** (1997), no. 4, 829–848.

[46] Ib Madsen and Michael Weiss, *The stable moduli space of Riemann surfaces: Mumford's conjecture*, Ann. of Math. (2), **165** (2007), no. 3, 843–941.

[47] Mark Mahowald and Stewart Priddy (eds.), *Algebraic topology*, Contemporary Mathematics, vol. 96, American Mathematical Society, Providence, RI, 1989.

[48] M. A. Mandell, J. P. May, S. Schwede, and B. Shipley, *Model categories of diagram spectra*, Proc. London Math. Soc. (3), **82** (2001), no. 2, 441–512.

[49] Randy McCarthy, *Relative algebraic K-theory and topological cyclic homology*, Acta Math., **179** (1997), no. 2, 197–222.

[50] J. E. McClure, *E_∞-ring structures for Tate spectra*, Proc. Amer. Math. Soc., **124** (1996), no. 6, 1917–1922.

[51] J. E. McClure and R. E. Staffeldt, *On the topological Hochschild homology of* $b\mathrm{u}$. *I*, Amer. J. Math., **115** (1993), no. 1, 1–45.

[52] Robert E. Mosher, *Some stable homotopy of complex projective space*, Topology, **7** (1968), 179–193.

[53] E. Noether, *Normalbasis bei Körpern ohne höhere Verzweigung*, J. Reine Angew. Math., **167** (1932), 147–152.

[54] Daniel Quillen, *Higher algebraic K-theory. I*, Algebraic K-theory, I: Higher K-theories (Proc. Conf., Battelle Memorial Inst., Seattle, Wash., 1972), Springer, Berlin, 1973.

[55] ———, *Higher algebraic K-theory*, Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 1, Canad. Math. Congress, Montreal, Que., 1975, pp. 171–176.

[56] John Rognes, *Characterizing connected K-theory by homotopy groups*, Math. Proc. Cambridge Philos. Soc., **114** (1993), no. 1, 99–102.

[57] ———, *Trace maps from the algebraic K-theory of the integers (after Marcel Bökstedt)*, J. Pure Appl. Algebra, **125** (1998), no. 1-3, 277–286.

[58] ———, *The product on topological Hochschild homology of the integers with mod* 4 *coefficients*, J. Pure Appl. Algebra, **134** (1999), no. 3, 211–218.

[59] ———, *Topological cyclic homology of the integers at two*, J. Pure Appl. Algebra, **134** (1999), no. 3, 219–286.

[60] ———, *Algebraic K-theory of the two-adic integers*, J. Pure Appl. Algebra, **134** (1999), no. 3, 287–326.

[61] ———, *Two-primary algebraic K-theory of pointed spaces*, Topology, **41** (2002), no. 5, 873–926.

[62] ———, *The smooth Whitehead spectrum of a point at odd regular primes*, Geom. Topol., **7** (2003), 155–184.

[63] ———, *Galois extensions of structured ring spectra. Stably dualizable groups*, Mem. Amer. Math. Soc., **192** (2008), no. 898, viii+137.

[64] ———, *Topological logarithmic structures*, New topological contexts for Galois theory and algebraic geometry (BIRS 2008), Geom. Topol. Monogr., vol. 16, Geom. Topol. Publ., Coventry, 2009, pp. 401–544.

[65] John Rognes, Steffen Sagave, and Christian Schlichtkrull, *Localization sequences for logarithmic topological Hochschild homology* (2014), available at http://arxiv.org/abs/1402.1317.

[66] ———, *Logarithmic topological Hochschild homology of topological K-theory spectra* (2014).

[67] J. Rognes and C. Weibel, *Two-primary algebraic K-theory of rings of integers in number fields*, J. Amer. Math. Soc., **13** (2000), no. 1, 1–54. Appendix A by Manfred Kolster.

[68] Steffen Sagave, *Spectra of units for periodic ring spectra and group completion of graded E_∞ spaces* (2013), available at http://arxiv.org/abs/1111.6731v2.

[69] ———, *Logarithmic structures on topological K-theory spectra*, Geom. Topol., **18** (2014), no. 1, 447–490.

[70] Steffen Sagave and Christian Schlichtkrull, *Diagram spaces and symmetric spectra*, Adv. Math., **231** (2012), no. 3-4, 2116–2193.

[71] R. Schwänzl, R. M. Vogt, and F. Waldhausen, *Adjoining roots of unity to E_∞ ring spectra in good cases—a remark*, Homotopy invariant algebraic structures (Baltimore, MD, 1998), Contemp. Math., vol. 239, Amer. Math. Soc., Providence, RI, 1999, pp. 245–249.

[72] J.-P. Serre, *Local class field theory*, Algebraic Number Theory (Proc. Instructional Conf., Brighton, 1965), Thompson, Washington, D.C., 1967, pp. 128–161.

[73] S. Smale, *On the structure of manifolds*, Amer. J. Math., **84** (1962), 387–399.

[74] Christophe Soulé, *On higher p-adic regulators*, Algebraic K-theory, Evanston 1980 (Proc. Conf., Northwestern Univ., Evanston, Ill., 1980), Lecture Notes in Math., vol. 854, Springer, Berlin, 1981, pp. 372–401.

[75] Andrei Suslin and Vladimir Voevodsky, *Bloch–Kato conjecture and motivic cohomology with finite coefficients*, The arithmetic and geometry of algebraic cycles (Banff, AB, 1998), NATO Sci. Ser. C Math. Phys. Sci., vol. 548, Kluwer Acad. Publ., Dordrecht, 2000, pp. 117–189.

[76] John Tate, *Duality theorems in Galois cohomology over number fields*, Proc. Internat. Congr. Mathematicians (Stockholm, 1962), Inst. Mittag-Leffler, Djursholm, 1963, pp. 288–295.

[77] Takeshi Torii, *Topological realization of the integer ring of local field*, J. Math. Kyoto Univ. **38** (1998), no. 4, 781–788.

[78] Vladimir Voevodsky, *Motivic cohomology with $\mathbb{Z}/2$-coefficients*, Publ. Math. Inst. Hautes Études Sci., **98** (2003), 59–104.

[79] ———, *On motivic cohomology with $\mathbb{Z}/l$-coefficients*, Ann. of Math. (2), **174** (2011), no. 1, 401–438.

[80] Friedhelm Waldhausen, *Algebraic K-theory of topological spaces. I*, Algebraic and geometric topology (Proc. Sympos. Pure Math., Stanford Univ., Stanford, Calif., 1976), Part 1, Proc. Sympos. Pure Math., XXXII, Amer. Math. Soc., Providence, R.I., 1978, pp. 35–60.

[81] ———, *Algebraic K-theory of spaces, localization, and the chromatic filtration of stable homotopy*, Algebraic topology, Aarhus 1982 (Aarhus, 1982), Lecture Notes in Math., vol. 1051, Springer, Berlin, 1984, pp. 173–195.

[82] ———, *Algebraic K-theory of spaces*, Algebraic and geometric topology (New Brunswick, N.J., 1983), Lecture Notes in Math., vol. 1126, Springer, Berlin, 1985, pp. 318–419.

[83] Friedhelm Waldhausen, Bjørn Jahren, and John Rognes, *Spaces of PL manifolds and categories of simple maps*, Annals of Mathematics Studies, vol. 186, Princeton University Press, Princeton, NJ, 2013.

[84] C. T. C. Wall, *Surgery on compact manifolds*, Academic Press, London, 1970. London Mathematical Society Monographs, No. 1.

[85] Michael Weiss and Bruce Williams, *Automorphisms of manifolds and algebraic K-theory. I*, K-Theory, **1** (1988), no. 6, 575–626.

[86] ———, *Automorphisms of manifolds*, Surveys on surgery theory, Vol. 2, Ann. of Math. Stud., vol. 149, Princeton Univ. Press, Princeton, NJ, 2001, pp. 165–220.

Department of Mathematics, University of Oslo, Norway
E-mail: rognes@math.uio.no

The topology of positive scalar curvature

Thomas Schick

Abstract. Given a smooth closed manifold M we study the space of Riemannian metrics of positive scalar curvature on M. A long-standing question is: when is this space non-empty (i.e. when does M admit a metric of positive scalar curvature)? More generally: what is the topology of this space? For example, what are its homotopy groups? Higher index theory of the Dirac operator is the basic tool to address these questions. This has seen tremendous development in recent years, and in this survey we will discuss some of the most pertinent examples. In particular, we will show how advancements of *large scale index theory* (also called *coarse index theory*) give rise to new types of obstructions, and provide the tools for a systematic study of the existence and classification problem via the K-theory of C^*-algebras. This is part of a program "mapping the topology of positive scalar curvature to analysis". In addition, we will show how advanced surgery theory and smoothing theory can be used to construct the first elements of infinite order in the k-th homotopy groups of the space of metrics of positive scalar curvature for arbitrarily large k. Moreover, these examples are the first ones which remain non-trivial in the moduli space of such metrics.

Mathematics Subject Classification (2010). Primary 53C21; Secondary 53C27, 58D17, 53C20, 58D27, 58B05, 53C23, 19K56, 58J22, 19K33, 46L80, 57R15, 57N16, 57R65.

Keywords. Positive scalar curvature, higher index theory, large scale index theory, coarse index theory, coarse geometry, C^*-index theory.

1. Introduction

One of the fundamental questions at the interface of geometry and topology concerns the relation between local geometry and global topology.

More specifically, given a compact smooth manifold M without boundary, what are the possibilities for Riemannian metrics on M? Even more specifically, can we find a metric of positive scalar curvature on M and if yes, what does the space of such metrics look like?

Recall the following definition of the scalar curvature function.

Definition 1.1. Given an n-dimensional smooth Riemannian manifold (M, g), the *scalar curvature* at x describes the volume expansion of small balls around x via

$$\frac{\mathrm{vol}(B_\epsilon(M,x))}{\mathrm{vol}(B_\epsilon(\mathbb{R}^n,0))} = 1 - \frac{\mathrm{scal}(x)}{6(n+2)}\epsilon^2 + O(\epsilon^4),$$

compare [3, 0.60]. In particular, this means that if $\mathrm{scal}(x) > 0$ then geodesic balls around x for small radius have smaller volume than the comparison balls in Euclidean space. Of

Proceedings of the International Congress of Mathematicians, Seoul, 2014

course, alternatively, $\mathrm{scal}(x)$ can be defined as a second contraction of the Riemannian curvature operator at x.

The most important tool to investigate these questions goes back to Erwin Schrödinger [30], rediscovered by André Lichnerowicz [22]: If M has positive scalar curvature and a spin structure then the Dirac operator on M is invertible. This forces its index (which is the super-dimension of the null space) to vanish.

Recall that a spin structure is a (global) differential geometric datum for a Riemannian manifold M which allows to construct a specific Riemannian vector bundle S, the spinor bundle, together with a specific differential operator of order 1, the Dirac operator D (compare e.g. [21] for a nice introduction).

On the other hand non-vanishing of the index follows from index theorems, giving rise to powerful *obstructions* to positive scalar curvature. For example, the Atiyah-Singer index theorem says that $\mathrm{ind}(D) = \hat{A}(M)$, where the $\hat{A}$-genus is a fundamental differential topological invariant (not depending on the metric!).

The most intriguing question around this method to rule out positive scalar curvature asks to which extent a sophisticated refinement of $\mathrm{ind}(D)$ the *Rosenberg index* $\alpha^{\mathbb{R}}(M)$, what takes values in the K-theory of the (real) C^*-algebra of the fundamental group Γ of M, is the only obstruction. This is the content of the (stable) Gromov-Lawson-Rosenberg conjecture.

Conjecture 1.2. *Let M be a connected closed spin manifold of dimension ≥ 5. The* Gromov-Lawson-Rosenberg conjecture *asserts that M admits a metric with positive scalar curvature if and only if $\alpha^{\mathbb{R}}(M) = 0 \in KO_*(C^*_{\mathbb{R}}\pi_1(M))$.*

The stable Gromov-Lawson-Rosenberg conjecture *claims that $\alpha^{\mathbb{R}}(M) = 0$ if and only if there is $k \in \mathbb{N}$ such that $M \times B^k$ admits a metric with positive scalar curvature. Here, B is any so-called* Bott manifold *i.e. a simply connected 8-dimensional spin manifold with $\hat{A}(M) = 1$.*

To put the stable version in context: given two closed manifolds M, B such that M admits a metric of positive scalar curvature, so does $M \times B$, simply using the product of a sufficiently scaled metric on M with any metric on B. Therefore the unstable Gromov-Lawson-Rosenberg conjecture implies the stable one.

Recall here that for a discrete group Γ the *maximal group C^*-algebra* $C^*_{max}\Gamma$ is defined as the completion of the group ring $\mathbb{C}[\Gamma]$ with respect to the maximal possible C^*-norm on $\mathbb{C}[\Gamma]$, and the *reduced group C^*-algebra* $C^*_r\Gamma$ is defined as the norm closure of $\mathbb{C}[\Gamma]$, embedded in $\mathcal{B}(l^2(\Gamma))$ via the regular representation. The *real* group C^*-algebras $C^*_{\mathbb{R},r}\Gamma$ and $C^*_{\mathbb{R},max}\Gamma$ replace $\mathbb{C}$ by $\mathbb{R}$ throughout. Using them gives more information, necessary in the Gromov-Lawson-Rosenberg conjecture. In this survey, for simplicity we will not discuss them but concentrate on the complex versions. For the Rosenberg index one can use them all, where a priori $\alpha_{max}(M) \in K_*(C^*_{max}\pi_1(M))$ is stronger than $\alpha_r(M) \in K_*(C^*_r\pi_1(M))$. The *Baum-Connes isomorphism conjecture*, compare [2], predicts the calculation of $K_*(C^*_r\Gamma)$ in terms of the equivariant K-homology of a suitable classifying space. The *strong Novikov conjecture* predicts that this equivariant K-homology at least embeds.

In celebrated work Stephan Stolz [32, 33] has established the following two partial positive results.

Theorem 1.3. *The Gromov-Lawson-Rosenberg conjecture is true for manifolds with trivial fundamental group. In other words, if M is a closed connected spin manifold of dimension*

≥ 5 with trivial fundamental group, then M admits a Riemannian metric with positive scalar curvature if and only if $\alpha^{\mathbb{R}}(M) = 0$.

More generally, if $\pi_1(M)$ satisfies the strong Novikov conjecture then the stable Gromov-Lawson-Rosenberg conjecture is true for M.

On the other hand, recall the counterexamples of [6, 28] which show that the unstable Gromov-Lawson-Rosenberg conjecture is not always true.

Theorem 1.4. *For $5 \leq n \leq 8$ there exist closed spin manifolds M^n of dimension n such that $\alpha(M^n) = 0$, but such that M^n does not admit a metric with positive scalar curvature.*

The manifolds M_n can be constructed with fundamental groups $\mathbb{Z}^{n-1} \times \mathbb{Z}/3\mathbb{Z}$ or with appropriately chosen torsion-free fundamental group, but not with a free abelian fundamental group [19]. It remains one of the most intriguing open questions whether the Gromov-Lawson-Rosenberg conjecture is true for all n-dimensional manifolds with fundamental group $(\mathbb{Z}/3\mathbb{Z})^n$.

The obstructions used in the counterexamples of Theorem 1.4 are not based on index theory, but on the minimal hypersurface method of Richard Schoen and Shing-Tung Yau [29] which we will not discuss further in this survey.

As a companion to the Gromov-Lawson-Rosenberg conjecture we suggest a slightly weaker conjecture about the strength of the Rosenberg index:

Conjecture 1.5. *Let M be a closed spin manifold. Every obstruction to positive scalar curvature for manifolds of dimension ≥ 5 which is based on index theory of Dirac operators can be read off the Rosenberg index $\alpha^{\mathbb{R}}(M) \in KO_*(C^*_{\mathbb{R}}\pi_1(M))$.*

This is vague because the statement "based on index theory of Dirac operators" certainly leaves room for interpretation.

By Stolz' Theorem 1.3, Conjecture 1.5 follows from the strong Novikov conjecture. On the other hand, every index theoretic obstruction which is not (yet) understood in terms of the Rosenberg index is particularly interesting. After all, it is a potential starting point to obtain counterexamples to the strong Novikov conjecture.

Around this question we discuss the following results [8, 10, 11].

Theorem 1.6. *Let M be an area-enlargeable spin manifold (which implies by the work of Mikhail Gromov and Blaine Lawson [7] that M does not admit a metric of positive scalar curvature).*

Then $\alpha_{max}(M) \neq 0 \in K_*(C^*_{max}\pi_1(M))$.

If M is even (length)-enlargeable, then $\alpha_r(M) \neq 0 \in K_*(C^*_r\pi_1(M))$.

Recall that a closed n-dimensional manifold M is called *enlargeable* if it admits a sequence of coverings $M_i \to M$ which come with compactly supported maps $f_i \colon M_i \to S^n$ of non-zero degree but such that $\sup_{x \in M_i} \|D_x f_i\|$ tends to 0 as $i \to \infty$. It is *area-enlargeable* if the same holds with $\left\|\Lambda^2 D_x f_i\right\|$ (a weaker condition). For the definition of the norms, we use a fixed metric on M and its pull-backs to M_i and a fixed metric on S^n.

As a potential counterexample to Conjecture 1.5 we describe a *codimension*-2 obstruction to positive scalar curvature (in a special form introduced by Mikhail Gromov and Blaine Lawson in [7, Theorem 7.5]) which is based on index theory of the Dirac operator, but which so far is not known to be encompassed by the Rosenberg index.

Theorem 1.7 (compare [9, Section 4]). *Let M be a closed connected spin manifold with vanishing second homotopy group. Assume that $N \subset M$ is a smooth submanifold of codimension 2 with trivial normal bundle and such that the inclusion induces an injection on the level of fundamental groups $\pi_1(N) \hookrightarrow \pi_1(M)$. Finally, assume that the Rosenberg index of the Dirac operator on the submanifold N does not vanish: $0 \neq \alpha(N) \in K_*(C^*\pi_1 N)$.*

Then M does not admit a Riemannian metric with positive scalar curvature.

(Secondary) index invariants of the Dirac operator can be used in the *classification* of metrics of positive scalar curvature, if applied to appropriately constructed examples. "Classification" means in particular to understand how many deformation classes of metrics of positive scalar curvature a given manifold carries, or more generally what the topology of the space of such metrics looks like.

A promising tool to systematically study the existence and classification problem is the "Stolz positive scalar curvature long exact sequence". It has the form

$$\cdots \to R_{n+1}(\pi_1(M)) \to \mathrm{Pos}_n(M) \to \Omega_n^{\mathrm{spin}}(M) \to R_n(\pi_1(M)) \to \cdots$$

Here the group we would like to understand is $\mathrm{Pos}_n(M)$, the structure group of metrics of positive scalar curvature (on the spin manifold M and related manifolds, and modulo a suitable bordism relation). The group $\Omega_n^{\mathrm{spin}}(M)$ is the usual spin bordism group from algebraic topology, which is very well understood. Finally, $R_*(\pi_1(M))$ indeed depends only on the fundamental group of the manifold in question. Note that this positive scalar curvature sequence is very similar in spirit to the surgery exact sequence coming up in the classification of manifolds.

Unfortunately we have not yet been able to fully compute all the terms in this exact sequence, even for the simplest possible case of trivial fundamental group. However, a lot of information can be gained using index theory by mapping out to more manageable targets. Here, we refer in particular to [23], joint with Paolo Piazza, where we construct a commutative diagram of maps, using *large scale index theory*, to the K-theory sequence of associated C^*-algebras

$$\cdots \to K_{n+1}(C_r^*\Gamma) \to K_{n+1}(D^*\tilde{M}^\Gamma) \to K_n(M) \xrightarrow{\alpha} K_n(C_r^*\Gamma) \to \cdots$$

We again abbreviate $\Gamma = \pi_1(M)$. This sequence was introduced by Nigel Higson and John Roe [14] and called there the *analytic surgery exact sequence*. A lot is known about this K-theory sequence: $K_n(M)$ is just the usual topological K-homology of M (an important generalized homology theory). Moreover, α is the Baum-Connes assembly map.

A good deal of this survey will discuss the *large scale index theory* underlying the constructions. In particular, we will explain two primary and secondary index theorems which play key roles in the application of this theory:

- A vanishing theorem for the large scale index of the Dirac operator under partial positivity of the scalar curvature, Theorem 6.1.
- A higher secondary index theorem, that shows how the large scale index of a manifold with boundary (which has positive scalar curvature near the boundary) determines a structure invariant of the boundary's metric of positive scalar curvature.

A fundamental problem in the use of (higher) index theory centers around the question whether there is a difference between *topological information* (which typically can be

computed much more systematically) and analytical information (which has the desired geometric consequences but often is hard to compute). This is answered (conjecturally) by the strong Novikov conjecture. This explains why these conjectures play such a central role in higher index theory. The appropriate version for large scale index theory is the *coarse Baum-Connes conjecture* (here, "with coefficients", for details compare Section 4).

Conjecture 1.8. *Given a locally compact metric space X of bounded geometry and an auxiliary coefficient C^*-algebra A, then in the composition*

$$K_*^{lf}(X;A) \to KX_*(X;A) \to K_*(C^*(X;A))$$

the second map is an isomorphism. Here, $K_*^{lf}(X)$ is the topologists' locally finite K-homology (a generalized homology theory) of the space X, and $K_*^{lf}(X;A)$ is a version with coefficients, still a generalized cohomology theory. Finally $KX_*(X;A)$, the coarse K-homology, is a variant which depends only on the large scale geometry of X.

This conjecture has many concrete applications. It implies the strong Novikov conjecture. However, I expect that counterexamples to these conjectures eventually will be found.

Concerning the classification question mentioned above, in the last part of the survey we will discuss a new construction method, based on advanced surgery theory and smoothing theory in the topology of manifolds.

Definition 1.9. We define $\mathrm{Riem}^+(M)$ to be the space of Riemannian metrics of positive scalar curvature on M, an infinite dimensional manifold.

The main result is that $\pi_k(\mathrm{Riem}^+(M))$ is very often non-trivial, even its image in the moduli space of such metrics remains non-trivial.

More precisely, we have the following theorem, derived in joint work with Bernhard Hanke and Wolfgang Steimle (compare [12, Theorem 1.1]).

Theorem 1.10. *For every $k \in \mathbb{N}$, there is $n_k \in \mathbb{N}$ such that, whenever M is a connected closed spin manifold with a metric g_0 of positive scalar curvature and with $\dim(M) > n_k$ and $k+\dim(M)+1 \equiv 0 \pmod 8$, then $\pi_k(\mathrm{Riem}^+(M), g_0)$ contains an element of infinite order.*

If M is a sphere, then the image of this element in $\pi_k(\mathrm{Riem}^+(M)/\mathrm{Diffeo}_{x_0}(M))$ also has infinite order, where the diffeomorphism group acts by pullback.

The second part of the theorem implies that the examples constructed do not rely on the homotopy properties of the diffeomorphism group of M. This is in contrast to all previous known cases, compare in particular [5, 17].

Here, $\mathrm{Diffeo}_{x_0}(M)$ is the subgroup of the full diffeomorphism group consisting of diffeomorphisms of M which fix the point $x_0 \in M$ and whose differential at x_0 is the identity. It is much more reasonable to use this subgroup instead of the full diffeomorphism group, because it ensures that the moduli space $\mathrm{Riem}^+(M)/\mathrm{Diffeo}_{x_0}(M)$ remains an infinite dimensional manifold, instead of producing a very singular space.

Remark 1.11. Most of the results mentioned so far display also how poorly the topology of positive scalar curvature is understood: the method relies on the index theory of the Dirac operator and the Schrödinger-Lichnerowicz formula. This is quite a miraculous relation which certainly is very helpful. But it requires the presence of a spin structure. Manifolds

without spin structure (and where not even the universal covering admits a spin structure) a priori shouldn't be very different from manifolds with spin structure, i.e. one would expect that many of them do not admit a metric of positive scalar curvature. But until now we have almost no tools to decide this (apart from the minimal surface method, which is only established in small dimensions).

Almost any progress in this direction would be a real breakthrough. An interesting recent contribution is the work of Dmitri Bolotov and Alexander Dranishnikov, which deals in particular with n-dimensional non-spin manifolds with fundamental group free abelian of rank n [4].

2. Index theory and obstructions to positive scalar curvature

The underlying principle how scalar curvature is coupled to the Dirac operator comes from a formula of Schrödinger [30], rediscovered and first applied by Lichnerowicz [22]. The starting point is a spin manifold (M, g), with spinor bundle S and *Dirac operator* D. Schrödinger's formula says

$$D^2 = \nabla^*\nabla + \frac{\mathrm{scal}}{4}.$$

The first term on the right is the "rough Laplacian", by definition a non-negative unbounded operator on the L^2-sections of S. The second term stands for point-wise multiplication with the scalar curvature function. If the scalar curvature is uniformly positive, this is a positive operator and hence D is invertible.

Let us recall the basics of the index theory of the Dirac operator, formulated in the language of operator algebras and K-theory. This is the most convenient setup for the generalizations we have in mind.

We start with a very brief introduction to the K-theory of C^*-algebras.

1. The assignment $A \mapsto K_*(A)$ is a functor from the category of C^*-algebras to the category of graded abelian groups.
2. We can (for a unital C^*-algebra A) define $K_0(A)$ as the group of equivalence classes of projectors in A and the matrix algebras $M_n(A)$.
3. We can (for a unital C^*-algebra A) define $K_1(A)$ as the group of equivalence classes of invertible elements in A and $M_n(A)$.
4. There is a natural Bott periodicity isomorphism $K_n(A) \to K_{n+2}(A)$.
5. For each short exact sequence of C^*-algebras $0 \to I \to A \to Q \to 0$ there is naturally associated a long exact sequence in K-theory

$$\cdots \to K_{n+1}(Q) \xrightarrow{\delta} K_n(I) \to K_n(A) \to K_n(Q) \to \cdots .$$

6. One can generalize K-theory for real and graded C^*-algebras. In the former case Bottperiodicity has period 8.
7. On can use extra symmetries based on Clifford algebras to give descriptions of $K_n(A)$ which are adapted to the treatment of n-dimensional spin manifolds.

On an even dimensional manifold, the spinor bundle canonically splits into $S = S^+ \oplus S^-$ and the Dirac operator is odd, i.e. has the form $D = \begin{pmatrix} 0 & D^- \\ D^+ & 0 \end{pmatrix}$. Let $\chi\colon \mathbb{R} \to \mathbb{R}$ be any continuous function which is odd (i.e. $\chi(-x) = \chi(x)$ for $x \in \mathbb{R}$) and with $\lim_{x\to\infty} \chi(x) = 1$. Functional calculus allows to define $\chi(D)$, which is an odd bounded operator acting on $L^2(S)$. Choosing an isometry $U\colon L^2(S^-) \to L^2(S^+)$ we form $U\chi(D)^+ \in \mathcal{B} := \mathcal{B}(L^2(S^+))$, the C^*-algebra of all bounded operators on $L^2(S^+)$. If M is *compact*, ellipticity of D implies that $U\chi(D)^+$ is invertible modulo the ideal $\mathcal{K} := \mathcal{K}(L^2(S^+))$ of compact operators. The short exact sequence of C^*-algebras $0 \to \mathcal{K} \to \mathcal{B} \to \mathcal{B}/\mathcal{K} \to 0$ gives rise to a long exact K-theory sequence and the relevant piece of this exact sequence for us is

$$\cdots \to K_1(\mathcal{B}) \to K_1(\mathcal{B}/\mathcal{K}) \xrightarrow{\delta} K_0(\mathcal{K}) \to \dots \tag{2.1}$$

As invertible elements in A represent classes in $K_1(A)$, the above spectral considerations yield a class $[U\chi(D)^+] \in K_1(\mathcal{B}/\mathcal{K})$ and we define the index to be $\operatorname{ind}(D) := \delta(U\chi(D)^+) \in K_0(\mathcal{K})$.

Of course, for the compact operators, $K_0(\mathcal{K})$ is isomorphic to $\mathbb{Z}$, generated by any rank 1 projector in $\mathcal{K}$. In our case $\delta(U\chi(D)^+)$ is represented by the projector onto the kernel of D^+ minus the projector onto the kernel of D^-, so that we arrive at the usual $\operatorname{ind}(D) := \dim\ker(D^+) - \dim\ker(D^-)$.

Analysing the situation more closely, the additional geometric information of positive scalar curvature implies invertibility of D which translates to the fact that $U\chi(D)^+$ is invertible already in $\mathcal{B}$ and then $[U\chi(D)^+] \in K_1(\mathcal{B}/\mathcal{K})$ in the sequence (2.1) has a lift to $K_1(\mathcal{B})$. Exactness implies $\operatorname{ind}(D) = 0$.

A second main ingredient concerning the index of the Dirac operator is the Atiyah-Singer index theorem [1]. A priori $\operatorname{ind}(D)$ (like the operator D) depends on the Riemannian metric on M. However, the index theorem expresses it in terms which are independent of the metric. More specifically, $\operatorname{ind}(D) = \hat{A}(M)$, the $\hat{A}$-genus of M.

This result has vast generalizations in many directions. A very important one (introduced by Jonathan Rosenberg [27]) modifies the Dirac operator by "twisting" with a smooth flat bundle E of (finitely generated projective) modules over an auxiliary C^*-algebra A. One then obtains an index in $K_*(A)$. Indeed, the construction is pretty much the same as above, with the important innovation that one replaces the scalars $\mathbb{C}$ by the more interesting C^*-algebra A, as detailed in Section 3.

The second generalization works for a non-compact manifold X. In this case the classical Fredholm property of the Dirac operator fails. To overcome this, *large scale index theory*, synonymously called *coarse index theory* is developing. Again this is based on C^*-techniques and pioneered by John Roe [24]. It is tailor-made for the non-compact setting. One obtains an index in the K-theory of the Roe algebra $C^*(X; A)$. In C^*-algebras, positivity implies invertibility, which finally implies that all the generalized indices vanish if one starts with a metric of uniformly positive scalar curvature.

The general pattern (from the point of view adopted in this article) of index theory is the following:

1. The geometry of the manifold M produces an interesting operator D.
2. This operators defines an element in an operator algebra A, which depends on the precise context.

3. The operator satisfies a Fredholm condition, which means it is invertible module an ideal I of the algebra A, again depending on the context.
4. The algebras in question are C^*-algebras. This implies that "positivity" of elements is defined, and moreover "positivity" implies invertibility.
5. A very special additional geometric input implies positivity and hence honest invertibility of our operator. For us, this special context will be the fact that we deal with a metric of uniformly positive scalar curvature.
6. Indeed, any element which is invertible in A modulo an ideal I defines an element in $K_{n+1}(A/I)$, where $n = \dim(M)$. (Instead of getting K_1, the fact that we deal with the Dirac operator of an n-dimensional manifold produces additional symmetries (related to actions of the Clifford algebra Cl_n) which give rise to the element in $K_{n+1}(A/I)$.)
7. We interpret the class defined by the Dirac operator as a fundamental class $[M] \in K_{n+1}(A/I)$. Homotopy invariance of K-theory implies that $[M]$ does not depend on the full geometric data which goes in the construction of the operator D, but only on the topology of M.
8. The K-theory exact sequence of the extension $0 \to I \to A \to A/I \to 0$ contains the boundary map δ. We call the image of $[M]$ under δ the index

$$\delta\colon K_{n+1}(A/I) \to K_n(I) \ ; \ [M] \mapsto \operatorname{ind}(D).$$

   Note that the degree arises from additional dimension-dependent symmetries which we do not discuss in this survey.
9. The additional *geometric* positivity assumption (uniformly positive scalar curvature) which implies invertibility already in A, gives rise to a canonical lift of $[M]$ to an element $\rho(M,g) \in K_{n+1}(A)$. Because of this, we think of $K_{n+1}(A)$ as a *structure group* and $\rho(M,g)$ is a *structure class*. It contains information about the underlying geometry.

Indeed, we want to advocate here the idea that the setup just described has quite a number of different manifestations, depending on the situation at hand. It can be adapted in rather flexible ways. The next section treats one example.

3. Large scale index theory

We describe "large scale index theory" for a complete Riemannian manifold of positive dimension.

Therefore, let (M,g) be such a complete Riemannian manifold. Fix a Hermitian vector bundle $E \to M$ of positive dimension. We first describe the operator algebras which are relevant. They are all defined as norm-closed subalgebras of $\mathcal{B}(L^2(M;E))$.

Definition 3.1. We need the following concepts.

- An operator $T\colon L^2(M;E) \to L^2(M;E)$ has *finite propagation* (namely $\leq R$) if $\phi T \psi = 0$ whenever $\phi, \psi \in C_c(M)$ are compactly supported continuous functions whose supports have distance at least R.

Here, we think of ϕ also as bounded operator on $L^2(M;E)$, acting by point-wise multiplication.

- T as above is called *locally compact* if ϕT and $T\phi$ are compact operators whenever $\phi \in C_c(M)$.
- T is called *pseudolocal* if $\phi T \psi$ is compact whenever $\phi, \psi \in C_c(M)$ with disjoint supports, i.e. such that $\phi\psi = 0$.
- The *Roe algebra* $C^*(M)$ is defined as the norm closure of the algebra of all bounded finite propagation operators which are locally compact. It is an ideal in the *structure algebra* $D^*(M)$ which is defined as the closure of the algebra of finite propagation pseudolocal operators.
- Assume that a discrete group Γ acts by isometries on M. Requiring in the above definitions that the finite propagation operators are in addition Γ-equivariant and then completing, we obtain the pair $C^*(M)^\Gamma \subset D^*(M)^\Gamma$.

Remark 3.2. For technical reasons, one actually should replace the bundle E by the bundle $E \otimes l^2(\mathbb{N})$ whose fibers are separable Hilbert spaces (or in the equivariant case by $E \otimes l^2(\mathbb{N}) \otimes l^2(\Gamma)$). Via the embedding $\mathcal{B}(L^2(M;E)) \to \mathcal{B}(L^2(M; E \otimes l^2(\mathbb{N})))$ implicitly we think of operators on $L^2(M;E)$ as operators in the bigger algebra without mentioning this. Using the larger bundle guarantees functoriality and independence on E, implicit in our notion $D^*(M)$.

Let A be an auxiliary C^*-algebra. Typical examples arise from a discrete group Γ, namely $C^*_{max}\Gamma$ or $C^*_r\Gamma$. An important role is then played by smooth bundles E over M with fibers finitely generated projective A-modules. These inherit fiberwise A-valued inner products (or more precisely Hilbert A-module structures in the sense of [20]). Integrating the fiberwise inner product then also defines a Hilbert A-module structure on the space of continuous compactly supported sections of E. By completion, we obtain the Hilbert A-module $L^2(M;E)$. The bounded, adjointable, A-linear operators on $L^2(M;E)$ form the Banach algebra $\mathcal{B}(L^2(M;A))$. The ideal of *A-compact operators* is defined as the norm closure of the ideal generated by operators of the form $s \mapsto v \cdot \langle s, w\rangle; L^2(M;E) \to L^2(M;E)$, with $v, w \in L^2(M;E)$.

We now define $C^*(M;A)$ and $D^*(M;A)$ mimicking Definition 3.1, but replacing the Hilbert space concepts by the Hilbert A-module concepts throughout. In particular, we use A-compact operators instead of compact operators. Also, the stabilization as discussed in Remark 3.2 is replaced suitably.

Example 3.3. Given a connected manifold M with fundamental group Γ, there is a canonical $C^*\Gamma$-module bundle, the *Mishchenko bundle* $\mathcal{L}$. It is the flat bundle associated to the left multiplication action of Γ on $C^*\Gamma$, where we treat $C^*\Gamma$ as the free right $C^*\Gamma$-module of rank 1, i.e. $\mathcal{L} = \tilde{M} \times_\Gamma C^*\Gamma$.

The construction of the large scale index is now based on two principles.

Proposition 3.4. *Let M be a complete Riemannian spin manifold, A an auxiliary C^*-algebra and $E \to M$ a smooth bundle of (finitely generated projective) A-modules with compatible connection. We form the twisted Dirac operator D_E as an unbounded operator on the Hilbert A-module of L^2-spinors on M with values in E.*

For the operator D_E, there exists a functional calculus. In particular, we can form $f(D_E)$ for $f\colon \mathbb{R} \to \mathbb{R}$ a continuous function which vanishes at ∞ or which has limits as $t \to \pm\infty$. Moreover, $f(D_E)$ depends only on the restriction of f to the spectrum of D_E. Then

(1) *if f has a compactly supported (distributional) Fourier transform then $f(D_E)$ has finite propagation.*

(2) *if f vanishes at infinity, then $f(D_E)$ is locally compact; if $f(t)$ converges as $t \to \pm\infty$ then $f(D_E)$ is at least pseudolocal.*

The first property is a rather direct consequence of unit propagation speed for the fundamental solution of the heat equation. The second one is an incarnation of ellipticity and local elliptic regularity.

These results are well known and have been used a lot in the literature (compare in particular [24]), indeed they form the basis of "large scale index theory". For the very general case needed in the proposition (with coefficients, arbitrary complete M), a complete proof is given in [9].

The construction of the large scale index is now rather straight-forward:

1. Take any continuous function $\psi\colon \mathbb{R} \to \mathbb{R}$ (later assumed to be odd) with $\lim_{t\to\infty} \psi(t) = 1$. Then Proposition 3.4 implies that $\psi(D_E)$ belongs to $D^*(M;A)$. Even better, $1-\psi^2$ vanishes at infinity, so that $1-\psi(D_E)^2$ belongs to $C^*(M;A)$.
2. By the principles listed at the end of Section 2, $\psi(D_E)$ gives rise to an element $[M;E]$ in $K_{n+1}(D^*(M;A)/C^*(M;A))$. (Here, we again avoid the discussion of the additional symmetries which raise the index by n). Homotopy invariance of C^*-algebra K-theory implies that this element is independent of the choice of ψ and depends only on the large scale features of the metric on M.
3. We define the large scale index (or synonymously "coarse index")
$$\operatorname{ind}(D_E) \in K_n(C^*(M;A))$$
as the image of $[M;E]$ under the boundary map of the long exact K-theory sequence.
4. If M has uniformly positive scalar curvature and E is flat, the Lichnerowicz-Weitzenböck formula implies that 0 is not in the spectrum of D_E. Then we can choose a function ψ which is equal to -1 on the negative part of the spectrum of D_E and equal to $+1$ on the positive part of the spectrum of D_E, so that $1-\psi^2$ vanishes on the spectrum of D_E, i.e. $\psi^2(D_E) = 1$. This means that $[M;E]$ lifts in a canonical way to $\rho(D;E) \in K_{n+1}(D^*(M;A))$ (this class depends on the metric of positive scalar curvature) and it implies that $\operatorname{ind}(D_E) = 0$.
5. A special feature is that $K_{n+1}(D^*(M;A)/C^*(M;A))$ indeed is homological in nature: it is canonically isomorphic to the locally finite K-homology $K_n^{lf}(M;A)$, satisfying the Eilenberg-Steenrod axioms of a (locally finite) generalized homology theory.

Example 3.5. If we apply this construction to a closed n-dimensional spin manifold M and the Mishchenko bundle $\mathcal{L}$ on M, we obtain $\operatorname{ind}(D_{\mathcal{L}}) \in K_n(C^*(M;C^*\Gamma))$.

However, there is a canonical isomorphism $K_*(C^*(M;C^*\Gamma)) \cong K_*(C^*\Gamma)$. Using this isomorphism, the Rosenberg index mentioned above is exactly $\operatorname{ind}(D_{\mathcal{L}})$:

$$\alpha(M) = \operatorname{ind}(D_{\mathcal{L}}) \in K_n(C^*\Gamma) \cong K_n(C^*(M;C^*\Gamma)).$$

The reduced C^*-algebra $C^*_r\Gamma$ of a discrete group is a canonical construction which captures many features of the group Γ, e.g. concerning its representation theory. However, it is very rigid. In particular, it is not functorial: a homomorphism $\Gamma_1 \to \Gamma_2$ will in general not induce a homomorphism $C^*_r\Gamma_1 \to C^*_r\Gamma_2$. As a consequence it is very hard to find homomorphisms out of $C^*_r\Gamma$ and also out of $K_*(C^*_r\Gamma)$.

Coarse geometry, however, immediately provides such a homomorphism (which allows one to detect elements in $K_*(C^*_r\Gamma)$). This is based on simple calculation: If a discrete group Γ isometrically acts freely and cocompactly on a metric space X, then C^*X^Γ is isomorphic to $C^*_r(\Gamma)\otimes\mathcal{K}$. "Forgetting equivariance" therefore gives the composite homomorphism

$$C^*_r\Gamma \hookrightarrow C^*_r\Gamma\otimes\mathcal{K} \cong C^*X^\Gamma \hookrightarrow C^*X.$$

The induced map in K-theory allows one to detect elements in $K_*(C^*_r\Gamma)$ using large scale index theory, as we will show in one case in Section 5.

4. The coarse Baum-Connes conjecture

Being the home of important index invariants, it is very important to be able to compute the K-theory of the Roe algebras $C^*(M;A)$ for arbitrary complete manifolds M and coefficient C^*-algebras A. It turns out that there are quite a number of tools to do this. Even better, at least conjecturally there is a purely homological answer to this task.

Let us start with the three most important computational tools.

1. $K_*(C^*(M;A))$ is invariant under *coarse homotopy*, compare [13].
2. There are powerful *vanishing theorems* for $K_*(C^*(M;A))$. An important one is valid if M is *flasque* [24, Proposition 9.4]. This means that M admits a shift map $f\colon M\to M$ such that, on the one hand, f is uniformly close to the identity (i.e. there is a constant C such that $d(f(x),x)<C$ for all $x\in M$). On the other hand, f moves everything to infinity in the sense that for each compact subset K of M, $\mathrm{im}(f^k)\cap K=\emptyset$ for all sufficiently large iterations f^k of f.
3. The group $K_*(C^*(M;A))$ satisfies a Mayer-Vietoris principle. For this, one needs a *coarsely excisive* decomposition $M=M_1\cup M_2$, which means that the intersection $M_0:=M_1\cap M_2$ captures all the large scale features of the relation between M_1 and M_2. The technical definition is that for each $R>0$ there is an $S>0$ such that the S-neighborhood of $M_1\cap M_2$ contains the intersection of the R-neighborhoods of M_1 and M_2.

   In this situation, there is a long exact Mayer-Vietoris sequence (compare [16, 31])

$$\begin{aligned}\cdots\to K_i(C^*(M_1;A))\oplus K_i(C^*(M_2;A))\to K_i(C^*(M;A))\to\\ K_{i-1}(C^*(M_0;A))\to K_i(C^*(M_1;A))\oplus K_i(C^*(M_2;A))\to\cdots\end{aligned}$$

One of the powerful principles for the K-theory of C^*-algebras is their close relation to purely topological quantities via isomorphism conjectures. Most prominent here is the Baum-Connes conjecture for the computation of $K_*(C^*_r\Gamma)$. The properties of $K_*(C^*(M;A))$ listed above indicate that a similar "topological expression" should be possible here. Indeed, we have the *coarse Baum-Connes conjecture (with coefficients)* [24, Conjecture 8.2], verified in many cases.

Conjecture 4.1. *Given a metric space X of bounded geometry, in the composition*

$$K_*^{lf}(X;A) \to KX_*(X;A) \to K_*(C^*(X;A))$$

the second map is an isomorphism.

Here $K_*^{lf}(X)$ is the usual locally finite K-homology of the space X, defined analytically as $K_{*+1}(D^*X/C^*X)$, and as we saw above it is no problem to introduce as coefficients a C^*-algebra A. The *coarse K-homology* KX_* is obtained as the limit of $K_*^{lf}(|\mathfrak{U}_i|)$, where the $\mathfrak{U}_i$ form a sequence of coverings of X which become coarser as $i \to \infty$. Here $|\mathfrak{U}_i|$ is the geometric realization of the associated Čech simplicial complex. If X is uniformly locally contractible, e.g. if X is the universal covering of a closed non-positively curved manifold, then the "coarsening map" $K_*^{lf}(X;A) \to KX_*(X;A)$ is an isomorphism.

Recall that (in the context of large scale geometry) "bounded geometry" means that X contains a discrete subset T such that on the one hand T coarsely fills the space (i.e. there is an $R > 0$ such that the R-neighborhood of T is all of X), but on the other hand T is uniformly discrete (i.e. for each $R > 0$ the number of elements of T contained in any R-ball is uniformly bounded from above).

The coarse Baum-Connes conjecture has a number of important consequences. Most notably, there is a principle of descent [24, Section 5] that uses the close relation between $C_r^*\Gamma$ and C^*X for any metric space X on which Γ acts properly and cocompactly. The principle of descent asserts that if such a space satisfies the coarse Baum-Connes conjecture, then the strong Novikov conjecture for Γ is true.

On the other hand, the "bounded geometry" condition of the coarse Baum-Connes conjecture is indispensable. Guoliang Yu has constructed a metric space which is a disjoint union of (scaled) spheres of growing dimension for which the analysis of the Dirac operator shows easily that the coarse assembly map is not injective. Despite its simplicity, this example remains intriguing. It is important to understand this better and to construct other examples. We believe that the coarse Baum-Connes conjecture will not be satisfied in full generality.

5. Enlargeability and index

Let M be an (area)-enlargeable closed spin manifold. Recall that this means that M comes with a sequence of (not necessarily compact) coverings M_i with compactly supported maps of non-zero degree $f_i\colon M_i \to S^n$ which are arbitrarily (area) contracting.

Mikhail Gromov and Blaine Lawson show in [7] that an enlargeable spin manifold does not admit a metric of positive scalar curvature. Theorem 1.6 verifies Conjecture 1.5 for this "enlargeability obstruction", i.e. shows that the Rosenberg index is non-zero in this situation. This was achieved in [11] by refining the construction of Gromov and Lawson as follows:

1. One constructs vector bundles E_i on M_i of small curvature which represent interesting K-theory classes.
2. Next one produces associated bundles $M(E_i)$ on M. If $M_i \to M$ is a finite covering, this is finite dimensional. If the covering $M_i \to M$ is infinite, we canonically obtain an associated "structure C^*-algebra" C_i such that $M(E_i)$ is a Hilbert C_i-module bundle with finitely generated projective fiber.

3. The crucial step is the construction of a bundle $E := \prod_i M(E_i)/\bigoplus_i M(E_i)$ which becomes a *flat* Hilbert A-module bundle where $A = \prod_i C_i/\bigoplus_i C_i$. Being flat, this corresponds to a representation of $\pi_1(M)$.
4. As E is flat, the Schrödinger-Lichnerowicz formula implies that $\mathrm{ind}(D_E) \in K_*(A)$ is an obstruction to positive scalar curvature.
5. On the other hand, the universal property of $C^*_{max}\pi_1(M)$ implies that the representation of $\pi_1(M)$ which gives rise to the bundle E induces a C^*-algebra homomorphism $C^*_{max}\pi_1(M) \to A$. Moreover, the induced map in K-theory sends $\alpha_{max}(M) \in K_*(C^*_{max}\pi_1(M))$ to $\mathrm{ind}(D_E) \in K_*(A)$.
6. Finally, an index theorem computes $\mathrm{ind}(D_E)$ in terms of the degrees of the maps f_i and in particular shows that $\mathrm{ind}(D_E) \neq 0$. It follows that $\alpha_{max}(D) \neq 0 \in K_*(C^*_{max}\pi_1(M))$.

A main innovation is the technically quite non-trivial construction of an associated honestly flat bundle, but with infinite dimensional fibers.

In [8] we relate enlargeability to the classical strong Novikov conjecture, which deals with $C^*_r\Gamma$ instead of $C^*_{max}\Gamma$.

The main idea here is to use the functoriality of the large scale index. The argument becomes technically easier if we assume that all the covering spaces which determine the enlargeability of M are the universal covering $\tilde{M}$. In this situation, the first main point is that the geometry allows us to combine all the maps $f_i\colon \tilde{M} \to S^n$ into one map $F\colon \tilde{M} \to B_\infty$, where B_∞ is the "infinite balloon space", sketched in Figure 5.1. It is defined using a collection of n-spheres of increasing radii $i = 1, 2, 3, \ldots$, with the sphere of radius i attached to the point $i \in [0, \infty)$ at the south pole of S^n, and is equipped with the path metric.

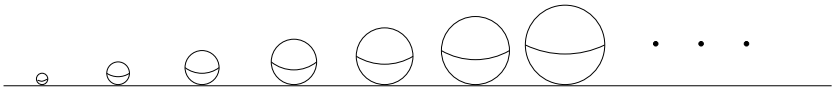


Figure 5.1. The connected balloon space B_∞

Using the Mayer-Vietoris sequence and induction on the dimension, one can calculate the coarse K-homology of B_∞ and the K-theory of its Roe algebra C^*B_∞. In particular, we obtain $KX_n(B_\infty) \cong \prod_{i\in\mathbb{N}} \mathbb{Z}/\bigoplus_{i\in\mathbb{N}} \mathbb{Z}$ and the direct calculation allows us to establish the coarse Baum-Connes conjecture for this space. We obtain a commutative diagram of K-homology and K-theory groups as follows:

$$
\begin{array}{ccccccc}
K_n(M) & \xrightarrow{\cong} & K_{n+1}(D^*\tilde{M}^\Gamma/C^*\tilde{M}^\Gamma) & \longrightarrow & K_n(C^*\tilde{M}^\Gamma) & \xrightarrow{\cong} & K_n(C^*_r\Gamma) \\
 & & \downarrow & & \downarrow & & \\
 & & K_{n+1}(D^*\tilde{M}/C^*\tilde{M}) & \longrightarrow & K_n(C^*\tilde{M}) & & \\
 & & \downarrow & & \downarrow & & \\
 & & K_{n+1}(D^*B_\infty/C^*B_\infty) & \longrightarrow & K_n(C^*B_\infty) & & \\
 & & \downarrow & & \downarrow{=} & & \\
\prod \mathbb{Z}/\bigoplus \mathbb{Z} & \xrightarrow{\cong} & KX_n(B_\infty) & \xrightarrow{\cong} & K_n(C^*B_\infty) & &
\end{array}
$$

A topological calculation allows to work out the image of the fundamental class of M in $KX_n(B_\infty) \cong \prod \mathbb{Z}/\bigoplus \mathbb{Z}$: it is the class represented by the sequence of degrees $(\deg(f_i))_{i\in\mathbb{N}}$ which by assumption is non-zero. Because of the coarse Baum-Connes conjecture, the image in the bottom right corner is also non-zero, which finally implies that also the image $\alpha(M) \in K_n(C^*_r\Gamma)$ is non-zero, as claimed by the theorem.

6. Vanishing of the index under partial positivity

The main reason why one can apply index theory to geometric and topological questions is that a special geometric situation implies vanishing results for the index. It is very important to develop further instances of such vanishing theorems, in order to widen the scope of the consequences of the index method. Here we present one of these, which is valid in the context of large scale index theory:

Theorem 6.1. *Let M be a complete non-compact connected Riemannian spin manifold. Let $E \to M$ be a flat bundle of Hilbert A-modules for a C^*-algebra A. Assume that the scalar curvature is uniformly positive outside a compact subset.*

Then the large scale index of the Dirac operator twisted with E vanishes.

For $A = \mathbb{C}$, this result has been stated by John Roe [24, 26]. A different proof, which covers the general case, is given by Bernhard Hanke, Daniel Pape and the author in [9, Theorem 3.11].

A concrete application of Theorem 6.1 to compact spin manifolds is the codimension-2 obstruction to positive scalar curvature of Theorem 1.7. In its proof in [9], a gluing and bending construction of an intermediate space gives positive scalar curvature outside of a neighborhood of the hypersurface.

7. The Stolz exact sequence

Stephan Stolz (compare [23, Proposition 1.27]) introduced a long exact sequence that makes systematic the bordism classification of metrics of positive scalar curvature. It is quite similar in spirit to the surgery exact sequence for the classification of closed manifolds.

Convention 7.1. Throughout the remainder of the article, a Riemannian metric on a manifold with boundary is assumed to have product structure near the boundary.

Definition 7.2. Fix a reference space X.

1. The group $\Omega^{\mathrm{spin}}_n(X)$ is the usual spin bordism group, consisting of cycles $f\colon M \to X$, with M a closed n-dimensional spin manifold. The equivalence relation is spin bordism.
2. The *structure group* $\mathrm{Pos}^{\mathrm{spin}}_n(X)$ is the group of bordism classes of metrics of positive scalar curvature on n-dimensional closed spin manifolds with reference map to X. Two such manifolds $(M_i, g_i, f_i\colon M_i \to X)$ are called *bordant* if there is a manifold W with boundary, with metric G of positive scalar curvature and with reference map $F\colon W \to X$ such that its boundary is $M_1 \amalg (-M_2)$ and G, F restrict to the given g_i, f_i at the boundary.

3. Finally, the group $R_n(X)$ is the group of equivalence classes of compact spin manifolds W with boundary, with reference map $f\colon W \to X$ and with a metric g of positive scalar curvature on ∂W. Again, the equivalence relation on such cycles is bordism, where a bordism between (W_1, f_1, g_1) and (W_2, f_2, g_2) is a manifold Y with boundary $\partial Y = W_1 \cup_{\partial W_1} Z \cup_{\partial W_2} -W_2$ (where Z is a spin bordism between W_1 and W_2) together with a continuous map $f\colon Y \to X$ and a positive scalar curvature metric g on Z. Of course, the restriction of f to W_j must be f_j and the restriction of g to ∂W_j must be g_j. It turns out that $R_n(X)$ only depends on $\pi_1(X)$ for $n > 5$.
4. The group structure in each of the three cases is given by disjoint union, and the inverse is obtained by reversing the spin structure and leaving all other data unchanged.
5. There are evident "forget structure" and "take boundary" maps between these groups. Using these, one obtains a long exact sequence, the *Stolz positive scalar curvature exact sequence*

$$\cdots \to R_{n+1}(\pi_1(X)) \to \mathrm{Pos}_n^{\mathrm{spin}}(X) \to \Omega_n^{\mathrm{spin}}(X) \to \dots \tag{7.3}$$

The most useful cases of this sequence arises if $X = M$ and $f = \mathrm{id}\colon M \to M$ or if $X = B\Gamma$ is the classifying space of a discrete group and $f\colon M \to B\Gamma$ induces the identity on the fundamental groups.

To get information about $\mathrm{Pos}_n^{\mathrm{spin}}(M)$ we want to use index theory systematically by mapping in a consistent way to the analytic exact sequence of Nigel Higson and John Roe. This sequence is simply the long exact K-theory sequence of the extension $0 \to C^*\tilde{M}^\Gamma \to D^*\tilde{M}^\Gamma \to D^*\tilde{M}^\Gamma/C^*\tilde{M}^\Gamma \to 0$, where $\tilde{M}$ is the universal covering of M, and $\Gamma = \pi_1(M)$.

Using that $K_*(C^*\tilde{M}^\Gamma) = K_*(C_r^*\Gamma)$ and that $K_{*+1}(D^*\tilde{M}^\Gamma/C^*\tilde{M}^\Gamma) = K_*(M)$, we obtain the following theorem, compare [23, Theorem 1.39]

Theorem 7.4. *Let X be a topological space with Γ-covering $\tilde{X}$. We have a natural canonical commutative diagram (if n is odd proved in detail)*

$$\begin{array}{ccccccc}
\longrightarrow & \Omega_{n+1}^{\mathrm{spin}}(X) & \longrightarrow & R_{n+1}(X) & \longrightarrow & \mathrm{Pos}_n^{\mathrm{spin}}(X) & \longrightarrow \Omega_n^{\mathrm{spin}}(X)\cdots \\
& \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle \mathrm{ind}} & & \downarrow{\scriptstyle \rho_\Gamma} & \downarrow{\scriptstyle \beta} \\
\longrightarrow & K_{n+1}(X) & \longrightarrow & K_{n+1}(C_r^*\Gamma) & \longrightarrow & K_{n+1}(D^*\tilde{X}^\Gamma) & \longrightarrow K_n(X)\cdots
\end{array}$$

Here, β is obtained by taking the large scale (equivariant) index of the Dirac operator on the covering of a cycle $f\colon N \to X$ and then use functoriality of large scale index theory to push forward via f_* from $K_*(C^*\tilde{N}^\Gamma)$ to $K_*(C^*\tilde{X}^\Gamma)$. It coincides with the Atiyah orientation as natural transformation from spin bordism to K-homology. Similarly, ρ_Γ is obtained by constructing the structure invariant of the positive scalar curvature metric of the covering of $(N, g, f\colon N \to X)$ and then use naturality to push forward along f_* from $K_*(D^*\tilde{N}^\Gamma)$ to $K_*(D^*\tilde{X}^\Gamma)$.

Finally, ind *assigns to a manifold with boundary and positive scalar curvature at the boundary an Atiyah-Patodi-Singer type index.*

Note that the assertion of Theorem 8.1 is that the (index based) maps all exist, that they are indeed well defined, i.e. invariant under bordism, and that the diagram is commutative. This means that we have to work (for the cycles and for the equivalence relation) throughout with manifolds with boundary. It turns out that large scale index theory can very elegantly and efficiently be used to carry out index theory on manifolds with boundary, as well.

8. Index theory on manifolds with boundary

Our method to do index theory on a manifold with boundary is simply to attach an infinite half-cylinder to the boundary. This produces a manifold without boundary, of course at the expense that the resulting manifold is never compact. However, large scale index theory can deal with such spaces.

To obtain the appropriate information, the construction must take the extra information into account coming from the fact that the metric of the boundary is assumed to have positive scalar curvature. Let us review the construction:

1. We start with a smooth manifold W with boundary, with a Riemannian metric g which has positive scalar curvature near the boundary (and a product structure there, by our general convention). As a metric space, W is assumed to be complete. Moreover, we fix an auxiliary C^*-algebra A and a flat Hilbert A-module bundle E on W (again with product structure near the boundary).
2. We now attach a half-cylinder $\partial W \times [0,\infty)$ to the boundary to obtain W_∞ and extend all the structures over W_∞. We obtain a complete manifold without boundary, with product end $\partial W \times [0,\infty)$.
3. As in Section 3, the Dirac operator D_E produces a bounded operator $\psi(D_E)$ in $D^*(W_\infty; A)$.
4. Now, however, we use the invertibility of D_E on $\partial M \times [0,\infty)$ coming from the Schrödinger-Lichnerowicz formula: for suitable ψ the element $1-\psi(D_E)^2$ does not only lie in $C^*(W_\infty; A)$ but in the smaller ideal $C^*(W \subset W_\infty; A)$. This is by definition generated by all locally compact finite propagation operators T which are *supported near* W, which means that there is $R > 0$ such that $T\phi = 0$ and $\phi T = 0$ whenever ϕ is a compactly supported function with $d(\mathrm{supp}(\phi), W) > R$.
5. Correspondingly, the fundamental class of W_∞ has a canonical lift to a class $[W, g|_{\partial W}]$ in $K_{n+1}(D^*(W_\infty; A)/C^*(W \subset W_\infty; A))$. This class does in general depend on the positive scalar curvature metric on the boundary.
6. As usual, we next define the "large scale Atiyah-Patodi-Singer index" by applying the boundary map of the long exact K-theory sequence, now for the extension

$$0 \to C^*(W \subset W_\infty; A) \to D^*(W_\infty; A) \to D^*(W_\infty; A)/C^*(W \subset W_\infty; A) \to 0,$$

   to obtain $\mathrm{ind}(D_W, g_{\partial W}) \in K_n(C^*(W \subset W_\infty; A)) \cong K_n(C^*(W; A))$. The latter isomorphism is induced by the inclusion $C^*(W; A) \hookrightarrow C^*(W \subset W_\infty; A)$ which just extends the operators by zero. Note that this construction of the index required the invertibility of the operator at the boundary and indeed depends in general on the metric of positive scalar curvature at ∂W.

Note that large scale index theory in the situation we just described produces two invariants which depend on the positive scalar curvature metric on ∂W, namely $\mathrm{ind}(D_W, g_{\partial W}) \in K_n(C^*(W \subset W_\infty; A))$, but also the secondary invariant $\rho(\partial W, g_{\partial W}) \in K_n(D^*(\partial W; A))$. A major result, which we consider a secondary higher Atiyah-Patodi-Singer index theorem, relates these two.

Theorem 8.1. *(compare [23, Theorem 1.22]) Let (W, g_W) be an even dimensional Riemannian spin-manifold with boundary ∂W such that $g_{\partial W}$ has positive scalar curvature. Assume that a group Γ acts isometrically on M. Then*

$$\iota_*(\mathrm{ind}_\Gamma(D_W)) = j_*(\rho(\partial W, g_{\partial W})) \text{ in } K_0(D^*(W)^\Gamma).$$

Here, we use $j\colon D^*(\partial W)^\Gamma \to D^*W^\Gamma$ induced by the inclusion $\partial W \to W$ and $\iota\colon C^*(W)^\Gamma \to D^*(W)^\Gamma$ the inclusion.

Remark 8.2. Above we apply the obvious generalization of the construction of the large scale index of Sections 3 and 8 to an equivariant situation, where subalgebras $D^*(W)^\Gamma$ and $C^*(W)^\Gamma$ generated by invariant operators are used. This works because the Dirac operator then itself is invariant under the group Γ.

Remark 8.3. The heart of the proof of Theorem 8.1 is an explicit secondary index calculation in a model (product) case which is surprisingly intricate.

Remark 8.4. The assertion of Theorem 8.1 should generalize to arbitrary (non-cocompact) spin manifolds, to Hilbert C^*-algebra coefficient bundles and to the K-theory of real C^*-algebras. In a preprint of Zhizhang Xie and Guoliang Yu [35] an argument is sketched which shows how to extend the result to arbitrary dimensions and to real C^*-algebras.

9. Constructions of new classes of metrics of positive scalar curvature

The fundamental idea in the construction of the "geometrically significant" homotopy classes of the space of metrics of positive scalar curvature of Theorem 1.10 is quite old and based on index theory:

Given a closed n-dimensional spin manifold B with $\hat{A}(B) \neq 0$, we know that B does not admit a metric of positive scalar curvature.

Remove an embedded disc from B. The result is a manifold W with boundary $\partial W = S^{n-1}$. Given any metric of positive scalar curvature on W (with product structure near the boundary), the corresponding boundary metric g_1 can not be homotopic to the standard metric on S^{n-1} because then one could glue in the standard disc (with positive scalar curvature) to obtain a metric of positive scalar curvature on B. Now, if g_1 is homotopic to $\psi^* g_{eucl}$ for a non-identity diffeomorphism we can glue the disc back in with ψ to obtain a new manifold B_ψ which is of positive scalar curvature. Note that B_ψ is not necessarily diffeomorphic to B, but using the Alexander trick there is a homeomorphism between B_ψ and B. As the rational Pontryagin classes and therefore the $\hat{A}$-genus are homeomorphism invariant, $\hat{A}(B_\psi) \neq 0$, also B_ψ can not carry a metric of positive scalar curvature.

Observe that exactly the same argument can be applied to a family situation: Let $Y \to S^k$ be a family (i.e. bundle) of manifolds with boundary, with boundary $S^n \times S^k$. Assume there is a family of metrics g_x of positive scalar curvature on Y (product near the boundary). If this family of metrics is homotopic to the constant family consisting of the standard metric (or to a pullback of that one along a family of diffeomorphisms ψ_x) we can glue in $D^{n+1} \times S^k$ to obtain a family of closed manifolds which admits a metric of positive scalar curvature (fiberwise, and then also the total space X admits such a metric). Note that this is only interesting if each g_x is in the component of the standard metric, which we therefore assume.

Alas: if the total space X has non-trivial $\hat{A}$-genus, this is a contradiction (and again, by the homeomorphism invariance and the Alexander trick the argument works modulo diffeomorphism).

Note that this requires two important ingredients:

1. the topological situation with the bundle Y (and X)
2. the geometric input of a family of metrics of positive scalar curvature on Y.

It turns out that already the topological input is surprisingly difficult to get. It means that (after the gluing) we have a fiber bundle $M \to X \to S^k$ of spin manifolds where M does admit a metric of positive scalar curvature, therefore $\hat{A}(M) = 0$, whereas $\hat{A}(X) \neq 0$. Note that this means that the $\hat{A}$-genus is not multiplicative in fiber bundles, even if the base is simply connected (in contrast to the L-genus).

In [12, Theorem 1.4] we use advanced differential topology, in particular surgery theory, Casson's theory of prefibrations and Hatcher's theory of concordance spaces to prove that the required fiber bundles X exist:

Theorem 9.1. *For sufficiently large n, there are $4n$-dimensional smooth closed spin manifolds X with non-vanishing $\hat{A}$-genus fitting into a smooth fiber bundle $F \to X \to S^k$ such that F admits a metric of positive scalar curvature, is highly connected and the bundle contains as subbundle $D^{4n-k} \times S^k$.*

How about the second ingredient, the existence of the family of metrics of positive scalar curvature on $Y = X \setminus D^{4n-k} \times S^k$?

The only tool known which can provide such metrics is the surgery method of Gromov and Lawson. In highly non-trivial work [34] this has been extended by Mark Walsh to families of the kind X as constructed in Theorem 9.1. To apply this, we use the high connectivity and results of Kiyoshi Igusa on Morse theory for fiber bundles [18]. As a consequence we obtain Theorem 1.10.

10. Open problems

The geometry of positive scalar curvature and the development and application of large scale index theory is a vibrant field of research, with a host of important open problems. Many of those were already mentioned above; here we want to highlight them and add a couple of further directions of research.

Gromov-Lawson-Rosenberg conjecture. We should find further obstructions to positive scalar curvature on spin manifolds, in particular for finite fundamental group $(\mathbb{Z}/p\mathbb{Z})^n$ for the so-called toral manifolds. We expect that this will require fundamentally new ideas.

On the other hand, can the class of fundamental groups for which the conjecture holds be described systematically?

Stable Gromov-Lawson-Rosenberg conjecture. The stable Gromov-Lawson-Rosenberg conjecture and its weaker cousin 1.5 which states that "the Rosenberg index sees everything about positive scalar curvature which can be seen using the Dirac operator" follow from the strong Novikov conjecture. It would be spectacular to find counterexamples to either of these

(they are expected to exist, but to find them will of course be very hard). It is necessary to investigate this question further. In this context, the role of the codimension-2 obstruction as discussed in Theorem 1.7 should be understood.

This theorem should extend to the signature operator, which will shed new light on its meaning. Vaguely, we conjecture the following.

Conjecture 10.1. *Let M_1, M_2 be two complete non-compact connected oriented Riemannian manifolds and $f\colon M_1 \to M_2$ a sufficiently well behaved map which is an "oriented homotopy equivalence near infinity".*

Let $E \to M_2$ be a flat bundle of Hilbert A-modules for a C^*-algebra A. Let D_E^{sgn} be the signature operator on M_1 twisted with the flat bundle E, and $D_{f^*E}^{sgn}$ the signature operator on M_2 twisted by f^*E. Then the large scale indices of these two operators should coincide, i.e.

$$f_*(\mathrm{ind}(D_{f^*E}^{sgn})) = \mathrm{ind}(D_E^{sgn}) \in K_*(C^*(M_1; A)).$$

Note that, in this conjecture, one has to work out the precise concept of "sufficiently well behaved" and of "homotopy equivalence at infinity".

Area based large scale geometry. Large scale geometry is based on the metric spaces and distances, viewed from a coarse perspective. Curvature, on the other hand, is a concept based on the bending of surfaces, where scalar curvature looks at the average over all possible surface curvatures through a given point.

This is reflected in the fact that area-enlargeability suffices to obstruct positive scalar curvature (Theorem 1.6). So far, this is not captured well by large scale index theory.

This suggests that a program should be developed for large scale geometry based on 2-dimensional areas instead of lengths. A possible starting point would be to work on a relative of the loop space and carry out the analysis there. This is interesting in its own right, with a host of potential further applications, but seems to require new analytical tools.

Note that the axiomatic abstraction from metric spaces to coarse structures as developed by John Roe [25] does not seem to apply here. Of course, this generalization is interesting in its own right and applications to positive scalar curvature should be developed further.

Coarse Baum-Connes conjecture. If the coarse Baum-Connes conjecture holds for the classifying space of a discrete group, then also the strong Novikov conjecture is true for this group. Moreover, the validity of the coarse Baum-Connes conjecture is a powerful computational tool. On the other hand, there are enigmatic counterexamples due to Guoliang Yu if one drops the "bounded geometry" condition on the space in question.

We expect that many new classes of metric spaces can be found where the coarse Baum-Connes conjecture can be established. But we also feel that the search for counterexamples (of bounded geometry) should be intensified.

Aspherical manifolds. A lot of attention has been given to the special class of aspherical manifolds.

Conjecture 10.2. *Let M be a closed manifold whose universal covering is contractible (i.e. M is* aspherical*). Then M does not admit a metric of positive scalar curvature.*

Often the geometry implies that a manifold is aspherical (e.g. if it admits a metric which is non-positively curved in the sense of comparison geometry). The conjecture states that

in a weak sense this is the only way a manifold can be aspherical. The strong Novikov conjecture for an aspherical spin manifold M implies that $\alpha(M) \neq 0$ because the Dirac operator of a manifold M always represents a non-zero K-homology class in $K_*(M)$, and here $M = B\pi_1(M)$. Of course, we now look for ways to directly use the asphericity in proofs of (special cases) of Conjecture 10.2.

Mapping surgery to analysis to homology. The program to map surgery to analysis has been fully carried out in [23] only for half the dimensions, and only for complex C^*-algebras, based on a delicate explicit index calculation. New developments, in particular the work of Zhizhang Xie and Guoliang Yu [35] extend this to all dimensions with a modified method. It remains to develop the details of this (or an alternative) approach and to carry it over to the more powerful real C^*-algebras.

K-theory of C^*-algebras is a very powerful tool. Most useful, however is its combination with homological tools (in particular Hochschild and cyclic (co)homology). To achieve this systematically and use it for the classification of metrics of positive scalar curvature, we propose a program to not only map the positive scalar curvature sequence to analysis, as described in Section 7, but then to map further to a corresponding long exact sequence of (cyclic) homology groups. There, one would then see primary and secondary numerical invariants of higher index theory.

The primary invariants are well developed. Rather not understood, however, is the theory related to the secondary invariants. Indeed, the relevant algebra D^*M is very large and the usual dense and holomorphically closed subalgebras on which explicit formulas for the Connes-Karoubi Chern character would make sense seem hard to come by. Exactly because of this we feel that the development of such a theory will shed important new light on the power of the secondary invariants of rho-type as proposed here.

Apart from this general program, the theory described above needs to be applied in more concrete situations. The analytic structure set $K_*(D^*M)$, despite its evident potential, so far has only been used in a small number of concrete contexts (compare in particular [15]). This must change before we will have a definite idea of its power.

Minimal surface obstructions to positive scalar curvature. The minimal surface method is the only known tool to obstruct the existence of a metric of positive scalar curvature which works for non-spin manifolds of dimension ≥ 5. So far, it is controversial how to extend the method if the dimension of the manifold in question is larger than 8 (due to singularities which develop in the minimal hypersurfaces one wants to use). Joachim Lohkamp has a program to achieve this.

The minimal surface technique has so far only been used together with the Gauss-Bonnet theorem (via an iterative approach, until the hypersurface is 2-dimensional). Are there other ways to exploit it and combine it with obstructions to positive scalar curvature (Seiberg-Witten, spin Dirac) and what are the relations?

Enlargeability and non-spin manifolds. As just one case of the question whether the known results for spin manifolds carry over to non-spin manifolds consider the following question:

Question 10.3. *Let M be an arbitrary (non-spin) closed n-dimensional manifold with coverings M_ϵ and with maps $f_\epsilon\colon M_\epsilon \to S^n$ which are constant outside a compact subset of M_ϵ, which have non-zero degree and which are ϵ-contracting.*

Can M admit a metric of positive scalar curvature?

Using the minimal surface method, Gromov and Lawson [7, Section 12] have shown that this is not the case if $n \leq 7$.

Topology of the space of metrics of positive scalar curvature. The stable Gromov-Lawson-Rosenberg conjecture shows that there is a stability feature in the topology of the space $\mathrm{Riem}^+(M)$ of metrics of positive scalar curvature: if index theory suggests that it should be non-empty, this might be violated by M itself, but after iterated product with B, eventually it is non-empty. Are there similar stability features concerning the (higher) homotopy groups of $\mathrm{Riem}^+(M)$? It would also be important to develop estimates on the stable range.

The space of metrics of positive scalar curvature and fundamental group. We know well that, for a spin manifold M with a complicated fundamental group Γ, the existence of a metric of positive scalar curvature is not only obstructed by $\hat{A}(M)$, but by $\alpha(M) \in K_*(C^*\Gamma)$, and many elements of $K_*(C^*\Gamma)$ are indeed realized as values of $\alpha(M)$.

Similarly, we should expect that the topology of $\mathrm{Riem}^+(M)$, if non-empty, should be governed by $K_*(C^*\Gamma)$. At the moment, a precise conjecture (e.g. about the homotopy groups) seems too far-fetched. Still, the methods of large scale and higher index theory should be developed to the point that they are available to detect new elements in $\pi_*(\mathrm{Riem}^+(M))$, and one should systematically construct non-trivial examples.

Acknowledgements. Supported by the Courant Research Center "Higher order structures in Mathematics" of Georg-August-Universität Göttingen.

References

[1] M. F. Atiyah and I. M. Singer, *The index of elliptic operators. III*, Ann. of Math. (2), **87**:546–604, 1968.

[2] Paul Baum, Alain Connes, and Nigel Higson, *Classifying space for proper actions and K-theory of group C^*-algebras*, In C^*-algebras: 1943–1993 (San Antonio, TX, 1993), volume 167 of Contemp. Math., pages 240–291. Amer. Math. Soc., Providence, RI, 1994.

[3] Arthur L. Besse, *Einstein manifolds*, volume 10 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*, Springer-Verlag, Berlin, 1987.

[4] Dmitry Bolotov and Alexander Dranishnikov, *On Gromov's conjecture for totally non-spin manifolds*, arXiv:1402.4510.

[5] Diarmuid Crowley and Thomas Schick, *The Gromoll filtration, KO-characteristic classes and metrics of positive scalar curvature*, Geom. Topol., **17**(3):1773–1789, 2013.

[6] William Dwyer, Thomas Schick, and Stephan Stolz, *Remarks on a conjecture of Gromov and Lawson*, In High-dimensional manifold topology, pp. 159–176, World Sci. Publ., River Edge, NJ, 2003.

[7] Mikhael Gromov and H. Blaine Lawson, Jr., *Positive scalar curvature and the Dirac operator on complete Riemannian manifolds*, Inst. Hautes Études Sci. Publ. Math., (58):83–196 (1984), 1983.

[8] Bernhard Hanke, Dieter Kotschick, John Roe, and Thomas Schick, *Coarse topology, enlargeability, and essentialness*, Ann. Sci. Éc. Norm. Supér. (4), **41**(3):471–493, 2008.

[9] Bernhard Hanke, Daniel Pape, and Thomas Schick, *Codimension two index obstructions to positive scalar curvature*, arXiv:1402.4094, 2014.

[10] Bernhard Hanke and Thomas Schick, Enlargeability and index theory, *J. Differential Geom.*, **74**(2):293–320, 2006.

[11] ———, *Enlargeability and index theory: infinite covers*, K-Theory, **38**(1):23–33, 2007.

[12] Bernhard Hanke, Thomas Schick, and Wolfgang Steimle, *The space of metrics of positive scalar curvature*, arXiv:1212.0068, to appear in Publ. Math. IHES.

[13] Nigel Higson and John Roe, *A homotopy invariance theorem in coarse cohomology and K-theory*, Trans. Amer. Math. Soc., **345**(1):347–365, 1994.

[14] ———, *Mapping surgery to analysis. I. Analytic signatures*, K-Theory, **33**(4):277–299, 2005.

[15] ———, *K-homology, assembly and rigidity theorems for relative eta invariants*, *Pure Appl. Math. Q.*, 6(2, Special Issue: In honor of Michael Atiyah and Isadore Singer):555–601, 2010.

[16] Nigel Higson, John Roe, and Guoliang Yu, *A coarse Mayer-Vietoris principle*, Math. Proc. Cambridge Philos. Soc., **114**(1):85–97, 1993.

[17] Nigel Hitchin, *Harmonic spinors*, Advances in Math., **14**:1–55, 1974.

[18] Kiyoshi Igusa, *The stability theorem for smooth pseudoisotopies*, K-Theory, **2**(1-2): vi+355, 1988.

[19] Michael Joachim and Thomas Schick, *Positive and negative results concerning the Gromov-Lawson-Rosenberg conjecture*, In Geometry and topology: Aarhus (1998), volume 258 of Contemp. Math., pp. 213–226. Amer. Math. Soc., Providence, RI, 2000.

[20] E. C. Lance, *Hilbert C^*-modules*, volume 210 of *London Mathematical Society Lecture Note Series*, Cambridge University Press, Cambridge, 1995, A toolkit for operator algebraists.

[21] H. Blaine Lawson, Jr. and Marie-Louise Michelsohn, *Spin geometry*, volume 38 of *Princeton Mathematical Series*, Princeton University Press, Princeton, NJ, 1989.

[22] André Lichnerowicz, *Spineurs harmoniques*, C. R. Acad. Sci. Paris, 257:7–9, 1963.

[23] Paolo Piazza and Thomas Schick, *Rho-classes, index theory and Stolz' positive scalar curvature sequence*, arXiv:1210.6892, to appear in Journal of Topology.

[24] John Roe, *Index theory, coarse geometry, and topology of manifolds*, volume 90 of *CBMS Regional Conference Series in Mathematics*, American Mathematical Society, Providence, RI, 1996.

[25] ———, *Lectures on coarse geometry*, volume 31 of *University Lecture Series*, American Mathematical Society, Providence, RI, 2003.

[26] ——— *Positive curvature, partial vanishing theorems, and coarse indices*, arXiv:1210.6100, 2012.

[27] Jonathan Rosenberg, *C^*-algebras, positive scalar curvature, and the Novikov conjecture*, Inst. Hautes Études Sci. Publ. Math., (58):197–212 (1984), 1983.

[28] Thomas Schick, *A counterexample to the (unstable) Gromov-Lawson-Rosenberg conjecture*, Topology, **37**(6):1165–1168, 1998.

[29] Richard Schoen and Shing-Tung Yau, *On the structure of manifolds with positive scalar curvature*, Manuscripta Math., **28**(1-3):159–183, 1979.

[30] Erwin Schrödinger, *Diracsches Elektron im Schwerefeld. I.*, Sitzungsber. Preuß. Akad. Wiss., Phys.-Math. Kl., 1932:105–128, 1932.

[31] Paul Siegel, *The Mayer-Vietoris sequence for the analytic structure group*, arXiv:1212.0241, 2012.

[32] Stephan Stolz, *Simply connected manifolds of positive scalar curvature*, Ann. of Math. (2), **136**(3):511–540, 1992.

[33] ———, *Manifolds of positive scalar curvature*, In Topology of high-dimensional manifolds, No. 1, 2 (Trieste, 2001), volume 9 of ICTP Lect. Notes, pp. 661–709, Abdus Salam Int. Cent. Theoret. Phys., Trieste, 2002.

[34] Mark Walsh, *Metrics of positive scalar curvature and generalised Morse functions, Part II.*, Trans. Amer. Math. Soc., **366**(1):1–50, 2014.

[35] Zhizhang Xie and Guoliang Yu, *Positive scalar curvature, higher rho-invariants and localization algebras*, arXiv:1302.4418.

Mathematisches Institut, Georg-August-Universität Göttingen, Bunsenstr. 3, 37073 Göttingen, Germany

E-mail: thomas.schick@mathematik.uni-goettingen.de

Gauge theory and mirror symmetry

Constantin Teleman

Abstract. Outlined here is a description of *equivariance* in the world of 2-dimensional extended topological quantum field theories, under a topological action of compact Lie groups. In physics language, I am gauging the theories — coupling them to a principal bundle on the surface world-sheet. I describe the data needed to gauge the theory, as well as the computation of the gauged theory, the result of integrating over all bundles. The relevant theories are 'A-models', such as arise from the Gromov-Witten theory of a symplectic manifold with Hamiltonian group action, and the mathematical description starts with a group action on the generating category (the Fukaya category, in this example) which is factored through the topology of the group. Their mirror description involves holomorphic symplectic manifolds and Lagrangians related to the Langlands dual group. An application recovers the complex mirrors of flag varieties proposed by Rietsch.

Mathematics Subject Classification (2010). Primary 57R56, 55N91; Secondary 18D05, 81T13.

Keywords. Gauge theory, holomorphic symplectic space, Toda system, flag variety.

1. Introduction

This paper tells the story of equivariance, under a compact Lie group, in the higher algebra surrounding *topological quantum field theory* (TQFT). Speaking in riddles, if 2-dimensional TQFT is a higher analogue of cohomology (the reader may think of the Fukaya-Floer theory of a symplectic manifold as refining ordinary cohomology), my story of gauged TQFTs is the analogue of equivariant cohomology. The case of finite groups, well-studied in the literature [30], provides a useful and easy reference point, but the surprising features of the continuous case, such as the appearance of holomorphic symplectic spaces and Langlands duality, are missing there.

From another angle, this is a story of the categorified representation theory of a compact Lie group G, with the provision that representations are *topological*: the G-action (on a linear category) factors through the topology of G. One floor below, where the group acts on vector spaces, these would be not the ordinary complex representations of G, but the local systems of vector spaces on the classifying space BG. There is no distinction for a finite group, but in the connected case, BG is simply connected, and we must pass to the derived category to see anything interesting. The same will hold in the categorified story, where simply connected groups will appear to have trivial representation theory, before deriving. This observation suggests a straightforward homological algebra approach to the investigation, worthy of featuring as an example in a graduate textbook. Pursuing that road, however, leads to faulty predictions, even in the simplest case of pure gauge theory of a

Proceedings of the International Congress of Mathematicians, Seoul, 2014

point (topological Yang-Mills theory). One reason for this failure is a curious predilection of interesting TQFTs to break the obvious $\mathbb{Z}$-grading information present, collapsing it to a $\mathbb{Z}/2$ grading, or encoding it in more labored form (as in the Euler field of Gromov-Witten theory [17]). The result is that homological algebra, which localizes the spectrum of a graded ring to its degree zero part, loses relevant information, which needs restoration by ulterior guesswork. In our example, we will see the homological information in the neighborhood of a Lagrangian within a certain holomorphic symplectic manifold, whereas most of the interesting 'physics' happens elsewhere.

The emerging geometric picture for this categorical topological representation theory is surprisingly attractive. Representations admit a character theory, but characters are now coherent sheaves on a manifold related to the conjugacy classes, instead of functions. The manifold in question, the *BFM space* of the Langlands dual Lie group $G^\vee$, introduced in [5], is closely related to the cotangent bundle to the space of conjugacy classes in the complex group $G^\vee_{\mathbb{C}}$. (For SU_2, it is the Atiyah-Hitchin manifold studied in detail in [4].) Multiplicity spaces of G-invariant maps between linear representations are now replaced by multiplicity categories, whose 'dimensions' are the Hom-spaces in the category of coherent sheaves. (In interesting examples, they are the Frobenius algebras underlying 2-dimensional TQFTs.) There is a preferred family of simple representations, which in a sense exhausts the space of representations: they foliate the BFM space. Every such representation is 'symplectically induced' from a one-dimensional representation of a certain Levi subgroup of G: more precisely, it is the Fukaya category of a flag variety of G. This is formally similar to the Borel-Weil construction of irreducible representations of G by holomorphic induction. Recall that in that world there is another kind of "L^2-induction" from closed subgroups, which is right adjoint to the restriction functor. The counterpart of naïve induction also exists in our world, and gives the (curved) *string topologies* [8] of the same flag varieties, instead of their Fukaya categories.

This story might seem a bit unhinged, were it not for the appearance of the governing structure in the work of Kapustin, Rozansky and Saulina [15]. Studied there are boundary conditions in the 3-dimensional TQFT associated to a holomorphic symplectic manifold X, known as Rozansky-Witten theory [23]. Among those are holomorphic Lagrangian submanifolds of X, or more generally, sheaves of categories over such sub-manifolds. (The full 2-category of all boundary conditions does not yet have a precise definition.) The relation to gauge theory is summarized by the observation that gaugeable 2-dimensional field theories are topological boundary conditions for pure 3-dimensional topological gauge theory. The reader may illustrate this with an easy example: the representations of a finite group F are the boundary conditions for pure F-gauge theory in 2 dimensions; yet these representations are exactly the 1-dimensional topological field theories (vector spaces) which admit F-symmetry. Modulo the [15] description of Rozansky-Witten theory, my entire story is underpinned by the following

Meta-Statement. *Pure topological gauge theory in* 3 *dimensions for a compact Lie group G is equivalent to the Rozansky-Witten theory for the BFM space of the Langlands dual Lie group $G^\vee$.*

I shall offer no elucidation of this, beyond its inspirational value; however, strong indications of this statement have been known in the physics literature, at least for special G [3, 19, 27]. Formulating this statement in a mathematically useable way will require an excursion through much preliminary material in §2-5. A small reward will come in §6, where we

illustrate how these ideas can lead to 'real answers'.

A closing warning is that the results in this paper are partly experimental: enough examples have been checked to rule out plausible alternatives, but I do not claim to know proofs in full generality. In fact, the status of Floer-Fukaya theory makes such claims difficult to sustain, and the author has no special expertise on that topic. In topological cases, such as for string topology (Fukaya theory of cotangent bundles), precise statements and proofs are possible (and easy). More generally, the results apply to the abstract setting of differential graded (or A_∞-categories) with topological G-action, the question being to what extent the Fukaya category of a symplectic manifold with Hamiltonian G-action qualifies. (For non-compact manifolds, this depends on the 'wrapping' condition at ∞.) If nothing else, the paper can be read as a template for what a nice world should look like.

2. Topological field theory

Topological field theory, introduced originally by Atiyah[2], Segal [25] and Witten [32], promised to systematize a slew of new 3-manifold invariants. The invariants of a 3-manifold M are thought to arise from *path integrals* over a space of maps from M to a target X. The latter is often a manifold, but in interesting cases, related to *gauge theory*, it is a stack. One example relevant for us will have X a holomorphic symplectic manifold, leading to *Rozansky-Witten theory* [23]. The 2-dimensional version of this notion quickly found application to the counting of holomorphic curves, the Gromov-Witten invariants of a symplectic manifold X: these are controlled by a family of TQFTs parametrized by the even cohomology space $H^{ev}(X)$.

2.1. Extended TQFTs. Both theories above have a bearing on my story, once they are *extended down to points*. In the original definition, a d-dimensional TQFT is a symmetric, strongly monoidal functor form the category whose objects are closed $(d-1)$-manifolds and whose morphisms are compact d-bordisms, to the category $\mathfrak{Vect}$ of complex finite-dimensional vector spaces; the monoidal structures are disjoint union and tensor product, respectively. (Some tangential structure on manifolds is chosen, as part of the starting datum.) Fully extending the theory means extending this functor to one from the *bordism d-category* $Bord_d$, whose objects are points and whose k-morphisms are compact k-manifolds with corners (and some tangential structure), to some de-looping of the category of vector spaces: a symmetric monoidal d-category whose top three layers are complex numbers, vector spaces and linear categories, or a differential graded (dg) version of this. When $d = 2$, which most concerns us, the target is usually the 2-category $\mathfrak{LCat}$ of linear dg categories, linear functors and natural transformations. The reader may consult Lurie [16], references therein and the wide following it inspired, for a precise setting of higher categories.

Example 2.1 (2-dimensional gauge theory with finite gauge group F). This theory is defined for unoriented manifolds; among others, the functor Z_F which sends a point $*$ to the category $\mathfrak{Rep}(F)$ of (finite-dimensional) linear representations of F, the half-circle bordism $\subset\colon \emptyset \to \{*, *'\}$ to the functor $\mathfrak{Vect} \to \mathfrak{Rep}(F) \otimes \mathfrak{Rep}(F)$ sending $\mathbb{C}$ to the (2-sided) regular representation of F, the opposite bordism $\supset\colon \{*, *'\} \to \emptyset$ to the functor $\mathfrak{Rep}(F) \otimes \mathfrak{Rep}(F) \to \mathfrak{Vect}$ sending $V \otimes W$ to the subspace of F-invariants therein. A closed surface gives a number, which is the (weighted) count of principal F-bundles. See

for instance [9] for a uniform construction of the complete functor and generalizations.

The first theorem of [16] is that an such extended TQFT $Z : Bord_d \to ??$ is determined by its value $Z(+)$ on the point, at least in the setting of *framed* manifolds. The object $Z(+)$, which we call the generator of Z, must satisfy some strong (*full dualizability*) conditions, but carries no additional structure, beyond being a member of an ambient d-category.

On the other hand, the ability to pass to surfaces with less structure than a framing on their tangent bundle forces additional structure on the generator $Z(+)$. The point (conceived together with an ambient germ of d-manifold) carries a d-framing, on which the group $\mathrm{O}(d)$ acts. Lurie's second theorem states that, given a tangential structure, encoded in a homomorphism $G \to \mathrm{O}(d)$, factoring the theory Z from $Bord_d$ through the category $Bord_d^G$ of d-folds with G-structure is equivalent to exhibiting $Z(+)$ as a fixed-point for the G-action on the image of TQFTs in the target d-category (more precisely, the sub-groupoid of fully dualizable objects and invertible morphisms).

The best-known case of oriented surfaces, when $G = \mathrm{SO}(2)$, requires a *Calabi-Yau* structure on $Z(+)$. This can be variously phrased: as a trivialization of the *Serre functor*, which is an automorphism of any fully dualizable linear dg category (see Remark 2.3 below); alternatively, as a linear functional on the cyclic homology of $Z(+)$ whose restriction to Hochschild homology $HH_*(Z(+))$ induces a perfect pairing on Hom spaces:

$$\mathrm{Hom}(x,y) \otimes \mathrm{Hom}(y,x) \to \mathrm{Hom}(x,x) \to HH_* \to \mathbb{C}.$$

This case of Lurie's theorem recovers earlier results of Costello, Kontsevich and Hopkins-Lurie [6, 13].

The Hochschild homology $HH_*(Z(+))$ is meaningful in a different guise: it is the space of states $Z(S^1)$ of the theory, for the circle with the radial framing. The circle is pictured here with a germ of surrounding surface, and therefore carries a $\mathbb{Z}$'s worth of framings, detected by a winding number. The Hochschild cohomology HH^* goes with the blackboard framing, and the space for the framing with winding number n is HH^* of the nth power of the Serre functor. (Of course, for oriented theories there is no framing dependence, and these spaces agree.)

2.2. Topological group actions. An important point is that the action of $\mathrm{O}(2)$ (and thus G) on the target category $Z(+)$ is *topological*, or factored through its topology. There are several ways to formulate this constraint, which is vacuous when G is discrete. The favored formulation will depend on the nature of the target category; in the linear case, and when G is connected, we will provisionally settle for the one in Theorem 2.5 below. Combined with Statement 2.9 below, this generalizes an old result of Seidel [26] on Hamiltonian diffeomorphism groups.

Here are some alternative definitions:

1. We can ask for a *local trivialization* of the action in a contractible neighborhood of $1 \in G$, an isomorphism with the trivial action of that same neighborhood (up to coherent homotopies of all orders).
2. Using the action to form a bundle of categories with fiber $Z(+)$ over the classifying stack BG, we ask for an integrable flat connection on the resulting bundle of categories. (Formulating the flatness condition requires some care, in light of the fiberwise automorphisms.)

3. Exploiting the contractibility of the group P_1G of paths starting at $1 \in G$, we can ask for a trivialization of the lifted P_1G-action.

   Now, the action of the based loop group ΩG (kernel of $P_1G \to G$) is already trivial (being factored through $1 \in G$), and the difference of trivializations defines a (topological) representation of ΩG by automorphisms of the identity functor in $Z(+)$.

   The group ΩG has an E_2 structure, seen from its equivalence with the second loop space $\Omega^2 BG$; and the representation on $\mathrm{Id}_{Z(+)}$ is the 2-holonomy, over spheres, of the flat connection in #2. Importantly, it is an E_2 representation.

Remark 2.2. When G is connected, description #3 above captures all the information for the action (up to contractible choices), because the space of trivializations of a trivial topological action of P_1G is contractible.

Example 2.3. A topological action of the circle on a category is given by a group homomorphism from $\mathbb{Z} = \pi_1 S^1 = \pi_0 \Omega S^1$ to the automorphisms of the identity: equivalently, a central (in the category) automorphism of each object. Because there is no higher topology in S^1, this also works when the target is a 2-category, such as the (sub-groupoid of fully dualizable objects in the) 2-category $\mathfrak{2Cat}$. The structural $\mathrm{SO}(2) \subset \mathrm{O}(2)$ action gives an automorphism of each category: this is the Serre functor.

Example 2.4. Endomorphisms of the identity in the linear category $\mathfrak{Vect}$ are the complex scalars, so that linear topological representations of a connected G on $\mathfrak{Vect}$ are 1-dimensional representations of $\pi_0 \Omega G \cong \pi_1 G$. These are the points in the center of the complexified Langlands dual group $G^\vee_{\mathbb{C}}$.

Recall that the endomorphisms of the identity in a category (the center) form the 0^{th} Hochschild cohomology. To generalize the above example to the derived world, we should include the entire Hochschild cochain complex.

Theorem 2.5. *Topological actions of a connected group G on a linear dg-category $\mathfrak{C}$ are captured (up to contractible choices) by the induced E_2 algebra homomorphism from the chains $C_*\Omega G$, with Pontrjagin product, to the Hochschild cochains of $\mathfrak{C}$.* □

Example 2.6. From a continuous action of G on a space X, we get a locally trivial action on the cochains C^*X. Indeed, we get an action of ΩG on the free loop space LX of X. The action is fiber-wise with respect to the bundle $\Omega X \to LX \to X$. Let $C^*\big(X; C_*\tilde{\Omega}X\big)$ be the cochain complex on X with coefficients in the fiber-wise chains for this bundle. With the fiber-wise Pontrjagin product, this is a model for the Hochschild cochains of the algebra $C^*(X)$, and the action of ΩG exhibits the E_2 homomorphism in the theorem.

Remark 2.7. The "E_2" in the statement is not jus a commutativity constraint, but can contain (infinite amounts of!) data; see Lesson 3.2.5.

Remark 2.8. One floor below, for 1-dimensional field theories, the category $Z(+)$ is replaced with a vector space (or a complex), and we recognize #2 above as defining a topological representation of G. The datum in Theorem 2.5 is replaced by an (E_1) algebra homomorphism from the chains C_*G, with Pontrjagin product, to $\mathrm{End}(Z(+))$; there is no connectivity assumption. Climbing to the higher ground of n-categories, we can extract an E_{n+1}-algebra homomorphism from $C_*\Omega^n G$ to the E_n Hochschild cohomology; but this misses the information from the homotopy of G below n.

The following key example captures the relevance of my story to real mathematics. (In fact, it contains *all* examples I know for topological group actions!)

Conjecture 2.9. *Let G act in Hamiltonian fashion action on a symplectic manifold X. Then, G acts topologically on the Fukaya category of X.*

Proof. A Hamiltonian action of G on X defines, in the category of symplectic manifolds and Lagrangian correspondences, an action of the group object T^*G.[1] This makes the Fukaya category of X into a module category over the wrapped Fukaya category $\mathfrak{WF}(T^*G)$. A theorem of Abouzaid [1] identifies the latter with that of $C_*\Omega G$-modules. The tensor structure is identified with the E_2 structure of the Pontrjagin product, by detecting it on generators of the category (the cotangent fibers). The resulting structure is equivalent to the datum in Theorem 2.5. □

Remark 2.10. It may seem strange to state a conjecture and then provide a proof. However, the reader will detect certain assumptions which have not been clearly stated in the conjecture: mainly, functoriality of Fukaya categories under Lagrangian correspondences. If X is non-compact, equivariance of the wrapping condition at ∞ is essential; the statement fails for the *infinitesimally wrapped* Fukaya category of Nadler and Zaslow [20], see below. (Another outline argument is more tightly connected to holomorphic disks and $G_{\mathbb{C}}$-bundles, but that relies on details of the construction of the Fukaya category.)

Remark 2.11. A closely related notion to the one discussed, but distinct from it, is that of an *infinitesimally trivialized* Lie group action. Here, we ask for the action to be differentiable, and the restricted action to the formal group $\hat{G}$ (equivalently, the Lie algebra $\mathfrak{g}$) should be homologically trivialized. An example is furnished by an action of G on a manifold X and the induced action on the algebra $\mathcal{D}(X)$ of differential operators: the Lie action of $\mathfrak{g}$ is trivialized in the sense that it is inner, realized by the natural Lie homomorphism from $\mathfrak{g}$ to the 1st order differential operators. Theorem 2.5 does *not* usually apply to such situations. With respect to the alternative definition #2 above, the relevant distinction is between *flat* and *integrable* connections over BG.

2.3. Gauging a topological theory. Given a guantum field theory and a (compact Lie) group G, physicists normally produce a G-gauged theory in two stages. The theory is first coupled to a 'classical gauge background', a principal G-bundle. (No connection is needed in the case of topological actions.[2]) Then, we 'integrate over all principal bundles' to quantize the gauge theory.

These two distinct stages are neatly spelt out in the setting of extended TQFTs. Lurie's theory already captures the first stage of gauging. Namely, we convert the principal G-bundle into a tangential structure by choosing the trivial homomorphism $G \to \mathrm{O}(2)$. (Of course, we may add any desired tangential structure, such as orientability, by switching to $G \times \mathrm{SO}(2) \to \mathrm{O}(2)$, by projection.) Making $Z(+)$ into a fixed point for the trivial G-action means defining a (topological) G-action on $Z(+)$. This is the input datum for a classically gauged theory.

Quantizing the gauge theory, or integrating over principal G-bundles, is tricky. It is straightforward for finite groups: integration of numbers is a weighted sum, and integration

[1] The moment map $\mu : X \to \mathfrak{g}^*$ appears in the requisite Lagrangian, $\{(g, \mu(gx), x, gx)\} \subset T^*G \times (-X) \times X$.
[2] Flat connections would be needed when G action does not factor through topology, as in B-model theories.

of vector spaces and categories is a finite limit or colimit. (The duality constraints require the limits and colimts to agree; working in characteristic 0 ensures that [9].) For Lie groups G, integration of the numbers requires a fundamental class on the moduli of principal bundles. For instance, the symplectic volume form is relevant to topological Yang-mills theory. A limited K-theoretic fundamental class was defined in [31], and cohomological classes, such as the one relevant to topological Yang-Mills theory, can be extracted from it. But this matter seems worthy of more subtle discussion than space allows here.

In fact, the gauge theory *cannot* always be fully quantized. The generating object for the quantum gauge theory is the invariant category $Z(+)^G$, which agrees with the co-invariant category $Z(+)_G$ under mild assumptions. In the framework of Theorem 2.5, we compute the generator $Z(+)_G$ as a tensor product

$$Z(+)_G = Z(+) \otimes_{C_*\Omega G} \mathfrak{Vect} \tag{2.1}$$

with the trivial representation. The 1-dimensional part of the field theory, and sometimes part of the surface operations, are well-defined; but the complete surface-level operations often fail to be defined. Thus, for the trivial $2D$ theory, $Z(+) = \mathrm{dg}{-}\mathfrak{Vect}$ with trivial G-action, and the fixed-points are local systems over BG. This generates a partially defined $2D$ theory, a version of string topology for the space BG. The space associated to the circle is the equivariant cohomology $H^*_G(G)$ for the conjugation action, and the theory is defined the subcategory of $Bord_2$ where all surfaces (top morphisms) have non-empty output boundaries for each component.

This example can be made more interesting by noting that the trivial action of G on $\mathrm{dg}{-}\mathfrak{Vect}$ has interesting topological deformations, in the $\mathbb{Z}/2$-graded world; the notable one comes from the quadratic Casimir in $H^4(BG)$, and gives topological Yang-Mills theory with gauge group G. When G is semi-simple, this theory is almost completely defined, and the invariants of a closed surface (of genus 2 or more) are the symplectic volumes of the moduli spaces of flat connections. (Further deformations exits, by the entire even cohomology of BG and relate to more general integrals over those spaces.) These should be regarded as twisted Gromov-Witten theories with target space BG. A starting point of the present work was the abject failure of the homological calculation (2.1) in these examples: for topological Yang-Mills theory, (2.1) gives the zero answer when G is simple.

2.4. The space of states. The space(s) of states of the gauged theory are well-defined, independently of good behavior of the fixed-point category $Z(+)^G$. More precisely, each $g \in G$ gives an autofunctor g_* of the category. The Hochshild cochain complexes $HCH^*(g_*; Z(+))$ assemble to a (derived) local system $\mathcal{H}(Z(+))$ over the group G, which is equivariant for the conjugation action, and the space of states for the (blackboard framed) circle in the gauge theory is the equivariant homology $H^G_*(G; \mathcal{H})$. It has a natural E_2 multiplication, using the Pontrjagin product in the group. When $Z(+) = \mathfrak{Vect}$, with the trivial G-action, we recover the string topology space $H^G_*(G)$ of BG by exploiting Poincaré duality on G.[3]

[3]The last space goes with the radially framed circle.

3. The 2-category of Kapustin-Rozansky-Saulina

As the image of the point, an object in the 3-dimensional bordism 3-category, Lurie's generator for pure 3-dimensional gauge theory should have categorical depth 2. My proposal for this generator is a 2-category associated to a certain holomorphic symplectic manifold, to be described in §5.

Fortunately, the existence of the requisite 2-category has already been conjectured, and a proposal for its construction has been outlined in [14, 15]. When X is compact, this 2-category should generate the Rozansky-Witten theory [23] of X. In particular, its Hochschild cohomology, which on general grounds is a 1-category with a braided tensor structure, should be (a dg refinement of) the derived category of coherent sheaves on X described in [24]. Just like Rozansky-Witten theory, the narrative takes place in a differential graded world, and in applications, the integer grading must be collapsed $\mathrm{mod}\ 2$ (the symplectic form needs to have degree 2, if the integral grading is to be kept). To keep the language simple, I will use 'sheaf' for 'complex of sheaves' and write $\mathfrak{Coh}$ for a differential graded version of the category of coherent sheaves, etc.

Remark 3.1. The 2-category may at first appear analogous to the deformation quantization of the symplectic manifold; but that is not so. That analogue — a double categorification — is $\mathfrak{Coh}(X)$ with its braided tensor structure. The category [15] is a 'square root' of that, and I will denote it $\sqrt{\mathfrak{Coh}}(X)$ or $KRS(X)$.

3.1. Simplified description. The following partial description of the KRS 2-category applies to a Stein manifold X, when deformations coming from coherent cohomology vanish.[4] In our example, X will be affine algebraic. Among objects of $\sqrt{\mathfrak{Coh}}(X)$ are smooth holomorphic Lagrangians $L \subset X$; more general objects are coherent sheaves of $\mathcal{O}_L$-linear categories on such L. (The object L itself stands for its dg category $\mathfrak{Coh}(L)$ of coherent sheaves, a generator for the above.) To make this even more precisee, $\sqrt{\mathfrak{Coh}}(X)$ is the sheaf of global sections of a coherent sheaf of $\mathcal{O}_X$-linear 2-categories, whose localization at any smooth L as above is equivalent the 2-category of module categories over the sheaf of tensor categories $(\mathfrak{Coh}(L), \otimes)$ on L; with a bit of faith, this pins down $\sqrt{\mathfrak{Coh}}(X)$, as follows.

For two Lagrangians $L, L' \in X$, $\mathrm{Hom}(L, L')$ will be a sheaf of categories supported on $L \cap L'$, and a $(\mathfrak{Coh}(L), \otimes) - (\mathfrak{Coh}(L'), \otimes)$ bi-module. Localizing at L, we choose a (formal) neighborhood identified symplectically with T^*L, so that we regard (locally) L' as the graph of a differential $d\Psi$, for a *potential* function $\Psi : L \to \mathbb{C}$. Locally where this identification is valid, $\mathrm{Hom}(L, L')$ becomes equivalent to the *matrix factorization* category $MF(L, \Psi)$. (See for instance [21] for a definition of the latter.)

3.2. Lessons. Several insights emerge from this important notion.

1. A familiar actor in mirror symmetry, a complex manifold L with potential Ψ, is really the object in $\sqrt{\mathfrak{Coh}}(T^*L)$ represented by the graph $\Gamma(d\Psi)$, masquerading as a more traditional geometric object. The matrix factorization category $MF(L, \Psi)$ is its Hom with the zero-section. This resolves the contradiction in which the restriction of the category $MF(L, \Psi)$ to a sub-manifold $M \subset L$ is commonly taken to be the matrix factorization category of $\Psi|_M$. That is clearly false in the 2-category of $(\mathfrak{Coh}(L), \otimes)$-

[4] My discussion is faulty in another way, failing to incorporate the Spin structures, which must be carried by the Lagrangians. I am grateful to D. Joyce for flagging their role.

module categories (the result of localizing to the zero-section $L \subset T^*L$). For instance, if the critical locus of Ψ does not meet M, Hom computed in $(\mathfrak{Coh}(L), \otimes)$-modules gives zero. Instead, M must be replaced by the object represented by its co-normal bundle in $\sqrt{\mathfrak{Coh}}(T^*L)$, whose Hom there with $\Gamma(d\Psi)$ computes precisely $MF(M, \Psi|_M)$.

2. The well-defined assignment sends $(\mathfrak{Coh}(L), \otimes)$-module categories to sheaves of categories with Lagrangian support in the cotangent bundle $\hat{T}^*L$, completed at the zero-section. Namely, the Hochschild cohomology of such a category $\mathfrak{K}$ is (locally on L) an E_2-algebra over the second (E_2) Hochschild cohomology of $(\mathfrak{Coh}(L), \otimes)$, which is an E_3 algebra. The spectrum of the latter is $\hat{T}^*L$, with E_3 structure given by the standard symplectic form. This turns $\operatorname{Spec} HH^*(\mathfrak{K})$ into a coherent sheaf with coisotropic support in $\hat{T}^*L$, and $\mathfrak{K}$ sheafifies over it. The *Lagrangian* condition is clearly related to a finiteness constraint, but this certainly shows the need to include singular Lagrangians in the KRS 2-category.
3. The deformation of a $(\mathfrak{Coh}(L), \otimes)$-module category $\mathfrak{M}$ by the addition of a potential ('curving') $\Psi \in \mathcal{O}(L)$ shifts the support of $\mathfrak{M}$ vertically by $d\Psi$ in T^*L. This allows one to move from formal to analytic neighborhoods of L, if the deformation theory under curvings is well-understood. For instance, one can compute the Hom between two objects that do not intersect the zero-section — such as two potentials without critical points — by drawing their intersection into L: $\operatorname{Hom}(\Gamma(d\Phi), \Gamma(d\Psi)) = MF(L, \Psi - \Phi)$.
4. More generally, Hamiltonian vector fields on $\hat{T}^*L$ give the derivations of $\sqrt{\mathfrak{Coh}}(\hat{T}^*L)$ defined from its E_2 Hochschild cohomology. Hamiltonians vanishing on the zero-section preserve the latter, and give first-order automorphisms of $(\mathfrak{Coh}(L), \otimes)$.
5. The KRS picture captures in geometric terms sophisticated algebraic information. For example, the category $\mathfrak{Vect}$ can be given a $(\mathfrak{Coh}(L), \otimes)$-module structure in many more ways in the $\mathbb{Z}/2$ graded world: any potential Ψ with a single, Morse critical point will accomplish that. The location of the critical point $p \in L$ misses an infinite amount of information, which is captured precisely by the graph of $d\Psi$; this is equivalent to an E_2 structure on the evaluation homomorphism $\mathcal{O}_L \to \mathbb{C}_p$ at the residue field (cf. Theorem 2.5).

Parts of this story can be made rigorous at the level of formal deformation theory, see for instance [10], and of course the outline in [14]. Lesson 3 also offers a working definition of the 2-category $\sqrt{\mathfrak{Coh}}(T^*L)$ as that of $(\mathfrak{Coh}(L), \otimes)$-modules, together with all their deformations by curvings. On a general symplectic manifold X, we can hope to patch the local definitions from here.[5] It is not my purpose to supply a construction of $\sqrt{\mathfrak{Coh}}(X)$ here — indeed, that is an important open question — but rather, to indicate enough structure to explain my answer to the mirror of (non-abelian) gauge theory. I believe that one important reason why that particular question has been troublesome is that the mirror holomorphic symplectic manifold, the BFM space of §5, *not* quite a cotangent bundle, so the usual description in terms of complex manifolds with potentials is inadequate.

Remark 3.2. If $X = T^*L$ for a manifold L, and we insist on integer, rather than $\mathbb{Z}/2$-gradings, then the cotangent fibers have degree 2 and all structure in the KRS category

[5] If X is not Stein, deformations will be imposed upon this story by coherent cohomology.

is invariant under the scaling action on T^*L. In that case, we are dealing precisely with $(\mathfrak{Coh}(L), \otimes)$-modules.

3.3. Boundary conditions and domain walls. The Hom category $\mathrm{Hom}(L, L')$ for two Lagrangians $L, L' \subset X$ with finite intersection supplies a 2-dimensional topological field theory for framed surfaces; this follows form its local description by matrix factorizations. Since X itself aims to define a $3D$ (Rozansky-Witten) theory and each of L, L' is a boundary condition for it, one should picture a sandwich of Rozansky-Witten filling between a bottom slice of L and a top one of L'. The formal description is that $L, L' : \mathrm{Id} \to RW_X$ are morphisms from the trivial $3D$ theory Id to Rozansky-Witten theory RW_X, viewed as functors from $Bord_2$ to the 3-category of linear 2-categories, and the category $\mathrm{Hom}(L, L')$ of natural transformations between these morphisms is the generator for this sandwich theory. Geometrically, it is represented by the interval, with RW_X in the bulk and L, L' at the ends, and is also known as the *compactification* of RW_X along the interval, with the named boundary conditions.

Factoring this theory through *oriented* surfaces requires a trace on the Hochschild homology HH_* (cf. §2.1). Now, the canonical description of the only non-zero group, $HH_{\dim L}$, turns out to involve the Spin square roots[6] of the canonical bundles ω, ω' of L, L' on their scheme-theoretic overlap:

$$HH_{\dim L}\,\mathrm{Hom}(L, L') \cong \Gamma\big(L \cap L'; (\omega \otimes \omega')^{1/2}\big). \tag{3.1}$$

A non-degenerate quadratic form on $HH_{\dim L}$ comes from the Grothendieck residue (and the symplectic volume on X). A non-degenerate trace on HH_* will thus be defined by choosing non-vanishing sections of $\omega^{1/2}, \omega'^{1/2}$ on L, L'.

Remark 3.3. A generalization of the notion of boundary condition is that of a *domain wall* between TQFTs. This is an adjoint pair of functors between the TQFTs meeting certain (dualizability) conditions, see [16], §4. A boundary condition is a domain wall with the trivial TQFT. Just as a holomorphic Lagrangian in X can be expected to define a boundary condition for RW_X, a holomorphic Lagrangian correspondence $X \leftarrow C \to Y$ should define a domain wall between RW_X and RW_Y. We shalll use these in §5 and §6, in comparing gauge theories for different groups.

4. The mirror of abelian gauge theory

This interlude recalls the mirror story of torus gauge theory; except for the difficulty mentioned in Lesson 1 of §3.2, this story is well understood and can be phrased as a categorified Fourier-Mukai transform. In fact, in this case we can indicate the other mirror transformation, from the gauged B-model to a family of A-models.

4.1. The $\mathbb{Z}$-graded story. We will need to correct this when abandoning $\mathbb{Z}$-gradings, in light of the wisdom of the previous section; nevertheless the following picture is nearly right.

[6] The *cohomology* is easy to pin down canonically, as the functions on $L \cap L'$.

Proposition 4.1.

(i) *Topological actions of the torus T on the category $\mathfrak{Vect}$ are classified by points in the complexified dual torus $T^\vee_{\mathbb{C}}$.*

(ii) *A topological action of T on a linear category $\mathfrak{C}$ is equivalent to a quasi-coherent sheafification of $\mathfrak{C}$ over $T^\vee_{\mathbb{C}}$.*

Proof. Both statements follow from Theorem 2.5, considering that the group ring $C_*(\Omega T)$ is quasi-isomorphic to the ring of algebraic functions on $T^\vee_{\mathbb{C}}$, and that a category naturally sheafifies over its center, the zeroth Hochschild cohomology. □

There emerges the following 0^{th} order approximation to abelian gauged mirror symmetry: if X is a symplectic manifold with Hamiltonian action of T, and $X^\vee$ is a mirror of X — in the sense that $\mathfrak{Coh}(X^\vee)$ is equivalent to the Fukaya category $\mathfrak{F}(X)$ — then the group action on X is mirrored into a holomorphic map $\pi : X^\vee \to T^\vee_{\mathbb{C}}$. This picture could be readily extracted from Seidel's result, [26].

Proposition 4.1 interprets the mirror map $X^\vee \to T^\vee_{\mathbb{C}}$ as a *spectral decomposition* of the category $\mathfrak{F}(X)$ into irreducibles $\mathfrak{Vect}_\tau$. One of the motivating conjectures of this program gives a geometric interpretation of this spectral decomposition, in terms of the original manifold X and the moment map $\mu : X \to \mathfrak{t}^*$.

Conjecture 4.2 (Torus symplectic quotients). *The multiplicity of $\mathfrak{Vect}_\tau$ in $\mathfrak{F}(X)$ is the Fukaya category of the symplectic reduction of X at the point $\operatorname{Re}\log\tau \in \mathfrak{t}^*$, with imaginary curving (B-field) $\operatorname{Im}\log\tau$.*

Remark 4.3. This is, for now, meaningless over singular values of the moment map, where there seems to be no candidate definition for the Fukaya category of the quotient.

Remark 4.4. The conjecture relies on using the *unitary mirror* of X, constructed from Lagrangians with unitary local systems. Otherwise, in the toric case, the algebraic mirror $X^\vee$ is $T^\vee_{\mathbb{C}}$, obviously having a point fiber over every point in $T^\vee_{\mathbb{C}}$; yet the symplectic reduction is empty for values outside the moment polytope. That polytope is precisely the cut-off prescribed for the mirror by unitarity.

Example 4.5 (Toric varieties). The following construction of mirrors for toric manifolds, going back to the work of Givental and Hori-Vafa, illustrates both the conjecture and the need to correct the picture by moving to the KRS category.

Start with the mirror of $X = \mathbb{C}^N$, with standard symplectic form, as the space $T^\vee_{\mathbb{C}} := (\mathbb{C}^\times)^N$ with potential $\Psi = z_1 + \dots z_N$. Here, $T^\vee$ is the dual of the diagonal torus acting on X, and the mirror map $X^\vee \to T^\vee_{\mathbb{C}}$ is the identity.[7] For a sub-torus $i : K \hookrightarrow T$, the mirror of the symplectic reduction $X_q := \mathbb{C}^N /\!/_q K$ at $q \in \mathfrak{k}^*$ is the (torus) fiber $X^\vee_q$ of the dual surjection $i^\vee : T^\vee_{\mathbb{C}} \twoheadrightarrow K^\vee_{\mathbb{C}}$, with restricted super-potential Ψ. The parameter q lives in the small quantum cohomology of X. We see here the familiar, but faulty restriction to the fiber of the matrix factorization category $MF(T^\vee_{\mathbb{C}}, \Psi)$ of Lesson 3.2, #1. The problem is glaring, because the original MF category is null.

The mirror $X^\vee_q$ projects isomorphically to the kernel $S^\vee_{\mathbb{C}}$ of $i^\vee$; this is the map π mirror to the action of $S = T/K$ on X.

[7] This is readily obtained from the SYZ picture, using coordinate tori as Lagrangians; the unitary mirror is cut off by $|z_k| < 1$.

4.2. Fourier transform. As can be expected in the abelian case, the spectral decomposition of Proposition 4.1 is formally given by a Fourier transform. Specifically, there is a 'categorical Poincaré line bundle'

$$\mathfrak{P} \to BT_{\mathbb{C}} \times T^{\vee}_{\mathbb{C}},$$

with an integrable flat connection along BT. (Of course, $\mathfrak{P}$ is the universal one-dimensional topological representation of T, and its fiber over $\tau \in T^{\vee}_{\mathbb{C}}$ is $\mathfrak{Vect}_{\tau}$.) Given a category $\mathfrak{C}$ with topological T-action, we form the bundle $\mathrm{Hom}(\mathfrak{P}, \mathfrak{C})$ and integrate along $BT_{\mathbb{C}}$ to obtain the spectral decomposition of $\mathfrak{C}$ laid out over $T^{\vee}_{\mathbb{C}}$.

Remark 4.6 (B to A). The interest in this observation stems from a related Fourier transformation, giving a "B to A" mirror symmetry. There is another Poincaré bundle $\mathfrak{Q} \to BT_{\mathbb{C}} \times T^{\vee}_{\mathbb{C}}$, with flat structure this time along $T^{\vee}_{\mathbb{C}}$. It may help to exploit flatness and descend to $B(T_{\mathbb{C}} \times \pi_1(T)^{\vee})$, in which case $\mathfrak{Q}$ is the line $\mathfrak{Vect}$ with action of the group $T \times \pi_1(T)^{\vee}$, defined by the Heisenberg $\mathbb{C}^{\times}$-central extension. (The extension is a multiplicative assignment of a line to every group element, and the action on $\mathfrak{Vect}$ tensors by that line.)

Fourier transform converts a category $\mathfrak{C}$ with (non-topological!) T-action into a local system $\tilde{\mathfrak{C}}$ of categories over $T^{\vee}_{\mathbb{C}}$. The fiber of $\tilde{\mathfrak{C}}$ over 1 is the fixed-point category $\mathfrak{C}^T$, and the monodromy action of $\pi_1(T^{\vee})$ comes from the natural action thereon of the category $\mathfrak{Rep}(T)$ of complex T-representations. For example, when $\mathfrak{C} = \mathfrak{Coh}(X)$, the (dg) category of coherent sheaves on a complex manifold with holomorphic T-action, $\mathfrak{C}^T$ is, almost by definition, the category of sheaves on the quotient stack $X/T_{\mathbb{C}}$. The analogue of Conjecture 4.2 is completely obvious here.

I do not know a non-abelian analogue of this "B to A" story.

4.3. The $\mathbb{Z}/2$-graded story. In light of Lesson 3.2.1 and Example 4.5, the only change needed to reach the true story is to replace the $(\mathfrak{Coh}(T^{\vee}_{\mathbb{C}}), \otimes)$-module category $\mathfrak{Coh}(X^{\vee})$, determined from $\pi : X^{\vee} \to T^{\vee}_{\mathbb{C}}$, by an object in the KRS category of $T^*T^{\vee}_{\mathbb{C}}$: the category with T-action sees precisely the germ of a KRS object near the zero-section.

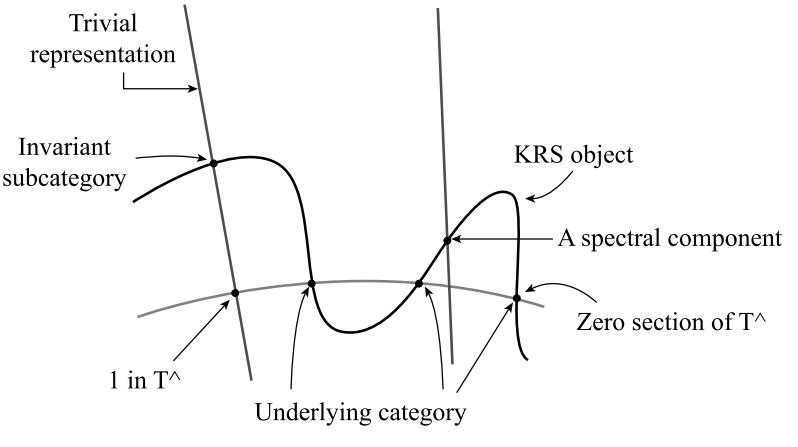


Figure 4.1. Pictorial representation of $\sqrt{\mathfrak{Coh}}(T^*T^{\vee}_{\mathbb{C}})$

This enhancement of information relies upon knowing not just the Fukaya category $\mathfrak{F}(X)$ with its torus action, but all of its curvings with respect to functions lifted from the mirror map $\pi : X^\vee \to T^\vee_{\mathbb{C}}$. However, we can expect in examples that a meaningful geometric construction of the mirror would carry that information. For instance, in Example 4.5, we replace $(X^\vee_q, \Psi|_{X^\vee_q})$ and its map to $S^\vee_{\mathbb{C}}$ by the graph of $d\Psi|_{X^\vee_q}$ in $T^*S^\vee_{\mathbb{C}}$; this is the result of intersecting the graph of $d\Psi$ with the cotangent space at $q \in K^\vee_{\mathbb{C}}$.

Figure 4.1 attempts to capture the distinction between $(\mathfrak{Coh}T^\vee_{\mathbb{C}}, \otimes)$-modules and their KRS enhancement. The squiggly line stands for (the support of) a general object; its germ at the zero-section is the underlying category, with topological T-action. In that sense, the zero-section represents the regular representation of T (its Hom category with any object recovers the underlying category.) The invariant category is the intercept with the trivial representation, the cotangent space at $1 \in T^\vee_{\mathbb{C}}$; other spectral components are intercepts with vertical axes. We see that the invariant subcategory is computed 'far' from the underlying category, and a homological calculation centered at the zero-section will fail.

5. The non-abelian mirror $BFM(G^\vee)$

For torus actions, the insight was that gauging a Fukaya category $\mathfrak{F}(X)$ amounted to enriching it from a $\mathfrak{Coh}(T^\vee_{\mathbb{C}})$ module to an object in $\sqrt{\mathfrak{Coh}}(T^*T^\vee_{\mathbb{C}})$. In a cotangent bundle, this promotion may seem modest. A non-abelian Lie group G will move us to a more sophisticated holomorphic algebraic manifold which is *not* a cotangent bundle. Let T be a maximal torus of G, W the Weyl group and B, B_+ two opposite (lower and upper triangular) Borel subgroups, N, N_+ their unipotent radicals; Fraktur letters will stand for the Lie algebras and $^\vee$ will indicate their counterparts in the Langlands dual Lie group $G^\vee$.

5.1. The home of $2D$ gauge theory. The space $BFM(G)$ was introduced and studied by Bezrukavnikov, Mirkovic and Finkelberg [5] in general, but special instances were known in many guises. Here are several descriptions. Call $T^*_{\mathrm{reg}}G_{\mathbb{C}} \subset T^*G_{\mathbb{C}}$ the Zariski-open subset comprising the *regular* cotangent vectors (centralizer of minimal dimension, the rank of G).

Theorem 5.1. *The following describe the same holomorphic symplectic manifold, denoted $BFM(G)$.*

(i) *The spectrum of the complex equivariant homology $H^{G^\vee}_*(\Omega G^\vee)$, with Pontrjagin multiplication.*

(ii) *The holomorphic symplectic reduction of $T^*_{\mathrm{reg}}G_{\mathbb{C}}$ by conjugation under $G_{\mathbb{C}}$.*

(iii) *The affine resolution of singularities of the quotient $T^*T_{\mathbb{C}}/W$, obtained by adjoining the functions $(e^\alpha - 1)/\alpha$. (α ranges over the roots of $\mathfrak{g}$, $e^\alpha - 1$ is the respective function on $T_{\mathbb{C}}$ and the denominator α is the linear function on $\mathfrak{t}^*$.)*

(iv) *$BFM\big(\mathrm{SU}_n\big)$ is the moduli space of SU_2 monopoles of charge n, and is a Zariski-open subset of the Hilbert scheme of n points in $T^*\mathbb{C}^\times$ [4].*

(v) $BFM(T) = T^*T_{\mathbb{C}}$

Remark 5.2. The moment map zero-fiber for the conjugation $G_{\mathbb{C}}$-action on $T^*_{\mathrm{reg}}G_{\mathbb{C}}$ is the (regular) *universal centralizer* $\mathcal{Z}_{\mathrm{reg}} = \{(g, \xi) \,|\, g\xi g^{-1} = \xi, \xi \text{ is regular}\}$. $\mathcal{Z}_{\mathrm{reg}}$ is smooth, and $BFM(G) = \mathcal{Z}_{\mathrm{reg}}/G_{\mathbb{C}}$, with stabilizer of constant dimension and local slices. This is the only one of the descriptions that makes the holomorphic symplectic structure evident.

The space $BFM(G^\vee)$ inherits two projections from $T^*_{\mathrm{reg}}G_{\mathbb{C}} : \pi_v$, to the space $(\mathfrak{g}^\vee)^*_{\mathbb{C}}/G^\vee_{\mathbb{C}} \cong \mathfrak{t}_{\mathbb{C}}/W$ of co-adjoint orbits, and π_h, to the conjugacy classes in $G^\vee_{\mathbb{C}}$. Both are Poisson-integrable with Lagrangian fibers. The projection π_v will have the more obvious meaning for gauge theory, capturing the $H^*(BG)$-module structure on fixed-point categories. The projection π_h is closely related to the restriction to T (and to the *string topology* of flag varieties.)

The symplectic structure on $BFM(G^\vee)$ relates to its nature as (an uncompletion of) the second Hochschild cohomology of the E_2-algebra $H_*(\Omega G)$.[8] In fact, $BFM(G)$ contains the zero-fiber of π_v, $Z := \mathrm{Spec}H_*(\Omega G)$, as a smooth Lagrangian; it comes from the part of $\mathcal{Z}_{\mathrm{reg}}$ with nilpotent ξ (cf. Remark 5.2).

Theorem 2.5 and Lesson 3.2.2 sheafify categories with topological G-action over the formal neighborhood of Z. However, it is the entire space $BFM(G^\vee)$ which is the correct receptacle for G-gauge theory: gauged TQFTs are objects in the 2-category $\sqrt{\mathfrak{Coh}}(BFM(G^\vee))$. Clearly, that requires a rethinking of the notion: the definition of 'topological category with G-action' as in §2 would complete the BFM space at the exceptional Lagrangian Z. Loosely speaking, we need to know a theory together with all its deformations of the group action.

The Lagrangian Z replaces the zero-section from the torus case, and plays the role of the regular representation of G: $\mathrm{Hom}(Z, L)$ gives the underlying category of the representation L. The formal calculation is $\mathrm{Hom}_{C_*\Omega G}(C_*\Omega G; L) = L$, if we use Theorem 2.5 to model representations. Figure 5.1 below sketches $BFM(\mathrm{PSU}_2)$.

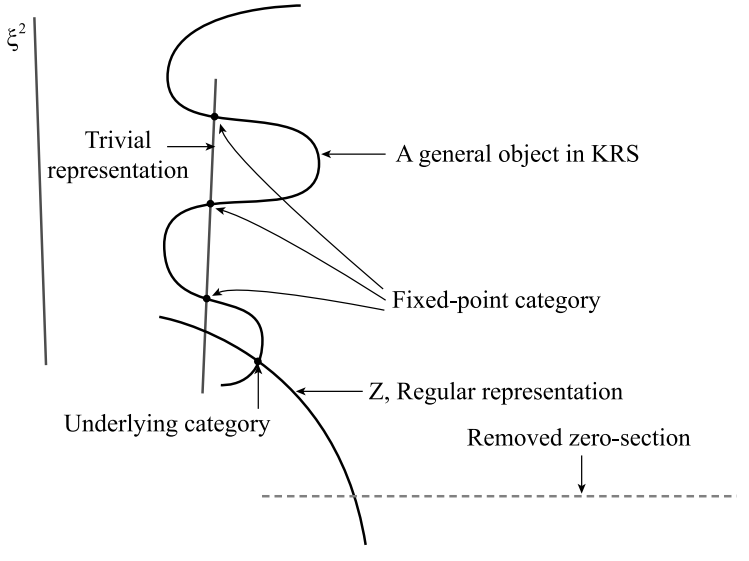


Figure 5.1. BFM space of PSU_2; the fiber of π_h at 1 is $Z \cup$trivial representation

5.2. Induction by String topology. No map relates $BFM(T^\vee) = T^*T^\vee_{\mathbb{C}}$ and $BFM(G^\vee)$, because of the blow-up, but a holomorphic Lagrangian correspondence is defined from the

[8] Of course, the E_2 structure is trivial over the complex numbers and the algebra is quasi-isomorphic to its underlying dg ring of chains.

branched cover

$$BFM(G^\vee) \longleftarrow BFM(G^\vee) \times_{\mathfrak{t}_\mathbb{C}/W} \mathfrak{t}_\mathbb{C} \longrightarrow T^*T^\vee_\mathbb{C}. \tag{5.1}$$

The right map is neither proper not open.[9] A holomorphic Lagrangian correspondences could give a pair of adjoint functors between the respective $\sqrt{\mathfrak{Coh}}$ 2-categories, thus a domain wall between T- and G- gauge theories (cf. §3.3). This is indeed the case, and we can identify the functors.

Theorem 5.3. *The correspondence* (5.1) *matches an adjoint pair of restriction-induction functors between categorical T- and G-representations. Induction from a category $\mathfrak{C}$ with topological T-action is effected by* string topology with coefficients *of the flag variety G/T:*

$$\mathrm{Ind}(\mathfrak{C}) = C_*\Omega G \otimes_{C_*\Omega T} \mathfrak{C}.$$

Restriction is the obvious functor. □

Remark 5.4.

(i) An alternative (slightly worse) description of induction is given by the category of (derived) global sections $\mathrm{R}\Gamma\big(G/T; \tilde{\mathfrak{C}}\big)$ for the associated local system $\tilde{\mathfrak{C}}$ of categories.

(ii) Neither description is quite correct. Just as the BFM spaces carry more information than the category and the action, so does induction.

(iii) For example, inducing from the representation $\mathfrak{Vect}_\tau$, for a point $\tau \in T^\vee_\mathbb{C}$ which is *not* central in G, by either method above, will appear to give zero. (This is what a homological algebra calculation of the curved string topology of G/T for a non-trivial curving $\tau \in H^2(G/T; \mathbb{C}^\times)$ gives.) However, geometric induction gives the fiber of $\pi_v^{-1}(\tau)$. The puzzle is resolved by noting that none of those fibers meet the regular representation Z, so the underlying categories are null. We are letting G act on categories without objects, and growing wiser.

(iv) The 'naïvely induced' representations can serve to probe the entire BFM space by abelianization. It is therefor not *conceptually* more difficult to understand non-abelian gauged mirrors than abelian ones. However, the symplectically induced representations of the next section are much nicer.

5.3. Alternative model for induction. I close with a new model for the correspondence (5.1), useful in a later mirror calculation. Call $\mathfrak{b}_{+,\mathrm{reg}} \subset \mathfrak{b}_+$ the open subset of regular elements. Identify $\mathfrak{b}_+ = \big(\mathfrak{g}_\mathbb{C}/\mathfrak{n}_+\big)^*$, B_+-equivariantly; the last space matches the fibers of the bundle, over $B_+ \subset G_\mathbb{C}$, of co-normals to the N_+-translation orbits. Using this to define the left map below and projection on the right gives a holomorphic Lagrangian correspondence

$$\begin{array}{ccc} & \dfrac{B_+ \times \mathfrak{b}_{+,\mathrm{reg}}}{B_+} & \\ \swarrow & & \searrow \\ T^*_{\mathrm{reg}} G_\mathbb{C} /\!/_{\mathrm{ad}} B_+ & & T^*T_\mathbb{C} \end{array}$$

[9] Z maps to $1 \in T^\vee$, but most of the zero-section in $T^*T^\vee_\mathbb{C}$ is missed by the map.

having divided by the conjugation action of $B_+^\vee$. We can also divide out by B_+ in the defining correspondence for $BFM(G)$,

$$BFM(G) \longleftarrow \mathcal{Z}_{\mathrm{reg}}/B_+ \longrightarrow T^*_{\mathrm{reg}}G_{\mathbb{C}}/\!/_{\mathrm{ad}}B_+.$$

The composition of these two can be shown to yield (5.1) (for the group G).

6. Mirrors of flag varieties

I will now explain the place of flag varieties in the mirror view of gauge theory. Lifting to the torus-equivariant picture will recover a construction of K. Rietsch [22].

6.1. Flag varieties as domain walls. Let $L \subset G$ be a Levi subgroup, centralizer of a dominant weight $\lambda : \mathfrak{l} \to \mathrm{i}\mathbb{R}$. The flag variety $X = G/L$ is a symplectic manifold with Hamiltonian G-action (the co-adjoint orbit of λ), and as such it should have a mirror holomorphic Lagrangian in $BFM(G^\vee)$. This will be true, but we forgot some structure relevant to gauge theory. Namely, we can use G/L to *symplectically induce* categorical representations from L to G.

A categorical representation $\mathfrak{C}$ of L gives the local system of categories $\tilde{\mathfrak{C}} = G \times_L \mathfrak{C} \to X$, and we can construct the Fukaya category of X with coefficients in $\tilde{\mathfrak{C}}$. (Objects would be horizontal sections of objects over Lagrangians, and Floer complexes can be formed in the usual way from the Hom-spaces over intersections.) In fact, the weight λ (or rather, its exponential e^λ in the center of $L_{\mathbb{C}}^\vee$) defines a topological representation $\mathfrak{Vect}_\lambda$ of L, and we can think of the ordinary Fukaya category $\mathfrak{F}(X, \lambda)$ as the symplectic induction from the latter. The precise meaning is that deforming λ in $\mathfrak{Vect}_\lambda$ achieves the same effect as the matching deformation of the symplectic form. An imaginary variation of λ (movement in the unitary group $L^\vee$) has the effect of adding a unitary B-field twist to the Fukaya category.

Remark 6.1. Left adjoint to the symplectic induction functor SInd_L^G is a *symplectic restriction* from G to L. This is not the ordinary (forgetful) restriction, which instead is adjoint to string topology induction (§5). For example, when $L = T$, the spectral decomposition under T of the symplectic restriction of $\mathfrak{C}$ would extract the multiplicities of the $\mathfrak{F}(X, \tau)$ in $\mathfrak{C}$, rather than those of the $\mathfrak{Vect}_\tau$.

This pair of functors is a new *domain wall* between pure 3-dimensional G- and L-gauge theories. On the mirror side, we can hope to represent a domain wall by a holomorphic Lagrangian correspondence between $BFM(L^\vee)$ and $BFM(G^\vee)$. We will be fortunate to identify this correspondence with an open embedding.

To recover the mirror of X in its various incarnations (as a symplectic manifold, or a G-equivariant symplectic one) we must apply boundary conditions to the two gauge theories, aiming for the 'sandwich picture' of a $2D$ TQFT, as in §3.3. For example, to find the underlying symplectic manifold (X, λ), we must apply the representation $\mathfrak{Vect}_\lambda$ of L and the regular representation Z of G. I shall carry out this (and a more general) exercise in the final section.

The study of symplectically induced representations can be motivated by the following conjecture, the evident non-abelian counterpart of Conjecture 4.2 (with the difference that it seems much less approachable).

Conjecture 6.2. *For a Hamiltonian G-action on the compact symplectic manifold X and a regular value μ of the moment map, the Fukaya category $\mathfrak{F}(X//G)$, reduced at the orbit of μ (and with unitary B-field $\mathrm{i}\nu$) is the multiplicity in $\mathfrak{X}$ of the representation symplectically induced from $\mathfrak{Vect}_{\mu+i\nu}$.*

6.2. The Toda isomorphism. The following isomorphism of holomorphic symplectic manifolds is mirror to symplectic induction. It fits within a broad range of related results ('Whittaker constructions') due to Kostant. Its relation to Fukaya categories of flag varieties is mysterious, and now only understood with reference to the appearance of the Toda integrable system in the Gromov-Witten theory of flag varieties [11, 12]. From that point of view, the isomorphism enhances the Toda system by supplying the conjugate family of commuting Hamiltonians, pulled back from conjugacy classes in the group, rather than orbits in the Lie algebra.

The mirror picture of G-gauge theory involves the Langlands dual group $G^\vee$ of G, but the notation is cleaner with G. With notation as in §5, call $\chi : \mathfrak{n} \to \mathbb{C}^\times$ the regular character (unique up to $T_{\mathbb{C}}$-conjugation) and consider the *Toda space*, the holomorphic symplectic quotient of $T^*G_{\mathbb{C}}$

$$T(G) := (N, \chi)\backslash\backslash T^*G_{\mathbb{C}}//(N, \chi)$$

under the left×right action of N, reduced at the point $(\chi, \chi) \in \mathfrak{n}^* \oplus \mathfrak{n}^*$.

Theorem 6.3. *We have a holomorphic symplectic isomorphism*

$$T(G) = (N, \chi)\backslash\backslash T^*G_{\mathbb{C}}//(N, \chi) \cong T^*_{\mathrm{reg}}G_{\mathbb{C}}//_{\mathrm{Ad}}G_{\mathbb{C}} = BFM(G)$$

induced from the presentation of the two manifolds as holomorphic symplectic reductions of the same manifold $T^*_{\mathrm{reg}}G_{\mathbb{C}}$.

Proof. The $N \times N$ moment fiber in $T^*G_{\mathbb{C}} \cong G_{\mathbb{C}} \times \mathfrak{g}^*_{\mathbb{C}}$ (by left trivialization) is

$$\mathcal{T} := \{(g, \xi) \in G_{\mathbb{C}} \times \mathfrak{g}^*_{\mathbb{C}} \,|\, \pi(\xi) = \pi(g\xi g^{-1}) = \chi\},$$

where $\pi : \mathfrak{g}^*_{\mathbb{C}} \to \mathfrak{n}^*$ is the projection. As $\pi^{-1}(\chi)$ consists of regular elements, we may use $T^*_{\mathrm{reg}}G_{\mathbb{C}}$ instead. Now, N acts freely on $\pi^{-1}(\chi)$, with Kostant's global slice, so the $N \times N$ action on $\mathcal{T}$ is free also and $T(G) = N\backslash\mathcal{T}/N$ is a manifold.

The moment map fibers $\mathcal{T}$ and $\mathcal{Z}_{\mathrm{reg}}$ (for the Ad-action of $G_{\mathbb{C}}$) provide holomorphic Lagrangian correspondences

$$\begin{array}{ccccc} & \mathcal{T} & & \mathcal{Z}_{\mathrm{reg}} & \\ \swarrow & & \searrow \quad \swarrow & & \searrow \\ T(G) & & T^*_{\mathrm{reg}}G_{\mathbb{C}} & & BFM(G) \end{array} \tag{6.1}$$

whose composition $\mathcal{T} \times_{T^*_{\mathrm{reg}}G_{\mathbb{C}}} \mathcal{Z}_{\mathrm{reg}}$, I claim, induces an isomorphism. Actually, the clean correspondence must mind the fact that the two actions on T^*G, of $N \times N$ and G, respectively, have in common the conjugation action of N (sitting diagonally in $N \times N$): so we must really factor through $T^*_{\mathrm{reg}}G_{\mathbb{C}}//_{\mathrm{Ad}}(N)$, within which the co-isotropics $\mathcal{T}/_{\mathrm{Ad}}N$ and $\mathcal{Z}_{\mathrm{reg}}/_{\mathrm{Ad}}N$ turn out to intersect transversally.

We check that the composition in (6.1) induces a bijection on points: preservation of the Poisson structure then supplies the Jacobian criterion. Choose $(g, \xi) \in \mathcal{T}$; then, $\xi, g\xi g^{-1} \in$

$\pi^{-1}(\chi)$ are in the same $G_{\mathbb{C}}$-orbit in $\mathfrak{g}^*_{\mathbb{C}}$. Kostant's slice theorem ensures that the two elements are then Ad-related by a unique $\nu \in N$, $\nu g \xi (\nu g)^{-1} = \xi$. There is then, up to right action of N, a unique $(g', \xi') \in \mathcal{Z}_{\mathrm{reg}}$ in the $N \times N$-orbit of (g, ξ). We thus get an injection $T(G) \hookrightarrow BFM(G)$. To see surjectivity, conjugate a chosen $(h, \eta) \in \mathcal{Z}_{\mathrm{reg}}$ to bring η into $\pi^{-1}(\chi)$. The result is in $\mathcal{T}$ (and is again unique up to N-conjugation). □

Remark 6.4. The space $T(G)$ has a hyperkähler structure; it comes from a third description, as a moduli space of solutions to Nahm's equations. This is closely related to a conjectural derivation of my mirror conjecture (6.5) below from Langlands (electric-magnetic) duality in 4-dimensional $N = 4$ Yang-Mills theory. (I am indebted to E. Witten for this explanation.)

6.3. The mirror of symplectic induction. Inclusion of the open cell $N \times w_0 \cdot T_{\mathbb{C}} \times N \subset G_{\mathbb{C}}$ leads to a holomorphic symplectic embedding $T^*T_{\mathbb{C}} \subset T(G)$. Sending a co-tangent vector to its co-adjoint orbit projects $T(G)$ to $\mathfrak{g}^*_{\mathbb{C}}//G^{ad}_{\mathbb{C}}$, and the functions on the latter space lift to the commuting Hamiltonians of the Toda integrable system; so the theorem completes the picture by providing a complementary set of Hamiltonians lifted from the conjugacy classes of G.

More generally, if $L \subset G$ is a Levi subgroup, with representative $w_L \in L$ of its longest Weyl element, and with unipotent group $N_L = N \cap L_{\mathbb{C}}$, then χ restrict to a regular character of N_L and the inclusion

$$N \times_{N_L} w_0 w_L^{-1} \cdot L_{\mathbb{C}} \times_{N_L} N \subset G_{\mathbb{C}}$$

determines an open embedding $T(L) \subset T(G)$. The following is, among others, a character formula for induced representations. It relies on too many wobbly definitions to be called a theorem, but assuming it is meaningful, its truth can be established form existing knowledge.

Conjecture 6.5. *Via the Toda isomorphism, the embedding $T(L^\vee) \subset T(G^\vee)$ is mirror to symplectic induction from L to G, representing the flag variety G/L as a domain wall between L- and G-gauge theories.*

Example 6.6. With the torus $L = T$, a one-dimensional representation of T is described by a point $q \in T^\vee$, represented in $\sqrt{\mathfrak{Coh}}(T^*T^\vee)$ by the cotangent space at q. Its image under the Toda isomorphism, a Lagrangian leaf $\Lambda(q) \subset BFM(G)$, is the symplectically induced representation, or the G-equivariant Fukaya category of the flag variety G/T with quantum parameter q. The analogue of the character is the structure sheaf $\mathcal{O}_{\Lambda(q)}$, whose algebra of global sections is the G-equivariant quantum cohomology of G/T [11].

Remark 6.7. It is difficult to prove the conjecture without a precise definitions (of equivariant Fukaya categories with coefficients and of the KRS 2-category). Nevertheless, accepting that $BFM(G^\vee)$ as the correct mirror of G-gauge theory, the conjecture follows from known results about the equivariant quantum cohomology of flag varieties [7, 11, 18]. The latter describe $qH^*_G(G/L)$ as a module over $H^*(BG) = \mathbb{C}[\mathfrak{g}]^G$, the algebra of Toda Hamiltonians, induced from the projection π_v. The symplectic condition turns out to pin the map uniquely.

6.4. Foliation by induced representations. Recall (Example 2.4) the one-dimensional representations of a Levi subgroup $L \subset G$, corresponding to the points in the center of $L^\vee_{\mathbb{C}}$. Let us call them *cuspidal*: they are not symplectically induced from a smaller Levi subgroup. (Such a symplectic induction produces representation of rank equal to the Euler characteristic of the flag variety.) The following proposition suggests that these induced representations are better suited to spectral theory that the naïvely induced ones of §5.

Theorem 6.8. *The space $BFM(G^\vee)$ is smoothly foliated by symplectic inductions of cuspidal representations: each leaf comes from a unique cuspidal representation of a unique Levi subgroup L, with $T \subset L \subset G$.*

Proof. The leaves are the fibers of $N\backslash\mathcal{T}^\vee/N \to N\backslash G^\vee_{\mathbb{C}}/N$, and induction on the semi-simple rank reduces us to checking that the part of $\mathcal{T}^\vee$ which does *not* come from any $T(L^\vee)$, for a proper $L \subset G$, lives over the center of $G^\vee_{\mathbb{C}}$.

Omit $^\vee$ from the notation and choose $(g, \xi) \in \mathcal{T}$. From $G_{\mathbb{C}} = \coprod_w N \cdot wT_{\mathbb{C}} \cdot N$, we may take $g \in wT_{\mathbb{C}}$ for some $w \in W$. Split $\mathfrak{g}^*_{\mathbb{C}} = \mathfrak{n}^* \oplus \mathfrak{t}^*_{\mathbb{C}} \oplus \mathfrak{n}^*_+$; then,

$$\begin{aligned} \xi &= \chi + \eta + \nu, && \text{for some } \eta \in \mathfrak{t}_{\mathbb{C}}, \nu \in \mathfrak{n}^*_+ \\ g\xi g^{-1} &= \chi + w(\eta) + \nu', && \text{for some } \nu' \in \mathfrak{n}^*_+ \end{aligned}$$

whence we see that w sends each simple negative root either to a simple negative root, or to a positive root. If $w = 1$, then $g \in T_{\mathbb{C}}$ centralizes $\chi \pmod{\mathfrak{b}^*_+}$ and thus lies in the center of $G_{\mathbb{C}}$. Otherwise, I claim that $w = w_0 w_L^{-1}$, for the Levi L whose negative simple roots stay negative. Equivalently, the unique simple root system of $\mathfrak{g}$ comprising the simple negative roots of L and otherwise only positive roots, is the w_L-transform of the positive root system. This can be seen by choosing a point $\zeta + \varepsilon$, with ζ generic on the L-fixed face of the dominant Weyl chamber, and ε a dominant regular displacement: $w_L(\zeta + \varepsilon)$ must be in the dominant chamber of the new root system. □

Example 6.9 ($G = \mathrm{SU}_2$). The dual complex group is $G^\vee_{\mathbb{C}} = \mathrm{PSL}_2(\mathbb{C})$, whose BFM space is the blow-up of $\mathbb{C} \times \mathbb{C}^\times/\{\pm 1\}$ at $(0, 1)$, with the proper transform of the zero-section $\{0\} \times \mathbb{C}^\times/\{\pm 1\}$ removed. This is the Atiyah-Hitchin manifold studied in [4]. The $\mathbb{Z}/2$-action identifies (ξ, z) with $(-\xi, z^{-1})$. Projection to the line of co-adjoint orbits is given by the Toda Hamiltonian ξ^2.

The Toda inclusion of $T^*T^\vee_{\mathbb{C}} \cong \mathbb{C} \times \mathbb{C}^\times$ sends a point (u, q) to

$$\xi^2 = u^2 - q, \qquad \frac{z + z^{-1}}{4} = \frac{u^2}{q} - \frac{1}{2}$$

(A match of signs is required between z and ξ.) The induced leaves of constant q are given by

$$\xi = q\frac{\sqrt{z} - \sqrt{z}^{-1}}{2},$$

after lifting to the coordinates $\xi, \sqrt{z}$ for the double-cover maximal torus in SL_2. We recognize here the (graph of the differentiated potential in the) S^1-equivariant mirror of the flag variety $\mathbb{P}^1$.

The one remaining leaf in $BFM(\mathrm{PSU}_2)$ is the trivial representation of SU_2; it is the proper transform of $T^*_1\mathbb{C}^\times/\{\pm 1\}$, the image in $\mathbb{C} \times \mathbb{C}^\times/\{\pm 1\}$ of) the cotangent fiber at 1. If we switch instead to $\mathrm{PSU}(2)$, the new BFM space (on the Langlands dual side) is a double cover of the former, and there is a new cuspidal leaf over the central point $(-\mathrm{I}_2) \in \mathrm{SU}_2$, corresponding to the sign representation of $\pi_1\mathrm{PSU}_2$.

6.5. Torus-equivariant flag varieties. Restricting the G-action to T, the flag manifold G/L is a transformation from L-gauge theory to T-gauge theory, given by composition of the symplectic induction and string topology domain walls:

$$T(L^\vee) \xrightarrow{\text{SInd}} T(G^\vee) \xrightarrow[\text{Toda}]{\sim} BFM(G) \xrightarrow{\text{ST}} BFM(T^\vee) = T^*T^\vee_{\mathbb{C}} \tag{6.2}$$

The equivariant mirror is a family of $2D$ TQFTs, which can be defined, for instance, by a family of complex manifolds with potentials parametrized by the Lie algebra $\mathfrak{t}_{\mathbb{C}}$. This family reflects the $H^*(BT)$-module structure on equivariant quantum cohomology. When $\mathfrak{F}(G/L)$ has been represented by an object $\Lambda \in \sqrt{\mathfrak{Coh}}(T^*T^\vee_{\mathbb{C}})$, the family comes from the projection of $T^*T^\vee_{\mathbb{C}}$ to the cotangent fiber, and the TQFTs are the fibers of Λ over $\mathfrak{t}_{\mathbb{C}}$, the Hom categories with the constant sections of $T^*T^\vee_{\mathbb{C}}$.

To recover this family of mirrors from the double domain wall (6.2), we must use it to pair two Lagrangians, in $T(L^\vee)$ and in $T^*T^\vee_{\mathbb{C}}$. The Lagrangians are

- the Lagrangian leaf $\Lambda(q) \subset BFM(L^\vee)$ over a point q in the center of $L^\vee_{\mathbb{C}}$, describing a cuspidal representation of L (q is also the quantum parameter for G/L);
- the constant Lagrangian section S_ξ of $T^*T^\vee_{\mathbb{C}}$, with fixed value $\xi \in \mathfrak{t}_{\mathbb{C}}$.

Note that S_ξ is the differential of a multi-valued character $\xi \circ \log : T^\vee_{\mathbb{C}} \to \mathbb{C}$.

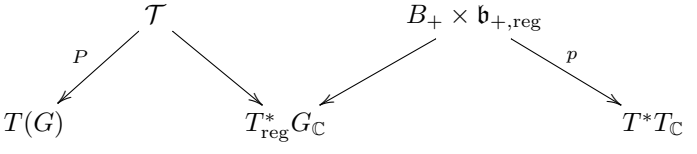
Remark 6.10. The relevant TQFT picture is a sandwich with triple-decker filling: the base slice is the representation $\mathfrak{Vect}_q$ of L corresponding to $\Lambda(q)$, a boundary condition for L-gauge theory. The filling of the sandwich is a triple layer of L, G, T gauge theories, separated by the SInd and string topology domain walls in (6.2). The sandwich is topped with the slice S_ξ, a boundary condition for T-gauge theory. Its underlying representation category is null, if $\xi \neq 0$; S_ξ is a deformation of the regular representation of T by the multi-valued potential $\xi \circ \log$.

6.6. Rietsch mirrors. Building on ideas of Peterson and earlier calculations of Givental-Kim, Ciocan-Fontanine, Kostant and Mihalcea [7, 11, 12, 18], Rietsch [22] proposed torus-equivariant complex mirrors for all flag varieties G/L.

Let us recover these from my story by computing the answer outlined above. Recall (§5.3) the Lagrangian correspondence

$$T^*T_{\mathbb{C}} \leftarrow B_+ \times \mathfrak{b}_{+,\text{reg}} \to T^*_{\text{reg}}G_{\mathbb{C}},$$

appearing in the alternate model for the string topology induction. Compose this with the Toda construction to define the following holomorphic Lagrangian correspondence between $T(G)$ and $BFM(T) = T^*T_{\mathbb{C}}$:

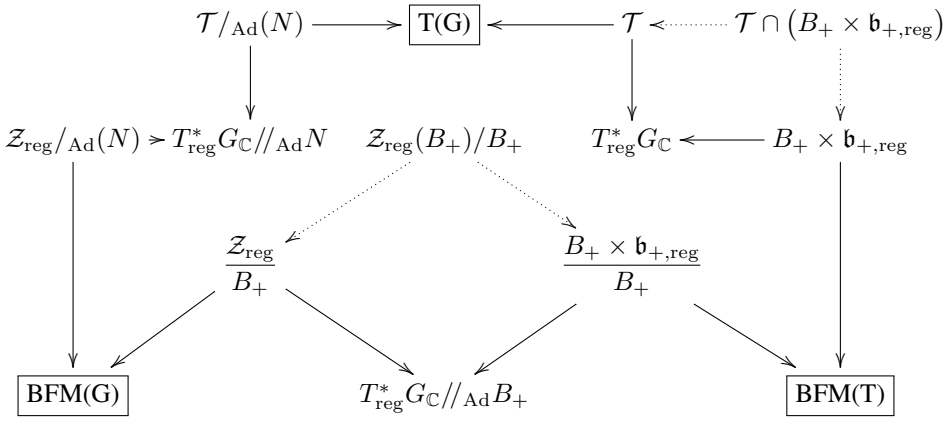


(6.3)

Proposition 6.11. *Correspondence* (6.3) *is the composition* $\text{ST} \circ \text{Toda}$ *of* (6.2).

Sketch of proof. In the jagged triangle of correspondences below, the left edge is the Toda isomorphism, the right edge the correspondence (6.3) and the bottom edge the string topology domain wall. The long, counterclockwise way from top to right involves division by the

complementary subgroups N and B_+ of $G_{\mathbb{C}}$; so it seems reasonable that the composition should agree with the undivided correspondence (6.3) on the right edge:



The argument exploits the regularity of the Lie algebra elements. The intersection in the upper right corner comprises the pairs $(b, \beta) \in B_+ \times \mathfrak{b}_+$ with b centralizing $\beta \in E + \mathfrak{t}_{\mathbb{C}}$. ($E = \chi$ under $\mathfrak{n}_+ \cong \mathfrak{n}^*$.) That is a slice for the conjugation B_+-action on the regular centralizer $\mathcal{Z}_{\mathrm{reg}}(B_+)$ in B_+, which makes clear the isomorphism with the fiber product in the center the triangle; and the map is compatible with the Toda isomorphism on the left edge. □

We now calculate the pairing $S_\xi \subset T(L^\vee)$ and $\Lambda(q) \subset T^*T^\vee_{\mathbb{C}}$ by the correspondence (6.3) for the dual group $G^\vee$. We do so by computing in $T^*_{\mathrm{reg}}G^\vee_{\mathbb{C}}$

$$\mathrm{Hom}\left(p^{-1}S_\xi, P^{-1}\Lambda(q)\right).$$

The two Lagrangians meet over the intersection

$$B^\vee_+ \cap \left(N^\vee \cdot w_0 w_L^{-1} L^\vee_{\mathbb{C}} \cdot N^\vee\right) \subset G^\vee_{\mathbb{C}}.$$

Lift $\xi \circ \log$ to $B^\vee_+$ by p; over $B^\vee_+$, $p^{-1}S_\xi$ is the conormal bundle to $B^\vee_+ \subset G^\vee_{\mathbb{C}}$ shifted by the graph of $d(\xi \circ \log)$. (The shifted bundle is well-defined, independently of any local extension of the function $\xi \circ \log$.)

The Lagrangian $P^{-1}\Lambda(q)$ lives over the open set $N^\vee \cdot w_0 w_L^{-1} L^\vee_{\mathbb{C}} \cdot N^\vee$ in $G^\vee_{\mathbb{C}}$, where it is the shifted co-normal bundle to the submanifold

$$M := N^\vee \cdot w_0 w_L^{-1} q \cdot N^\vee \cong \frac{N^\vee \times N^\vee}{\mathrm{diag}(N^\vee \cap L^\vee_{\mathbb{C}})},$$

shifted into $\mathcal{T}$ by the graph of the differential of the following function f:

$$f : n_1 \cdot w_0 w_L^{-1} l \cdot n_2 \mapsto \chi(\log n_1 + \log n_2).$$

Now, $B^\vee_+$ and M meet transversally in $G^\vee_{\mathbb{C}}$, in a manifold isomorphic to a Zariski-open in the flag variety $G^\vee/L^\vee$; this is the $\mathcal{R}_{w_0,w_L}$ of [22]-. Transversality permits us to dispense

with the conormal bundles, and identify $\mathrm{Hom}(S_\xi, \Lambda(q))$ with the pairing, in the cotangent bundles, between graphs of the restricted functions to $B_+^\vee \cap M$

$$\mathrm{Hom}_{T^*(B_+^\vee \cap M)}\left(\Gamma(d(\xi \circ \log)), \Gamma(df)\right);$$

this is the matrix factorization category $MF\big(B_+^\vee \cap M; f - \xi \circ \log\big)$. This is the Rietsch mirror of G/L.

The last mirror comes with a volume form, which defines the trace on HH_*. In the Lagrangian correspondence, we need instead a half-volume form on each leaf. The two leaves S_ξ and $\Lambda(q)$ do in fact carry natural half-volumes, translation-invariant for the groups (B and $N \times N$) and along the cotangent fibers. Rietsch's volume form on the mirror $\mathcal{R}_{w_0,w_L}$ comes from the product of these half-volumes.

Acknowledgements. The author thanks the MSRI for its hospitality during the writing of this paper. The work was partially supported by NSF grant DMS-1007255. I thank M. Abouzaid, D. Ben-Zvi, K. Fukaya, K. Hori, A. Kapustin, A. Neitzke, C. Woodward for helpful comments and conversation, and am especially indebted to E. Witten for explaining the relation to 4-dimensional gauge theory and the Nahm equations. Many thanks are due to the Geometry group at UT Austin for the invitation to lecture there, where a primitive version of this material was first outlined [28]; for later developments, see [29].

References

[1] M. Abouzaid, *On the wrapped Fukaya category and based loops*, J. Symplectic Geom., **10** (2012), 27–79.

[2] M.F. Atiyah, *Topological quantum field theories*, Inst. Hautes Études Sci. Publ. Math., **68** (1988), 175–186.

[3] P.C. Argyres and A.E. Farragi, *The vacuum structure and spectrum of $N = 2$ supersymmetric $\mathrm{SU}(n)$ gauge theory*, Phys. Rev. Letter, **74** (1995), 3931–3934.

[4] M.F. Atiyah and N. Hitchin, *The geometry and dynamics of magnetic monopoles*, M.B. Porter Lectures, Princeton University Press (1988).

[5] R. Bezrukavnikov, M. Finkelberk, and I. Mirkovic, *Equivariant homology and K-theory of affine Grassmannians and Toda lattices*, Compos. Math., **141** (2005), 746–768.

[6] K. Costello, *Topological quantum field theories and Calabi-Yau categories*, Adv. Math., **210** (2007), 165–214.

[7] I. Ciocan-Fontanine, *On quantum cohomology rings of partial flag varieties*, Duke Math. J., (1999), 485–523.

[8] M. Chas and D. Sullivan, *String topology.*, arXiv:math/9911159.

[9] D.S. Freed, J. Lurie, M.J. Hopkins, and C. Teleman, *Topological quantum field theories from compact Lie groups*.

[10] J. Francis, *The tangent complex and Hochschild cohomology of E_n-rings*, Compos. Math., **149** (2013), 430–480.

[11] A. Givental and B. Kim, *Quantum cohomology of flag manifolds and Toda lattices*, Comm. Math. Phys., **168** (1995), 609–641.

[12] B. Kostant, *Flag manifold quantum cohomology, the Toda lattice, and the representation with highest weight ρ*, Selecta Math. (N.S.), **2** (1996), 43–91.

[13] M. Kontsevich and Y. Soibelman, *Notes on A_∞-algebras, A_∞-categories and non-commutative geometry. In*, Homological mirror symmetry, 153–219, Lecture Notes in Phys., **757**, Springer 2009.

[14] A. Kapustin and L. Rozansky, *Three-dimensional topological field theory and symplectic algebraic geometry II.*, Commun. Number Theory Phys., **4** (2010), 463–549.

[15] A. Kapustin, L. Rozansky, and N. Saulina, *Three-dimensional topological field theory and symplectic algebraic geometry*, Nuclear Phys. B, **816** (2009), 295–355.

[16] J. Lurie, *On the classification of topological field theories*, Current developments in mathematics, 2008, 129–180, Int. Press, Somerville, MA 2009.

[17] Yu.I. Manin, *Frobenius manifolds, quantum cohomology and moduli spaces*, AMS Colloquium Publications, **47**. AMS, Providence, 1999.

[18] L. Mihalcea, *On equivariant quantum cohomology of homogeneous spaces*, Chevalley formulae and algorithms. Duke Math. J, **140** (2007), 321–350.

[19] E. Martinec and N. Warner, *Integrable systems and supersymmetric gauge theory*, Nuclear Phys. B, **459** (1996), 97–112.

[20] D. Nadler and E. Zaslow, *Constructible sheaves and the Fukaya category*, J. Amer. Math. Soc., **22** (2009), 233–286.

[21] D.O. Orlov, *Triangulated categories of singularities and D-branes in Landau-Ginzburg models. (Russian)*, Tr. Mat. Inst. Steklova, **246** (2004), 240–262, translation in Proc. Steklov Inst. Math., 2004, (246), 227–248

[22] K. Rietsch, *A mirror symmetric construction of $qH^*_T(G/P)_{(q)}$*, Adv. Math., **217** (2009), 2401–2442.

[23] L. Rozansky and E. Witten, *Hyper-Kähler geometry and invariants of three-manifolds*, Selecta Math. (N.S.), **3** (1997), 401–458.

[24] J. Roberts and S. Willerton, *On the Rozansky-Witten weight systems*, Algebr. Geom. Topol., **10** (2010), 1455–1519.

[25] G.B. Segal, *Stanford notes. 1: Topological field theories*, www.cgtp.duke.edu/ITP99/\segal/stanford/lect1.pdf.

[26] P. Seidel, *π_1 of symplectic automorphism groups and invertibles in quantum homology rings*, Geom. Funct. Anal., **7** (1997), 1046–1095.

[27] N. Seiberg and E. Witten, *Gauge dynamics and compactification to three dimensions*, The mathematical beauty of physics (Saclay, 1996), 333–366, Adv. Ser. Math. Phys., **24**, World Sci., 1997.

[28] C. Teleman, *Lectures on Langlands duality in mirror symmetry*, Available at www.ma.utexas.edu/rtgs/geomtop/rtg/perspectives.html.

[29] ———, *Mirror symmetry, Langlands duality and gauge theory*, Lecture at CIRM, Luminy, Slides at http://math.berkeley.edu/~teleman/math/luminy.pdf.

[30] V. Turaev, *Homotopy quantum field theory*, Tracts in Mathematics, **10**, EMS Publishing House, Zürich, 2010.

[31] C. Teleman and C. Woodward, *The index formula for the moduli of G-bundles on a curve*, Ann. Math. (2), **170** (2009), 495–527.

[32] E. Witten, *Topological quantum field theory*, Comm. Math. Phys., **117** (1988).

Department of Mathematics, UC Berkeley, CA 94720, USA
E-mail: teleman@berkeley.edu

Author Index