Section 1

Logic and Foundations

Justin Tatch Moore
The Proper Forcing Axiom 3
André Nies
Interactions of Computability and Randomness 30
Ya'acov Peterzil* **and Sergei Starchenko***
Tame Complex Analysis and o-minimality 58

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

The Proper Forcing Axiom

Justin Tatch Moore*

This article is dedicated to Stephanie Rihm Moore.

Abstract

The Proper Forcing Axiom is a powerful extension of the Baire Category Theorem which has proved highly effective in settling mathematical statements which are independent of ZFC. In contrast to the Continuum Hypothesis, it eliminates a large number of the pathological constructions which can be carried out using additional axioms of set theory.

Mathematics Subject Classification (2010). Primary 03E57; Secondary 03E75.

Keywords. Forcing axiom, Martin's Axiom, OCA, Open Coloring Axiom, PID, P-ideal Dichotomy, proper forcing, PFA

1. Introduction

Forcing is a general method introduced by Cohen and further developed by Solovay for generating new *generic* objects. While the initial motivation was to generate a counterexample to the Continuum Hypothesis, more sophisticated forcing notions can be used both to generate morphisms between structures and also obstructions to morphisms between structures.

Forcing axioms assert that the universe of all sets has some strong degree of closure under the formation of such generic objects by forcings which are sufficiently *non pathological*. Since forcings can in general add generic bijections between countable and uncountable sets, *non pathological* should include *preserves uncountability* at a minimum. The exact quantification of *non pathological* yields the different strengths of the forcings axioms. The first and among

*The author's preparation of this article and his travel to the 2010 meeting of the International Congress of Mathematicians was supported by NSF grant DMS–0757507. I would like to thank Ilijas Farah, Stevo Todorcevic, and James West for reading preliminary drafts of this article and offering suggestions.

Department of Mathematics, Cornell University, Ithaca, NY 14853–4201, USA.
E-mail: justin@math.cornell.edu.

the weakest of these axioms is *Martin's Axiom* which was abstracted by Martin from Solovay and Tennenbaum's proof of the independence of Souslin's Hypothesis [59]. Progressively stronger axioms were formulated and proved consistent as advances were made in set theory in the 1970s. The culmination of this progression was [22], where the strongest forcing axiom was isolated.

Forcing axioms have proved very effective in classifying and developing the theory of objects of an uncountable or non separable nature. More generally they serve to reduce the complexity of set-theoretic difficulties to a level more approachable by the non specialist. The central goal in this area is to establish the consistency of a structure theory for uncountable sets while at the same time working within a single axiomatic framework.

In this article, I will focus attention on the *Proper Forcing Axiom (*PFA*)*:

> If Q is a proper forcing and $\mathscr{A}$ is a collection of maximal antichains in Q with $|\mathscr{A}| \leq \aleph_1$, then there is a filter $G \subseteq Q$ which meets each element of $\mathscr{A}$.

This axiom was formulated and proved consistent relative to the existence of a supercompact cardinal by Baumgartner using Shelah's Proper Forcing Iteration Lemma. The details of the formulation of this axiom need not concern us at the moment (see Section 5 below). I will begin by mentioning two applications of PFA.

Theorem 1.1. *[7] Assume* PFA*. Every two $\aleph_1$-dense sets of reals are isomorphic.*

Theorem 1.2. *[57] Assume* PFA*. If Φ is an automorphism of the Boolean algebra $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$, then Φ is induced by a function $\phi : \mathbb{N} \to \mathbb{N}$.*

The role of PFA in these two theorems is quite different. In the first case, PFA is used to build isomorphisms between $\aleph_1$-dense sets of reals (here a linear order is *κ-dense* if each of its proper intervals is of cardinality κ). The procedure for doing this can be viewed as a higher cardinal analog of Cantor's back-and-forth argument which is used to establish that any two $\aleph_0$-dense linear orders are isomorphic. As we will see in Section 3.1, however, the situation is fundamentally more complicated than in the countable case since there are many non isomorphic $\aleph_1$-dense linear orders.

In the second theorem, PFA is used to build an obstruction to any *non trivial* automorphism of $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$. This grew out of Shelah's seminal result in which he established the consistency of the conclusion of Theorem 1.2 [56, Ch. IV]. The difference between Theorems 1.1 and 1.2 is that one can generically introduce new elements to the quotient $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$. This can moreover be done in such a way that it may be impossible to extend the function Φ to these new generic elements of the domain.

In both of the above theorems, there is a strong contrast with the influence of the Continuum Hypothesis (CH). CH implies that there are $2^{2^{\aleph_0}}$ isomorphism

types of $\aleph_1$-dense sets of reals [14] and $2^{2^{\aleph_0}}$ automorphisms of $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$ [52] (notice that there are only $2^{\aleph_0}$ functions from $\phi : \mathbb{N} \to \mathbb{N}$). This is in fact a common theme in the study of forcing axioms.

I will finish the introduction by saying that there was a great temptation to title this article *Martin's Maximum*. Martin's Maximum (MM) is a natural strengthening of PFA in which *proper* is replaced by *preserves stationary subsets of* ω_1. This is the broadest class of forcings for which a forcing axiom is consistent. This axiom was proved consistent relative to the existence of a supercompact cardinal in [22].

I have chosen to focus on PFA instead for a number of reasons. First, when applying forcing axioms to problems arising outside of set theory, experience has shown that PFA is nearly if not always sufficient for applications. Second, we have a better (although still limited) understanding of how to apply PFA. (Of course this is an equally strong argument for why we need to develop the theory of MM more completely and understand its advantages over PFA.)

Finally, and most importantly, a wealth of new mathematical ideas and proofs have come out of reducing the hypothesis of MM in existing theorems to that of PFA. In every instance in which this has been possible, there have been significant advances in set theory of independent interest. For example the technical accomplishments of [45] led to the solution of the basis problem for the uncountable linear orders in [46] soon after. Thus while MM is trivially sufficient to derive any consequence of PFA, working within the more limited framework of PFA has led to the discovery of new consequences of MM and new consistency results.

The reader is referred to [22] for the development of MM and to [9, pp. 57–60] for a concise account of the typical consequences of MM which do not follow from PFA. An additional noteworthy example can be found in [31]. Finally, the reader is referred to [87] for a somewhat different axiomatic framework due to Woodin for achieving some of the same end goals. It should be noted that reconciling this alternate framework with MM (or even PFA) is a major open problem in set theory.

This article is organized as follows. After reviewing some notation, I will present a case study of how PFA was used to give a complete classification of a certain class of linear orderings known as *Aronszajn lines*. After that, I will present two combinatorial consequences of PFA and illustrate how they can be applied through several different examples. These principles both have a diverse array of consequences and at the same time are simple enough in their formulation so as to be usable by a non specialist. In Section 5, I will formulate PFA and illustrate Todorcevic's method of building proper forcings. I will utilize the combinatorial principles from the previous section as examples to illustrate this technique. Section 6 presents some examples of how PFA has been successfully used to solve problems arising outside of set theory. The role of the equality $2^{\aleph_0} = \aleph_2$, which follows from PFA, will be discussed in Section 7. Section 8 will give some examples of how the mathematics developed in the

study of PFA has been used to prove theorems in ZFC. I will close the article with some open problems.

With the possible exception of Section 5, I have made an effort to keep the article accessible to a general audience with a casual interest in the material. Needless to say, details are kept to a minimum and the reader is encouraged to consult the many references contained throughout the article. In a number of places I have presented examples and stated lemmas simply to hint at the mathematics which is being omitted due to the nature of the article. It is my hope that the curious reader will take a pen and paper and try to fill in some of the details or else use this as an impetus to head to the library and consult some of the many references.

2. Notation and Background

The reader with a general interest in set theory should consult [32]. Further information on linear orders, trees, and coherent sequences can be found in [67] and [79], respectively. Information on large cardinals and the determinacy of games can be found in [29]. Further information on descriptive set theory can be found in [30].

For the most part I will follow the conventions of [32]. $\mathbb{N} = \omega$ will be taken to include 0. An *ordinal* is a set α linearly ordered by $\in$ such that if β is in α, then $\beta \subseteq \alpha$. Thus an ordinal is the set of its predecessors. In particular ω_1, which is the first uncountable ordinal, is the set of all countable ordinals. A *cardinal* is the least ordinal of its cardinality. While $\aleph_\alpha$ is a synonym for ω_α, the former is generally used to measure cardinality while the latter is generally used to measure length and when there is a need to refer to the set itself. Lower case Greek letters will be used to denote ordinals, with κ, λ, μ, and θ denoting cardinals.

If X is a set, then $[X]^k$ denotes all subsets of X of cardinality k. In particular, $[X]^2$ is the set of all unordered pairs of elements of X. A *graph* is a pair (G, X) where X is a set and $G \subseteq [X]^2$ (X is the *vertex set* and G is the *edge set*).

A *tree* is a partial order $(T, \leq)$ in which the predecessors of each element of T are well ordered by $<$. The ordertype of the set of strict predecessors of a t in T is the *height* of t; the collection of all elements of T of a fixed height is a *level* of T. All trees will also be assumed to be *Hausdorff*: if s and t have limit height and the same sets of predecessors, then they are equal. In particular, trees are equipped with a well defined meet operation $\wedge : T \times T \to T$. A subset of a tree is an *antichain* if it consists of pairwise incomparable elements (in the setting of trees this coincides with the notion of antichain in Section 5 below).

Generally a superscript of $*$ on a relation symbol is taken to mean "with only a finite number of exceptions" (in a context in which this makes sense). In particular, $A \subseteq^* B$ means that $A \setminus B$ is finite.

An *ideal* $\mathscr{I}$ on a set S is a subset of $\mathscr{P}(S)$ which is closed under subsets and finite unions. To avoid trivialities, all ideals in this article will be assumed

to contain all of the finite subsets of the underlying set. Fin is the ideal of all finite subsets of $\mathbb{N}$. $\mathrm{Fin} \times \emptyset$ is the collection of all subsets of $\mathbb{N} \times \mathbb{N}$ in which all but finitely many vertical sections are empty. $\emptyset \times \mathrm{Fin}$ is the collection of all subsets of $\mathbb{N} \times \mathbb{N}$ in which all vertical sections are finite. An ideal $\mathscr{I}$ is a *P-ideal* if $(\mathscr{I}, \subseteq^*)$ is countably directed (i.e. every countable subset has an upper bound). $\emptyset \times \mathrm{Fin}$ is a P-ideal; $\mathrm{Fin} \times \emptyset$ is not. If $\mathscr{I}$ is a collection of subsets of S, then $\mathscr{I}^\perp$ is the ideal of all subsets of S which have finite intersections with all elements of $\mathscr{I}$. Observe that $(\emptyset \times \mathrm{Fin})^\perp = \mathrm{Fin} \times \emptyset$ and $(\mathrm{Fin} \times \emptyset)^\perp = \emptyset \times \mathrm{Fin}$.

Throughout this article, all topological spaces are assumed to be T_3. When discussing Banach spaces, *basis* will always refer to a Schauder basis. A *Polish space* is a separable, completely metrizable topological space. A subset of a Polish space is *analytic* if it is the continuous image of a Borel set in a Polish space. The σ-algebra generated by the analytic sets will be denoted by $\mathscr{C}$.

3. Classification and $\aleph_1$

3.1. The basis problem for uncountable linear orders: a case study. In order to illustrate the influence of PFA and how it plays a role in classification problems for uncountable structures, I will begin with an example of a recent success in this area. Consider the following problem.

Problem 3.1. *Do the uncountable linear orders have a finite basis?*

That is, is there a finite set of uncountable linear orders such that every other contains an isomorphic copy of one from this finite set?

Observe that any such basis must contain a set of reals of minimum possible cardinality — namely $\aleph_1$. The following theorem, which actually predates PFA, shows that under PFA a single set of reals of cardinality $\aleph_1$ is sufficient to form a basis for the uncountable separable linear orders.

Theorem 3.2. *[7] Assume* PFA*. Every two $\aleph_1$-dense sets of reals are isomorphic. In particular any set of reals of cardinality $\aleph_1$ embeds into any other.*

This is in stark contrast to the situation under CH.

Theorem 3.3. *[14] If $X \subseteq \mathbb{R}$ with $|X| = |\mathbb{R}|$, then there is a $Y \subseteq X$ with $|Y| = |X|$ such that no two distinct subsets of Y of cardinality $|\mathbb{R}|$ are isomorphic. In particular if $|\mathbb{R}| = \aleph_1$, then there is no basis for the uncountable suborders of $\mathbb{R}$ of cardinality less than $|\mathscr{P}(\mathbb{R})|$.*

In fact this is part of general phenomenon: it is typically not possible to classify arbitrary structures of cardinality $2^{\aleph_0}$. (This statement is not meant to be applied to objects such as manifolds which, while of cardinality $2^{\aleph_0}$, are really coded by a countable — or even finite — mathematical structure.)

How does one reconcile Theorems 3.2 and 3.3? Baumgartner's result actually shows that given any model of ZFC, it is possible to go into a forcing extension

in which uncountability is preserved and every two $\aleph_1$-dense sets of reals are isomorphic. In particular, two $\aleph_1$-dense sets of reals which may not have been isomorphic are made isomorphic by Baumgartner's forcing. Thus while CH implies that there are many non-isomorphic $\aleph_1$-dense sets of reals, the reason for this is simply that there is an inadequate number of embeddings between such orders, rather than some intrinsic property of the sets of reals which prevents them from being isomorphic.

Now we return to our basis problem. Since ω_1 can not be embedded into $\mathbb{R}$ and since ω_1 is isomorphic to each of its uncountable suborders, any basis for the uncountable linear orders must also contain ω_1 and $-\omega_1$. The following classical construction of Aronszajn and Kurepa shows that any basis for the uncountable linear orders must have at least four elements (see [67, 5.15] for a historical discussion).

Theorem 3.4. *There is an uncountable linear order which does not contain an uncountable separable suborder and does not contain ω_1 or $-\omega_1$.*

Such linear orders are commonly known as *Aronszajn lines* or *A-lines*. Regardless of the value of $2^{\aleph_0}$, A-lines necessarily have cardinality $\aleph_1$. Like uncountable suborders of $\mathbb{R}$, every A-line contains an $\aleph_1$-dense suborder. Following Theorem 3.2, there was an effort to prove an analogous result for the class of A-lines. It turned out that the answer to this pursuit lay in the following question of R. Countryman.

Question 3.5. *Does there exist an uncountable linear order C such that $C \times C$, equipped with the coordinatewise partial order, is the union of countably many chains?*

Such linear orders are known as *Countryman lines* or *C-lines*. Clearly every uncountable suborder of a C-line is a C-line. It was observed by Galvin that such linear orders are necessarily Aronszajn. Their most remarkable property is that if C is a C-line, then no uncountable linear order can be embedded into both C and $-C$. Indeed, if $f : L \to C$ and $g : L \to -C$ were to witness such embeddings, then the range of $f \times g$, regarded as a subset of $C \times C$, would be the graph of a strictly decreasing function. As such a graph can intersect every chain in $C \times C$ in at most a singleton, L must be countable. Thus, unlike the situation with uncountable suborders of $\mathbb{R}$ under CH, there is a fundamental obstruction preventing an embedding of C into $-C$ if C is a C-line. The following theorem of Shelah, therefore, ruled out an analog of Baumgartner's result for $\aleph_1$-dense A-lines.

Theorem 3.6. *[55] There is a Countryman line.*

It was in this paper that precursors of Problem 3.1 began to be considered. Shelah made two conjectures at the end of [55]:

1. It is consistent that every two Countryman lines contain uncountable suborders which are either isomorphic or reverse isomorphic.

2. It is consistent that every Aronszajn line contains a Countryman suborder.

Shelah's construction led Todorcevic to prove the following theorem, indicating that such linear orders occur quite naturally.

Theorem 3.7. *[69] If e_β $(\beta < \omega_1)$ is a coherent sequence such that for each β, e_β is a finite-to-one function from β into ω, then the lexicographical order on $\{e_\beta : \beta < \omega_1\}$ is a Countryman line.*

Here a sequence e_β $(\beta < \omega_1)$ with $e_\beta : \beta \to \omega$ is *coherent* if whenever $\beta < \gamma$, $e_\beta =^* e_\gamma \restriction \beta$. Given such a sequence, we can also form an *Aronszajn tree* $T = \{e_\beta \restriction \alpha : \alpha \leq \beta < \omega_1\}$. Here an *Aronszajn tree* (or A-tree) is an uncountable tree in which all levels and chains are countable. An A-tree which is the set of restrictions of a coherent sequence is said to be *coherent*.

Before proceeding, I will mention the method from [69] for explicitly constructing such a coherent sequence. Let $\langle C_\alpha : \alpha < \omega_1 \rangle$ be a sequence such that $C_{\alpha+1} = \{\alpha\}$ and if α is a limit ordinal then C_α is a cofinal subset of α isomorphic to ω. Such a sequence is known as a *C-sequence*. Given a C-sequence, there is a canonical "walk" between any two ordinals $\alpha < \beta$ in ω_1:

$$\beta_i = \begin{cases} \beta & \text{if } i = 0 \\ \min(C_{\beta_{i-1}} \setminus \alpha) & \text{if } i > 0 \text{ and } \beta_{i-1} > \alpha \end{cases}$$

The walk starts at β and stops once α is reached at some stage l (l is always finite since otherwise we would have defined an infinite descending sequence of ordinals). Such walks have a number of associated statistics:

$$\varrho_0(\alpha, \beta) = \langle |C_{\beta_i} \cap \alpha| : i < l \rangle$$

$$\varrho_1(\alpha, \beta) = \max \varrho_0(\alpha, \beta)$$

$$\varrho_2(\alpha, \beta) = l = |\varrho_0(\alpha, \beta)|$$

If we set $e_\beta(\alpha) = \varrho_1(\alpha, \beta)$, then this defines a coherent sequence satisfying the hypothesis of Theorem 3.7. In fact if we define (for $i = 0, 1, 2$)

$$C(\varrho_i) = (\{\varrho_i(\cdot, \beta) : \beta < \omega_1\}, \leq_{\text{lex}})$$

$$T(\rho_i) = (\{\varrho_i(\cdot, \beta) \restriction \alpha : \alpha \leq \beta < \omega_1\}, \subseteq)$$

then $C(\varrho_i)$ is a C-line and $T(\varrho_i)$ is an A-tree. Not only does the above construction yield an informative example of a C-line and an A-tree, it is the simplest instance of a widely adaptable technique of Todorcevic for building combinatorial objects both at the level of $\aleph_1$ and on higher cardinals. A modern account of this can be found in [79].

Again we return to our analysis of Problem 3.1. The following theorem of Todorcevic shows that, under PFA, C-lines are indeed canonical objects.

Theorem 3.8. *(see [48]) Assume* $\mathrm{MA}_{\aleph_1}$*. If* C *and* C' *are Countryman lines which are* $\aleph_1$*-dense and non stationary, then either* $C \simeq C'$ *or* $-C \simeq C'$*.*

Here an A-line A is *non stationary* if $A = \bigcup \mathscr{C}$ where $\mathscr{C} \subseteq [A]^\omega$ is a $\subseteq$-chain which is closed under countable unions and is such that if X is in $\mathscr{C}$, then the convex components of $A \setminus X$ contain no first or last elements. It is routine to show that every A-line contains an $\aleph_1$-dense non stationary suborder. On the other hand, there are $2^{\aleph_1}$ isomorphism types of $\aleph_1$-dense stationary C-lines [71].

The following theorem reduced Problem 3.1 to a purely Ramsey theoretic statement about A-trees.

Theorem 3.9. *[1] (see [79, §4.4] for a proof) Assume* PFA*. The following are equivalent:*

1. *Every Aronszajn line has a Countryman suborder;*
2. *For every Aronszajn tree* T *and every partition* $T = K_0 \cup K_1$*, there is an uncountable antichain* $A \subseteq T$ *and an* $i < 2$ *such that* $s \wedge t$ *is in* K_i *for all* $s \neq t$ *in* A*;*
3. *For some Aronszajn tree* T*, if* $T = K_0 \cup K_1$ *then there is an uncountable antichain* $A \subseteq T$ *and an* $i < 2$ *such that* $s \wedge t$ *is in* K_i *for all* $s \neq t$ *in* A*.*

Progress on Problem 3.1 then stopped until [64], where a number of additional properties of A-trees were discovered, assuming PFA.

Theorem 3.10. *[64] Assume* $\mathrm{MA}_{\aleph_1}$*. If* T *is a coherent Aronszajn tree, then*

$$\mathscr{U}(T) = \{K \subseteq \omega_1 : \exists A \in \mathscr{A}(T)(\wedge(A) \subseteq K)\}$$

is an ultrafilter, where $\mathscr{A}$ *is the collection of all uncountable antichains of* T *and* $\wedge(A) = \{s \wedge t : s \neq t \in A\}$*.*

Theorem 3.11. *[64] If* $S \leq T$ *denotes the existence of a strictly increasing map from* S *into* T*, then the class of all Aronszajn trees contains a* $\leq$*-antichain of cardinality* $2^{\aleph_1}$ *and an infinite* $<$*-descending chain.*

Theorem 3.12. *[64] Assume* PFA*. The coherent Aronszajn trees are linearly ordered by* $\leq$ *without a first or last element. Furthermore* $S \leq T$ *holds if and only if there is an increasing function* $f : \omega_1 \to \omega_1$ *such that*

$$U \in \mathscr{U}(T) \text{ if and only if } f^{-1}(U) \in \mathscr{U}(S)$$

(i.e. $\beta f(\mathscr{U}(S)) = \mathscr{U}(T)$*).*

Finally, the following theorem was proved, thus solving Problem 3.1. This was accomplished by proving that PFA implies (3) of Theorem 3.9.

Theorem 3.13. *[46] Assume PFA. Every Aronszajn line contains a Countryman suborder.*

Corollary 3.14. *Assume* PFA. *If $X \subseteq \mathbb{R}$ has cardinality $\aleph_1$ and C is a Countryman line, then X, ω_1, $-\omega_1$, C, and $-C$ form a basis for the uncountable linear orders.*

In the wake of Theorem 3.13, two additional results were obtained which completely clarified our understanding of the A-lines assuming PFA.

Theorem 3.15. *[48] If C is a Countryman line, then the direct limit η_C of the alternating lexicographic products $C \times (-C) \times \cdots \times (-C)$ is universal for the class of Aronszajn lines.*

Theorem 3.16. *[42] The Aronszajn lines are well quasi-ordered by embeddability: if A_i $(i \in \mathbb{N})$ are Aronszajn lines, then there are $i < j$ such that A_i embeds into A_j.*

These results draw a strong analogy between the A-lines and the countable linear orderings: C and $-C$ play the roles of $\mathbb{N}$ and $-\mathbb{N}$ and η_C plays the role of $\mathbb{Q}$. Theorem 3.15 is analogous to Cantor's theorem that all countable dense linear orders are isomorphic; Theorem 3.16 should be compared to the following theorem of Laver.

Theorem 3.17. *[34] The countable linear orders are well quasi-ordered by embeddability.*

3.2. The Ramsey Theory of ω_1.

The study of the Ramsey theory of ω_1 has played a central role in the development of PFA (see, e.g., [70]). It was noticed early on by Sierpinski that the analog of Ramsey's theorem for ω_1 is false.

Theorem 3.18. *[58] There is a partition $[\omega_1]^2 = K_0 \cup K_1$ such that if $X \subseteq \omega_1$ is uncountable, $[X]^2 \cap K_i \neq \emptyset$ for each $i < 2$.*

This was strengthened considerably by Todorcevic, using the method of minimal walks discussed above.

Theorem 3.19. *[69] There is a partition $[\omega_1]^2 = \bigcup_{\xi<\omega_1} K_\xi$ such that if $X \subseteq \omega_1$ is uncountable, then $[X]^2 \cap K_\xi \neq \emptyset$ for each $\xi < \omega_1$.*

Still, many problems in set theory boil down to Ramsey theoretic statements about ω_1 for restricted classes of partitions or where weaker notions of homogeneity are required. Theorem 3.9 is a typical instance of this. Another important example is the reformulation of the S and L space problems in terms of Ramsey theoretic statements [51]. These problems were eventually solved with different outcomes.

Theorem 3.20. *[65] [70] Assume* PFA. *Every non Lindelöf space contains an uncountable discrete subspace.*

Theorem 3.21. *[47] There is a non separable space without an uncountable discrete subspace. Moreover, there is no basis for the uncountable topological spaces of cardinality less than* $\aleph_2$.

I will finish the section by mentioning another classification result under PFA which is closely aligned with the study of the Ramsey theory of ω_1.

Theorem 3.22. *[68] Assume* PFA. *Every directed system of cardinality at most* $\aleph_1$ *is cofinally equivalent to one of the following:* 1, ω, ω_1, $\omega \times \omega_1$, $[\omega_1]^{<\omega}$.

This classification was extended to the transitive relations on ω_1 in [73]. It is interesting to note that it is unknown whether a similar classification of relations on ω_2 is possible under any axiomatic assumptions. Such a classification would require that $2^{\aleph_0} > \aleph_2$ and in particular that PFA fails (see [68]).

4. Combinatorial Principles

While direct applications of PFA require specialized knowledge of set theory, there are an increasing number of combinatorial principles that follow from PFA which are at the same time powerful and approachable by the non specialist. Both applying these principles and isolating new and useful ones is an important theme in set-theoretic research (it should be stressed that one must always hold utility as paramount here).

Two prominent examples are the *P-Ideal Dichotomy* [76] and Todorcevic's formulation of the *Open Coloring Axiom* [70]:

PID: If S is a set and $\mathscr{I} \subseteq [S]^\omega$ is a P-ideal, then either

1. there is an uncountable $Z \subseteq S$ such that $[Z]^\omega \subseteq \mathscr{I}$ or
2. S can be covered by countably many sets in $\mathscr{I}^\perp$.

OCA: If G is a graph on a separable metric space X whose edge set is topologically open, then either

1. there is an uncountable $H \subseteq X$ such that $[H]^2 \subseteq G$ (i.e. G contains an uncountable complete subgraph) or
2. X can be covered by countably many sets Y such that $[Y]^2 \cap G = \emptyset$ (i.e. G is *countably chromatic*).

I will now present a number of typical examples of graphs and ideals to which these principles can be applied.

Example 4.1. [2] Let G be the graph on $\mathbb{R}^2$ consisting of all edges $\{(x, y), (x', y')\}$ such that $x < x'$ and $y < y'$. Observe that G is open. If X is a complete subgraph of G, then X is the graph of a partial strictly increasing function from $\mathbb{R}$ to $\mathbb{R}$. If A and B are uncountable subsets of $\mathbb{R}$, then the subgraph of G induced by $A \times B$ is never countably chromatic and therefore OCA implies that there is an uncountable partial increasing function from A to B.

Example 4.2. [70] Recall that if f and g are in $\mathbb{N}^{\mathbb{N}}$, then $f <^* g$ means that $f(i) < g(i)$ for all but finitely many i. It is well known that this is a countably directed partial order. If $f \neq g$ are in $\mathbb{N}^{\mathbb{N}}$, define $\{f, g\} \in G$ if there are i and j such that $f(i) < g(i)$ and $f(j) > g(j)$. This defines an open graph. Subsets $E \subseteq \mathbb{N}^{\mathbb{N}}$ such that $[E]^2 \cap G = \emptyset$ are quite sparse. For example such an E can not contain an uncountable $<^*$-well ordered set.

In [70, 0.7] it is shown that if $X \subseteq \mathbb{N}^{\mathbb{N}}$ consists of increasing functions and is unbounded and countably $<^*$-directed, then there are $f \neq g$ in X such that $f \leq g$ (i.e. $\{f, g\}$ is not in G). This can be used to argue that OCA implies every subset of $\mathbb{N}^{\mathbb{N}}$ of cardinality $\aleph_1$ is $<^*$-bounded. This is among the simplest applications of the phenomenon of *oscillation* which is explored further in [44], [70] and in different contexts in [47], [79].

Example 4.3. Let $\sigma\mathbb{Q}$ denote the collection of all subsets of $\mathbb{Q}$ which are well ordered in the usual order on $\mathbb{Q}$. $\sigma\mathbb{Q}$ is a tree with the order defined by $a \leq b$ if a is an initial part of b. This is a separable metric space with the topology inherited from $\mathscr{P}(\mathbb{Q})$. Let G denote the set of all pairs $\{a, b\}$ which are comparable in the tree order on $\sigma\mathbb{Q}$. This is a *closed* graph on $\sigma\mathbb{Q}$. Observe that if $H \subseteq \sigma\mathbb{Q}$ is a complete subgraph, then $\cup H$ is in $\sigma\mathbb{Q}$ and every element of H is an initial part of $\cup H$. In particular, G has no uncountable complete subgraphs. On the other hand, if $E \subseteq \sigma\mathbb{Q}$ satisfies that $[E]^2 \cap G$ is empty, then E is an antichain. Since $\sigma\mathbb{Q}$ is not a countable union of antichains [33], this example shows that the asymmetry in the statement of OCA is necessary, even for graphs on vertex sets which are nicely definable. By contrast, it is a ZFC theorem that the conclusion of OCA holds for every open graph on an analytic subset of a Polish space [21]. Furthermore, OCA holds for open graphs on projective sets as well under appropriate large cardinal assumptions.

For the next two examples, suppose that $\mathscr{J}$ is a P-ideal on a set S and ϕ_J ($J \in \mathscr{J}$) is a collection of functions such that ϕ_J is a function from J into some countable set C and whenever J and J' are in $\mathscr{J}$

$$\{s \in J \cap J' : \phi_J(s) \neq \phi_{J'}(s)\}$$

is finite. Such a family of functions is said to be *coherent*. A coherent family of functions is *trivial* if there is a single $\Phi : S \to C$ such that $\{s \in J : \phi_J(s) \neq \Phi(s)\}$ is finite for all J in $\mathscr{J}$.

Example 4.4. [70, 8.7] If S is countable, then define a graph G on the set of pairs of elements of $\mathscr{J}$ by $\{J, J'\} \in G$ if and only if there is an s in $J \cap J'$ such that $\phi_J(s) \neq \phi_{J'}(s)$. If we topologize $\mathscr{J}$ by identifying it with the subspace $\{(J, \phi_J) : J \in \mathscr{J}\}$ of $\mathscr{P}(S) \times \mathscr{P}(S \times S)$, then G is an open graph in a separable metric topology. If G is countably chromatic, then the coherent family is trivial. If $\mathscr{H} \subseteq \mathscr{I}$ is uncountable and satisfies that $[\mathscr{H}]^2 \subseteq G$, then $\mathscr{H}$ is unbounded in $(\mathscr{J}, \subseteq^*)$. Notice that any such $\mathscr{H}$ contains such a subset of cardinality $\aleph_1$ and therefore this alternative of OCA implies that $(\mathscr{J}, \subseteq^*)$ contains an unbounded

subset of cardinality $\aleph_1$. Such an $\mathscr{H}$ is quite closely related to the *obstruction* to non trivial automorphisms of $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$ mentioned in the introduction.

An important instance of this example is when $S = \mathbb{N}\times\mathbb{N}$ and $\mathscr{J} = \emptyset\times\mathrm{Fin}$. If every subset of $(\mathbb{N}^{\mathbb{N}}, <^*)$ of cardinality $\aleph_1$ is bounded (this is a consequence of OCA), then every uncountable $\mathscr{H} \subseteq \mathscr{J}$ contains an uncountable $\mathscr{H}'$ whose union is in $\mathscr{J}$. Thus OCA implies every coherent family indexed by $\emptyset \times \mathrm{Fin}$ is trivial.

Remark 4.5. In [38] it is shown that the triviality of coherent families of functions indexed by $\emptyset \times \mathrm{Fin}$ has an influence on the computation of the *strong homology* of certain locally compact subspaces of $\mathbb{R}^n$. Specifically, non-trivial coherent families indexed by $\emptyset \times \mathrm{Fin}$ coincide with the 1-cocycles in a certain cochain complex. This is used to show that, assuming CH, strong homology is not *additive* [38]. In [13] it is pointed out that PFA can be used rule out such 1-cocycles.

The existence of non-trivial n-cocyles in this cochain complex for any n, however, implies that strong homology fails to be additive [38, Theorem 8]. Unlike with 1-cocycles, very little is known what hypotheses entail the non existence of n-cocycles beyond Goblot's Vanishing Theorem (see [38]). For instance it is entirely possible that it is a theorem of ZFC that either 1-cocycles or 2-cocycles exist in this cochain complex. Additionally, while it is known that there are no $\mathscr{C}$-measurable 1-cocycles in this cochain complex, it is unclear whether the same can be said for n-cocycles for $n > 1$.

The body of work surveyed in [37, Ch. 11–14] has not yet been developed from a set-theoretic perspective (although see [62], [63], [74]). Recasting this material in set-theoretic language and developing it to the level of [79] would likely be a rewarding endeavor.

Example 4.6. [76] Given a coherent family ϕ_J ($J \in \mathscr{J}$) of functions mapping into $\{0, 1\}$, we can define $\mathscr{I}$ to be the collection of all countable $I \subseteq \mathscr{J}$ such that for some J in $\mathscr{J}$,

$$\{J' \in I : |\{s \in J \cap J' : \phi_J(s) = 0 \wedge \phi_{J'}(s) = 1\}| \leq n\}$$

is finite for each n in $\mathbb{N}$. If $\mathscr{H} \subseteq \mathscr{J}$ is uncountable and satisfies that $[\mathscr{H}]^\omega \subseteq \mathscr{I}$, then $\mathscr{H}$ is unbounded in $(\mathscr{J}, \subseteq^*)$. As noted above, this implies that $(\mathscr{J}, \subseteq^*)$ contains an unbounded subset of cardinality $\aleph_1$. If $\mathscr{J}$ is a countable union of sets in $\mathscr{I}^\perp$, then the coherent sequence is trivial.

Example 4.7. [3] Suppose that T is an ω_1-*tree* (i.e. an uncountable tree in which every level is countable). Define $\mathscr{I}$ to be the collection of all countable subsets I of T such that if t is in T, then $\{s \in I : s \leq t\}$ is finite. The assumption that the levels of T are countable implies that $\mathscr{I}$ is a P-ideal. If $Z \subseteq T$ is uncountable and $[Z]^\omega \subseteq \mathscr{I}$, then it follows that Z contains an uncountable antichain. If $T = \bigcup_n S_n$ where S_n is in $\mathscr{I}^\perp$, then it follows that T is a countable union of chains. Since neither of these alternatives is compatible with T being a Souslin tree, PID implies Souslin's Hypothesis.

Example 4.8. [76] Recall that if κ is a regular cardinal, then $\square(\kappa)$ is the assertion that there is a sequence $\langle C_\alpha : \alpha < \kappa \rangle$ with the following properties:

1. $C_\alpha \subseteq \alpha$ is closed and unbounded for each $\alpha < \kappa$ and $C_{\alpha+1} = \{\alpha\}$;
2. if α is a limit point of C_β, then $C_\alpha = C_\beta \cap \alpha$;
3. there is no closed unbounded $C \subseteq \kappa$ such that for every limit point α of C, $C_\alpha = C \cap \alpha$.

As in Section 3.1, we can define $\varrho_2 : [\kappa]^2 \to \omega$ using a $\square(\kappa)$-sequence: $\varrho_2(\alpha, \beta)$ is the length of the walk from β down to α. If $\beta < \kappa$ and $n \in \omega$, set

$$K_{\beta,n} = \{\alpha < \beta : \varrho_2(\alpha, \beta) \leq n\}.$$

One can argue that if $\mathscr{I}$ is the collection of all countable I which have finite intersection with every $K_{\beta,n}$, then $\mathscr{I}$ is a P-ideal which does not satisfy either alternative of PID. In fact, κ is not the union of countably many sets in $\mathscr{I}^\perp$, even though each $\beta < \kappa$ has this property (as witnessed by $\{K_{\beta,n}\}_n$). The failure of $\square(\kappa)$ for all κ is known to have considerable large cardinal strength (see [53]).

In fact the properties of the family $\mathscr{K} = \{K_{\beta,n} : (\beta < \kappa^+) \wedge (n < \omega)\}$ which violate PID can be abstracted so as to be applied to more general situations. For instance this argument can be adapted to prove that PID implies that $2^\mu = \mu^+$ whenever μ is a *singular strong limit cardinal* [84].

5. Proper Forcings and How to Construct Them

We will now turn to the task of formulating PFA. Recall that a *forcing* is a partial order Q with a greatest element. Elements of a forcing are generally referred to as *conditions* and $q \leq p$ is generally taken to mean q is an *extension* of p. Two conditions are *compatible* if they have a common extension and *incompatible* otherwise. A *filter* is a collection of conditions which is upward closed and downward directed. An *antichain* is a collection of pairwise incompatible conditions. A forcing Q is *c.c.c.* if every antichain is countable.

A completely general example of a forcing is the collection of non empty open sets in a compact topological space, with $U \leq V$ defined to mean that $\overline{U} \subseteq V$. In this setting, points correspond to maximal filters and antichains are families of pairwise disjoint open sets. If U is dense and open in a topological space, then U is the union of a maximal antichain $\mathscr{A}$ of open sets V such that $V \leq U$. This allows one to translate forcing axioms into statements about Baire category.

Now we turn to formulating *properness*, which is a weakening of being *c.c.c.*. Unless specified otherwise, θ will always be used to denote a regular uncountable cardinal. Recall that $H(\theta)$ is the collection of all sets of hereditary cardinality at most θ. In this case $(H(\theta), \in)$ satisfies all of the axioms of ZFC except possibly

the powerset axiom. $M \subseteq H(\theta)$ is an *elementary submodel* of $H(\theta)$ if whenever $\phi(x_1, \ldots, x_n)$ is a formula in the language of set theory and $a_1, \ldots, a_n$ are in M, $(M, \in)$ satisfies $\phi(a_1, \ldots, a_n)$ if and only if $(H(\theta), \in)$ satisfies $\phi(a_1, \ldots, a_n)$.

If Q is a forcing, then a *suitable model* for Q is a countable elementary submodel of $H(\theta)$ for some θ such that $\mathscr{P}(Q)$ is in M. If M is a suitable model for Q, then a condition in q is (M, Q)-*generic* if whenever $A \subseteq Q$ is a maximal antichain which is in M, every extension of q is compatible with an element of $A \cap M$. Finally, Q is *proper* if whenever M is a suitable model for Q, every condition in $Q \cap M$ has an (M, Q)-generic extension. We are now in a position to understand the formulation of PFA given in the introduction:

> If Q is a proper forcing and $\mathscr{A}$ is a collection of maximal antichains in Q with $|\mathscr{A}| \leq \aleph_1$, then there is a filter $G \subseteq Q$ such that $G \cap A \neq \emptyset$ for every A in $\mathscr{A}$.

Thus PFA is just the statement obtained by replacing "c.c.c." by "proper" in the formulation of $\mathrm{MA}_{\aleph_1}$. It is not difficult to verify that in fact every c.c.c. forcing is proper and hence that PFA implies $\mathrm{MA}_{\aleph_1}$. While proper forcings necessarily preserve uncountability, they may collapse cardinals above $\aleph_1$. To a large extent, this is where PFA derives its additional strength.

In situations where there is a need to apply PFA directly, Todorcevic has developed a general approach for building proper forcings to accomplish a given task such as introducing an uncountable complete subgraph to a given graph or an embedding between two structures. This method was introduced in [66] and further detailed in [70] and [73]. Typically the conditions in the forcing Q consist of pairs $q = (X_q, \mathscr{N}_q)$ where X_q is a finite approximation of the desired object and $\mathscr{N}_q$ is a finite $\in$-chain of elementary substructures of some $(H(\theta), \in)$ for θ suitably large. In all cases, there are additional requirements placed on the pairs which are specific to the application at hand. One verifies properness by proving that if M is a suitable model for Q and $M \cap H(\theta)$ is in $\mathscr{N}_q$, then q is (M, Q)-generic. In situations in which this construction results in a proper forcing, the forcing Q can usually be regarded as a two step iteration of a forcing which collapses $|H(\theta)|$ to $\aleph_1$ by covering it with an $\in$-chain of countable substructures, followed by a *c.c.c.* forcing of finite approximations to the desired object.

I will illustrate this method of construction by defining forcings which can be used to show that PFA implies OCA and PID. These examples are relatively simple in terms of the interaction between the finite working part and the chain of models. Still, they contain all of the important features of other examples built using these methods.

5.1. The OCA forcing. Let G be a fixed open graph on a separable metric space X and let $\mathscr{E}$ denote the collection of all $E \subseteq X$ such that $[E]^2 \cap G = \emptyset$. Define Q_G to be the collection of all pairs $q = (H_q, \mathscr{N}_q)$ such that:

1. $H_q \subseteq X$ is finite and $[H_q]^2 \subseteq G$;

2. $\mathscr{N}_q$ is a finite $\in$-chain of countable elementary submodels of $H(2^{\aleph_0{}^+})$, each containing X and G;
3. if $x \neq y$ are in H_q, then there is an N in $\mathscr{N}_q$ such that $|N \cap \{x, y\}| = 1$ (i.e. $\mathscr{N}_q$ *separates* H_q);
4. if N is in $\mathscr{N}_q$ and x is in $H_q \setminus N$, then x is not in E for any E in $\mathscr{E} \cap N$.

The order on Q_G is defined by $q \leq p$ if $H_p \subseteq H_q$ and $\mathscr{N}_p \subseteq \mathscr{N}_q$.

The following is the key lemma in establishing the properness of this forcing.

Lemma 5.1. *Suppose that N_i $(i < k)$ is a finite $\in$-chain of suitable models for X and G and that x is an element of X^k such that if $i < k$, then x_i is not an element of any E in $\mathscr{E} \cap N_i$ and x_i is in N_{i+1} if $i < k-1$. If $D \subseteq X^k$ is an element of N_0 which has x as an accumulation point, then there is an open $U \subseteq X$ in N_0 satisfying:*

- *$x(k-1) \notin \overline{U}$ and $\{x(k-1), y\}$ is in G whenever y is in U;*
- *$\{y \restriction k-1 : (y \in D) \wedge (y(k-1) \in U)\}$ accumulates to $x \restriction k-1$.*

5.2. The PID forcing. We will now turn to a class of forcings which can be used to force instances of PID. Suppose that $\mathscr{I}$ is a P-ideal on a set S. Let θ be sufficiently large such that $\mathscr{I}$ is in $H(\theta)$ and for each countable $N \prec H(\theta)$, let I_N be an element of $\mathscr{I}$ such that $I \subseteq^* I_N$ whenever I is in $\mathscr{I} \cap N$ (this is possible since N is countable and $\mathscr{I}$ is a P-ideal). Define $Q_{\mathscr{I}}$ to be the collection of all pairs $q = (Z_q, \mathscr{N}_q)$ such that:

1. $Z_q \subseteq S$ is finite;
2. $\mathscr{N}_q$ is a finite $\in$-chain of suitable models for $\mathscr{I}$ which separates Z_q;
3. if N is in $\mathscr{N}_q$ and x is in $Z_q \setminus N$, then x is not in J for any J in $\mathscr{I}^\perp \cap N$.

The order on $Q_{\mathscr{I}}$ is slightly more complicated than in the case of Q_G. Define $q \leq p$ if $Z_p \subseteq Z_q$, $\mathscr{N}_p \subseteq \mathscr{N}_q$, and whenever N is in $\mathscr{N}_p$

$$N \cap (Z_q \setminus Z_p) \subseteq I_N.$$

This last condition ensures that if $G \subseteq Q_{\mathscr{I}}$ is a filter, then every countable subset of $\bigcup_{q \in G} Z_q$ is in $\mathscr{I}$.

The following is the key combinatorial lemma which is used in the proof that $Q_{\mathscr{I}}$ is proper (see [73, 7.8]).

Lemma 5.2. *Suppose that $\mathscr{J}$ is a σ-ideal on a set S, N_i $(i < k)$ is a finite $\in$-chain of suitable models for $\mathscr{J}$, and x is in S^k such that x_i is not in any element of $\mathscr{J} \cap N_i$ and x_i is in N_{i+1} if $i < k-1$. If $D \subseteq S^k$ is in N_0 and contains x, then there is a $T \subseteq D$ in N_0 which contains x and is $\mathscr{J}^+$-splitting:*

$$\{x \in S : \exists t \in T((u \restriction i)\hat{\ }x \subseteq t)\}$$

is not in $\mathscr{J}$ whenever u is in T and $i < k$.

6. Some Applications of PFA

I will now mention some applications PFA. The focus will be on applications outside of set theory and on those which are more recent. Two other applications of note are Shelah's solution to Whitehead's Problem [54] (which required only $\mathrm{MA}_{\aleph_1}$) and Woodin's resolution of Kaplanski's Conjecture concerning automatic continuity of homomorphisms of $C([0,1])$ into commutative Banach algebras [86]. In addition, an extensive list of applications of $\mathrm{MA}_{\aleph_1}$ can be found in [23].

6.1. Automorphisms of the Calkin algebra. Let H be a separable infinite dimensional Hilbert space and let $\mathscr{B}(H)$ and $\mathscr{K}(H)$ be the bounded and compact operators on H, respectively. The *Calkin algebra* is the quotient $\mathscr{C}(H) = \mathscr{B}(H)/\mathscr{K}(H)$, regarded as a C^*-algebra.

Every unitary operator in $\mathscr{C}(H)$ gives rise to an automorphism of $\mathscr{C}(H)$ via conjugation; such automorphisms are said to be *inner*. In [11], Brown, Douglas, and Filmore asked whether there are any other automorphisms of $\mathscr{C}(H)$. This turns out to be independent of ZFC:

Theorem 6.1. *[50] Assume CH. There is an outer automorphism of $\mathscr{C}(H)$.*

Theorem 6.2. *[16] Assume OCA. Every automorphism of $\mathscr{C}(H)$ is inner.*

At the core of Farah's proof of Theorem 6.2 is the construction and the analysis of *coherent families of unitaries* which are derived from a given automorphism of $\mathscr{C}(H)$. Such families are analogs of the coherent families of functions from Example 4.4.

Theorem 6.2 is a new direction in a natural progression of theorems concerning automorphisms and homomorphisms of quotient structures which began with Shelah's work on the automorphism group of $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$ in [56, IV]. The reader is referred to [17], [18] for a detailed account of the work in this area prior to [16]. Also Farah, Weaver, and others have recently begun an investigation into how PFA and other set-theoretic hypotheses and methods can be applied to operator algebras; see [20], [85].

6.2. Bases in quotients of Banach spaces. The following problem in Banach space theory has its roots in Banach's original monograph [6] (the problem appears explicitly only sometime later; see, e.g., [49]).

Problem 6.3. *Does every infinite dimensional Banach space have an infinite dimensional quotient with a basis?*

Johnson and Rosenthal proved that the answer to this problem is positive in the class of separable Banach spaces [28]. Whether it is true in general has become known as the *Separable Quotient Problem* (so called because it is equivalent to asking whether every infinite dimensional Banach space has an infinite dimensional separable quotient). In fact, the proof of [28] yields the following stronger result.

Theorem 6.4. *Assume that every subset of* $\mathbb{N}^{\mathbb{N}}$ *of cardinality at most* θ *is* $<^*$*-bounded. Every Banach space of density at most* θ *has an infinite dimensional quotient with a basis.*

In this vein it is also natural to ask whether a non separable Banach space has a non separable quotient with a basis. This question was addressed in part by the following result.

Theorem 6.5. *[78] Assume* $\mathrm{MA}_{\aleph_1}$ *and* PID*. Every Banach space of density* $\aleph_1$ *has a quotient with a basis of length* ω_1*.*

6.3. Von Neumann's problem on the existence of strictly positive measures. Given a complete Boolean algebra $\mathscr{B}$, it is natural to ask under what circumstances $\mathscr{B}$ admits a strictly positive probability measure. Two necessary requirements are that $\mathscr{B}$ be *c.c.c.* and that it be *weakly distributive*. Von Neumann asked whether these conditions are also sufficient.

Problem 6.6. *[43, Problem 163] Does every complete Boolean algebra which is* c.c.c. *and weakly distributive necessarily support a strictly positive measure?*

A positive answer implies Souslin's Hypothesis and therefore is not provable in ZFC [36]. Maharam divided von Neumann's problem into two complementary problems.

Problem 6.7. *[36] Does every weakly distributive* c.c.c. *complete Boolean algebra support a strictly positive continuous submeasure?*

Problem 6.8. *[36] Does every complete Boolean algebra equipped with a strictly positive continuous submeasure admit a strictly positive measure?*

This division was significant in part because it was possible to show that, unlike Souslin's Hypothesis, the answer to Problem 6.8 could not be changed by forcing and therefore was unlikely to be independent of ZFC. This is analogous to the division of Theorem 1.2 discussed in Section 8 below.

Recently two results completely resolved the situation.

Theorem 6.9. *[5] Assume PID. If* $\mathscr{B}$ *is a complete Boolean algebra which is c.c.c. and weakly distributive, then* $\mathscr{B}$ *supports a strictly positive continuous submeasure.*

Theorem 6.10. *[61] There is a complete Boolean algebra supporting a strictly positive continuous submeasure which does not support a measure.*

This application of PFA also demonstrates the merits of its large cardinal strength. While the conclusion of Theorem 6.9 does not apparently have any relationship to large cardinals, it was demonstrated after the fact that the conclusion of Theorem 6.9 does entail the existence of an inner model which satisfies a large cardinal hypothesis.

Theorem 6.11. *[19] Assume that every complete Boolean algebra which is c.c.c. and weakly distributive necessarily supports a strictly positive continuous submeasure. Then there is an inner model with a measurable cardinal κ such that $o(\kappa) = \kappa^{++}$.*

6.4. The determinacy of Gale-Stewart games. An application of PFA of a rather different nature is derived entirely through its consistency strength. Recall that in a *Gale-Stewart game*, two players play natural numbers alternately, resulting in an infinite sequence n_i $(i < \infty)$ of elements of $\mathbb{N}$. The winner of the game is determined based on whether the resulting sequence is in a predetermined set $\Gamma \subseteq \mathbb{N}^{\mathbb{N}}$. The principle question, in this level of abstraction, is under what circumstances such a game is *determined* — i.e. when does one of the two players have a strategy to win the game? The Axiom of Choice implies that there are sets $\Gamma \subseteq \mathbb{N}^{\mathbb{N}}$ which specify undetermined games. On the other hand, by a classical theorem of Gale and Stewart, closed games are determined.

The interest in such games arises from the fact that the regularity properties of subsets of $\mathbb{R}^n$ — such as Lebesgue measurability and the Baire Property — can be reformulated in terms of the determinacy of games (see [30, §20-21]). The assertion that the conclusion of OCA holds for open graphs on a given set of reals X can also be regarded as a regularity property of X and has a corresponding game associated to it [21]. In fact the determinacy of games for a *point class* has come to be regarded as the ultimate form of a regularity property. The first major success in understanding which games could be determined was the following result.

Theorem 6.12. *[39] Assume there is a measurable cardinal. Then every analytic game is determined.*

With a considerably more complicated proof, it was possible to prove Borel determinacy within ZFC.

Theorem 6.13. *[40] Every Borel game is determined.*

Unlike Borel games, however, the determinacy of analytic games does require a large cardinal assumption (see [29, §31]).

While there are natural examples of definable subsets of Polish spaces which are not Borel (see [8]), all simply definable sets tend to be *projective*. Here the class of projective sets in a Polish space X is the smallest algebra of subsets of X which contain the open sets and which is closed under continuous images. In a major breakthrough, Martin and Steel were able to prove projective determinacy from what turned out to be an optimal large cardinal hypothesis.

Theorem 6.14. *[41] If there are infinitely many Woodin cardinals, then all projective games are determined.*

While PFA does not imply the existence of large cardinals, it does entail the existence of inner models which satisfy substantial large cardinal hypotheses. This allowed for the proof of the following result.

Theorem 6.15. *[60] Assume* PFA*. The inner model $L(\mathbb{R})$ satisfies that all sets $\Gamma \subseteq \mathbb{N}^{\mathbb{N}}$ are determined. In particular, all projective sets are Lebesgue measurable and have the Baire Property.*

7. The Role of $2^{\aleph_0} = \aleph_2$

One of the important early results on PFA was that it implies $2^{\aleph_0} = \aleph_2$ [9] [82]. This is significant in part because it provides a natural limitation to the number of maximal antichains one can expect to meet in a proper forcing.[1] Since then a number of different proofs have been given that PFA implies $2^{\aleph_0} = \aleph_2$ [12] [44] [45]. In each case new ideas where required which were of independent interest. The most significant example of this is the isolation of the principle MRP in [45] which in turn played a key role in the solution of the basis problem for the uncountable linear orders [46] and which has since found other applications [12] [83].

What is clear from experience is that in order to prove structural results at the level of $\aleph_1$, one must deal with combinatorics similar to that involved in proofs that $2^{\aleph_0} = \aleph_2$. What is less clear is to what extent this connection can be made more explicit.

Problem 7.1. *Is there a consistent classification of structures of cardinality $\aleph_1$ which implies $2^{\aleph_0} = \aleph_2$?*

The classification of A-lines presented in Section 3.1 provides an intriguing test question. It is also an open problem whether the combinatorial principles presented in Section 4 already entail that $2^{\aleph_0} \leq \aleph_2$. (While OCA implies $\mathfrak{b} = \aleph_2$, it is known that PID is consistent with CH, relative to the existence of a supercompact cardinal [76].)

Problem 7.2. *Does* OCA *imply $2^{\aleph_0} = \aleph_2$?*

Problem 7.3. *Does* PID *imply $2^{\aleph_0} \leq \aleph_2$?*

Both OCA and PID can be used to classify gaps and therefore do imply that $\mathfrak{b} \leq \aleph_2$. Recall that a pair of sequences f_ξ $(\xi < \kappa)$, g_η $(\eta < \lambda)$ form a (κ, λ^*)-*gap* in $\mathbb{N}^{\mathbb{N}}/\mathrm{Fin}$ if:

- whenever $\xi < \xi' < \kappa$ and $\eta < \eta' < \lambda$, then $f_\xi <^* f_{\xi'} <^* g_{\eta'} <^* g_\eta$ and
- there does not exist an h in $\mathbb{N}^{\mathbb{N}}$ such that if $\xi < \kappa$ and $\eta < \lambda$, then $f_\xi <^* h <^* g_\eta$.

[1] It was known before the proof that PFA implies $2^{\aleph_0} = \aleph_2$ that $\aleph_1$ can not be replaced by $\aleph_2$ in the formulation of PFA. It had also already been known that the stronger MM implies $2^{\aleph_0} = \aleph_2$ [22].

Theorem 7.4. *[27] There is an (ω_1, ω_1^*)-gap.*

Theorem 7.5. *[27] The following are equivalent for a regular cardinal κ:*

- *There is a (κ, ω^*)-gap.*
- *There is an (ω, κ^*)-gap.*
- *There is an unbounded chain in $(\mathbb{N}^{\mathbb{N}}, <^*)$ of ordertype κ.*

Theorem 7.6. *[70] [76] Assume either* OCA *or* PID*. If κ and λ are regular cardinals and there is a (κ, λ^*)-gap, then either $\kappa = \omega$, $\lambda = \omega$, or $\kappa = \lambda = \omega_1$. In particular, $\mathfrak{b} \leq \aleph_2$.*

In [44], it was shown that the conjunction of OCA and the initial formulation of OCA presented in [2] does imply $2^{\aleph_0} = \aleph_2$.

8. The Role of PFA in Proving Theorems in ZFC

One of the remarkable features of the study of forcing axioms and their consequences is that one often obtains ZFC theorems of independent interest as byproducts. One instance of this is the following result which is implicit in Shelah's original proof of the consistency of the conclusion of Theorem 1.2 [56, IV], but which was first made explicit in [81].

Theorem 8.1. *If Φ is an automorphism of $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$, then either Φ is induced by a map $\phi : \mathbb{N} \to \mathbb{N}$ or else Φ does not have a $\mathscr{C}$-measurable lifting.*

We also have the following analogous result for the Calkin algebra.

Theorem 8.2. *[16] If Φ is an automorphism of $\mathscr{C}(H)$, then either Φ is inner or else Φ does not have a $\mathscr{C}$-measurable lifting.*

This is part of a more general phenomenon: one can show in ZFC that certain objects or morphisms must fail to have nice regularity properties and PFA can then be used to build regularity properties into such objects or morphisms. For instance, Theorem 1.2 can be viewed as the combination of Theorem 8.1 above and the following theorem.

Theorem 8.3. *Assume* PFA*. If Φ is an automorphism of $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$, then Φ has a $\mathscr{C}$-measurable lifting.*

The reader is referred to [18] for a detailed discussion of this phenomenon in quotients.

The following *Analytic Gap Theorem* was directly inspired by the influence of OCA on gaps in $\mathbb{N}^{\mathbb{N}}/\mathrm{Fin}$ and also closely parallels the formulation of PID.

It says that the pair $\mathscr{A} = \emptyset \times \mathrm{Fin}$, $\mathscr{B} = \mathrm{Fin} \times \emptyset$ is essentially the only analytic gap occurring in $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$.

Theorem 8.4. *[72] Suppose that $\mathscr{A} \subseteq \mathscr{P}(\mathbb{N})$ is analytic and closed under taking subsets. If $\mathscr{B} \subseteq \mathscr{A}^{\perp}$ then either there is a countable $\mathscr{A}_0 \subseteq \mathscr{B}^{\perp}$ such that every element of $\mathscr{A}$ is contained in an element of $\mathscr{A}_0$, or else there is tree $T \subseteq \mathbb{N}^{<\omega}$ such that*

1. *if t is in T, then $\{i \in \mathbb{N} : t\hat{\ }i \in T\}$ is an infinite element of $\mathscr{B}$ and*
2. *every branch through T is an element of $\mathscr{A}$.*

Remark 8.5. While there are many similarities between $\mathscr{P}(\mathbb{N})/\mathrm{Fin}$ and $\mathscr{C}(H)$, there are important differences as well. For instance recent work of Zamora-Aviles [88] shows that there are analytic gaps in $\mathscr{C}(H)$ in which both sides are countably directed (in an appropriate analog of $\subseteq^*$).

One application of this theorem is the following result concerning the metrizability of separable Fréchet groups.

Theorem 8.6. *[80] Suppose that G is a countable topological group which is Fréchet. If the topology on G is analytic as a subset of $\mathscr{P}(G)$, then G is metrizable.*

The Ramsey theoretic approach to applications of set theory which developed simultaneously with the theory of PFA also played a role in the results of [75].

Theorem 8.7. *[75] Suppose that K is a compact subset of the Baire class 1 functions on a Polish space X. The following are true:*

1. *K contains a dense metrizable subspace. In particular if K satisfies the countable chain condition, then it is separable.*
2. *If K does not contain an uncountable discrete subspace, then K admits an at most 2-to-1 map onto a compact metric space.*
3. *If K is non metrizable, then either K contains an uncountable discrete subspace or else K contains a homeomorphic copy of $[0,1] \times \{0,1\}$ with the interval topology.*
4. *If K is separable and x is a point in K, then either x has a countable neighborhood base or else there is a discrete subset of K of cardinality $2^{\aleph_0}$ which has x as its unique accumulation point.*

The Analytic Gap Theorem is especially important in the proof of (4), where it is used to bring the Ramsey theory of perfect sets of reals into this context. This has been further exploited in the following result which solves a special case of the Separable Quotient Problem.

Theorem 8.8. *[4] If X is an infinite dimensional Banach space, then X^* has an infinite dimensional separable quotient.*

In some cases, whether these results can be generalized to arbitrary compact spaces in the presence of PFA remains open (see [25] for a survey of related problems, including Problem 9.6 below).

Problem 8.9. *[24] Assume* PFA*. If K is compact and does not contain an uncountable discrete subspace, must K admit an at most 2-to-1 map onto a metric space?*

Finally, I will mention the following effective analog of Theorem 6.9 above.

Theorem 8.10. *[77] If a complete Boolean algebra satisfies the σ-bounded chain condition and is weakly distributive, then it supports a strictly positive continuous submeasure.*

9. Open Problems

In closing, I have collected a number of open problems. When possible I have included a reference to either recent progress or a survey of the problem.

Problem 9.1. *(Efimov [15]; see [26]) Is it consistent that every infinite compact space contains either a convergent sequence or a copy of $\beta\mathbb{N}$? Does this follow from* PFA*?*

Problem 9.2. *(Todorcevic; see [18]) Assume PFA. If $\mathscr{I}$ and $\mathscr{J}$ are analytic ideals on $\mathbb{N}$ such that $\mathscr{P}(\mathbb{N})/\mathscr{I} \simeq \mathscr{P}(\mathbb{N})/\mathscr{J}$, must the isomorphism be induced by a map $\phi : \mathbb{N} \to \mathbb{N}$?*

Problem 9.3. *(Todorcevic; see [44]) Does either OCA or PID imply $2^{\aleph_0} \leq \aleph_2$?*

Problem 9.4. *(Moore [48]) Suppose the following are true: (a) every two $\aleph_1$-dense non-stationary Countryman lines are isomorphic or reverse isomorphic, (b) every Aronszajn line can be embedded into η_C, and (c) the Aronszajn lines are well quasi-ordered. Does it follow that $2^{\aleph_0} = \aleph_2$?*

Problem 9.5. *(see [25] [35]) Assume* PFA*. If a compact convex set does not contain an uncountable discrete subspace, must it be metrizable?*

Problem 9.6. *(Gruenhage [24]; see [25]) Assume* PFA*. Do the uncountable first countable spaces have a three element basis consisting of a set of reals of cardinality $\aleph_1$ with the separable metric, the Sorgenfrey, and the discrete topologies?*

Problem 9.7. *[49] Does every infinite dimensional Banach space have an infinite dimensional quotient with a basis?*

Problem 9.8. *(Todorcevic [73]) Is there a consistent classification of the cofinal types of directed sets of cardinality at most $\aleph_2$ which is comparable to the classification of directed sets of cardinality at most $\aleph_1$ given in [68]?*

Problem 9.9. *(see [13] [38] [37]) Is it consistent that strong homology is additive?*

References

[1] U. Abraham, S. Shelah. Isomorphism types of Aronszajn trees. *Israel J. Math.*, 50(1-2):75–113, 1985.

[2] U. Abraham, M. Rubin, S. Shelah. On the consistency of some partition theorems for continuous colorings, and the structure of $\aleph_1$-dense real order types. *Ann. Pure Appl. Logic*, 29(2):123–206, 1985.

[3] U. Abraham, S. Todorcevic. Partition properties of ω_1 compatible with CH. *Fund. Math.*, 152(2):165–181, 1997.

[4] S. A. Argyros, P. Dodos, V. Kanellopoulos. Unconditional families in Banach spaces. *Math. Ann.*, 341(1):15–38, 2008.

[5] B. Balcar, T. Jech,, T. Pazák. Complete CCC Boolean algebras, the order sequential topology, and a problem of von Neumann. *Bull. London Math. Soc.*, 37(6):885–898, 2005.

[6] S. Banach. *Théorie des Opérations Linéaires.* Warszawa, 1932.

[7] J. E. Baumgartner. All $\aleph_1$-dense sets of reals can be isomorphic. *Fund. Math.*, 79(2):101–106, 1973.

[8] H. Becker. Descriptive set-theoretic phenomena in analysis and topology. In *Set theory of the continuum (Berkeley, CA, 1989)*, volume 26 of *Math. Sci. Res. Inst. Publ.*, pages 1–25. Springer, New York, 1992.

[9] M. Bekkali. *Topics in Set Theory.* Springer-Verlag, Berlin, 1991. Lebesgue measurability, large cardinals, forcing axioms, ρ-functions, Notes on lectures by Stevo Todorčević.

[10] M. Bell, J. Ginsburg, S. Todorcevic. Countable spread of $\exp Y$ and λY. *Topology Appl.*, 14(1):1–12, 1982.

[11] L. G. Brown, R. G. Douglas, P. A. Fillmore. Extensions of C^*-algebras and K-homology. *Ann. of Math. (2)*, 105(2):265–324, 1977.

[12] A. E. Caicedo, B. Veličković. The bounded proper forcing axiom and well orderings of the reals. *Mathematical Research Letters*, 13(3):393–408, 2006.

[13] A. Dow, P. Simon, J. E. Vaughan. Strong homology and the proper forcing axiom. *Proc. Amer. Math. Soc.*, 106(3):821–828, 1989.

[14] B. Dushnik, E. W. Miller. Concerning similarity transformations of linearly ordered sets. *Bull. Amer. Math. Soc.*, 46:322–326, 1940.

[15] B. Efimov. The imbedding of the Stone-Čech compactifications of discrete spaces into bicompacta. *Dokl. Akad. Nauk SSSR*, 189:244–246, 1969.

[16] I. Farah. All automorphisms of the Calkin algebra are inner. ArXiv preprint 0705.3085v9 (9/29/2009).

[17] ———. Analytic quotients: theory of liftings for quotients over analytic ideals on the integers. *Mem. Amer. Math. Soc.*, 148(702), 2000.

[18] ———. Rigidity conjectures. In *Logic Colloquium 2000*, volume 19 of *Lect. Notes Log.*, pages 252–271. Assoc. Symbol. Logic, Urbana, IL, 2005.

[19] I. Farah, B. Veličković. von Neumann's problem and large cardinals. *Bull. London Math. Soc.*, 38(6):907–912, 2006.

[20] I. Farah, E. Wofesy. Set theory and operator algebras. Lecture notes from Appalacian Set Theory Workshop, Feb. 2008.

[21] Q. Feng. Homogeneity for open partitions of pairs of reals. *Trans. Amer. Math. Soc.*, 339(2):659–684, 1993.

[22] M. Foreman, M. Magidor, S. Shelah. Martin's Maximum, saturated ideals, and nonregular ultrafilters. I. *Ann. of Math. (2)*, 127(1):1–47, 1988.

[23] D. H. Fremlin. *Consequences Of Martin's Axiom*. Cambridge University Press, 1984.

[24] G. Gruenhage. Perfectly normal compacta, cosmic spaces, and some partition problems. In *Open Problems in Topology*, pages 85–95. North-Holland, Amsterdam, 1990.

[25] G. Gruenhage, J. Tatch Moore. Perfect compacta and basis problems in topology. In Elliott Pearl, editor, *Open Problems in Topology II*. Elsevier, 2007.

[26] K. P. Hart. Efimov's problem. In Elliott Pearl, editor, *Open Problems in Topology II*. Elsevier, 2007.

[27] F. Hausdorff. Die graduierung nach dem endverlauf. *Abhandlun. König. Sc̈hsis. Gessellsch. Wissenschaften, Math.-Phys. Kl.*, pages 296–334, 1909.

[28] W. B. Johnson, H. P. Rosenthal. On ω^*-basic sequences and their applications to the study of Banach spaces. *Studia Math.*, 43:77–92, 1972.

[29] A. Kanamori. *The Higher Infinite*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, second edition, 2003. Large cardinals in set theory from their beginnings.

[30] A. S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[31] B. König, Y. Yoshinobu. Kurepa trees and Namba forcing. preprint, 2005.

[32] K. Kunen. *An Introduction to Independence Proofs*, volume 102 of *Studies in Logic and the Foundations of Mathematics*. North-Holland, 1983.

[33] Dj. Kurepa. Sur les fonctins réels dans la famille des ensembles bien ordonnés de nombres rationells. *Bull. Int. Acad. Yougoslave*, 4:35–42, 1954.

[34] R. Laver. On Fraïssé's order type conjecture. *Ann. of Math. (2)*, 93:89–111, 1971.

[35] J. Lopez-Abad, S. Todorcevic. Generic Banach spaces and generic simplices. preprint 3/2010.

[36] D. Maharam. An algebraic characterization of measure algebras. *Ann. of Math. (2)*, 48:154–167, 1947.

[37] S. Mardešić. *Strong shape and homology*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2000.

[38] S. Mardešić, A. V. Prasolov. Strong homology is not additive. *Trans. Amer. Math. Soc.*, 307(2):725–744, 1988.

[39] D. A. Martin. Measurable cardinals and analytic games. *Fund. Math.*, 66:287–291, 1969/1970.

[40] ———. Borel determinacy. *Ann. of Math. (2)*, 102(2):363–371, 1975.

[41] D. A. Martin, J. R. Steel. A proof of projective determinacy. *J. Amer. Math. Soc.*, 2(1):71–125, 1989.

[42] C. Martinez. It is consistent that the Aronszajn lines are well quasi-ordered. preprint, 2009.

[43] D. Mauldin, editor. *The Scottish Book*. Birkhäuser, 1981.

[44] J. Tatch Moore. Open colorings, the continuum and the second uncountable cardinal. *Proc. Amer. Math. Soc.*, 130(9):2753–2759, 2002.

[45] ———. Set mapping reflection. *J. Math. Log.*, 5(1):87–97, 2005.

[46] ———. A five element basis for the uncountable linear orders. *Ann. of Math. (2)*, 163(2):669–688, 2006.

[47] ———. A solution to the L space problem. *Jour. Amer. Math. Soc.*, 19(3):717–736, 2006.

[48] ———. A universal Aronszajn line. *Math. Res. Lett.*, 16(1):121–131, 2009.

[49] A. Pełczyński. Some problems on bases in Banach and Fréchet spaces. *Israel J. Math.*, 2:132–138, 1964.

[50] N. C. Phillips, N. Weaver. The Calkin algebra has outer automorphisms. *Duke Math. J.*, 139(1):185–202, 2007.

[51] J. Roitman. A reformulation of S and L. *Proc. Amer. Math. Soc.*, 69(2):344–348, 1978.

[52] W. Rudin. Homogeneity problems in the theory of Čech compactifications. *Duke Math. J.*, 23:409–419, 1956.

[53] E. Schimmerling, M. Zeman. Square in core models. *Bull. Symbolic Logic*, 7(3):305–314, 2001.

[54] S. Shelah. Infinite abelian groups, Whitehead problem and some constructions. *Israel J. Math.*, 18:243–256, 1974.

[55] ———. Decomposing uncountable squares to countably many chains. *J. Combinatorial Theory Ser. A*, 21(1):110–114, 1976.

[56] ———. *Proper forcing*, volume 940 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1982.

[57] S. Shelah, J. Steprāns. PFA implies all automorphisms are trivial. *Proc. Amer. Math. Soc.*, 104(4):1220–1225, 1988.

[58] W. Sierpiński. Sur un problème de la thèorie des relations. *Ann. Scuola Norm. Sup. Pisa*, 2(2), 1933.

[59] R. Solovay, S. Tennenbaum. Iterated Cohen extensions and Souslin's problem. *Ann. of Math.*, 94:201–245, 1971.

[60] J. R. Steel. PFA implies $\mathrm{AD}^{L(\mathbb{R})}$. *J. Symbolic Logic*, 70(4):1255–1296, 2005.

[61] M. Talagrand. Maharam's problem. *Ann. of Math. (2)*, 168(3):981–1009, 2008.

[62] D. Talayco. Applications of cohomology to set theory. I. Hausdorff gaps. *Ann. Pure Appl. Logic*, 71(1):69–106, 1995.

[63] D. Talayco. Applications of cohomology to set theory. II. Todorčević trees. *Ann. Pure Appl. Logic*, 77(3):279–299, 1996.

[64] S. Todorcevic. Lipschitz maps on trees. *J. Inst. Math. Jussieu*, 6(3):527–556, 2007.

[65] _____. Forcing positive partition relations. *Trans. Amer. Math. Soc.*, 280(2):703–720, 1983.

[66] _____. A note on the proper forcing axiom. In *Axiomatic set theory (Boulder, Colo., 1983)*, volume 31 of *Contemp. Math.*, pages 209–218. Amer. Math. Soc., Providence, RI, 1984.

[67] _____. Trees and linearly ordered sets. In *Handbook of set-theoretic topology*, pages 235–293. North-Holland, Amsterdam, 1984.

[68] _____. Directed sets and cofinal types. *Trans. Amer. Math. Soc.*, 209:711–723, 1985.

[69] _____. Partitioning pairs of countable ordinals. *Acta Math.*, 159(3–4):261–294, 1987.

[70] _____. *Partition Problems In Topology*. Amer. Math. Soc., 1989.

[71] _____. Aronszajn orderings. *Publ. Inst. Math. (Beograd) (N.S.)*, 57(71):29–46, 1995. Đuro Kurepa memorial volume.

[72] _____. Analytic gaps. *Fund. Math.*, 150(1):55–66, 1996.

[73] _____. A classification of transitive relations on ω_1. *Proc. London Math. Soc. (3)*, 73(3):501–533, 1996.

[74] _____. The first derived limit and compactly F_σ sets. *J. Math. Soc. Japan*, 50(4):831–836, 1998.

[75] _____. Compact subsets of the first Baire class. *J. Amer. Math. Soc.*, 12(4):1179–1212, 1999.

[76] _____. A dichotomy for P-ideals of countable sets. *Fund. Math.*, 166(3):251–267, 2000.

[77] _____. A problem of von Neumann and Maharam about algebras supporting continuous submeasures. *Fund. Math.*, 183(2):169–183, 2004.

[78] _____. Biorthogonal systems and quotient spaces via Baire category methods. *Math. Ann.*, 335(3):687–715, 2006.

[79] _____. *Walks on ordinals and their characteristics*, volume 263 of *Progress in Mathematics*. Birkhäuser, 2007.

[80] S. Todorcevic, C. Uzcátegui. Analytic k-spaces. *Topology Appl.*, 146/147:511–526, 2005.

[81] B. Veličković. Definable automorphisms of $\mathscr{P}(\omega)/\mathrm{fin}$. *Proc. Amer. Math. Soc.*, 96(1):130–135, 1986.

[82] ———. Forcing axioms and stationary sets. *Adv. Math.*, 94(2):256–284, 1992.

[83] M. Viale. The proper forcing axiom and the singular cardinal hypothesis. *Jour. Symb. Logic*, 71(2):473–479, 2006.

[84] ———. A family of covering properties. *Math. Res. Lett.*, 15(2):221–238, 2008.

[85] N. Weaver. Set theory and C^*-algebras. *Bull. Symbolic Logic*, 13(1):1–20, 2007.

[86] W. H. Woodin. *Discontinuous homorphisms of $C(\omega)$ and set theory*. PhD thesis, University of California at Berkeley, 1984.

[87] ———. *The Axiom of Determinacy, Forcing Axioms, and the Nonstationary Ideal*. Logic and its Applications. de Gruyter, 1999.

[88] B. Zamora-Aviles. *The structure of order ideals and gaps in the Calkin algebra*. PhD thesis, York University, 2009.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Interactions of Computability and Randomness

André Nies*

Abstract

We survey results relating the computability and randomness aspects of sets of natural numbers. Each aspect corresponds to several mathematical properties. Properties originally defined in very different ways are shown to coincide. For instance, lowness for ML-randomness is equivalent to K-triviality. We include some interactions of randomness with computable analysis.

Mathematics Subject Classification (2010). 03D15, 03D32.

Keywords. Algorithmic randomness, lowness property, K-triviality, cost function.

1. Introduction

We will study sets of natural numbers. We refer to them simply as *sets*. Sets can be identified with infinite sequences of bits. Co-infinite sets can also be identified with real numbers in $[0, 1)$ via the binary representation.

We consider two aspects of a set, its *computational complexity* and its *randomness*. The principal observation is that these two aspects interact closely with one another.

The traditional interaction is from computability to randomness. One uses algorithmic methods to define and study randomness notions [43, 26, 27]. We will show that notions introduced in very different computability-theoretic ways coincide.

The converse interaction was discovered later. Concepts originating from randomness enrich computability theory [4, 19, 34]. We will give examples of this interaction through the study of *lowness properties* of a set A. Such a property specifies a sense in which A is close to being computable. Often this

*Partially supported by the Marsden Fund of New Zealand, grant no. 08-UOA-184
André Nies, Dept. of Computer Science, University of Auckland, Private Bag 92019, Auckland, New Zealand. E-mail: andre@cs.auckland.ac.nz.

is understood via the *weak-as-an-oracle* paradigm: A is weak in a specific sense when used as an oracle set in a Turing machine computation. Randomness-related concepts have led to two new paradigms of lowness [37, 31, 13].

The *Turing-below-many* paradigm says that A is close to being computable because it is easy for an oracle set to compute it, in the sense that the class of oracles computing A is large. Here, a class of oracles is considered large if it contains random sets of a certain kind. So far, all the sets that satisfy an instance of the Turing-below-many paradigm are Δ^0_2.

The *inertness* paradigm says that a set A is close to computable because it is computably approximable with a small number of changes. In particular, such a set is Δ^0_2 (see the Limit Lemma 2.1 below). To formalize the inertness paradigm, we use so-called *cost functions*. They measure the total number of changes of a Δ^0_2 set, and especially that of a computably enumerable set. Most examples of cost functions are based on randomness-related concepts.

In Sections 4-6, we will show that various lowness properties coincide. We introduce the K-trivial sets, and the strongly jump-traceable sets. For each class we give characterizations via all three lowness paradigms.

For some more motivation and background in non-technical language see [32]. For detailed background see [37, 8]. Most sections end with a summary and some interesting further facts. The keen student may want to prove some of these facts as exercises.

2. Some Background from Computability Theory

We assume that the reader knows the basics of computability theory, such as the notions of a computable set, a computably enumerable (c.e.) set, Turing reducibility $\leq_{\mathrm{T}}$, relativization, and (to some extent) Turing functionals. See [41] or [37, Ch. 1].

The capital letters A, B, X, Y, Z denote sets of natural numbers, simply called *sets* in what follows. For an "oracle" set A, we let $J^A(x)$ be the value on input x of a universal partial A-computable function. For instance, let $(\Phi_e)_{e\in\mathbb{N}}$ be an effective listing of all Turing functionals and let $J^A(x) = \Phi^A_x(x)$ (equality extends to the value 'undefined'). The domain of J^A is denoted A'. Thus, A' is the set of x such that $J^A(x)$ is defined, and $\emptyset'$ is (a version of) the halting problem. We say that a set A is Δ^0_2 if $A \leq_{\mathrm{T}} \emptyset'$. The following basic result of Shoenfield will be used frequently.

Lemma 2.1 (Limit Lemma).
A is $\Delta^0_2 \Leftrightarrow A(x) = \lim_s f(x,s)$ for some computable $0,1$-valued function f.

Usually we write $A_s(x)$ instead of $f(x,s)$.

Recall that $X \leq_{\mathrm{tt}} Y$ (X is truth-table below Y) if $X \leq_{\mathrm{T}} Y$ via a Turing functional Γ such that $\Gamma(Z)$ is total for all oracles Z. A variant of the Limit

Lemma says that $A \leq_{\mathrm{tt}} \emptyset'$ iff the number of changes in some computable approximation of A is computably bounded in x. Such a set is called *ω-c.e.*

Recall that a *lowness property* specifies a sense in which a set $A \subseteq \mathbb{N}$ is close to being computable. Such a property is closed downwards under $\leq_{\mathrm{T}}$. For instance, A is *low* if $A' \leq_{\mathrm{T}} \emptyset'$, that is, the Turing degree of A' is as low as possible. A is *superlow* if in fact A' is truth-table below $\emptyset'$.

Another example of a lowness property is the following. We say that a set A is *computably dominated* (or of hyperimmune-free degree) if each function f that can be computed with A as an oracle is dominated by a computable function. Outside the computable sets, this lowness property is not compatible with being low in the usual sense. In fact, the only computably dominated Δ^0_2 sets are the computable sets.

Cantor space. Finite sequences of bits will be called *strings*. The set of strings is denoted $\{0,1\}^*$. The variables x, y, z, σ, τ range over strings. We identify strings with natural numbers via a computable bijection $\{0,1\}^* \to \mathbb{N}$ (related to the binary presentation of a number).

Subsets of $\mathbb{N}$ are identified with infinite sequences of bits. They form the *Cantor space* $2^{\mathbb{N}}$, which is equipped with the product topology. For each string σ,

$$[\sigma] = \{X \colon \sigma \prec X\}$$

is the class of sets extending the string σ. The clopen classes $[\sigma]$ form a basis for the product topology. Thus, an *open class* (or set) has the form $\bigcup_{\sigma \in C}[\sigma]$ for some set C. Such a class is called *computably enumerable*, or Σ^0_1, if one can choose the set C computably enumerable.

The complements of computably enumerable open classes are called Π^0_1 *classes*. A Π^0_1 class is given as the set of paths through a computable binary tree. There are many examples of non-empty Π^0_1 classes without a computable member.

Basis Theorems for Π^0_1 classes. A *basis theorem* (for Π^0_1 classes) says that each non-empty Π^0_1 class has a member with a particular property, usually a lowness property.

Theorem 2.2 (Jockusch and Soare [16]). *Let $\mathcal{P}$ be a non-empty Π^0_1 class. Then $\mathcal{P}$ has a low member, and a computably dominated member.*

The proof of the first statement actually shows that $\mathcal{P}$ has a superlow member. In the second statement one obtains a computably dominated member A of $\mathcal{P}$ such that $A'' \leq_{\mathrm{tt}} \emptyset''$ (see [37, 1.8.38, 1.8.43]).

3. Randomness

In this section we consider the interaction between computability and randomness. We use algorithmic tools to introduce tests concepts, which determine

formal randomness notions. The tools are not only taken from computability theory on sets of natural numbers, but also from computable analysis, where the basic objects are continuous functions. We show that differentiability of certain computable functions defined on the unit interval can be used as a test notion.

3.1. Finite objects. Recall that $\{0,1\}^*$ denotes the set of strings over $\{0,1\}$. A *machine* is a partial computable function $M : \{0,1\}^* \mapsto \{0,1\}^*$. If $M(\sigma) = x$ we say that σ is an *M-description* of x.

We say that a machine M is *prefix-free* if no M-description is a proper initial segment of any other M-description. To build a prefix-free machine M, one usually specifies a set of *requests* $\langle r, y\rangle \in \mathbb{N} \times \{0,1\}^*$. Via such a request one asks that M can describe the string y with r bits. An important technical fact is that any consistent c.e. set of requests can be turned into a prefix-free machine. This result is often referred to as the Kraft-Chaitin theorem, but is called the Machine Existence Theorem in [37, 2.2.17].

Theorem 3.1. *Let L be a computably enumerable set of requests such that*

$$1 \geq \sum\{2^{-r} : \langle r, y\rangle \in L\}.$$

Then there is a prefix-free machine M such that for each request $\langle r, y\rangle \in L$, the machine M can describe y with at most r bits.

Let $(M_d)_{d\in\mathbb{N}}$ be an effective listing of the prefix-free machines. We define a prefix-free machine $\mathbb{U}$ by $\mathbb{U}(0^d 1\sigma) = M_d(\sigma)$. The machine $\mathbb{U}$ is *universal* in the sense that, if a string y has an M_d-description σ, it has a $\mathbb{U}$-description that is only longer by a constant. For a string x, we let $K(x)$ denote the length of a shortest $\mathbb{U}$-description of x:

$$K(x) = \min\{|\sigma| : \mathbb{U}(\sigma) = x\}.$$

The definitions given above can be generalized to the case that the computation model includes queries to an oracle set X. In this way, we define $\mathbb{U}^X(\sigma)$, $K^X(y)$, etc.

We list some facts about K. Let "$\leq^+$ denote "$\leq$" up to a constant. (For instance, we write $2n+5 \leq^+ n^2$.) Let $|x| \in \{0,1\}^*$ denote the length of a string x written in binary. The following bounds are proved by constructing appropriate prefix-free machines: for each computable function f we have $K(f(x)) \leq^+ K(x)$. In particular, we have the *lower bound* $K(|x|) \leq^+ K(x)$. Further, we have the *upper bound* $K(x) \leq^+ |x| + K(|x|)$. Since $K(|x|) \leq^+ 2\log|x|$, this upper bound is not much larger than $|x|$.

If $|x| \leq^+ K(x)$ we think of x as *incompressible*. This formalizes the intuitive notion of randomness for strings (see Section 2.5 of [37] for details).

3.2. Measure, tests, and Martin-Löf randomness. The *product measure* λ on Cantor space $2^{\mathbb{N}}$ is given by

$$\lambda[\sigma] = 2^{-|\sigma|}$$

for each string σ. If a class $\mathcal{G} \subseteq 2^{\mathbb{N}}$ is open then $\lambda\mathcal{G} = \sum_{\sigma \in B} 2^{-|\sigma|}$ where B is a prefix-free set of strings such that $\mathcal{G} = \bigcup_{\sigma \in B}[\sigma]$.

A class $\mathcal{C} \subseteq 2^{\mathbb{N}}$ is called *null* if $\mathcal{C}$ is contained in some Borel class $\mathcal{D}$ such that $\lambda\mathcal{D} = 0$. We discuss the connection of null classes and randomness. The intuition is that an object is random if it satisfies no exceptional properties. We give two examples of exceptional properties of a set Y. The first is that every other bit is zero. The second is that in the limit, there are at least twice as many zeros as ones:

$$2/3 \leq \liminf_n |\{i < n : Y(i) = 0\}|/n.$$

We would like to formalize "exceptional property" by "null class". The examples above are null classes, so they should not contain a random set. The problem is that if we do this, no set Z is random, because $\{Z\}$ itself is a null class. The solution is to consider only effective null classes. By specifying a particular notion of effectivity, we specify a notion of *tests*. To be random in this particular algorithmic sense, Z has to avoid these effective null classes, that is, to pass these tests. Since there are only countably many null classes of this type, the class of random sets in this sense will have measure 1.

Frequently test notions are based on the following fact from measure theory.

Fact 3.2. *The class $\mathcal{C} \subseteq 2^{\mathbb{N}}$ is null $\Leftrightarrow$ $\mathcal{C} \subseteq \bigcap \mathcal{G}_m$ for some sequence $(\mathcal{G}_m)_{m\in\mathbb{N}}$ of open sets such that $\lambda\mathcal{G}_m$ converges to* 0.

We obtain a type of effective null class (or test) by adding effectivity requirements to this condition characterizing null classes. We can require an effective presentation of $(\mathcal{G}_m)_{m\in\mathbb{N}}$; further, we can require fast convergence of $\lambda\mathcal{G}_m$ to 0. In this way, we obtain for instance the central randomness notion introduced by Martin-Löf in 1966 [26].

Definition 3.3. A *Martin-Löf test* (or ML-test) is a uniformly computably enumerable sequence $(\mathcal{G}_m)_{m\in\mathbb{N}}$ of open subclasses of $2^{\mathbb{N}}$ such that $\lambda\mathcal{G}_m \leq 2^{-m}$ for each m. A set Z is *Martin-Löf random* (or ML-random) if Z passes each ML-test $(\mathcal{G}_m)_{m\in\mathbb{N}}$, in the sense that Z is not a member of some $\mathcal{G}_m$.

The two properties given above (every other bit is zero, or in the limit there are at least twice as many zeros as ones) determine effective null classes in this sense. So a ML-random set does not have either of these properties.

In the following, we identify co-infinite sets with real numbers in $[0, 1)$ via the binary presentation. A natural example of a ML-random set was given by Chaitin. Consider the halting probability of the universal prefix-free machine $\mathbb{U}$:

$$\Omega = \sum \{2^{-|\sigma|} : \ \mathbb{U} \text{ halts on input } \sigma\}.$$

Note that this sum converges because the machine $\mathbb{U}$ is prefix-free. Chaitin proved that Ω is Martin-Löf random.

The left cut $\{q \in \mathbb{Q}\colon q < \Omega\}$ is computably enumerable. Since any real number is Turing equivalent to its left cut, this implies that $\Omega \leq_{\mathrm{T}} \emptyset'$. It is also not hard to show that $\emptyset' \leq_{\mathrm{T}} \Omega$. Thus Ω determines a ML-random set that is Turing equivalent to the halting problem.

Given a set Z and $n \in \mathbb{N}$, let $Z\restriction_n$ denote the initial segment $Z(0)\dots Z(n-1)$. Schnorr's 1972 Theorem [40] says that Z is ML-random if and only if each of its initial segments is incompressible.

Theorem 3.4. *Z is ML-random $\Leftrightarrow$*
there is $b \in \mathbb{N}$ such that $\forall n\, K(Z\restriction_n) > n - b$.

Levin [24] proved the analogous theorem for a variant of K called monotone string complexity.

Schnorr's Theorem yields a *universal* ML-test: Let

$$\mathcal{R}_b = \{X\colon \exists n\, [K(X\restriction_n) \leq n - b]\}.$$

The relation "$K(x) \leq r$" is computably enumerable, so the sequence of open classes $\mathcal{R}_b$ is uniformly computably enumerable. One shows that $\lambda\mathcal{R}_b \leq 2^{-b}$. Thus, using this notation, Schnorr's Theorem says that

$$Z \text{ is ML-random} \;\Leftrightarrow\; Z \text{ passes the ML-test } (\mathcal{R}_b)_{b\in\mathbb{N}}.$$

So, the single test $(\mathcal{R}_b)_{b\in\mathbb{N}}$ suffices to emulate all the others.

This fact can be used to obtain ML-random sets with lowness properties. The complement of $\mathcal{R}_1$ is $\{X\colon \forall n\, K(X\restriction_n) \geq n\}$. This is a Π^0_1 class of measure at least $1/2$. By Schnorr's Theorem, it consists entirely of ML-random sets. So we can apply the Jockusch-Soare Basis Theorems 2.2 to obtain ML-random sets satisfying lowness properties:

Example 3.5. *(i) There is a low ML-random set.*
(ii) There is a computably dominated ML-random set.

3.3. Randomness and differentiability.

A well-known theorem from analysis states that every function $f\colon [0,1] \to \mathbb{R}$ of bounded variation is differentiable almost everywhere (with respect to Lebesgue measure λ). In particular, this holds for every monotonic function. In the following we identify co-infinite subsets of $\mathbb{N}$ with reals in $[0,1]$ via the binary representation (we identify the set $\mathbb{N}$ with the real 1). If one also requires an effectiveness condition on the function, the reals at which it is not differentiable form a type of effective null class, and hence a test notion for reals. In the 1970s Demuth had a program to show that effective functions are well-behaved, and in particular differentiable, at random reals. For instance, in his own constructivist language he proved that if a real x is Martin-Löf random then each constructive function of bounded variation is differentiable at x [6].

We will describe a similar coincidence due to Brattka, Miller and Nies [2]. We characterize computable randomness using differentiability of non-decreasing computable functions on the unit interval. First we explain the notions involved.

Computable randomness. Martin-Löf tests are c.e. objects. For this reason Schnorr [39] maintained the point of view that ML-tests are already too powerful to be considered algorithmic. He proposed a more restricted notion of a test. His tests formalize computable betting strategies. A test in Schnorr's sense is a computable function M from $\{0,1\}^*$ to the non-negative rationals. When the player has seen $z = Z\restriction_n$, she can make a bet q where $0 \leq q \leq M(z)$ on the next bit $Z(n)$. If she is right she wins q, otherwise she loses q. Thus M must satisfy the fairness condition $M(z0) + M(z1) = 2M(z)$ for each string z. She wins on Z if $M(Z\restriction_n)$ is unbounded. We call a set Z *computably random* if no computable betting strategy wins.

Choose $c \in \mathbb{N}$ such that the start capital $M(\emptyset)$ is at most 2^c. Let $\mathcal{G}_r$ be the class of Z such that $M(Z\restriction_k) \geq 2^{r+c}$ for some k. It is not hard to see that $(\mathcal{G}_r)_{r\in\mathbb{N}}$ forms a ML-test. If $M(Z\restriction_n)$ is unbounded then $Z \in \bigcap_m \mathcal{G}_m$. This shows that computable betting strategies induce a type of effective null class. Further, each ML-random set is computably random.

Computable functions on the unit interval. A Cauchy representation of a real $x \in \mathbb{R}$ is a sequence $(q_i)_{i\in\mathbb{N}}$ of rationals converging to x such that $|q_k - q_i| \leq 2^{-i}$ for each $k \geq i$. A function $f\colon [0,1] \to \mathbb{R}$ is called *computable* if there is an effective method (i.e., a Turing functional) to transform each Cauchy representation of an $x \in [0,1]$ into a Cauchy representation of $f(x)$. Such a function is necessarily continuous. Functions from analysis such as e^x, $\sqrt{x}$ etc. are computable.

We are now able to state the result of Brattka, Miller and Nies in [2].

Theorem 3.6. *Let $x \in [0,1]$. Then x is computably random if and only if $f'(x)$ exists for each computable non-decreasing function f.*

Further research in [2] indicates that x is Martin-Löf random if and only if each computable function of bounded variation is differentiable at x. The forward implication is a variant of the aforementioned result of Demuth [6].

3.4. A notion stronger than Martin-Löf randomness. One can also argue that ML-randomness is too weak to be viewed as a formal counterpart of our intuitive idea of randomness for sets. For instance, the ML-random real Ω has a computably enumerable left cut, and is Turing equivalent to the halting problem $\emptyset'$. These properties may contradict our intuition on randomness: the halting problem is not random at all, so a random set should not match its computational strength. In fact, the set should be Turing incomparable with the halting problem. The following stronger notion was proposed by Kurtz [23].

Definition 3.7. We say that Z is *2-random* if Z is ML-random relative to the halting problem.

Clearly, a set $Z \leq_{\mathrm{T}} \emptyset'$ is not 2-random: let $\mathcal{G}_m = [Z\restriction_m]$, then $(\mathcal{G}_m)_{m\in\mathbb{N}}$ is a ML-test relative to $\emptyset'$ which Z fails. It is also not hard to show that a set $Z \geq_T \emptyset'$ is not 2-random: given a Turing functional Φ, the halting problem can for each m compute k such that the measure of $\mathcal{G}_m = \{Y\colon \forall i < k\,[\Phi^Y(i) = \emptyset'(i)]\}$ is at most 2^{-m}. If $\Phi(Z) = \emptyset'$ then Z fails $(\mathcal{G}_m)_{m\in\mathbb{N}}$, which is a ML-test relative to $\emptyset'$.

For an example, note that the real $\Omega^{\emptyset'}$ is 2-random. Kurtz [23] showed, among other things, that no 2-random set Z is computably dominated (see Section 2 for the definition). In fact he obtained the stronger result that Z is c.e. relative to some set $Y <_T Z$.

Let $C(x)$ be the plain Kolmogorov complexity of a string x, without restriction to prefix-free machines. Clearly $C(x) \leq^+ |x|$. A string is incompressible in the sense of C if $C(x) > |x| - b$ for some (small) constant b. One can show that for some constant slightly larger than b, all prefixes of such a string are incompressible in the sense of K.

Our next coincidence result characterizes 2-randomness in terms of C-incompressibility of initial segments. This can be seen as a variant of Schnorr's Theorem 3.4. However, we merely need C-incompressibility of *infinitely* many initial segments to arrive at the stronger notion of 2-randomness. This suggests that C-incompressibility of a string is a condition much stronger than K-incompressibility.

The coincidence result is due to Nies, Stephan and Terwijn [36]; the harder implication "$\Rightarrow$" was also independently (and slightly earlier) obtained by Miller [28].

Theorem 3.8. *Z is 2-random $\Leftrightarrow$*
there is $b \in \mathbb{N}$ such that $C(Z\restriction_n) > n - b$ for infinitely many n.

Sketch of Proof (for the details see [37, Thm. 3.6.10]).
$\Leftarrow$: Recall that for each oracle X the domain of $\mathbb{U}^X$ is prefix-free. The plain machine M on input σ searches for a splitting $\sigma = \tau z$ such that $y = \mathbb{U}^{\emptyset'}(\tau)$ converges in $|\sigma|$ steps with the approximation of the oracle $\emptyset'$ at stage $|\sigma|$. In this case, it outputs yz. That is, M prints y followed by the rest of σ.

Now suppose that Z is not 2-random. Then by Theorem 3.4 relative to $\emptyset'$, for each d there is $r \in \mathbb{N}$ such that $K^{\emptyset'}(Z\restriction_r) \leq r - d$. Let n_0 be so large that the final computation $\mathbb{U}^{\emptyset'}(\tau) = Z\restriction_r$ converges in n_0 steps for some τ such that $|\tau| \leq r - d$. For each $n \geq n_0$, if the string y contains the bits of Z from position r to $n-1$, then $M(\tau y) = Z\restriction_n$. Thus M can describe $Z\restriction_n$ with at most $n - d$ bits for each $n \geq n_0$, whence $C(Z\restriction_n) \leq n - d + O(1)$ for each $n \geq n_0$.

$\Rightarrow$: A function $F\colon \{0,1\}^* \to \{0,1\}^*$ is called a *compression function* if F is one-one and $|F(x)| \leq C(x)$ for each x. By the Low Basis Theorem 2.2 there is

a compression function F such that $F' \equiv_T \emptyset'$. Using this lowness of F, if Z is 2-random one can show that there is b such that $|F(Z\restriction_n)| > n - b$ for infinitely many n. This implies that $C(Z\restriction_n) > n - b$ for infinitely many n. $\square$

As a corollary, in [36] we obtained a simple new proof of Kurtz's result [23] that no 2-random set is computably dominated.

Summary of Section 3. For binary strings x, we introduce plain descriptive string complexity $C(x)$, and prefix-free string complexity $K(x)$. The intuitive notion of randomness for strings can be formalized by incompressibility. One formal version of incompressibility is $|x| \leq^+ K(x)$. A stronger one is $|x| \leq^+ C(x)$.

For an infinite sequence of bits (i.e., a set), the intuitive notion of randomness corresponds to a hierarchy of mathematical randomness notions. The central one is Martin-Löf randomness; computable randomness is a weaker notion where the tests formalize the idea of a computable betting strategy; 2-randomness is the relativization of ML-randomness to $\emptyset'$.

ML-randomness and 2-randomness can be characterized via incompressibility of initial segments of the sequence. Computable randomness is implied by ML-randomness. It can be characterized by the condition that each non-decreasing computable function on the unit interval is differentiable at the corresponding real number.

Some further facts. The implications between randomness notions are proper:
2-random $\Rightarrow$ Martin-Löf random $\Rightarrow$ computably random. See [37, 3.6.2, 7.4.8].

Suppose that the set A is computable. If Z satisfies a randomness notion, then the symmetric difference $Z\Delta A$ satisfies the same notion. Further, if ρ is a computable permutation of $\mathbb{N}$, then $\rho(Z)$ satisfies the same notion (see [37, 7.6.24] for the case of computable randomness).

4. For Δ^0_2, Close to Computable = Far From Random

We will introduce a lowness property via relativized ML-randomness. Further, we will introduce the K-trivial sets which are far from random. Recall from Subsection 3.1 that $K(x)$ is the length of a shortest prefix-free description of a string x.

Definition 4.1. *Let $A \subseteq \mathbb{N}$.*

(i) A is *low for ML-randomness* (Zambella 1990, [44]) if each ML-random set is already ML-random relative to A.

(ii) A is *K-trivial* (Chaitin 1975, [4]) if each initial segment of A has prefix-free complexity no greater than the complexity of its length. That is, there is $b \in \mathbb{N}$ such that, for each n, $K(A\restriction_n) \leq K(n) + b$. (Here n is written in binary.)

We will see that these two properties of sets are equivalent.

4.1. Background on the two properties.

Lowness for ML-randomness. Zambella asked whether lowness for ML-randomness implies being computable. Kučera and Terwijn [22] answered this in the negative. They proved that in fact some incomputable c.e. set is low for ML-randomness.

Kjos-Hanssen [17] characterized being low for ML-randomness using only effective topology and the uniform measure on Cantor space: A is low for ML-randomness $\Leftrightarrow$ each open class $G \subseteq 2^{\mathbb{N}}$ that is c.e. in A and has measure $\lambda\mathcal{G}$ less than 1 is contained in an open class $\mathcal{S}$ that is c.e. (without the oracle A) and still of measure $\lambda\mathcal{S}$ less than 1.

J. Miller observed that for an incomputable set A, Kjos-Hanssen's result is not constructive. An *index* for a c.e. open set $\mathcal{R}$ is a number e such that $\mathcal{R}$ is the class of sets extending some string in W_e. Assume that from an index relative to A for $\mathcal{G}$ we can effectively obtain an index for the covering class $\mathcal{S} \supseteq \mathcal{G}$. To compute $A\restriction_n$, let $\mathcal{G} = 2^{\mathbb{N}} - [A\restriction_n]$. Compute the index for $\mathcal{S}$. Wait for a stage when all strings y of length n except for one satisfy $[y] \subseteq \mathcal{S}$. Then $A\restriction_n$ must be the remaining string.

K-triviality. This property of sets is the opposite of ML-randomness: K-trivial sets are "antirandom". For, by Schnorr's Theorem 3.4, Z is ML-random iff all values $K(Z\restriction_n)$ are near their upper bound $n + K(n)$; on the other hand Z is K-trivial if the values $K(Z\restriction_n)$ are at their lower bound $K(n)$ (all within constants).

Chaitin [4] was the first to study K-triviality. He showed that the number of strings of a fixed length with minimal K-complexity up to a constant b is bounded by $O(2^b)$.

Theorem 4.2 (Counting Theorem [4]). *For each $b \in \mathbb{N}$, at most $O(2^b)$ strings of length n satisfy $K(x) \leq K(n) + b$. Thus, at most $O(2^b)$ sets are K-trivial with constant b.*

The following is an easy consequence.

Theorem 4.3 ([4]). *Each K-trivial set is Δ^0_2.*

Proof. A is K-trivial via the constant b iff A is a path on the Δ^0_2 tree of strings z such that $K(x) \leq K(|x|) + b$ for each $x \preceq z$. This tree has only $O(2^b)$ paths. Therefore A is Δ^0_2 as an isolated path on a Δ^0_2 tree. $\square$

Instigated by Chaitin, in 1975 Solovay [42] built an incomputable K-trivial set. His set was merely Δ^0_2. Calude and Coles [3] modified Solovay's construction in order to make the set c.e. In 2002, Downey, Hirschfeldt, Nies and Stephan [9] gave an easier construction of a c.e. incomputable K-trivial set. It is similar to the 1999 Kučera-Terwijn construction of a set that is low for ML-randomness.

These constructions gave rise to the cost function method described in Section 5. In Proposition 5.4 we will explain how to obtain a c.e. incomputable K-trivial set via the general Existence Theorem 5.3.

For sets A and B, let $A \oplus B$ denote the set $2A \cup 2B + 1$, namely the set which is A on the even bit positions and B on the odd positions. The K-trivial sets are closed under $\oplus$ by the following result of Downey, Hirschfeldt, Nies and Stephan.

Theorem 4.4 ([9]). *If A and B are K-trivial via b, then $A \oplus B$ is K-trivial via $3b + O(1)$.*

Proof. It is sufficient to describe each string $A \oplus B \restriction_{2n}$ with $K(n)+3b+O(1)$ bits. To do this, we need to describe n only once; if we have a shortest description of n we also know its length $r = K(n)$.

We (somewhat generously) use $b+1$ bits to describe b itself, by putting the string 0^b1 at the beginning of our description of $A \oplus B \restriction_{2n}$. Next, we put the prefix-free description of n. The set of strings x of length n such that $K(x) \leq r + b$ is uniformly c.e. and has size $O(2^b)$ by Chaitin's Counting Theorem 4.2. So we only need to put $b + O(1)$ further bits each to describe the positions of $A \restriction_n$ and $B \restriction_n$ in its enumeration. □

4.2. Coincidence of the two properties.

Theorem 4.5. *A is low for ML-randomness $\Leftrightarrow$ A is K-trivial.*

Known since 2002, this result published in [34] is now considered fundamental in the area. Nies [34] proved "$\Rightarrow$". The converse implication has a complicated history. Downey, Hirschfeldt, Nies and Stephan [9] showed that each K-trivial set is Turing incomplete. These ideas were later explained through the decanter model [10]. Nies combined this model with a new technique called the golden run method in order to show that the K-trivial sets are closed downward under $\leq_T$. Hirschfeldt and Nies together used the golden run method to show the stronger result that K-triviality implies being low for ML-randomess; see [34, 37].

Conceptually, lowness for ML-randomness and K-triviality are quite far apart: the former is a lowness property defined in terms of randomness, while the latter expresses being far from random. So it come at no surprise that the proof of their coincidence is hard. On the other hand, this makes the coincidence quite beneficial, because properties that are easily obtained via one definition can be very hard to obtain directly via the other. For instance, it is easy to see from the definition that each set A that is low for ML-randomness is generalized low_1, i.e., $A' \leq_{\mathrm{T}} A \oplus \emptyset'$ [22], while it takes the golden run method to see this for the K-trivials. On the other hand, for the K-trivial sets, containment in the Δ^0_2 sets and closure under $\oplus$ is not very hard to see (Theorems 4.3, 4.4). If one takes the definition via lowness for ML-randomness, containment in the

Δ^0_2 sets is much harder [33], and no direct proof is even known for the closure under $\oplus$.

Outline of the Proof. It is easiest to introduce two further properties and show the coincidence of the theorem via these properties. The implication from left to right is proved via the notion of a base for ML-randomness. The converse implication is proved via the notion of being low for K. These two notions are of independent interest.

$\Rightarrow$: Bases for ML-randomness were introduced by Kučera [21] in a different terminology.

Definition 4.6. *We say that A is a base for ML-randomness if $A \leq_{\mathrm{T}} Z$ for some set Z that is ML-random relative to A.*

Each set A that is low for ML-randomness is a base for ML-randomness. For, by the Kučera-Gács Theorem (see [37, Thm. 3.3.2]) there is a ML-random set Z such that $A \leq_T Z$. Then Z is ML-random relative to A.

It is now sufficient to show that each base for ML-randomness is K-trivial. This is a result of Hirschfeldt, Nies and Stephan [15] whose proof we follow. Suppose there are a set Z and a Turing functional Φ such that $\Phi^Z = A$ and Z is ML-random relative to A. We will build a prefix-free machine N_d for each $d \in \mathbb{N}$. We want to ensure that there is a d such that N_d can describe each $\tau \prec A$ with $K(|\tau|) + d + 2$ bits. Of course, A is unknown. Thus, given the limitation that the total measure of the N_d-descriptions must not exceed 1, we have to be judicious in deciding which strings τ receive such a description. The idea is to build uniformly c.e. open classes $\mathcal{C}^\tau_d \subseteq 2^\omega$ for $d \in \mathbb{N}$ and $\tau \in 2^{<\omega}$. Their purpose is to test whether a string τ is likely to be an initial segment of A. Roughly, τ fulfills this test if sufficiently many σ satisfy $\tau \preceq \Phi^\sigma$.

For each fixed d, the $\mathcal{C}^\tau_d$ are pairwise disjoint. If we let $\mathcal{G}_d = \bigcup_{\tau \prec A} \mathcal{C}^\tau_d$, then the following hold.

- $(\mathcal{G}_d)_{d\in\mathbb{N}}$ is a Martin-Löf test relative to A.
- If $Z \notin \mathcal{G}_d$ then $\lambda\mathcal{C}^\tau_d = 2^{-K(|\tau|)-d}$ for all $\tau \prec A$.

For a c.e. open class $\mathcal{C}$ and a stage s, let $\mathcal{C}[s] \subseteq \mathcal{C}$ denote the clopen class approximating $\mathcal{C}$ at stage s. We define N_d by enumerating a description of length $K_s(|\tau|)+d+2$ of τ at stage s whenever we have not previously enumerated such a description and $\lambda\mathcal{C}^\tau_d[s] \geq 2^{-K_s(|\tau|)-d-1}$. Since the $\mathcal{C}^\tau_d$ are disjoint for fixed d, we don't run out of descriptions. For the formal definition of the N_d we apply the Machine Existence Theorem 3.1.

Since Z is ML-random relative to A, we have $Z \notin \mathcal{G}_d$ for some d and hence $\lambda\mathcal{C}^\tau_d = 2^{-K(|\tau|)-d}$ for all $\tau \prec A$. This implies that there is an N_d-description of length $K(|\tau|) + d + 2$ of τ for all $\tau \prec A$, as desired.

To build the $\mathcal{C}_d^\tau$, as long as at a stage s we have $\lambda\mathcal{C}_d^\tau[s] < 2^{-K_s(|\tau|)-d}$, we look for strings σ such that $\tau \preceq \Phi^\sigma$ and $\lambda\mathcal{C}_d^\tau[s] + 2^{-|\sigma|} \leq 2^{-K_s(|\tau|)-d}$, and put $[\sigma]$ into $\mathcal{C}_d^\tau[s+1]$. To keep our open classes pairwise disjoint, we then ensure that no $[\sigma']$ such that σ' is compatible with σ is later put into $\mathcal{C}_d^\nu$ for any string ν.

If $Z \notin \mathcal{G}_d$, then no $[\sigma]$ with $\sigma \prec Z$ is ever put into any $\mathcal{C}_d^\tau$. This means that the measure of each $\mathcal{C}_d^\tau$ with $\tau \prec A = \Phi^Z$ must eventually exceed $2^{-K(|\tau|)-d-1}$.

$\Leftarrow$: Recall that $\mathbb{U}^A$ is the universal prefix-free machine with oracle A, and $K^A(y)$ is the length of a shortest $\mathbb{U}^A$-description of y. In general, enhancing the computational power of the universal machine by an oracle A decreases $K(y)$. We say that A is *low for* K if this is not so:

Definition 4.7. *A is* low for K *if $K(y) \leq^+ K^A(y)$ for each string y.*

This property was introduced by Andrej Muchnik Jr. in a 1999 Moscow seminar. He showed that some incomputable c.e. set is low for K. Among the properties discussed in this section, it is the most well-behaved. For instance, if a c.e. set A is low for K, we can, effectively in the constant for being low for K and the c.e. index for A, find an index for a truth table reduction showing $A' \leq_{\mathrm{tt}} \emptyset'$, that is, the superlowness of A [37, 5.1.3].

Also, lowness for K easily implies the other properties we have discussed. By Schnorr's Theorem relative to A, being low for K implies being low for ML-randomness: if Z is ML-random, then $n \leq^+ K(Z\restriction_n) \leq^+ K^A(Z\restriction_n)$ for each n, so Z is ML-random relative to A. To show that lowness for K implies K-triviality, one uses the finitary methods common in algorithmic information theory (see [25]): $K(A\restriction_n) \leq^+ K^A(A\restriction_n) \leq^+ K^A(n) \leq^+ K(n)$. The hypothesis is only used in the first inequality.

To prove that K-triviality implies being low for ML-randomness, it now suffices to show that conversely, each K-trivial set is low for K. From the formulation of this implication, one could hope that it can also be proved using finitary methods, such as manipulating inequalities involving K and K^A. However, so far no one has found such a proof.

The difficulty of proving the implication "K-trivial $\Rightarrow$ low for K" is in part explained by the fact that it is not constructive: from a constant for K-triviality and a c.e. index for a set A, one can not compute a constant via which A is low for K. See [37, 5.5.6], which goes back to [9]. In fact, one cannot even compute an index of a Turing reduction for $A' \leq_{\mathrm{T}} \emptyset'$ [37, 5.5.5]. This shows that the original implication "$\Leftarrow$" in the theorem is also not constructive.

We give an outline of the proof that K-triviality implies being low for K, using the decanter model and the golden run method; for more details see [37, Sections 5.4-5]. We already know from Theorem 4.3 that each K-trival set A is Δ^0_2, and hence has a computable approximation $(A_s)_{s\in\mathbb{N}}$ in the sense of the Limit Lemma 2.1. We now have to understand why K-triviality of A can be seen as an inertness (in particular, a lowness) property. Roughly speaking, whenever $A\restriction_n$ changes, say at a stage s, a $\mathbb{U}$-description of length at most $K_s(n) + b$ of

the new version of $A\restriction_n$ is needed. The measure of possible descriptions is at most 1, so this restricts the changes of A.

Turing incompleteness. First we will discuss the result of [9] that a K-trivial set A is Turing incomplete. The proof is by contradiction. We build an auxiliary c.e. set B. If A is Turing complete, then by the Recursion Theorem we are given a Turing reduction Γ such that $B = \Gamma(A)$. Let $\gamma^A(m)[s]$ denote the use u, i.e., $u-1$ is the largest oracle question asked in the computation $\Gamma^A(m)[s]$. If we put m into B then A must change below u in order to maintain $B = \Gamma(A)$ at input m.

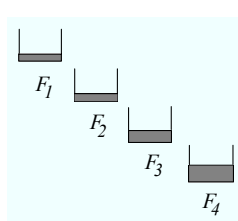
We also build a bounded request set L as in Theorem 3.1. Putting a request $\langle r, n\rangle$ into L causes $K(n) \leq r + d$, where d is the constant for the machine obtained from L (which is again known ahead of the construction by the Recursion Theorem). Hence $K(A\restriction_n) \leq r + b + d$ where A is K-trivial via b.

Let $\boldsymbol{k} = 2^{b+d+1}$. If we can force $A\restriction_n$ to change to a new configuration for more than $\boldsymbol{k}/2$ times, then for this n, our investment into L is overmatched by the opponent's investment into descriptions of $A\restriction_n$. The idea is to do this for so many numbers n that he does not have enough resources to match us.

We can cause these changes if we have $\boldsymbol{k}/2$ numbers m with use $\gamma^A(m) \leq n$ to put into B. If Γ is a weak truth-table reduction, i.e., the use is bounded by a computable function g, we can arrange this by choosing $n \geq g(\boldsymbol{k})$. In the general case, the opponent will simply change A "early" and then redefine the use $\gamma^A(m)$ with a value beyond n. This deprives us of the possibility to cause further A-changes when we need them.

Our solution to this problem is to pool numbers n together, so that a single A-change will let us make progress on lots of them. Further, we already make partial progress based on the A-changes the opponent relies on to move up $\gamma^A(m)$. For $i \leq \boldsymbol{k}$ let us say that a set E is an *i-set* if for each $n \in E$, we put a request $\langle r_n, n\rangle$ into L, and then see descriptions of length $r_n + b + d$ of i different $A\restriction_n$ configurations. The *weight* of such a set is $\sum_{n\in E} 2^{-r_n}$. If n is in a $\boldsymbol{k}$-set, then for each of the $\boldsymbol{k}$ different versions $A\restriction_n$ there is a $\mathbb{U}$-description of length at most $r_n + b + d$. Hence the weight of a $\boldsymbol{k}$-set cannot exceed $1/2$.

We visualize a set of numbers n associated with requests $\langle r_n, n\rangle$ as a quantity of precious wine of the corresponding weight. The *decanter model* consists of decanters $F_1, \ldots, F_{\boldsymbol{k}}$. For instance, in the case $\boldsymbol{k} = 4$ it looks like this:



Precious wine is first poured into F_1. Decanter F_{i-1} can be emptied into decanter F_i. At any stage the content of each F_i must form an i-set. We want as much wine as possible to reach $F_{\boldsymbol{k}}$, because from $F_{\boldsymbol{k}}$ we can pour it into a glass

and drink it. Under certain circumstances we cannot ensure that the content of F_{i-1} is promoted to F_i, so we have to spill it on the floor.

When we put $\langle r_n, n\rangle$ into L, we also put n into F_1, which means that we pour a quantity 2^{-r_n} of precious wine into F_1. At a stage s, all elements n of F_{i-1}, $i \leq \boldsymbol{k}$, satisfy $n \geq \gamma^A(m)[s]$ for a specific number m associated with F_i (to be explained shortly). Once the weight of F_{i-1} passes a certain quota, we put m into B and empty F_{i-1} into F_i. Since $A\restriction_{\gamma^A(m)}$ has to change to keep $B = \Gamma(A)$ correct, the content of F_i including the wine just added remains an i-set, as required.

Now we can get around the problem of an "early" A-change that would move $\gamma^A(m)$ beyond n. If A changes early, then the wine that has already reached F_{i-1} is still promoted to F_i. The only wine lost is the one currently in $F_1, \ldots, F_{i-2}$: the content of these decanters is spilled onto the floor. But the quotas of these decanters are chosen smaller and smaller as i decreases, so we can ensure that the total quantity of wine spilled has a weight of less than $1/4$.

In the construction we have many *runs* of procedures associated with a decanter F_i. Each run has a parameter m such that $\Gamma^A(m)$ converges, and a weight quota p called its *goal*. For $i > 1$ it will call a run associated with F_{i-1} with smaller quota for as many times as needed for F_{i-1} to fill to weight p. Then it puts m into B, empties F_{i-1} into F_i, and returns. If $A\restriction_{\gamma^A(m)}$ changes prematurely then the current content of F_{i-1} is poured into F_i, but the run for F_i continues.

The construction starts out by running $F_{\boldsymbol{k}}$ with a quota of $3/4$. It calls $F_{\boldsymbol{k}-1}$ with a smaller quota for a number of times, and so on down to F_1.

Since Γ^A is total we can force all the A changes needed for runs to return. Hence, the single run associated with $F_{\boldsymbol{k}}$ returns. This yields a $\boldsymbol{k}$-set of weight $3/4$, which is a contradiction.

The full result. We now discuss the full result that a K-trivial set A is low for K. The basic approach is to build a bounded request set W (see Theorem 3.1) as follows: if $\mathbb{U}^A(\sigma) = y$, there is a request $\langle |\sigma| + O(1), y\rangle$ in W. Similar to the proof of the implication "$\Rightarrow$" of the present theorem, we have to judiciously choose the computations $\mathbb{U}^A(\sigma)$ existing at a stage s for which we want to issue a request. The set W has bounded resources, so we have to limit the situation that, after a computation is chosen, A changes to destroy it. We will use such an A-change to promote numbers.

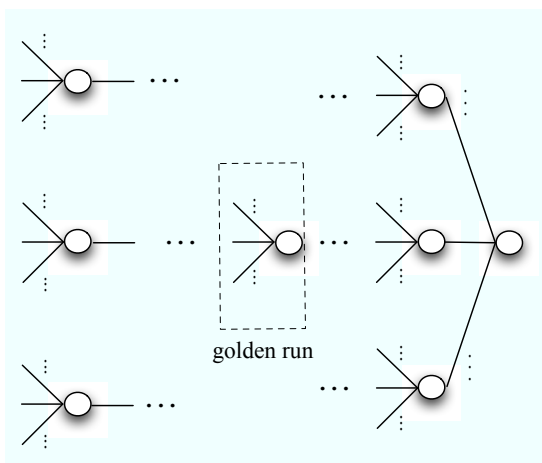
To exploit the hypothesis that A is K-trivial, as before we build a global bounded request set L. Numbers n go through levels $1, \ldots, \boldsymbol{k}$. The decanters are now arranged on a tree. While trying to fill, each decanter at a level greater than 1 builds its own bounded request set W in an attempt to show that A is low for K.

Suppose F_i is a decanter at level i where $1 < i \leq \boldsymbol{k}$. When $\mathbb{U}^A(\sigma)$ converges, F_i calls a decanter $F_{i-1,\sigma}$ at level $i-1$ that can be emptied into F_i. Its goal is

$2^{-|\sigma|}\alpha$, where α is a non-negative rational called the *garbage quota* of the run of F_i (to be explained shortly). When $F_{i-1,\sigma}$ reaches its goal, it returns. It now remains inactive, until possibly A changes below the use of $\mathbb{U}^A(\sigma)$. In this case the content of $F_{i-1,\sigma}$ becomes an i-set, so $F_{i-1,\sigma}$ can be emptied into F_i. We say that the run of $F_{i-1,\sigma}$ is *released*.

If A changes below the use of $\mathbb{U}^A(\sigma)$ before the run returns, then this run is *cancelled*, but we still can empty the current content of $F_{i-1,\sigma}$ into F_i, because the A-change turned it into an i-set.

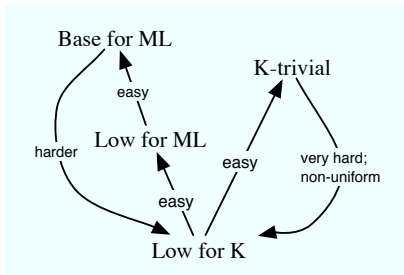
If A does not change at all, a quantity $2^{-|\sigma|}\alpha$ of garbage has been created in the form of wine that is forever stuck at the now defunct decanter $F_{i-1,\sigma}$. If we choose the α values small enough, we can make the amount of garbage tolerable.



To start the construction, we call the decanter at level $\boldsymbol{k}$ with goal $3/4$ and an appropriate small garbage quota. A *golden run* is a run of a decanter F_i that is never cancelled and never returns, while all the runs of $F_{i-1,\sigma}$ it calls are cancelled or return. A golden run exists, for otherwise the decanter at level $\boldsymbol{k}$ would reach its goal $3/4$, which is a contradiction.

At the golden run node F_i we can build a bounded request set W that succeeds in showing that A is low for K. Suppose the golden run of F_i has goal p and garbage quota α. Let $u \in \mathbb{N}$ be least such that $p/\alpha \leq 2^u$. When $F_{i-1,\sigma}$ returns we put $\langle |\sigma| + u + 1, y\rangle$ into W. To see that W is a bounded request set, note that we can bound by 2^u the sum of all $2^{|\sigma|}$ where σ is a $\mathbb{U}^A$-description at some stage, $F_{i-1,\sigma}$ is called, then $F_{i-1,\sigma}$ returns, and later on it is released by an A-change. If this sum exceeds 2^u then the run of F_i reaches its goal $p \leq 2^u\alpha$. So these descriptions contribute at most $1/2$ to W. The weight of the descriptions σ where A does not change after the run of $F_{i-1,\sigma}$ returns is at most the measure of the domain of $\mathbb{U}^A$, whence their contribution is at most $1/2$. Hence W is a bounded request set. $\square$

Summary of Section 4. We introduce a lowness property, being low for ML-randomness, and a far-from-randomness property, K-triviality. Each K-trivial set is Δ^0_2. The K-trivials are closed under $\oplus$.



We show the equivalence of the two properties. To do so we introduce two further properties, being a base for ML-randomness and being low for K, which are also of independent interest. The diagram summarizes the implications discussed. To show that each K-trivial is low for K, we need the decanter and golden run methods.
Some further facts. We say that A is C-trivial if $C(A\restriction_n) \leq^+ C(n)$. Each C-trivial set is computable (Chaitin; see [37, 5.2.20]).

Directly from the definition one can see that each set A that is low for ML-randomness is GL_1, namely $A' \leq_{\mathrm{T}} A \oplus \emptyset'$ [22]. As mentioned above, it is not hard to see from the definition that each c.e. set A that is low for K is superlow, namely $A' \leq_{\mathrm{tt}} \emptyset'$ [37, 5.1.3]. A golden run construction shows directly that each K-trivial set is superlow [37, p. 208].

5. The Inertness Paradigm and Cost Functions

In Section 4.1 we described four properties of Δ^0_2 sets that were introduced by different groups of researchers. For each property, the researchers gave a construction of a c.e. incomputable set with the property. All these constructions looked similar, which is not too surprising given that the properties later turned out to be equivalent. From 1999 on, the language of cost function was developed to formulate these constructions [22, 9, 34].

Nowadays cost functions are an indispensable tool for understanding the class of K-trivial sets and its subclasses [37, Section 5.3], [14, 35]. For instance, each K-trivial set is Turing below a c.e. K-trivial set (see Corollary 5.6 below). The only known proof of this result relies on a cost function.

5.1. Basics on cost functions. Recall the Limit Lemma 2.1: $A \leq_{\mathrm{T}} \emptyset'$ iff $A(x) = \lim_s A_s(x)$ for some 0,1-valued computable approximation $(A_s)_{s\in\mathbb{N}}$. Cost functions are used to measure the total of changes, taken over all numbers, of a computable approximation. In this way we have a formal version of the inertness paradigm from the introduction: a Δ^0_2 set is close to computable if it can be computably approximated with a small total amount of changes.

Definition 5.1. A *cost function* is a computable function

$$c : \mathbb{N} \times \mathbb{N} \to \{x \in \mathbb{Q} \colon\ x \geq 0\}.$$

We say that a cost function c satisfies the *limit condition* if

$$\lim_x \sup_s c(x, s) = 0.$$

When building a computable approximation of a Δ^0_2 set A, we view $c(x, s)$ as the cost of changing $A(x)$ at stage s. We now express that the *total* cost of changes, taken over all x, is finite [37, Section 5.3].

Definition 5.2. We say that a computable approximation $(A_s)_{s\in\mathbb{N}}$ *obeys* a cost function c if

$$\infty > \textstyle\sum_{x,s} c(x, s)\, [\![x < s \ \wedge\ x \text{ is least such that } A_{s-1}(x) \neq A_s(x)]\!].$$

We say that A obeys c if *some* computable approximation of A obeys c.

Mostly we use this to construct some auxiliary object of finite "weight", such as a bounded request set in the sense of 3.1, or a so-called Solovay test in the proof of Theorem 5.10 below.

The analytic approach to restricting changes is more powerful than most combinatorial approaches. For example, call a Δ^0_2 set A *slow* if for each non-decreasing unbounded computable function h, there is a computable approximation $(A_s)_{s\in\mathbb{N}}$ of A such that $A_s \restriction_n$ changes at most $h(n)$ times. Is it not hard to build a slow c.e. set that is Turing complete.

A co-infinite c.e. set is called *simple* [38] if it meets each infinite c.e. set. Clearly no such set is computable. The following theorem can be traced back to [22, 9].

Theorem 5.3. *If a cost function c satisfies the limit condition, then some simple set A obeys c.*

Proof. Let $(W_e)_{e\in\mathbb{N}}$ be an effective listing of the c.e. sets. To make A simple we meet the requirements $S_e :\ |W_e| = \infty \Rightarrow A \cap W_e \neq \emptyset$. Requirement S_e is allowed to spend at most 2^{-e}. Because of the limit condition, S_e can wait for an x to appear in W_e that is so large that S_e can afford it.

At stage s, if S_e is not satisfied yet, we look for an x, $2e \leq x < s$, such that $x \in W_{e,s}$ and

$$c(x, s) \leq 2^{-e}.$$

If so, we put the least such x into A and declare S_e satisfied.

Since a requirement S_e spends at most 2^{-e}, the total cost of changes is bounded by $\sum_e 2^{-e} = 2$. Hence A obeys c.

Suppose that W_e is infinite. As explained above, since c satisfies the limit condition, each S_e is met. A is co-infinite because we choose $x \geq 2e$. So A is simple. $\square$

We say that a cost function $c(x,s)$ is *monotonic* if $c(x,s)$ is non-increasing in x, and non-decreasing in s. Thus, at the same stage a smaller number can only be more expensive, and the same number can only get more expensive at later stages. Most cost functions given below will be monotonic.

5.2. Applications of cost functions.

We analyze some lowness properties and their corresponding constructions, using cost functions.

5.2.1. *K*-triviality. Recall that a set A is K-trivial if there is a $b \in \mathbb{N}$ such that $\forall n\ K(A\restriction_n) \leq K(n) + b$. We introduce a cost function $c_{\mathcal{K}}$ satisfying the limit condition such that any set obeying $c_{\mathcal{K}}$ is K-trivial. Then, by Theorem 5.3, there is a simple K-trivial set. In Theorem 5.5 we will prove that obeying $c_{\mathcal{K}}$ actually characterizes K-triviality.

To show that A is K-trivial we build an appropriate prefix-free machine M via the Machine Existence Theorem 3.1.

(a) Let $K_s(i)$ be the value of $K(i)$ at stage s. Whenever there is a new value $K_s(i)$, we give an M-description of $A_s\restriction_i$ with length $K_s(i)+1$. The combined weight of such descriptions is at most $1/2$.

(b) If $A(x)$ changes at stage s then, for all i such that $s \geq i > x$, the initial segment $A\restriction_i$ gets a new M-description of length $K_s(i)+1$. If we let

$$c_{\mathcal{K}}(x,s) = \textstyle\sum_{i=x+1}^{s} 2^{-K_s(i)},$$

then the measure of the new M-descriptions needed is $c_{\mathcal{K}}(x,s)/2$. If A obeys $c_{\mathcal{K}}$ and the total cost of changes is at most 1, this contributes a weight of at most $1/2$ in M-descriptions, so we build the desired machine. More generally, if the total cost of changes is at most 2^d for $d \in \mathbb{N}$, we choose the M-descriptions in (b) of length $K_s(i)+d+1$. We have shown the following.

Proposition 5.4. *Suppose that A obeys the cost function $c_{\mathcal{K}}$. Then A is K-trivial.*

Note that $\sup_s c_{\mathcal{K}}(x,s) = \sum_{i>x} 2^{-K(i)}$ is bounded above by the measure of the set of strings σ such that $\mathbb{U}(\sigma) > x$. Therefore $c_{\mathcal{K}}$ satisfies the limit condition, and by Theorem 5.3 some simple set is K-trivial.

By the implication "$\Rightarrow$" of the following result, any possible construction of a K-trivial set will be similar to the one in the proof of Theorem 5.3.

Theorem 5.5 (Nies [34]). *A is K-trivial $\Leftrightarrow$ A obeys $c_{\mathcal{K}}$.*

The implication "$\Leftarrow$" is Proposition 5.4. The implication "$\Rightarrow$" is not too hard for c.e. sets ([37, 5.3.27]). For Δ^0_2 sets in general, apparently it requires the full power of the golden run method (see [37, 5.5.2]).

As an application, we show that K-triviality is closely tied to being c.e.

Corollary 5.6. *For each K-trivial set A, there is a c.e. K-trivial set $D \geq_T A$.*

Proof. Let $D = \{\langle x, i\rangle \colon A(x) \text{ changes at least } i \text{ times}\}$. Thus, when $A(x)$ changes, we put the next element in the x-th column of $\mathbb{N}$ into D. Clearly, $D(\langle x, i\rangle)$ can only change at a stage s when $A(x)$ also changes. Now $x \leq \langle x, i\rangle$ and $c_{\mathcal{K}}(y, s)$ is non-increasing in y. Thus, if A obeys $c_{\mathcal{K}}$ then D obeys $c_{\mathcal{K}}$ as well. (Note that $c_{\mathcal{K}}$ can be replaced by any monotonic cost function in this argument.) □

In [35] we introduce a cost function c_Ω simpler than $c_{\mathcal{K}}$, and show that it also characterizes K-triviality. For each stage t, let Ω_t be the measure of the domain of $\mathbb{U}$ at stage t. Now let $c_\Omega(x, s)$ be the measure of $\mathbb{U}$–descriptions converging from stage x to s, that is, $c_\Omega(x, s) = \Omega_s - \Omega_x$.

Theorem 5.7. *A is K-trivial $\Leftrightarrow$ A obeys the cost function c_Ω.*

Outline of the proof. $\Leftarrow$: Clearly $c_{\mathcal{K}}(x, s+1) - c_{\mathcal{K}}(x, s) \leq \Omega_{s+1} - \Omega_s$, which implies that $c_{\mathcal{K}}(x, s) \leq c_\Omega(x, s)$ by induction on $s \geq x$. Thus, if a set A obeys c_Ω it also obeys $c_{\mathcal{K}}$. Therefore A is K-trivial by Proposition 5.4.
$\Rightarrow$: This is a further application of the golden run method. It can also be proved directly from the foregoing Theorem. □

We say that a monotonic cost function c is *additive* if $c(x, y) + c(y, z) = c(x, z)$ for each $x < y < z$. Clearly c_Ω is additive (while $c_{\mathcal{K}}$ is not). An additive cost function c is completely determined by the non-decreasing sequence of rationals $(c(0, s))_{s\in\mathbb{N}}$ approximating the real $\sup_s c(0, s)$. Nies [35] proved that A is K-trivial iff A obeys all additive cost functions. This characterizes K-triviality of a Δ^0_2 set A purely based on effective approximations of A, and on left-c.e. reals. In contrast, the characterizations in Section 4 used machines, measure, or relativization.

5.2.2. Strongly jump traceable sets. We discuss a lowness property which is defined by purely computability-theoretic means following the weak-as-an-oracle paradigm. It properly implies K-triviality for c.e. sets. It is much stronger than slowness mentioned after Definition 5.2. We will show how it can be characterized by obeying all so-called benign cost functions.

The property is an instance of the meta-concept of traceability. The idea behind traceability is the following. The set A is computationally weak because for certain functions ψ computed with oracle A, the possible values $\psi(n)$ are contained in a finite set T_n of small size. The sets T_n are obtained effectively from n (not using A as an oracle).

Traces for functions $\omega \to \omega$ also appear in combinatorial set theory, especially forcing results related to cardinal characteristics. They are called slaloms there, and were introduced by T. Bartoszyński (see [1]).

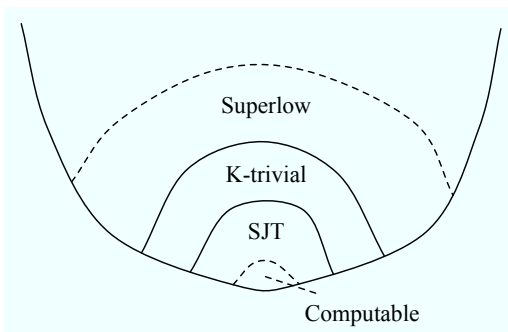
Recall that $J^A(x)$ is the value on input x of a universal A-partial computable function. We say that a computable function h with only positive values is an *order function* if it is non-decreasing and unbounded.

Definition 5.8. A *computably enumerable trace with bound* h is a uniformly computably enumerable sequence $(T_x)_{x\in\mathbb{N}}$ of finite sets such that $|T_x| \leq h(x)$ for each x.

We say that a set A is *strongly jump traceable* (SJT) if for each order function h, there is a c.e. trace $(T_x)_{x\in\mathbb{N}}$ with bound h such that, whenever $J^A(x)$ is defined, we have $J^A(x) \in T_x$.

Strong jump traceability was introduced by Figueira, Nies, and Stephan [11]. They built a simple strongly jump traceable set. Further, they show that A is SJT iff the relative Kolmogorov complexity $C^A(y)$ of a string y is not far below $C(y)$ (for each order function g we have $C(y) \leq^+ C^A(y) + g(C^A(y))$). This makes the notion an analog of being low for K.

It matters that we require *each* order function h as a bound for some trace. A much weaker notion is jump traceability, where one merely requires that there is a c.e. trace for J^A with *some* computable bound h. There is a perfect class of sets that are jump traceable as shown in [37, 8.4.4], while each SJT set is Δ^0_2 by [7].



The c.e. strongly jump traceable sets form a *proper* subclass of the c.e. K-trivial sets by Cholak, Downey, and Greenberg [5]. It is interesting to compare the two classes. Both are closed downward under $\leq_{\mathrm{T}}$. Both are closed under $\oplus$. By definition the c.e. K-trivials have a Σ^0_3 index set; in contrast, the c.e. SJTs have a Π^0_4-complete index set by Ng [30]. Thus, as already indicated by the definition, within the c.e. sets SJT is more complicated than the K-trivials as a class, even though its members are closer to being computable. Recent research of Downey and Greenberg [7] shows that in fact each SJT set (c.e. or not) is K-trivial.

Greenberg and Nies [14] characterized the c.e. SJTs according to the inertness paradigm. They specified the right class of cost functions to gauge how inert a c.e. set must be so that it is SJT. A monotonic cost function c is called *benign* if there is a computable bound $g(n)$ on the length of any finite sequence $x_0 < x_1 < \ldots < x_k$ such that $c(x_i, x_{i+1}) \geq 2^{-n}$ for each $i < k$. For instance, the cost function $c_{\mathcal{K}}$ characterizing K-triviality is benign via $g(n) = 2^n$. Further, any additive cost function is benign.

Theorem 5.9 ([14]). *Let A be c.e. Then*

$$A \text{ is strongly jump traceable} \Leftrightarrow A \text{ obeys each benign cost function.}$$

Because of Proposition 5.4, the harder implication "$\Rightarrow$" generalizes the result of Cholak, Downey, and Greenberg [5] that SJT implies K-triviality for c.e. sets.

5.2.3. Kučera's injury free solution to Post's problem. Post [38] asked whether a c.e. set can be incomputable but also Turing incomplete. Both Friedberg and Muchnik solved the problem in 1955 by building a pair of Turing incomparable c.e. sets. To do so, they introduced the finite injury method. A further solution to Post's problem is to build a low simple set (see [41]). This construction again uses the finite injury method, because it has injury to lowness requirements. In contrast, Kučera in 1986 [20] obtained an injury-free proof of the following result, and then used it for an injury-free solution to Post's problem.

Theorem 5.10. *Suppose Y is a ML-random Δ^0_2 set. Then some simple set A is Turing below Y.*

Now let Y be the bits of Ω in the even positions. An easy direct argument involving van Lambalgen's theorem on relative randomness shows that Y is low and ML-random ([36] or [37, 3.4.10]). Therefore one can build without injury a low simple set A.

We formulate the proof of Kučera's theorem in the language of cost functions. This argument is due to Greenberg and Nies [14], and indirectly also Hirschfeldt and Miller (2006, unpublished; see Section 6 of this paper).

Proof of Theorem 5.10. Fix a computable approximation $(Y_s)_{s\in\mathbb{N}}$ of Y. We define a cost function c_Y such that, if $e < x$ and $Y_t \upharpoonright_e$ does not change for $x \le t \le s$, then $c_Y(x,s) < 2^{-e}$. In more detail, let $c_Y(x,s) = 2^{-x}$ for each $x \ge s$. If $x < s$, and e is least such that $Y_{s-1}(e) \ne Y_s(e)$, let $c_Y(x,s) = \max(c_Y(x,s-1), 2^{-e})$.

Fact 5.11. *If a Δ^0_2 set A obeys c_Y, then $A \le_{\mathrm{T}} Y$ with use function bounded by the identity.*

A *Solovay test* $\mathcal{S}$ is given by an effective enumeration of strings $\sigma_0, \sigma_1, \ldots$, such that $\sum_i 2^{-|\sigma_i|} < \infty$. If Y is ML-random and $\sigma_0, \sigma_1, \ldots$ is a Solovay test, then for almost all i the string σ_i is not a prefix of Y (see [37, 3.2.19]).

To see that $A \le_{\mathrm{T}} Y$, we enumerate a Solovay test as follows. When $A_{s-1}(x) \ne A_s(x)$ and $c_Y(x,s) = 2^{-e}$, we put the string $Y_s \upharpoonright_e$ into $\mathcal{S}$. Since A obeys c_Y, $\mathcal{S}$ is indeed a Solovay test.

Choose s_0 such that $\sigma \not\prec Y$ for any σ enumerated into $\mathcal{S}$ after stage s_0. Given an input $x \ge s_0$, using Y as an oracle, compute $t > x$ such that $Y_t \upharpoonright_x = Y \upharpoonright_x$. Then $x \in A$ implies $x \in A_t$. For, by the definition of the cost function c, at each stage $s > t$, if $c(x,s) = 2^{-e}$ (where $e \le x$), then $Y_s \upharpoonright_e$ still has the same

value as at stage t, which is the true $Y\restriction_e$. Thus, if $A_{s-1}(x) \neq A_s(x)$ we will put a prefix σ of Y into $\mathcal{S}$, contradiction. This shows Fact 5.11.

Since Y is Δ^0_2, the cost function c_Y satisfies the limit condition. Hence some simple set A obeys c_Y. So $A \leq_{\mathrm{T}} Y$. $\square$

5.2.4. Adaptive cost functions and injury. In Theorem 5.3 we assumed that the cost function c was given in advance. In a more complicated variant, the cost function c may be defined during the construction. Such a variant is needed for the Kučera-Terwijn construction of a set that is low for ML-randomness [22], and also for Muchnik's direct construction of a set that is low for K. In the latter construction, say, $c(x,s)$ is the measure of all descriptions at stage $s-1$ such that a change at x would destroy the corresponding computation of $\mathbb{U}^A$ at stage $s-1$; that is, $c(x,s) = \sum_\sigma 2^{-|\sigma|}\, [\![\mathbb{U}^A(\sigma)[s-1]\downarrow \ \wedge\ x < \mathsf{use}\ \mathbb{U}^A(\sigma)[s-1]]\!]$. Extra care has to be taken now to ensure that c satisfies the limit condition. Note that this cost function is not monotonic.

If the cost function is defined during the construction, then the construction must be regarded as having injury. For instance, during the construction of a low simple set, the lowness requirements $L_e\colon\ \exists^\infty s\, J^A(e)[s-1]\downarrow \Rightarrow\ J^A(e)\downarrow$ are injured. The following cost function encodes the restraint imposed by L_e: if $J^A(e)$ newly converges at stage $s-1$, define $c(x,s) = \max\{c(x,s-1), 2^{-e}\}$ for each $x < \mathsf{use}\ J^A(e)[s-1]$. If A is enumerated in such a way that the total cost of changes is finite, then L_e is injured only finitely often. Thus A is low.

In contrast, a cost function c given in advance cannot be used to hide injury, because to encode a restraint that is in force at the beginning of stage s we have to know A_{s-1}.

Summary of Section 5. Cost functions arose to uniformize the constructions of Δ^0_2 sets with lowness properties. Nowadays they have turned into an important tool for understanding these lowness properties. We formulate in terms of cost functions the construction of a simple K-trivial set, and Kučera's construction of a simple set below a Δ^0_2 ML-random. We characterize the K-trivial sets and the strongly jump traceable sets in terms of obeying a class of cost functions with simple combinatorial properties: being additive for the K-trivials, and benign for the SJTs. For the K-trivials there is a universal cost function c_Ω.

Some further facts. If c is a monotonic cost function and sets A and B obey c, then $A \oplus B$ obeys c. The class of sets obeying c is closed downward under Turing reduction with use bounded by the input [35].

There is a computable enumeration $(A_s)_{s\in\mathbb{N}}$ of $\mathbb{N}$ in the order $0,1,2,\ldots$ such that $(A_s)_{s\in\mathbb{N}}$ does not obey $c_{\mathcal{K}}$ [37, Ex. 5.3.7]. Thus it matters in the Definition 5.2 of obedience that we require a finite total cost of changes only for *some* computable approximation.

The converse of Theorem 5.3 holds for a monotonic cost function c: if a computable approximation $(A_s)_{s\in\mathbb{N}}$ of an incomputable set A obeys c, then c satisfies the limit condition [37, Ex. 5.3.8].

6. The Turing-below-many Paradigm

In Theorem 5.10 we discussed the result of Kučera [20] that for every ML-random Δ^0_2 set Y there is an incomputable c.e. set $A \leq_{\mathrm{T}} Y$. If Y is Turing incomplete (i.e. $\emptyset' \not\leq_{\mathrm{T}} Y$), then A must be a base for randomness, and hence K-trivial by [15] (also see [37, 3.4.13]). Thus, for c.e. sets, being below a Turing incomplete ML-random set is a lowness property implying K-triviality. A major open question in the area is whether this property coincides with K-triviality [29, Question 4.6], [37].

Question 6.1. *Is each K-trivial set Turing below an incomplete ML-random?*

By Corollary 5.6, it is not necessary to require that the given set be c.e.

Kučera's result is our starting point for studying lowness properties of a set A according to the Turing-below-many lowness paradigm. To obtain lowness properties stronger than the ones mentioned in the previous paragraph, we strengthen the condition related to Kučera's result that $A \leq_{\mathrm{T}} Y$ for some Turing incomplete random set Y. There are two interrelated approaches:

(a) Replace the single oracle set Y by a null class $\mathcal{C} \subseteq 2^{\mathbb{N}}$ containing a ML-random set $Y \not\geq_T \emptyset'$, and require that $A \leq_{\mathrm{T}} Z$ for each ML-random set $Z \in \mathcal{C}$.

(b) Stay with a single oracle set Y, but require that it satisfy a randomness property stronger than Martin-Löf-randomness.

Both approaches lead to similar results related to strong jump traceability.

To carry out (a) the following notation is useful. For a class $\mathcal{C} \subseteq 2^{\mathbb{N}}$, let $\mathcal{C}^{\diamond}$ denote the collection of c.e. sets that are computable from all ML-random sets in $\mathcal{C}$. This "infimum" operator was implicitly introduced in unpublished work of Hirschfeldt and Miller. Each class of the form $\mathcal{C}^{\diamond}$ induces an ideal in the c.e. Turing degrees. Via cost functions Hirschfeldt and Miller showed that $\mathcal{C}^{\diamond}$ contains a simple set for each null Σ^0_3 class $\mathcal{C}$ (see [37, 5.3.15]). Since $\{Y\}$ is a Σ^0_3 class for each Δ^0_2 set Y, this strengthens Kučera's result.

A strengthening of ML-randomness as required in (b) is Demuth randomness, a notion between 2-randomness and Martin-Löf randomness that is still compatible with being Turing below $\emptyset'$ (but no longer with being above $\emptyset'$). We show that each c.e. set that is Turing below a Demuth random is strongly jump traceable. We leave open the question whether being below a Demuth random actually characterizes strong jump traceability for c.e. sets.

We give some more detail on the two approaches above.

Approach (a). By definition, the strongly jump traceable (SJT) sets are weak as an oracle. In Theorem 5.9 we discussed how to characterize the c.e. SJT sets via the inertness paradigm. Now we will characterize them via the Turing-below-many paradigm. Recall that a Δ^0_2 set Y is ω-c.e. if Y has a computable approximation with a computable bound on the number of times $Y(n)$ changes. It is

easy to obtain a ML-random ω-c.e. set. Examples are Chaitin's number Ω, or a superlow ML-random set. Thus, the following theorem of Greenberg, Hirschfeldt and Nies [13] says that a c.e. set A is strongly jump traceable iff it is Turing below many ML-random oracles.

Theorem 6.2. *Let A be c.e. Then*
A is strongly jump traceable $\Leftrightarrow$ A is Turing below each ω-c.e. ML-random set.

The implication "$\Rightarrow$" follows from Theorem 5.9: if Y is ω-c.e. then its associated cost function c_Y defined in the proof of Theorem 5.10 is benign. Since A obeys c_Y and Y is ML-random, we obtain $A \leq_{\mathrm{T}} Y$.
The implication "$\Leftarrow$" is harder. Given an order function h we want to build a c.e. trace for J^A with bound h. We threaten to build an ω-c.e. ML-random set Y such that $A \not\leq_{\mathrm{T}} Y$.

Let $(\Phi_e)_{e\in\mathbb{N}}$ be an effective list of Turing functionals. We have a tree of runs of procedures similar to the golden run method in Subsection 4.2. However, now the tree has infinitely many levels. At stage s, there is a procedure S^e_x at each level e, for each x such that $y = J^A(x)$ converges at s with use u. This procedure either shows that $A\restriction_u$ is not a prefix of $\Phi_e(Y)$, or places y into a trace set T_x of size at most $h(x)$. Since $A \leq_{\mathrm{T}} Y$, at some level e there is a golden run node which always succeeds via tracing. At this node we obtain the required trace for J^A with bound h.

Since diamond classes induce ideals, as a corollary the c.e. SJT sets are closed under $\oplus$. This result was first obtained by Cholak, Downey, and Greenberg [5] who used a direct construction.

The techniques in the proof of Theorem 6.2 are very adaptable. A variant shows that the c.e. sets in SJT coincide with $\mathcal{C}^\diamond$ when $\mathcal{C}$ is the class of superlow sets. A more complex variant shows that the c.e. sets in SJT also coincide with $\mathcal{C}^\diamond$ when $\mathcal{C}$ is the class of superhigh sets Z (namely, Z' is truth-table above $\emptyset''$).

In proving the implication "$\Leftarrow$", the hypothesis is actually not needed that the given set A be c.e. for the case of superlow (and hence ω-c.e.) sets. We conclude that the same hypothesis can be discarded from the implication "$\Leftarrow$" of Theorem 5.9: suppose A obeys all benign cost functions. Then, for each ω-c.e. set Y, A obeys the benign cost function c_Y defined in the proof of Theorem 5.10. Hence $A \leq_{\mathrm{T}} Y$. Thus A is strongly jump traceable.

By [7] each SJT set is K-trivial, and hence obeys $c_{\mathcal{K}}$. However, it is not known whether the implication "$\Rightarrow$" of Theorem 5.9 works for arbitrary sets, that is, whether each SJT set obeys each benign cost function.

Approach (b). Demuth tests generalize Martin-Löf tests $(G_m)_{m\in\mathbb{N}}$ in that one can change the m-th component (a Σ^0_1 set of measure at most 2^{-m}) for a computably bounded number of times. Z fails a Demuth test if Z is in infinitely many final versions of the G_m. (For a formal definition see [37, Section 3.6].)

Greenberg [12] built a Δ^0_2 Martin-Löf random set Y such that every c.e. set computable from Y is strongly jump traceable. Subsequently, Kučera and Nies [18] showed that any Demuth random Δ^0_2 set Y serves this purpose.

Theorem 6.3. *Let Y be Demuth random. Let A be a c.e. set such that $A \leq_{\mathrm{T}} Y$. Then A is strongly jump traceable.*

The following open problem is analogous to Question 6.1.

Question 6.4. *Is each strongly jump traceable c.e. set Turing below a Demuth random?*

Acknowledgments. I thank Rod Downey, Asher Kach, Justin Moore, Eamonn O'Brien, and Christopher Porter for comments on earlier drafts of this paper.

References

[1] T. Bartoszyński. Combinatorial aspects of measure and category. *Fund. Math.*, 127(3):225–239, 1987.

[2] V. Brattka, J. Miller, and A. Nies. Computable randomness and differentiability. To appear.

[3] C. Calude and Richard J. Coles. Program-size complexity of initial segments and domination reducibility. In *Jewels are forever*, pages 225–237. Springer, Berlin, 1999.

[4] G. Chaitin. A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.*, 22:329–340, 1975.

[5] P. Cholak, R. Downey, and N. Greenberg. Strongly jump-traceability I: the computably enumerable case. *Adv. in Math.*, 217:2045–2074, 2008.

[6] O. Demuth. The differentiability of constructive functions of weakly bounded variation on pseudo numbers. *Comment. Math. Univ. Carolin.*, 16(3):583–599, 1975. Russian.

[7] R. Downey and N. Greenberg. Strong jump traceability II: the general case. To appear.

[8] R. Downey and D. Hirschfeldt. *Algorithmic randomness and complexity.* Springer-Verlag, Berlin. To appear.

[9] R. Downey, D. Hirschfeldt, A. Nies, and F. Stephan. Trivial reals. In *Proceedings of the 7th and 8th Asian Logic Conferences*, pages 103–131, Singapore, 2003. Singapore University Press.

[10] R. Downey, D. Hirschfeldt, A. Nies, and S. Terwijn. Calibrating randomness. *Bull. Symbolic Logic*, 12(3):411–491, 2006.

[11] S. Figueira, A. Nies, and F. Stephan. Lowness properties and approximations of the jump. *Ann. Pure Appl. Logic*, 152:51–66, 2008.

[12] N. Greenberg. A Δ^0_2 random set which only computes strongly jump-traceable c.e. sets. To appear.

[13] N. Greenberg, D. Hirschfeldt, and A. Nies. Characterizing the strongly jump traceable sets via randomness. To appear.

[14] N. Greenberg and A. Nies. Benign cost functions and lowness properties. To appear.

[15] D. Hirschfeldt, A. Nies, and F. Stephan. Using random sets as oracles. *J. Lond. Math. Soc. (2)*, 75(3):610–622, 2007.

[16] C. Jockusch, Jr. and R. Soare. Π^0_1 classes and degrees of theories. *Trans. Amer. Math. Soc.*, 173:33–56, 1972.

[17] B. Kjos-Hanssen. Low for random reals and positive-measure domination. *Proc. Amer. Math. Soc.*, 135(11):3703–3709, 2007.

[18] A. Kučera and A. Nies. Demuth randomness and computational complexity. To appear.

[19] A. Kučera. Measure, Π^0_1-classes and complete extensions of PA. In *Recursion theory week (Oberwolfach, 1984)*, volume 1141 of *Lecture Notes in Math.*, pages 245–259. Springer, Berlin, 1985.

[20] A. Kučera. An alternative, priority-free, solution to Post's problem. In *Mathematical foundations of computer science, 1986 (Bratislava, 1986)*, volume 233 of *Lecture Notes in Comput. Sci.*, pages 493–500. Springer, Berlin, 1986.

[21] A. Kučera. On relative randomness. *Ann. Pure Appl. Logic*, 63:61–67, 1993.

[22] A. Kučera and S. Terwijn. Lowness for the class of random sets. *J. Symbolic Logic*, 64:1396–1402, 1999.

[23] S. Kurtz. *Randomness and genericity in the degrees of unsolvability.* Ph.D. Dissertation, University of Illinois, Urbana, 1981.

[24] L. A. Levin. The concept of a random sequence. *Dokl. Akad. Nauk SSSR*, 212:548–550, 1973.

[25] M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications.* Graduate Texts in Computer Science. Springer-Verlag, New York, second edition, 1997.

[26] P. Martin-Löf. The definition of random sequences. *Inform. and Control*, 9:602–619, 1966.

[27] Per Martin-Löf. On the notion of randomness. In *Intuitionism and Proof Theory (Proc. Conf., Buffalo, N.Y., 1968)*, pages 73–78. North-Holland, Amsterdam, 1970.

[28] J. Miller. Every 2-random real is Kolmogorov random. *J. Symbolic Logic*, 69:907–913, 2004.

[29] J. Miller and A. Nies. Randomness and computability: Open questions. *Bull. Symbolic Logic*, 12(3):390–410, 2006.

[30] K. Ng. On strongly jump traceable reals. *Ann. Pure Appl. Logic*, 154:51–69, 2008.

[31] A. Nies. Applying randomness to computability. Series of three lectures at the ASL summer meeting, Sofia, 2009.

[32] A. Nies. Computability and randomness: Five questions. To appear.

[33] A. Nies. Low for random sets: the story. Preprint, available at `http://www.cs.auckland.ac.nz/nies/papers/`, 2005.

[34] A. Nies. Lowness properties and randomness. *Adv. in Math.*, 197:274–305, 2005.

[35] A. Nies. Calculus of cost functions. To appear.

[36] A. Nies, F. Stephan, and S. Terwijn. Randomness, relativization and Turing degrees. *J. Symbolic Logic*, 70(2):515–535, 2005.

[37] A. Nies. *Computability and randomness*, volume 51 of *Oxford Logic Guides*. Oxford University Press, Oxford, 2009.

[38] E. Post. Recursively enumerable sets of positive integers and their decision problems. *Bull. Amer. Math. Soc.*, 50:284–316, 1944.

[39] C.P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*. Springer-Verlag, Berlin, 1971. Lecture Notes in Mathematics, Vol. 218.

[40] C.P. Schnorr. Process complexity and effective random tests. *J. Comput. System Sci.*, 7:376–388, 1973. Fourth Annual ACM Symposium on the Theory of Computing (Denver, Colo., 1972).

[41] R. Soare. *Recursively Enumerable Sets and Degrees*. Perspectives in Mathematical Logic, Omega Series. Springer–Verlag, Heidelberg, 1987.

[42] R. Solovay. Handwritten manuscript related to Chaitin's work. IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 215 pages, 1975.

[43] R. von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Math. Zeitschrift*, 5:52–99, 1919.

[44] D. Zambella. On sequences with simple initial segments. ILLC technical report ML 1990-05, Univ. Amsterdam, 1990.

Contents

1 Logic and Foundations

Justin Tatch Moore
The Proper Forcing Axiom 3

André Nies
Interactions of Computability and Randomness 30

Ya'acov Peterzil* **and Sergei Starchenko***
Tame Complex Analysis and o-minimality 58

2 Algebra

Paul Balmer
Tensor Triangular Geometry 85

David J. Benson
Modules for Elementary Abelian p-groups 113

Sergey Fomin
Total Positivity and Cluster Algebras 125

Nikita A. Karpenko
Canonical Dimension 146

Zinovy Reichstein
Essential Dimension 162

V. Suresh
Quadratic Forms, Galois Cohomology and Function Fields of p-adic Curves 189

In case of papers with several authors, invited speakers at the Congress are marked with an asterisk.

3 Number Theory

Christophe Breuil
The Emerging p-adic Langlands Programme 203

Ralph Greenberg
Selmer Groups and Congruences 231

D.R. Heath-Brown
Artin's Conjecture on Zeros of p-adic Forms 249

Kiran Sridhara Kedlaya
Relative p-adic Hodge Theory and Rapoport-Zink Period Domains 258

Chandrashekhar Khare* **and Jean-Pierre Wintenberger***
Serre's Modularity Conjecture 280

Mark Kisin
The Structure of Potentially Semi-stable Deformation Rings 294

Sophie Morel
The Intersection Complex as a Weight Truncation and an Application to Shimura Varieties 312

Takeshi Saito
Wild Ramification of Schemes and Sheaves 335

K. Soundararajan
Quantum Unique Ergodicity and Number Theory 357

Akshay Venkatesh* **and Jordan S. Ellenberg**
Statistics of Number Fields and Function Fields 383

4 Algebraic and Complex Geometry

Prakash Belkale
The Tangent Space to an Enumerative Problem 405

Christopher D. Hacon* **and James M^c^Kernan***
Boundedness Results in Birational Geometry 427

Daniel Huybrechts
Hyperkähler Manifolds and Sheaves 450

D. Kaledin
Motivic Structures in Non-commutative Geometry 461

Chiu-Chu Melissa Liu
Gromov-Witten Theory of Calabi-Yau 3-folds 497

Christopher D. Hacon* **and James M^c^Kernan***
Flips and Flops 513

Mihai Păun
Quantitative Extensions of Twisted Pluricanonical Forms and Non-vanishing . . . 540

Shuji Saito
Cohomological Hasse Principle and Motivic Cohomology of Arithmetic Schemes . . . 558

Frank-Olaf Schreyer* and David Eisenbud
Betti Numbers of Syzygies and Cohomology of Coherent Sheaves . . . 586

Vasudevan Srinivas
Algebraic Cycles on Singular Varieties . . . 603

Richard P. Thomas
An Exercise in Mirror Symmetry . . . 624

Jean-Yves Welschinger
Invariants Entiers en Géométrie Énumérative Réelle . . . 652

5 Geometry

Anna Erschler
Poisson-Furstenberg Boundaries, Large-scale Geometry and Growth of Groups . . . 681

Jixiang Fu
On non-Kähler Calabi-Yau Threefolds with Balanced Metrics . . . 705

William M. Goldman
Locally Homogeneous Geometric Manifolds . . . 717

Larry Guth
Metaphors in Systolic Geometry . . . 745

Sergei Ivanov
Volume Comparison via Boundary Distances . . . 769

Xiaonan Ma
Geometric Quantization on Kähler and Symplectic Manifolds . . . 785

Fernando Codá Marques
Scalar Curvature, Conformal Geometry, and the Ricci Flow with Surgery . . . 811

Isabel Fernández* and Pablo Mira*
Constant Mean Curvature Surfaces in 3-dimensional Thurston Geometries . . . 830

Alexander Nabutovsky
Morse Landscapes of Riemannian Functionals and Related Problems . . . 862

Frank Pacard
Constant Scalar Curvature and Extremal Kähler Metrics on Blow ups . . . 882

Takao Yamaguchi
Reconstruction of Collapsed Manifolds . . . 899

6 Topology

Denis Auroux
Fukaya Categories and Bordered Heegaard-Floer Homology . . . 917

Kevin Costello
A Geometric Construction of the Witten Genus, I . . . 942

David Gabai
Hyperbolic 3-manifolds in the 2000's . . . 960

Jesper Grodal
The Classification of p–compact Groups and Homotopical Group Theory . . . 973

Ursula Hamenstädt
Actions of the Mapping Class Group . . . 1002

Michael Hutchings
Embedded Contact Homology and Its Applications . . . 1022

Marc Lackenby
Finite Covering Spaces of 3-manifolds . . . 1042

Wolfgang Lück
K- and L-theory of Group Rings . . . 1071

Jacob Lurie
Moduli Problems for Ring Spectra . . . 1099

Maryam Mirzakhani
On Weil-Petersson Volumes and Geometry of Random Hyperbolic Surfaces . . . 1126

Jongil Park
A New Family of Complex Surfaces of General Type with $p_g = 0$ . . . 1146

András I. Stipsicz
Ozsváth-Szabó Invariants and 3-dimensional Contact Topology . . . 1159

Author Index . . . 1179

7 Lie Theory and Generalizations

Alex Eskin* **and David Fisher**
Quasi-isometric Rigidity of Solvable Groups 1185

Iain G. Gordon
Rational Cherednik Algebras 1209

Shrawan Kumar
Tensor Product Decomposition 1226

Erez M. Lapid
Some Applications of the Trace Formula and the Relative Trace Formula 1262

Ivan Losev
Finite W-algebras 1281

Hee Oh
Dynamics on Geometrically Finite Hyperbolic Manifolds with Applications to Apollonian Circle Packings and Beyond 1308

Nimish A. Shah
Equidistribution of Translates of Curves on Homogeneous Spaces and Dirichlet's Approximation 1332

Catharina Stroppel
Schur-Weyl Dualities and Link Homologies 1344

T. N. Venkataramana
Cohomology of Arithmetic Groups and Representations 1366

8 Analysis

Giovanni Alberti, Marianna Csörnyei*, and David Preiss
Differentiability of Lipschitz Functions, Structure of Null Sets, and Other Problems 1379

Alexander R. Its
Asymptotic Analysis of the Toeplitz and Hankel Determinants via the Riemann-Hilbert Method 1395

Pekka Koskela
Regularity of the Inverse of a Sobolev Homeomorphism 1411

Arno B.J. Kuijlaars
Multiple Orthogonal Polynomials in Random Matrix Theory 1417

Gaven J. Martin
Quasiregular Mappings, Curvature & Dynamics 1433

Fedor Nazarov* **and Mikhail Sodin***
Random Complex Zeroes and Random Nodal Lines 1450

Tatiana Toro
Potential Analysis Meets Geometric Measure Theory 1485

9 Functional Analysis and Applications

Damien Gaboriau
Orbit Equivalence and Measured Group Theory 1501

Masaki Izumi
Group Actions on Operator Algebras 1528

Assaf Naor
L_1 Embeddings of the Heisenberg Group and Fast Estimation of Graph Isoperimetry 1549

Mark Rudelson* **and Roman Vershynin***
Non-asymptotic Theory of Random Matrices: Extreme Singular Values 1576

Dimitri Shlyakhtenko
Free probability, Planar algebras, Subfactors and Random Matrices 1603

Stefaan Vaes
Rigidity for von Neumann Algebras and Their Invariants 1624

10 Dynamical Systems and Ordinary Differential Equations

Marie-Claude Arnaud
Green Bundles and Related Topics 1653

Patrick Bernard
Arnold's Diffusion: From the *a priori* Unstable to the *a priori* Stable Case 1680

Xavier Buff* **and Arnaud Chéritat***
Quadratic Julia Sets with Positive Area 1701

Chong-Qing Cheng
Variational Construction of Diffusion Orbits for Positive Definite Lagrangians 1714

Gonzalo Contreras
Generic Dynamics of Geodesic Flows 1729

Manfred Einsiedler
Applications of Measure Rigidity of Diagonal Actions 1740

Federico Rodriguez Hertz
Measure Theory and Geometric Topology in Dynamics 1760

Omri M. Sarig
Unique Ergodicity for Infinite Measures 1777

Dmitry Turaev
Richness of Chaos in the Absolute Newhouse Domain 1804

Amie Wilkinson
Conservative Partially Hyperbolic Dynamics 1816

11 Partial Differential Equations

Nalini Anantharaman
A Hyperbolic Dispersion Estimate, with Applications to the Linear Schrödinger Equation 1839

Nicolas Burq
Random Data Cauchy Theory for Dispersive Partial Differential Equations 1862

Shuxing Chen
Study of Multidimensional Systems of Conservation Laws: Problems, Difficulties and Progress 1884

E. N. Dancer
Finite Morse Index and Linearized Stable Solutions on Bounded and Unbounded Domains 1901

Camillo De Lellis
Almgren's Q-valued Functions Revisited 1910

Manuel del Pino
New Entire Solutions to Some Classical Semilinear Elliptic Problems 1934

Nils Dencker
The Solvability of Differential Equations 1958

Nicola Fusco* and Massimiliano Morini
Equilibrium Configurations of Epitaxially Strained Elastic Films: Existence, Regularity, and Qualitative Properties of Solutions 1985

Nikolai Nadirashvili* and Serge Vlăduţ
Weak Solutions of Nonvariational Elliptic Equations 2001

12 Mathematical Physics

Anton Kapustin
Topological Field Theory, Higher Categories, and Their Applications 2021

Antti Kupiainen
Origins of Diffusion 2044

Matilde Marcolli
Noncommutative Geometry and Arithmetic 2057

Vieri Mastropietro
Universality, Phase Transitions and Extended Scaling Relations 2078

Gregory A. Seregin
Weak Solutions to the Navier-Stokes Equations with Bounded Scale-invariant Quantities 2105

Herbert Spohn
Weakly Nonlinear Wave Equations with Random Initial Data 2128

Katrin Wendland
On the Geometry of Singularities in Quantum Field Theory 2144

Author Index 2171

13 Probability and Statistics

Itai Benjamini
Random Planar Metrics 2177

Alexei Borodin
Growth of Random Surfaces 2188

Arup Bose*, **Rajat Subhra Hazra, and Koushik Saha**
Patterned Random Matrices and Method of Moments 2203

David Brydges* **and Gordon Slade**
Renormalisation Group Analysis of Weakly Self-avoiding Walk in Dimensions Four and Higher 2232

Frank den Hollander
A Key Large Deviation Principle for Interacting Stochastic Systems 2258

Steven N. Evans
Time and Chance Happeneth to Them all: Mutation, Selection and Recombination 2275

Claudia Neuhauser
Coevolution in Spatial Habitats 2297

Jeremy Quastel
Weakly Asymmetric Exclusion and KPZ 2310

Qi-Man Shao
Stein's Method, Self-normalized Limit Theory and Applications 2325

Sara van de Geer
ℓ_1-regularization in High-dimensional Statistical Models 2351

Aad van der Vaart
Bayesian Regularization 2370

14 Combinatorics

Louis J. Billera
Flag Enumeration in Polytopes, Eulerian Partially Ordered Sets and Coxeter Groups 2389

Henry Cohn
Order and Disorder in Energy Minimization 2416

Sergei K. Lando
Hurwitz Numbers: On the Edge Between Combinatorics and Geometry 2444

Bernard Leclerc
Cluster Algebras and Representation Theory 2471

Brendan D. McKay
Subgraphs of Random Graphs with Specified Degrees 2489

J. Nešetřil* and P. Ossona de Mendez
Sparse Combinatorial Structures: Classification and Applications 2502

Eric M. Rains
Elliptic Analogues of the Macdonald and Koornwinder Polynomials 2530

Oliver Riordan
Percolation on Sequences of Graphs 2555

Benny Sudakov
Recent Developments in Extremal Combinatorics: Ramsey and Turán Type Problems 2579

15 Mathematical Aspects of Computer Science

Peter Bürgisser
Smoothed Analysis of Condition Numbers 2609

Cynthia Dwork
Privacy Against Many Arbitrary Low-sensitivity Queries 2634

Venkatesan Guruswami
Bridging Shannon and Hamming: List Error-correction with Optimal Rate 2648

Subhash Khot
Inapproximability of NP-complete Problems, Discrete Fourier Analysis, and Geometry 2676

Daniel A. Spielman
Algorithms, Graph Theory, and Linear Equations in Laplacian Matrices . . . 2698

Salil Vadhan
The Unified Theory of Pseudorandomness . . . 2723

16 Numerical Analysis and Scientific Computing

Bernardo Cockburn
The Hybridizable Discontinuous Galerkin Methods . . . 2749

Peter A. Markowich
Numerical Analysis of Schrödinger Equations in the Highly Oscillatory Regime . . . 2776

Ricardo H. Nochetto
Why Adaptive Finite Element Methods Outperform Classical Ones . . . 2805

Zuowei Shen
Wavelet Frames and Image Restorations . . . 2834

Mary F. Wheeler*, Mojdeh Delshad, Xianhui Kong, Sunil Thomas, Tim Wildey and Guangri Xue
Role of Computational Science in Protecting the Environment: Geological Storage of CO_2 . . . 2864

Jinchao Xu
Fast Poisson-based Solvers for Linear and Nonlinear PDEs . . . 2886

17 Control Theory and Optimization

Hélène Frankowska
Optimal Control under State Constraints . . . 2915

Satoru Iwata
Submodular Functions: Optimization and Approximation . . . 2943

Yurii Nesterov
Recent Advances in Structural Optimization . . . 2964

Alexander Shapiro
Computational Complexity of Stochastic Programming: Monte Carlo Sampling Approach . . . 2979

Robert Weismantel
A Cutting Plane Theory for Mixed Integer Optimization . . . 2996

Xu Zhang
A Unified Controllability/Observability Theory for Some Stochastic and Deterministic Partial Differential Equations . . . 3008

18 Mathematics in Science and Technology

Ellen Baake
Deterministic and Stochastic Aspects of Single-crossover Recombination . . . 3037

Freddy Delbaen
BSDE and Risk Measures . . . 3054

Kazufumi Ito and Karl Kunisch*
Novel Concepts for Nonsmooth Optimization and their Impact on Science and Technology . . . 3061

Philip K. Maini*, Robert A. Gatenby and Kieran Smallbone
Modelling Aspects of Tumour Metabolism . . . 3091

Natasa Djurdjevac, Marco Sarich, and Christof Schütte*
On Markov State Models for Metastable Processes . . . 3105

Nizar Touzi
Second Order Backward SDEs, Fully Nonlinear PDEs, and Applications in Finance . . . 3132

Zongben Xu
Data Modeling: Visual Psychology Approach and $L_{1/2}$ Regularization Theory . . . 3151

Xun Yu Zhou
Mathematicalising Behavioural Finance . . . 3185

19 Mathematics Education and Popularization of Mathematics

Jill Adler
Professional Knowledge Matters in Mathematics Teaching . . . 3213

20 History of Mathematics

Tinne Hoff Kjeldsen
History of Convexity and Mathematical Programming: Connections and Relationships in Two Episodes of Research in Pure and Applied Mathematics of the 20th Century . . . 3233

Norbert Schappacher
Rewriting Points . . . 3258

Author Index . . . 3293

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Tame Complex Analysis and o-minimality

Ya'acov Peterzil and Sergei Starchenko*

Abstract

We describe here a theory of holomorphic functions and analytic manifolds, restricted to the category of definable objects in an o-minimal structure which expands a real closed field R. In this setting, the algebraic closure K of the field R, identified with R^2, plays the role of the complex field. Although the ordered field R may be non-Archimedean, o-minimality allows to develop many of the basic results of complex analysis for definable K-holomorphic functions even in this non-standard setting. In addition, o-minimality implies strong theorems on removal of singularities for definable manifolds and definable analytic sets, even when the field R is $\mathbb{R}$. We survey some of these results and several examples.

We also discuss the definability in o-minimal structures of several classical holomorphic maps, and some corollaries concerning definable families of abelian varieties.

Mathematics Subject Classification (2010). Primary 03C64, 32B15, 32C20; Secondary: 32B25, 14P15, 03C98

Keywords. o-minimality, real closed fields, non-Archimedean analysis, complex analytic sets, Weierstrass function, theta functions, Abelian varieties

1. Introduction

Consider a real closed field R and its algebraic closure $K = R(\sqrt{-1})$. After fixing $\sqrt{-1}$, we can identify K with R^2, and then view subsets of K^n as subsets of R^{2n}. Under this identification polynomial functions from K^n into K become R-polynomial maps from R^{2n} into R^2.

*The second author was partially supported by NSF
Department of Mathematics, U. of Notre Dame, Notre Dame, In., USA.
E-mail: sstarche@nd.edu.
Department of Mathematics, U. of Haifa, Haifa, Israel. E-mail: kobi@math.haifa.ac.il.

When the fields are $\mathbb{R}$ and $\mathbb{C}$, the order topology of the reals endows the complex numbers, through the product topology, with the structure of a topological locally compact field. This is of course the setting of classical complex analysis, and local analytic theory is usually developed using convergent power series and integration (here and below, when we say "classical" we refer to the case $R = \mathbb{R}$ and $K = \mathbb{C}$). When R is an arbitrary real closed field then its order topology still endows K with the structure of a topological field but, since R could be non-Archimedean, this topology may be far from locally compact. In this case, the tools of integration and power series are often not available for the development of complex analysis over K.

While analysis in a non-Archimedean setting is also tackled in rigid analytic geometry we present here a different approach. The main idea is to consider only a limited collection of sets and maps, namely those which are definable in an o-minimal expansion $\mathcal{R} = \langle R, <, +, \cdot, \cdots \rangle$ of the field R. Recall that $\mathcal{R}$ is called o-minimal if every definable (with parameters) subset of R is a finite union of R-intervals whose endpoints are in $R \cup \{\pm\infty\}$. Real closed fields are the standard example but we are going to consider below much richer o-minimal structures (see [6], [38], [8], [19] and [20]).

It turns out (see [7] and [10]) that almost all basic theorems of real differential calculus hold for functions definable in $\mathcal{R}$, even though the field R may be non-Archimedean and as a topological space could be totally disconnected. As we will show, the same is true for many of the basic theorems of complex analysis.

When the field R equals $\mathbb{R}$, the category of definable sets in an o-minimal structure can be viewed as a natural candidate for Grothendieck's vision of "tame topology" (see discussion in [36]). The exclusion of wild topological phenomena from the tame setting of o-minimality implies that definable holomorphic functions cannot have essential singularities. This is easy to see, for if f is a holomorphic function on the punctured unit disc and 0 is an essential singularity then there exist $c \in \mathbb{C}$ with $f^{-1}(c)$ an infinite discrete subset of $\mathbb{C}$. But then, either $\{Im(z) : f(z) = c\}$ or $\{Re(z) : f(z) = c\}$ is an infinite discrete subset of $\mathbb{R}$, so f cannot be definable in an o-minimal structure. At first sight, this seems to exclude too much of classical analytic theory, but as we will see, it is still possible to define in o-minimal structures many classical holomorphic functions on properly chosen domains in a way which permits rich mathematical constructions.

Thus, the theory of holomorphic functions in o-minimal structures allows on one hand to develop analytic-like theory for an arbitrary algebraically closed field of characteristic zero K with respect to a maximal real closed field $R \subseteq K$ and an o-minimal expansion of R. On the other hand, when we specialize the investigation to the classical setting of the complex and real fields, we obtain, in addition, new results on holomorphic functions, complex manifolds and analytic sets, when these are definable in some o-minimal expansion of the real field. The

treatment of both of these settings is uniform and independent of the particular fields in questions.

Our goal here is to present the main definitions and a survey of results, accompanied with examples from both the standard and the nonstandard settings. The paper is structured as follows: In Section 2 we give the basic definition of a K-holomorphic function and discuss a variety of examples. In Section 3 we show how analogues of basic results from complex analysis can be obtained for definable K-holomorphic functions in arbitrary o-minimal structures. In Section 4 we discuss analogues of complex manifolds and analytic sets in o-minimal structures and in particular, in 4.2 and in the Appendix expand on how compact complex manifolds can be viewed within the o-minimal structure $\mathbb{R}_{an}$. In 4.3 we present a more general, o-minimal, version of Chow's theorem on analytic subvarieties of projective space (which in particular implies the classical version). We also consider definable families of manifolds, the particular case of complex tori and point out the connection between such families and non-standard tori. In this section we discuss how Riemann's Existence Theorem can fail (or hold) in the category of definable manifolds in an o-minimal structure. In Section 5 we present several results on what is probably the main feature of tame complex analysis: the theory on removal of singularities. Finally, in Section 6 we discuss theorems on the definability in o-minimal structures of certain classical holomorphic functions such as Schwarz-Christoffel maps, the Weierstrass $\wp$-functions and Riemann's theta functions. We also mention connections to arithmetical questions in algebraic geometry.

We assume here basic knowledge of definability, and o-minimality (see [7] and [10] for a presentation aimed at non-logicians).

Remark. Some work on complex analytic geometry restricted to semiagebraic and subanalytic sets can be found in [11] and [12]. In the non-standard setting of an arbitrary real closed field, such work was carried out, from a different point of view than ours, in [15].

2. K-holomorphic Functions

We start with the basic definitions. Let $\mathcal{R} = \langle R, <, +, \cdot, \ldots \rangle$ be an o-minimal expansion of a real closed field, and $K = R(\sqrt{-1})$ the algebraic closure of R. After fixing $i = \sqrt{-1}$, we can identify K with R^2, as in the classical case, and view subsets of K^n and maps from K^n into K as subsets of R^{2n} and maps from R^{2n} into R^2, respectively. The field operations of K become definable in the ordered field R. We have the order topology on R, the product topology on R^k, and with respect to this topology the field K, identified with R^2, is a topological field. We therefore have a natural notion of $\lim_{x \to a} f(x)$ for functions $f : K^n \to K$, where the limit is taken with respect to the topologies of K^n and of K.

Definition 2.1. Let $U \subseteq K$ be an open set. A function $f\colon U \to K$ is *K-differentiable at $z_0 \in U$* if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \text{ exists in } K.$$

The limit, if exists, is called the *K-derivative of f at z_0* and is denoted by $f'(z_0)$. If f is K-differentiable at every $z \in U$ then it is called *K-holomorphic on U*.

For $U \subseteq K^n$ an open set and $f : U \to K$ a continuous function, f is called *K-holomorphic on U* if it is K-differentiable in each of the variables separately.

The above definitions coincide with the classical definitions of holomorphic functions in one and several variables when $R = \mathbb{R}$ and $K = \mathbb{C}$. As pointed out above, in the general case the topology on R is not well-behaved and very far from locally compact or separable. Hence, although the definitions make sense for arbitrary functions, we are going to restrict our attention to K-holomorphic functions which are in addition definable in the o-minimal structure $\mathcal{R}$.

Note: Although every algebraically closed field of characteristic zero K contains a maximal real closed field R, the choice of R is far from being unique and even the field $\mathbb{C}$ contains maximal real closed subfields which are not isomorphic to $\mathbb{R}$ and are non-Archimedean.

Here are some examples of K-holomorphic functions which are definable in o-minimal structures.

Classical examples

• Let $\overline{\mathbb{R}} = \langle \mathbb{R}, <, +, \cdot \rangle$ (so $K = \mathbb{C}$). By Tarski's work, $\overline{\mathbb{R}}$ is o-minimal. Every complex polynomial is $\mathbb{C}$-holomorphic and definable in $\overline{\mathbb{R}}$.

• Consider the o-minimal structure

$$\mathbb{R}_{an} = \langle \mathbb{R}, <, +, \cdot, \{f|[-1,1]^n : f \text{ real analytic on open } U \supseteq [-1,1]^n\} \rangle$$

(see [6]). Using the real and imaginary parts, every power series convergent in a neighborhood of $0 \in \mathbb{C}^n$ can be represented by a definable $\mathbb{C}$-holomorphic function in $\mathbb{R}_{an}$.

If $V \subseteq \mathbb{C}^n$ is an open bounded set, and $f : V \to \mathbb{C}$ is a holomorphic function, which can be holomorphically extended to an open set $U \supseteq Cl(V)$ (where $Cl(V)$ is the topological closure of V), then $f|V$ is definable in $\mathbb{R}_{an}$. Indeed, $Cl(V)$ can be covered by finitely many open sets on each of which f is definable, hence $f|V$ is definable.

• Let $\mathbb{R}_{an,exp}$ be the o-minimal expansion of $\mathbb{R}_{an}$ by the real exponential function (see [38], [10], [8]). The restriction of the complex exponential function

e^z to any horizontal strip $\{a < Im(z) < b\}$, $a < b \in \mathbb{R}$, is definable in $\mathbb{R}_{an,exp}$, using the real exponential function and restricted $\sin, \cos$. It follows that every branch of $\ln z$ is definable in $\mathbb{R}_{an,exp}$. However, e^z is not definable on the whole of $\mathbb{C}$ because of its infinite discrete kernel, and in fact (see [23], Claim 2.1), if e^z is definable in some o-minimal structure on a set $U \subseteq \mathbb{C}$ then necessarily $Im(z)$ is bounded on U.

• Let $\mathbb{R}_{exp} = \langle \mathbb{R}, <, +, \cdot, e^x \rangle$. It follows from [3], that every germ of an n-variable holomorphic function which is definable in $\mathbb{R}_{exp}$ is already definable in $\overline{\mathbb{R}}$, namely semi-algebraic.

Non-standard examples

• Let $\overline{R} = \langle R, <, +, \cdot \rangle$, where R is a real closed field: Every polynomial over $K = R(\sqrt{-1})$ is K-holomorphic and definable in $\overline{R}$. In fact, in [25], Theorem 2.17, we prove a converse statement:

Theorem 2.2. *If $f : K^n \to K$ is definable and K-holomorphic then it is a polynomial over K.*

• Let $\mathcal{R}$ be a proper extension of $\mathbb{R}_{an,exp}$: If $\alpha \in R^{>0}$ is infinitesimally close to 0 (by that we mean that $0 < \alpha < 1/n$ for every $n \in \mathbb{N}$) then $e^{\alpha z}$ is K-holomorphic and definable on "infinitely wide" strip $-1/\alpha < Im(z) < 1/\alpha$.

These two non-standard examples of o-minimal structures are elementary extensions of structures over the field of reals. The example below does not arise from any structure over the reals (this is made precise in [14]):

• **Divergent power series as K-holomorphic functions.** Consider the real closed field of formal Puiseux series over $\mathbb{R}$, denoted by $R = \mathbb{R}((t^*))$, and its algebraic closure, $K = \mathbb{C}((t^*))$. The field R admits a natural valuation (with $v(t) = 1$) and the infinitesimal elements of R, denoted by μ, are all those of positive valuation. The valuation topology coincides in this case with the order topology of R.

Every formal power series $a(\bar{x}) \in \mathbb{R}[[x_1, \ldots, x_n]]$ can be computed on μ^n and hence defines a function $\bar{a} : \mu^n \to R$. Clearly, if we expand the field R by such a function, the expanded structure will not be o-minimal because μ^n is not definable in any o-minimal structure. However, consider the interval $I = [-t, t]$ in R and the structure

$$\mathcal{R} = \langle R, <, +, \cdot, \bar{a}|I^n \rangle_{a(\bar{x}) \in \mathbb{R}[[\bar{x}]]}.$$

It is proved in [20] that $\mathcal{R}$ is o-minimal.

Now, every formal power series $a(\bar{z}) \in \mathbb{C}[[z_1, \ldots, z_n]]$ (even if the series diverges in the complex field) determines a K-holomorphic function on the poly-disc of radius t in K^n, a map which is definable in $\mathcal{R}$.

3. Analogues of Classical Results in Non-Archimedean Fields

We assume here that $\mathcal{R}$ is an arbitrary o-minimal expansion of a real closed field R and $K = R(\sqrt{-1})$. All definability is assumed to take place in $\mathcal{R}$.

Although the classical tools of power series and integration are not available in this general setting, it is still possible to develop analogues of the classical theory for K-holomorphic functions which are definable in $\mathcal{R}$, by using methods of topological analysis, together with o-minimality. In the 1-variable case we followed the work of Whyburn from [37], and then extended it to functions of several variables. Almost all classical results go through in this case. When we specialize to the classical case, i.e. when R equals the field $\mathbb{R}$ and K equals $\mathbb{C}$, results of this type contribute no new information. However, even in this case model theory allows us to obtain new uniformity results for definable families of holomorphic functions.

3.1. The one-variable case.

All references are to [24].

Fact 3.1 (**The Cauchy-Riemann equations**). *If $U \subseteq K$ is an open definable set and $f : U \to K$ is a definable function then f is K-holomorphic if and only if, as a map $(x, y) \mapsto (v(x, y), w(x, y))$ from $U \subseteq R^2$ into R^2, it is R-differentiable and its R-derivatives satisfy*

$$\frac{\partial v}{\partial x} = \frac{\partial w}{\partial y} \quad \textit{and} \quad \frac{\partial v}{\partial y} = -\frac{\partial w}{\partial x}.$$

(see Fact 2.27)

We let $D \subseteq K$ denote the closed unit disc and C its boundary. For $z = a + b\sqrt{-1} \in K$, we use $|z| = a^2 + b^2 \in R$.

Theorem 3.2.

1. (**Maximum Principle**) *If $f : D \to K$ is a definable continuous function which is K-holomorphic on $Int(D)$ then $|f|$ attains its maximum on C (Theorem 2.31).*

2. (**Open mapping theorem**) *If $U \subseteq K$ is open, definable and $f : U \to K$ is a definable K-holomorphic, non-constant function then f is an open map (Corollary 2.34).*

3. (**Infinite differentiability**) *If $U \subseteq K$ is open, definable and $f : U \to K$ is a definable K-holomorphic map then $f'(z)$ is also K-holomorphic on U (Theorem 2.40).*

4. (**Identity Theorem**) *If $f : U \to K$ is definable and K-holomorphic in a neighborhood of $0 \in K$, and if $f^{(k)}(0) = 0$ for all $k \in \mathbb{N}$ then f vanishes in a neighborhood of 0.*

We re-emphasize that (4) is true although there is no available theory of converging power series (indeed, if the underlying o-minimal structure is sufficiently saturated then there are no converging sequences in R other than the eventually constant ones). One corollary of (4) is that "raising to an infinite power" is not possible for elements in K. The situation is different in the case of R-variables: Consider $\mathcal{R}$ a nonstandard elementary extension of $\mathbb{R}_{exp}$ and let $\alpha > 0$ be an element greater than all $n \in \mathbb{N}$. The function

$$h_\alpha(x) = \begin{cases} & x^\alpha & x \geq 0 \\ - & x^\alpha & x < 0 \end{cases}$$

is definable in $\mathcal{R}$ by $x^\alpha = e^{\alpha \ln x}$ for all $x \in R$. It is infinitely R-differentiable at 0 and all of its R-derivatives are 0 there.

Given a definable K-holomorphic function $f : U \to K$ in a neighborhood $U \subseteq K$ of 0, we let $ord_0(f)$ be the minimal $k \geq 0$ such that $f^{(k)}(0) \neq 0$, or ∞ if there is no such k. The Identity Theorem implies that if f does not vanish in a neighborhood of 0 then $ord_0(f) < \infty$. Moreover, since the above result holds in arbitrary o-minimal structures, we get a uniform version which is interesting over $\mathbb{R}$ as well:

Given a definable open $0 \in U \subseteq K$, we say that *a family $\mathcal{F}$ of functions from U to K is definable in $\mathcal{R}$* if there are definable sets $T \subseteq R^n$ and $F \subseteq U \times K \times T$, such that for every $t \in T$, the set $\{(z, y) \in U \times K : (z, y, t) \in F\}$ is the graph of a function, call it f_t, and $\mathcal{F} = \{f_t : t \in T\}$. Assume now that every $f_t \in \mathcal{F}$ is K-holomorphic on U and does not vanish in a neighborhood of 0. Then, we claim that there is a bound k on $ord_0(f_t)$ as t varies in T. Indeed, if not then by logical compactness we would be able to realize (possibly in an elementary extension) a K-holomorphic non-vanishing f_{t_0} such that $f_{t_0}^{(k)}(0) \neq 0$ for all $k \in \mathbb{N}$. A contradiction. We therefore proved:

Theorem 3.3. *For $U \subseteq K$ a definable neighborhood of 0, let $\mathcal{F} = \{f_t : t \in T\}$ be a definable family of K-holomorphic maps $f_t : U \to K$. Then there is $k \in \mathbb{N}$ such that for every $t \in T$, if $f^{(i)}(0) = 0$ for all $i = 0, \ldots, k$ then f_t vanishes in a neighborhood of 0.*

3.2. Functions of several variables.

Definable K-holomorphic functions of several variables also share many common properties with classical holomorphic functions (see [25]). We limit ourselves here to several results about the ring of germs at 0 of definable K-holomorphic functions.

Definition 3.4. For definable functions f, g in a neighborhood of $0 \in K^n$, we say that *f and g have the same germ at* 0 if there is an open neighborhood $U \ni 0$ such that $f(z) = g(z)$ for all $z \in U$. Let $\mathcal{O}_n(\mathcal{R})$ be the ring of germs at $0 \in K^n$ of all K-holomorphic functions near $0 \in K^n$ which are definable in $\mathcal{R}$.

Here are some results about $\mathcal{O}_n(\mathcal{R})$ (see [25] for all references).

Theorem 3.5. *1. The map from $\mathcal{O}_n$ into $K[[\bar{z}]]$, which sends a germ $f \in \mathcal{O}_n$ to its formal Taylor expansion at 0, is injective. Said differently, if all derivatives of a definable K-holomorphic f vanish at $0 \in K^n$ then f itself vanishes in a neighborhood of 0 (Theorem 2.30 (2)).*

2. $\mathcal{O}_n$ is a local ring.

3. The ring $\mathcal{O}_n$ satisfies the Weierstrass preparation and division theorems (see Theorem 2.20 and Theorem 2.23).

4. The ring $\mathcal{O}_n$ is Noetherian, (Theorem 2.30).

4. Definable K-manifolds and K-analytic Sets

4.1. Basic definitions. Once we have the notion of a K-holomorphic function in several variables we may define the notions of a manifold and an analytic set, with respect to the field K. We restrict our attention only to definable functions and definable sets in a fixed o-minimal expansion $\mathcal{R}$ of a real closed field R, with $K = R(\sqrt{-1})$.

Definition 4.1. *A definable n-dimensional K-manifold* is a definable set M (living in some R^k), equipped with a finite cover of definable sets $M = \bigcup_i U_i$, each of which is in definable bijection $\phi_i : U_i \to V_i$ with a definable open set $V_i \subseteq K^n$, such that the transition maps

$$\phi_j \phi_i^{-1} \,:\, \phi_i(U_i \cap U_j) \to \phi_j(U_j)$$

are K-holomorphic (as maps between open subsets of K^n). The collection $\{\langle U_i, \phi_i \rangle : i \in I\}$ is called a definable atlas for M.

Let M be a definable n-dimensional K-manifold. A definable $N \subseteq M$ is called *a d-dimensional K-submanifold of M* if every $a \in N$ has a definable open neighborhood $U \subseteq M$ and a definable K-holomorphic $f : U \to K^{n-d}$ such that $N \cap U = f^{-1}(0)$ and such that the K-differential of f at a (which is defined exactly as in the classical case) has K-rank $n - d$.

In [28], Lemma 3.3, we show that every definable K-submanifold of a definable manifold is itself a definable K-manifold, namely has a definable *finite* atlas.

If M and N are definable K-manifolds then a definable map $f : M \to N$ is called *K-holomorphic* if, when read through the charts of M and N, becomes a (definable) K-holomorphic map.

Definition 4.2. A definable $A \subseteq M$ is called *a K-analytic subset of M* if at every $z \in M$, the set A is given, locally near z, as the zero set of finitely

many definable K-holomorphic functions. The set $A \subseteq M$ is called *a locally K-analytic subset of M* if the same is true for every $z \in A$.

The *K-dimension of a K-analytic set A* is defined to be the maximal d such that A contains a d-dimensional K-submanifold of M.

We use $\dim_K A$ to denote the dimension of A as a K-analytic set and $\dim_R A$ to denote its o-minimal dimension. As we show in [28], $\dim_R A = 2 \dim_K A$.

When the underlying real closed field is the field of real numbers then definable $\mathbb{C}$-manifolds and definable $\mathbb{C}$-analytic subsets are just complex manifolds and complex analytic subsets, respectively, which are in addition definable in the underlying o-minimal structure $\mathcal{R}$.

We now review several examples of definable K-manifolds and K-analytic sets in o-minimal structures.

4.2. Compact complex manifolds. An important collection of definable manifolds in o-minimal structures is that of compact complex manifolds.

Every compact complex analytic manifold is isomorphic, as a complex manifold, to a definable $\mathbb{C}$-manifold in the structure $\mathbb{R}_{an}$. More explicitly, assume that $\{\langle U_i, \phi_i\rangle : i \in I\}$ is a finite atlas for an n-dimensional real analytic compact manifold M. Then, as we show in the Appendix, the atlas can be replaced by a new finite atlas $\{\langle B_x, \phi_{i(x)}|B_x\rangle : x \in X\}$, with each B_x an open subset of $U_{i(x)}$ for some $i(x) \in I$, and such that: (i) each $\phi_{i_x}(B_x)$ is a definable subset of $\mathbb{R}^n$ in $\mathbb{R}_{an}$, and (ii) for all $x, y \in X$, the transition maps $\phi_{y,x} := \phi_{i(y)}\phi_{i(x)}^{-1}$ are definable on $\phi_{i(x)}(B_x \cap B_y)$ in $\mathbb{R}_{an}$. It is not hard now to realize M as a definable quotient and, using definable choice in o-minimal expansions of fields (see 6.1.2 in [7]), as a definable set, with a definable atlas.

If M is a compact complex manifold then we use the same process as above. Since the transition maps we obtain are just restrictions of the original maps, we get in this manner a complex manifold which is definable in $\mathbb{R}_{an}$.

If M is a compact complex manifold which is already definable in $\mathbb{R}_{an}$ then every complex analytic subset of M is definable in $\mathbb{R}_{an}$.

Compact complex manifolds were studied elsewhere in model theory after Zil'ber ([39]) proved that, when endowed with all analytic subsets, they admit quantifier elimination and produce a stable structure of finite Morley rank. One may then study the many-sorted structure, denoted sometimes by CCM, given by the category of all compact complex manifolds (up to an isomorphism), with all their analytic subsets and with the analytic maps between them. For a survey of this work see [21] (see also in [13] and in [34]).

Our above discussion implies that the category CCM is interpretable in the o-minimal $\mathbb{R}_{an}$. However, this hides a subtlety that we wish to address here. Note that a compact complex manifold can be realized in many different ways, depending on the underlying set and the choice of atlas. Since we sometimes wish to examine the definability of a particular presentation of a manifold, or the definability of a particular holomorphic function on this manifold, it is often not sufficient to study the manifold "up to an isomorphism". As the following

claim shows, it is possible that the underlying topological space of compact complex manifold is semialgebraic, and yet a complex atlas for the manifold is only definable in $\mathbb{R}_{exp}$.

Claim 4.3. *There is an $\mathbb{R}_{exp}$-definable complex manifold structure $\mathcal{S}$ on the unit sphere S_2 in $\mathbb{R}^3$ such that $\mathcal{S}$ does not have an atlas definable in $\mathbb{R}_{an}$.*

Proof. Let $S_2 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \colon x_1^2 + x_2^2 + x_3^2 = 1\}$ be the unit sphere in $\mathbb{R}^3$, $p_n = (0, 0, 1), p_s = (0, 0, -1)$, and $S_2^* = S_2 \setminus \{p_n, p_s\}$. It is easy to see that S_2^* is semalgebraically homeomorphic to the cylinder $S_1 \times \mathbb{R}$, where $S_1 = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 = 1\}$, and we fix such a homeomorphism $h \colon S_2^* \to S_1 \times \mathbb{R}$. Let $\varphi \colon S_1 \times R \to \mathbb{C}^*$ be the map $\varphi : ((x, y), r) \mapsto (x + iy)e^r$. It is not difficult to see that φ is a homeomorphism definable in $\mathbb{R}_{exp}$. The map $\varphi \circ h \colon S_2^* \to \mathbb{C}^*$ extends to a homeomorphism $\Phi \colon S_2 \to \mathbb{P}_1(\mathbb{C})$ by mapping p_s to 0 and p_N to ∞. Obviously, Φ is definable in $\mathbb{R}_{exp}$.

We use Φ to pull-back the complex structure from $\mathbb{P}_1(\mathbb{C})$ to S_2, and obtain a complex manifold structure $\mathcal{S}$ on S_2. With respect to this structure the map Φ is a biholomorphism. Since $\mathbb{P}^1(\mathbb{C})$ has a semialgebraic atlas, the complex manifold $\mathcal{S}$ has an atlas definable in $\mathbb{R}_{exp}$.

We claim that $\mathcal{S}$ does not have a complex manifold atlas definable in $\mathbb{R}_{an}$. Indeed, if $\mathcal{S}$ admits a definable complex atlas in $\mathbb{R}_{an}$ then Φ should be definable in $\mathbb{R}_{an}$ as well the map $\varphi \colon S_1 \times R \to \mathbb{C}^*$, contradicting the fact that the real exponential function is not definable in $\mathbb{R}_{an}$. $\square$

Remark 4.4. In the above example, since S_2^* is an open subset of S_2, it has an induced complex manifold structure, call is $\mathcal{S}^*$. It is not hard to see that $\mathcal{S}^*$ has an atlas definable in the structure $\mathbb{R}_{an}$(but, as we saw above, this atlas cannot be extended definably in $\mathbb{R}_{an}$ to an atlas for $\mathcal{S}$)

4.3. K-algebraic and K-analytic sets. For every real closed field R and its algebraic closure K, the sets K^n and $\mathbb{P}^n(K)$ are naturally K-manifolds definable in $\langle R, <, +, \cdot \rangle$. More generally, every non-singular algebraic subvariety of K^n or $\mathbb{P}^n(K)$ can be naturally endowed with a semialgebraic K-manifold structure. Algebraic subvarieties of K^n or $\mathbb{P}^n(K)$ are K-analytic subsets of K^n or $\mathbb{P}^n(K)$, respectively.

In fact, using Theorem 2.2 above, we also have the converse (see [28], Theorem 5.1):

Theorem 4.5. *Let $\mathcal{R}$ be an o-minimal expansion of a real closed field R, with K its algebraic closure. If V is a definable K-analytic subset of K^n or of $\mathbb{P}^n(K)$ then V is an algebraic variety over K.*

When we specialize the above theorem to the o-minimal structure $\mathbb{R}_{an}$, we obtain that every definable analytic subset of $\mathbb{P}^n(\mathbb{C})$ is an algebraic variety. However, as we pointed out earlier, every analytic subset of a compact complex

manifold is definable in $\mathbb{R}_{an}$ so we obtain the classical theorem of Chow: *Every analytic subset of* $\mathbb{P}^n(\mathbb{C})$ *is algebraic.*

Similar results for semialgebraic complex analytic sets can be found in [11], and in the "isoalgebraic" setting in [15].

4.4. Definable families of K-manifolds. If X, Y, F are sets with $F \subseteq X \times Y$ then for $x \in X$, we will denote by F_x the fiber $F_x = \{y \in Y : (x, y) \in F\}$. We say that a family $\mathcal{F} = \{F_x : x \in X\}$ of subsets of Y is definable if X, Y and $F \subseteq X \times Y$ are definable sets.

If $\mathcal{R}$ is an o-minimal expansion of $\overline{\mathbb{R}}$ and $\mathcal{R}^*$ is an elementary extension of $\mathcal{R}$ then every K-manifold M which is definable in $\mathcal{R}^*$ is obtained as a fiber in a definable family $\mathcal{F}$ of complex manifolds in the structure $\mathcal{R}$ (by that we mean that the underlying sets of the manifolds as well as their atlases are given by definable families in $\mathcal{R}$). Thus, first order properties of the manifold M reflect uniform properties of manifolds in the family $\mathcal{F}$. Let us consider one such property:

As we know by Riemann's work, every one-dimensional compact complex manifold M is biholomorphic with an algebraic nonsingular projective curve $\mathcal{C}$. If M is definable in $\mathbb{R}_{an}$ then the graph of this biholomorphism is an analytic subset of the definable compact manifold $M \times \mathcal{C}$ and therefore is itself definable in $\mathbb{R}_{an}$.

Assume now that we are given a definable family of compact one-dimensional complex-manifolds $\{M_t : t \in T\}$ in some o-minimal structure $\mathcal{R}$ over $\mathbb{R}$. Is there a definable family of biholomorphisms of these manifolds with projective varieties? Or, equivalently, consider an elementary extension $\mathcal{R}^*$ of $\mathcal{R}$ and a member M_{t_0} of the family, for a parameter t_0 from $\mathcal{R}^*$. Is the K-manifold M_{t_0} definably K-biholomorphic with a projective algebraic variety over K? Our original motivation for asking this question was an analogous theorem of Moosa (see [22]) stating that if the family $\mathcal{F}$ is definable in CCM then indeed there is in CCM a definable family of such biholomorphisms with projective algebraic varieties. It turns out that in the o-minimal setting the answer is negative, as we now describe.

4.5. The family of complex tori. For $\bar{\omega} = (\omega_1, \ldots, \omega_{2n})$ a tuple of $2n$ vectors in $\mathbb{C}^n$ which are linearly independent over $\mathbb{R}$, let $\Lambda_{\bar{\omega}} \subset \mathbb{C}^n$ be the lattice $\mathbb{Z}\omega_1 + \cdots + \mathbb{Z}\omega_{2n}$. Since $\Lambda_{\bar{\omega}}$ is a discrete subgroup of $(\mathbb{C}^n, +)$, the quotient group $\mathcal{E}_{\bar{\omega}} = (\mathbb{C}^n, +)/(\Lambda_{\bar{\omega}}, +)$ inherits a complex-analytic structure, and with respect to this structure, $\mathcal{E}_{\bar{\omega}}$ is a connected compact complex Lie group of dimension n, i.e. *an n-dimensional complex torus.*

Although the lattice $\Lambda_{\bar{\omega}}$ is an infinite discrete set and thus is not definable in any o-minimal structure, we are going to view these tori definably as follows: The underlying set of $\mathcal{E}_{\bar{\omega}}$ is identified with the definable parallelogram

$$E_{\bar{\omega}} = \left\{ t_1\omega_1 + \cdots + t_{2n}\omega_{2n} : \bigwedge_{i=1}^{2n} 0 \le t_i < 1 \right\}, \tag{1}$$

and then it is not hard to produce a semialgebraic atlas on $E_{\bar{\omega}}$ with semialgebraic transition maps, corresponding to the complex analytic structure of $\mathcal{E}_{\bar{\omega}}$. Therefore each $\mathcal{E}_{\bar{\omega}}$ can be viewed as a $\mathbb{C}$-manifold definable in the field $\overline{\mathbb{R}}$, and moreover these definable charts and transition maps can be constructed uniformly in $\bar{\omega}$, thus obtaining a semi-algebraic family of all n-dimensional complex tori. It follows that in every real closed field R, if we take a tuple $\bar{\omega}$ of $2n$ vectors in K^n $(K = R(\sqrt{-1}))$ which are linearly independent over R, we have a corresponding definable K-manifold $\mathcal{E}_{\bar{\omega}}$, which we call a K-torus.

In [26] we considered the family, call it $\mathcal{F}$, of all one-dimensional complex tori in various o-minimal expansions of $\overline{\mathbb{R}}$. Each member of $\mathcal{F}$ is an elliptic curve, i.e. biholomorphic with a smooth projective cubic curve. The biholomorphism between these two compact complex manifolds is definable in $\mathbb{R}_{an}$. However, as we show in [26], Corollary 5.6, a full family of such biholomorphisms is not definable in an o-minimal structure. Formulated in the language of non-standard o-minimal structures we have:

Theorem 4.6. *Let $\mathcal{R}$ be an arbitrary o-minimal expansion of $\mathbb{R}_{an,exp}$ and let $\mathcal{R}^* = \langle R, <, +, \cdot, \cdots \rangle$ be a non-Archimedean elementary extension of $\mathcal{R}$, with $K = R(\sqrt{-1})$. If $\tau \in K$ is such that $Im(\tau) > 0$ and $Re(\tau)$ greater than all standard $n \in \mathbb{N}$, then the K-torus $\mathcal{E}_{1,\tau}$ is not definably K-biholomorphic, in the structure $\mathcal{R}^*$, with any algebraic curve.*

We thus showed the failure of the definable analogue to Riemann's Existence Theorem, for definably compact one-dimensional K-manifolds in o-minimal structure (a "definably compact manifold" here can be taken to mean a $\mathcal{R}^*$-fiber in an $\mathcal{R}$-definable family of compact real manifolds).

In Section 4.6 below and Section 6.2 we discuss some positive cases of Riemann's theorem.

4.6. Mild manifolds.

We let $\mathcal{R}$ be an o-minimal expansion of a real closed field R and $K = R(\sqrt{-1})$

Let M be a definable K-manifold, and let $\mathcal{A}(M)$ be the structure whose universe is M and its atomic relation are all the definable K-analytic subsets of M^n, $n \in \mathbb{N}$. In [27] we called M *a mild manifold* if $\mathcal{A}(M)$ admits quantifier elimination. Examples are compact complex manifolds (by Zil'ber's work [39]), definably compact K-manifolds (see Theorem 8.3 in [28]), the set of K-regular points of an algebraic variety over K (projective or affine). On the other hand, the open unit disc in $\mathbb{C}$ is a definable complex-manifold which is not mild in any o-minimal structure.

In an attempt to understand better the previous example of a non-algebraic one-dimensional K-torus we proved the following result (see Theorem 6.0.1, and Theorem 4.4.3 [27]), which can be seen as a conditional Riemann Existence Theorem.

Theorem 4.7. *Let M be a definable K-manifold which is mild and also strongly minimal (namely, in the structure $\mathcal{A}(M)$ every definable subset of M is finite or co-finite). Then the following are equivalent:*

1. *$\mathcal{A}(M)$ is non locally modular.*
2. *There is a finite $F \subseteq M$ and a definable non-constant K-holomorphic function $\phi : M \smallsetminus F \to K$ (we call ϕ* a K-mermomorphic function *on M).*
3. *There is a definable K-biholomorphism between M and a non-singular algebraic curve over K.*

In particular, the non algebraic one-dimensional K-torus $\mathcal{E}_{1,\tau}$ of Theorem 4.6 admits no definable nonconstant K-meromorphic map into K and $\mathcal{A}(\mathcal{E})$ is locally modular. If we translate the above theorem to definable families of compact complex one-dimensional manifolds (which are all mild and strongly minimal) then we get some uniform version of Riemann's theorem:

Corollary 4.8. *Let $\mathcal{R}$ be an o-minimal structure over $\mathbb{R}$ and let $\mathcal{F} = \{M_t : t \in T\}$ be a a definable family of one-dimensional compact complex manifolds, given together with a definable family $\phi_t : M_t \to \mathbb{C}$ of nonconstant meromorphic maps.*

Then, there is in $\mathcal{R}$ a definable family of complex algebraic curves $\{C_t : t \in T\}$ and a definable family of complex biholomorphisms $\sigma_t : M_t \to C_t$.

5. Theorems on Removal of Singularities

One of the most useful features of working with analytic objects which are definable in o-minimal structures is the theory of removal of singularities: Start with a complex manifold M, an open set $U \subseteq M$ and consider an analytic subset A of U. In general, the topological closure of A in M is not analytic in M. A great deal of attention has been given classically to conditions under which $Cl(A)$ is analytic in M. Assuming that M, U and A are definable in an o-minimal structure one obtains strong results in both the standard and non-standard settings.

5.1. Characterizing K-analytic sets

Definition 5.1. Given a definable K-manifold M, and a definable $A \subseteq M$, we define *the set of K-regular points of A*, denoted by $Reg_K(A)$, as the set of all points $a \in A$ such that in some neighborhood of a, the set A is a K-submanifold of M. We let $Sing_K(A) = A \smallsetminus Reg_K(A)$.

We call a definable $A \subseteq M$ *a finitely K-analytic subset of M* if M can be covered by finitely many definable open sets $M = \bigcup_j W_j$ and for each j there is a definable K-holomorphic map $\psi_j : W_j \to K^{m_j}$, such that $A \cap W_j = \psi^{-1}(0)$.

Clearly, every finitely K-analytic set in K-analytic. As for the converse, note that if M is a compact complex manifold, definable in $\mathbb{R}_{an}$, then every $\mathbb{C}$-analytic subset of M is $\mathbb{C}$-finitely analytic. It turns out that o-minimality can replace the role of compactness and that in the o-minimal setting this converse is always true. Here is one of the main theorems characterizing definable K-analytic sets (see[28], Corollary 4.14):

Theorem 5.2. *Let M be a definable K-manifold and $A \subseteq M$ a definable closed set. Then the following are equivalent:*

1. *A is a K-analytic subset of M.*
2. *A is a finitely K-analytic subset of M.*
3. *For every open $W \subseteq K^n$, $\dim_R(Sing_K(A \cap W)) \leq \dim_R(A \cap W) - 2$.*

Another strong variant of Remmert-Stein's Theorem is (see [28], Theorem 4.1.3):

Theorem 5.3. *Let M be a definable K-manifold and E a definable K-analytic subset of M. If A is a definable K-analytic subset of $M \setminus E$ then $Cl(A)$ is K-analytic in M.*

If we specialize to complex manifolds then the above theorem follows from Remmert-Stein when we assume that A is of pure dimension and $\dim_{\mathbb{C}} E < \dim_{\mathbb{C}} A$.

Remark 5.4. 1. Note that the implication (3) $\Rightarrow$ (1) in Theorem 5.2 fails without the definability assumption: Take $M = \mathbb{C}^3$ and let

$$A = \{(x, e^{1/x}, 1) \in \mathbb{C}^3 : x \neq 0\} \cup \{(0, y, z) \in \mathbb{C}^3\}.$$

The set A is a closed subset of $\mathbb{C}^3$ and its set of singular points is $\{(0, y, 1) \in \mathbb{C}^3\}$. For every open $W \subseteq \mathbb{C}^3$, either $W \cap \{(0, y, 1)\} = \emptyset$, in which case $Sing_{\mathbb{C}}(W \cap A) = \emptyset$ or $Sing_{\mathbb{C}}(W \cap A) = W \cap \{(0, y, 1) : y \in \mathbb{C}\}$, in which case the real dimension of this set is 2 while the real dimension of $W \cap A$ is 4. However, A is not an analytic subset of $\mathbb{C}^3$.

2. Clause *(3)* of Theorem 5.2 can be expressed in a first-order way, after showing that $Reg_K(A)$ is definable, uniformly in families, for $A \subseteq K^n$. Working in the charts of M, it then follows from Theorem 5.2 that if $\{A_t : t \in T\}$ is a definable family of subsets of M, then the collection

$$\{t \in T : A_t \text{ is an analytic subset of } M\}$$

is definable.

Putting this last observation together with Theorem 4.5, we obtain the following interesting result:

Theorem 5.5. *Let $\{X_t : t \in T\}$ be a definable family of subsets of K^n. Then the set of all $t \in T$ such that X_t is an algebraic subset of K^n is definable.*

5.2. Definable K-holomorphic maps. Let us now consider the implications of the above results on definable K-holomorphic maps. The main results here are (see [29], Corollary 6.3, and [28], Theorem 7.3)

Theorem 5.6. *Let $f : M \to N$ be a definable K-holomorphic map between definable K-manifolds, and $A \subseteq M$ a definable K-analytic subset of M. Then*

1. *There is a closed definable set $E \subseteq N$, with $\dim_R(E) \leq \dim_R f(A) - 2$, and with $\dim_R(f^{-1}(E) \cap A) \leq \dim_R(A) - 2$, such that $f(A) \smallsetminus E$ is a locally K-analytic subset of N.*

2. *If $f(A)$ is a closed subset of N then it is a K-analytic subset of N.*

Clause *(2)* is a strong variant of Remmert's proper mapping theorem. Again, it fails without the definability assumptions. Indeed, the projection of the analytic set $\{(n, 1/n) \in \mathbb{C} \times \mathbb{C} : n \geq 1\} \cup \{(0,0)\}$ on its first coordinate is the closed set $\{1/n : n \geq 1\} \cup \{0\}$ which is clearly not an analytic subset of $\mathbb{C}$.

5.3. Compactification of analytic spaces. Consider an action of an infinite discrete group Γ on a complex manifold M. Under various assumptions one can endow the quotient $\Gamma\backslash M$ with the structure of a complex analytic space or even that of quasi-affine or quasi-projective variety (see the seminal work [2] on arithmetic quotients). We note here how one may apply the theory on removal of singularities in order to prove results of similar flavor, assuming the existence of a partially definable holomorphic Γ-periodic map ϕ from M into another manifold N. Note that even if M and N are definable in some o-minimal structure the map ϕ is generally not definable there, because of the infinite period Γ. However, as we demonstrate in sections 6.2 and 6.3, we can sometimes prove the definability of ϕ on a definable $U \subseteq M$, with $\phi(U) = \phi(M)$ and, as the following result shows, for certain purposes this is sufficient (for a proof, see Appendix).

Theorem 5.7. *Let $\mathcal{R}$ be an o-minimal expansion of the real field. Let $\phi : U \to N$ be a definable finite-to-one holomorphic map from an open $U \subseteq \mathbb{C}^n$ into a definable complex manifold N. Assume that there is a set $D \subseteq U$ (not necessarily definable) which is closed in $\mathbb{C}^n$, such that $\phi(U) = \phi(D)$. Then the topological closure of $\phi(U)$ in N, call it A, is a complex analytic subset of N, and $\dim_{\mathbb{R}}(A \smallsetminus \phi(U)) \leq 2n - 2$.*

6. Classical Holomorphic Functions in an o-minimal Setting

Although all germs of holomorphic maps are definable in $\mathbb{R}_{an}$, if one wishes to apply o-minimal techniques to classical mathematical questions, it is necessary to consider certain holomorphic functions on their natural domains, or on

sufficiently large sub-domains, and prove their definability in some o-minimal structure. In this section we consider several such cases.

6.1. The Riemann mapping. The Riemann mapping theorem says that if $\Omega \subseteq \mathbb{C}$ is a non-empty simply connected open set which is not equal to $\mathbb{C}$ then there is a biholomorphism $f : \Omega \to D$ with the open unit disc in $\mathbb{C}$. The map is unique up to a biholomorphism of D. What can be said about the definability of f in some o-minimal structure, assuming that Ω is definable there?

In [17] Kaiser shows that when Ω is a polygon (in which case f is known as the Schwarz-Chirstoffel map), the map f is indeed definable in the o-minimal structure $\mathbb{R}_{an}^{\mathbb{R}}$, the expansion of $\mathbb{R}_{an}$ by all power functions x^α, $\alpha \in \mathbb{R}$. In [18] he also shows:

Theorem 6.1. *There is an o-minimal structure $\mathcal{R}$ with the following property. Let $\Omega \subset \mathbb{C}$ be a bounded simply connected domain that is definable in $\mathbb{R}_{an}$ and assume that for every x which is a singular boundary point of Ω, the angle of the boundary at x is an irrational multiple of π. Then the biholomorphic map $f : \Omega \to D$ which is given by Riemann's theorem is definable in $\mathcal{R}$.*

The o-minimal structure in the theorem is constructed in [19].

6.2. The Weierstrass $\wp$-function and elliptic curves. We return to the family of one dimensional tori discussed in Section 4.5. For the classical facts mentioned here, see [35].

Every one dimensional torus is bi-holomorphic with a torus $\mathbb{C}/\Lambda$, with $\Lambda = \mathbb{Z} + \tau\mathbb{Z}$ and τ in the upper half plain $\mathcal{H} = \{\tau \in \mathbb{C} \colon \mathrm{Im}(\tau) > 0\}$. We denote the corresponding torus by $\mathcal{E}_\tau$, and its underlying set defined in Section 4.5 (1) by E_τ.

The group of $SL(2,\mathbb{Z})$ acts on $\mathcal{H}$ via

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \mapsto \frac{az+b}{cz+d},$$

and two tori $\mathcal{E}_\tau, \mathcal{E}_{\tau'}$ are biholomorphic if and only if $\tau = A\tau'$ for some $A \in SL(2,\mathbb{Z})$.

Recall that the Weierstrass $\wp$-function is a meromorphic map $\wp(\tau, z)$ from $\mathcal{H} \times \mathbb{C}$ into $\mathbb{C}$, so that for each $\tau \in \mathcal{H}$ the map $\wp_\tau \colon z \mapsto \wp(z,\tau)$ is a Λ_τ-periodic meromorphic map on the whole of $\mathbb{C}$, and the map $g_\tau \colon z \mapsto (1 \colon \wp_\tau(z) \colon \wp'_\tau(z))$ induces an embedding of $\mathcal{E}_\tau$ into $\mathbb{P}_2(\mathbb{C})$. We also have $\wp(\tau, z) = \wp(A\tau, z)$ for any $A \in SL_2(2,\mathbb{Z}), \tau \in \mathcal{H}, z \in \mathbb{C}$.

The function $\wp(z,\tau)$ cannot be definable in an o-minimal structure on all of $\mathbb{C} \times \mathcal{H}$ because of the periodicity in z and in τ. We consider, instead of the whole of $\mathcal{H}$, the set

$$\mathfrak{F} = \{\tau \in \mathcal{H} : -1/2 \le \mathrm{Re}(\tau) < 1/2 \quad \text{and} \quad |\tau| \ge 1\},$$

and the family of tori $\mathcal{E}^{\mathfrak{F}} = \{\mathcal{E}_\tau : \tau \in \mathfrak{F}\}$. The choice of the subfamily $\mathcal{E}^{\mathfrak{F}}$ is quite standard, since $\mathfrak{F}$ contains a representative of every orbit of $SL(2,\mathbb{Z})$, and therefore every one-dimensional torus is biholomorphic with some $\mathcal{E}_\tau$ for $\tau \in \mathfrak{F}$.

We have (see [26], Theorem 4.1):

Theorem 6.2. *The restriction of $\wp(\tau, z)$ to the set*

$$\{(z,\tau) \in \mathbb{C} \times \mathcal{H} : \tau \in \mathfrak{F} \text{ and } z \in E_\tau\}$$

is definable in the structure $\mathbb{R}_{an,exp}$.

Since $\mathcal{H}$ and $\mathfrak{F}$ are semialgebraic sets, they can be interpreted in any real closed field R. We denote these by $\mathcal{H}(R)$ and $\mathfrak{F}(R)$. As a corollary to the theorem above we have (see [26], theorem 5.4):

Theorem 6.3. *Let $\mathcal{R} = \langle R, <, +, \cdots\rangle$ be an arbitrary model of $\mathbb{R}_{an,exp}$, $K = R(\sqrt{-1})$.*

(i) If $\tau \in \mathfrak{F}(R)$ then $\mathcal{E}_\tau$ is definably K-biholomorphic to a nonsingular cubic curve in $\mathbb{P}^2(K)$.

(ii) If $\mathcal{C} \subseteq \mathbb{P}^n(K)$ is a nonsingular algebraic curve of genus one then there is a $\tau \in \mathfrak{F}(K)$ and a K-biholomorphism of $\mathcal{C}$ and $\mathcal{E}_\tau$ which is definable in $\mathcal{R}$.

Thus in all models of $\mathbb{R}_{an,exp}$, every projective curve of genus one over K is definably K-biholomorphic to a one-dimensional K-torus $\mathcal{E}_\tau$ with $\tau \in \mathfrak{F}(R)$. But as we showed in Section 4.5, it is not true that every one dimensional K-torus is definably K-biholomorphic to an algebraic curve.

As before, the last theorem can be stated in the language of $\mathbb{R}_{an,exp}$-definable families of complex curves and holomorphic maps.

6.2.1. O-minimality and arithmetic. Several articles in recent years make connections between o-minimality and arithmetical questions in complex algebraic geometry. The starting point of this analysis is a theorem of Pila and Wilkie concerning the distribution of rational points on subsets of $\mathbb{R}^n$ which are definable in o-minimal structures (see [33]). Given a complex algebraic variety, Pila and Zannier, [32], used transcendental holomorphic functions on bounded sets, definable in $\mathbb{R}_{an}$, to translate questions about torsion points in complex abelian varieties into questions on rational points of $\mathbb{R}_{an}$-definable subsets of $\mathbb{C}^n$. Having that, they use the Pila-Wilkie result, together with number theoretic considerations and o-minimality to give a new proof for the Manin-Mumford conjecture.

More recently, Pila, [30], [31], used Theorem 6.2 above to translate questions about special points in the moduli space of elliptic curves into questions on quadratic points in $\mathbb{R}_{an,exp}$-definable subsets of $\mathbb{C}^n$. Using a variant of his theorem with Wilkie, together with number theoretic results and o-minimality, he was able to prove certain open cases of the André-Oort conjecture.

6.3. The theta functions and abelian varieties. In this section we describe a recent, still unpublished work.

As was pointed out in Section 4.5, the family of n-dimensional complex tori $\mathcal{E}_{\bar{\omega}}$ can be viewed as a definable family of complex manifolds in the structure $\overline{\mathbb{R}}$. A torus $\mathcal{E}_{\bar{\omega}}$ is called *an abelian variety* if it is biholomorphic to a projective algebraic variety. We first review briefly the relevant information regarding abelian varieties (see [5] [16]).

We already discussed the fact that every 1-torus is an abelian variety. When $n > 1$, the family of abelian varieties is a proper sub-collection of the family of all n-tori, given as countable union of definable subfamilies $\mathcal{F}_D$, where D runs over all $n \times n$ diagonal matrices

$$D = Diag(d_1, d_2, \ldots, d_n),$$

with $d_1|d_2|\ldots|d_n$ positive integers. Each $\mathcal{F}_D$ is defined as follows:

We denote by $\mathcal{H}_n$ the Siegel half space of all $n \times n$ complex symmetric matrices with a positive definite imaginary part. We now fix D as above (called the polarization type). For $\tau \in \mathcal{H}_n$ we denote by $\Lambda_{\tau,D}$ the lattice which is generated by the columns of the $n \times 2n$ complex matrix (τ, D). We let $\mathcal{E}_{\tau,D}$ denote the corresponding torus. Let $\mathcal{F}_D = \{\mathcal{E}_{\tau,D} : \tau \in \mathcal{H}_n\}$ be the family of all polarized tori with polarization type D. It is known that a complex n-torus is an abelian variety if and only if it is biholomorphic to a torus from one of the families $\mathcal{F}_D$ (but each abelian variety appears in more than one such family).

Let $Sp(D, \mathbb{Z})$ be the group of $2n \times 2n$ integral matrices preserving the alternating form

$$\begin{pmatrix} 0 & D \\ -D & 0 \end{pmatrix}.$$

The group $Sp(D, \mathbb{Z})$ acts on $\mathcal{H}_n$ and any two polarized varieties $\mathcal{E}_{\tau_1,D}$ and $\mathcal{E}_{\tau_2,D}$ in $\mathcal{F}_D$ are isomorphic (as polarized varieties) if and only if they are in the same orbit of $Sp(D, \mathbb{Z})$.

There is a natural number k such that every $\mathcal{E}_{\tau,D}$ can be embedded, via a map which we denote by $\Theta_{\tau,D}$, into $\mathbb{P}^k(\mathbb{C})$. We are interested in uniform definability of these embeddings.

As in the case $n = 1$, although each $\Theta_{\tau,D}$ is definable in $\mathbb{R}_{an}$, the whole family $\Theta_{\tau,D}(z), \tau \in \mathcal{H}_n$, can not be defined in any o-minimal structure because of periodicity in τ, and we need to choose an appropriate subfamily. It follows from Siegel's reduction theory (see [16], p. 189-197), that there is a a semi-algebraic set $\mathfrak{F}_n^D \subseteq \mathcal{H}_n$ containing finitely many representatives for each orbit of $Sp(D, \mathbb{Z})$.

Theorem 6.4. *For every polarization type D the family of embeddings $\{\Theta_{\tau,D} : \tau \in \mathfrak{F}_n^D\}$ is definable in the structure $\mathbb{R}_{an,exp}$.*

The above theorem is equivalent to definability of certain theta functions which we now describe.

We use $\widetilde{\Theta}_{\tau,D} : \mathbb{C}^n \to \mathbb{P}^k(\mathbb{C})$ to denote the pullback of $\Theta_{\tau,D}$, i.e. $\widetilde{\Theta}_{\tau,D}$ is a $\Lambda_{\tau,D}$-periodic map which induces $\Theta_{\tau,D}$, and let $\widetilde{\Theta}_D : \mathcal{H}_n \times \mathbb{C}^n \to \mathbb{P}^k(\mathbb{C})$ be $\widetilde{\Theta}_D(\tau, z) = \widetilde{\Theta}_{\tau,D}(z)$. The map $\widetilde{\Theta}_D$ can be obtained as the composition $\pi \circ \widetilde{\vartheta}_D$, where $\pi\colon \mathbb{C}^{k+1} \to \mathbb{P}^k(\mathbb{C})$ is the canonical projection, and $\widetilde{\vartheta}_D\colon \mathcal{H}_n \times \mathbb{C}^n \to \mathbb{C}^{k+1}$ is a map whose coordinate functions are given by theta functions $\vartheta_{a,b}(z, \tau)$, for various $a, b \in \mathbb{Q}^n$. The theta functions are given explicitly by the following formula:
For $a, b \in \mathbb{R}^n$, $z \in \mathbb{C}^n$ column vectors and a matrix $\tau \in \mathcal{H}_n$ (we use ${}^t z$ to denote the transpose of a column vector z),

$$\vartheta_{a,b}(\tau, z) = \sum_{m \in \mathbb{Z}^n} e^{i\pi({}^t(m+a)\tau(m+a) + 2\,{}^t(m+a)(z+b))}.$$

We define

$$\Omega_n = \{(\tau, z) \in \mathcal{H}_n \times \mathbb{C}^n \colon \tau \in \mathfrak{F}_n^{I_n} \text{ and } z \in E_{\tau.I_n}\}.$$

Theorem 6.4 can be deduced from the following result.

Theorem 6.5. *For every $a, b \in \mathbb{R}^n$, the map $(\tau, z) \mapsto \vartheta_{a,b}(\tau, z)$ restricted to Ω_n is definable in the o-minimal structure $\mathbb{R}_{an,exp}$.*

We end this section by observing how o-minimality can be be used in the construction of moduli spaces of polarized abelian varieties. We assume that D is a polarization type with d_1 divisible by 4.

We need the following fact (see Theorem V.4 in [16])

Fact 6.6. *There is a subgroup $\Gamma < Sp(D, \mathbb{Z})$ of finite index and a holomorphic map $\varphi\colon \mathcal{H}_n \to \mathbb{P}^N(\mathcal{C})$, whose coordinates are given by maps $\tau \mapsto \vartheta_{a,b}(\tau, 0)$ such that $\varphi(\tau) = \varphi(\tau')$ if and only if τ and τ' are in the same orbit of Γ.*

The map φ from the above fact induces a map from $\Gamma \setminus \mathcal{H}_n$ into $\mathbb{P}^N(\mathbb{C})$, and an important issue in the theory of moduli spaces is the nature of the image of this map. The main result is that this image is dense inside some algebraic subvariety of $\mathbb{P}^N(\mathbb{C})$ (see Theorem V.8 in [16]). Let us see how o-minimality yields an alternative proof of this fact.

Since Γ has finite index in $Sp(D, \mathbb{Z})$, we can choose a semi-algebraic F consisting of finitely many translates of $\mathfrak{F}_n^D$ such that $\varphi(\mathcal{H}_n) = \varphi(F)$.

Using Theorem 6.5 and transformation formulas for theta functions, we can get φ to be definable on an open set $U \subseteq \mathcal{H}_n$ containing the closure of F. We now view $\varphi\colon U \to \mathbb{P}^N(\mathbb{C})$ as a definable holomorphic map from an open subset of $\mathbb{C}^\ell$ (with $\ell = dim(\mathcal{H}_n)$) into the definable manifold $\mathbb{P}^N(\mathbb{C})$. We can therefore apply Theorem 5.7 and deduce that the closure of $\varphi(F)$ is an analytic subvariety of $\mathbb{P}^N(\mathbb{C})$, so by Chow's Theorem must be algebraic. It immediately follows that the closure of image of $\Gamma \setminus \mathcal{H}_n$ under φ is algebraic as well.

7. Appendix

7.1. Definability of compact real analytic manifolds in $\mathbb{R}_{an}$.

We prove here the result claimed in Section 4.2.

Proposition 7.1. *Let M be a n-dimensional compact real analytic manifold with a given finite atlas $\{\langle U_i, \phi_i\rangle : i \in I\}$. Then there is a finite open cover $M = \bigcup_{x\in X} B_x$ with the properties:*
(i) For each $x \in X$ there is an $i(x) \in I$ with $B_x \subseteq U_{i(x)}$, such that $\phi_{i(x)}(B_x)$ is a subset of $\mathbb{R}^n$ which is definable in $\mathbb{R}_{an}$.
(ii) For all $x, y \in X$, the sets $\phi_{i(x)}(B_x \cap B_y)$ and the restriction of the transition map $\phi_{i(y)}\phi_{i(x)}^{-1}$ to this set are definable in $\mathbb{R}_{an}$.

Proof. Without loss of generality each $\phi_i(U_i)$ is a bounded subset of $\mathbb{R}^n$. We denote by ϕ_{ij} the real analytic transition map

$$\phi_i\phi_j^{-1} : \phi_j(U_i \cap U_j) \to \phi_i(U_i \cap U_j).$$

As was pointed out in the examples of Section 2, if $B \subseteq \phi_j(U_i \cap U_j)$ is a definable set whose closure is contained $\phi_j(U_i \cap U_j)$ then the restriction of ϕ_{ij} to B is definable in $\mathbb{R}_{an}$.

By compactness, for each $i \in I$ there is an open $V_i \subseteq Cl(V_i) \subseteq U_i$ such that $M = \bigcup_{i\in I} V_i$. Now, for every $x \in M$, we choose a neighborhood B_x of x such that

$$B_x \subseteq Cl(B_x) \subseteq \bigcap_{x\in V_i} V_i \cap \bigcap_{x\in U_j} U_j \cap \bigcap_{x\in Cl(V_t)^c} Cl(V_t)^c, \tag{2}$$

and

$$\text{for every } i \in I \text{ for which } x \in U_i, \text{ the set } \phi_i(B_x) \text{ is definable in } \mathbb{R}_{an}. \tag{3}$$

Indeed, this is possible to do since we only we need to choose B_x small enough to satisfy (2) and in addition require that for some fixed $U_j \ni x$, the set $\phi_j(B_x)$ is an open rectangular box in $\mathbb{R}^n$. To verify (2), by our choice of B_x, if $x \in U_i$ then $Cl(B_x) \subseteq U_i \cap U_j$ and hence $\phi_j(Cl(B_x))$ is a closed rectangular box inside $\phi_j(U_i \cap U_j)$. Therefore, as we observed already, the restriction of ϕ_{ij} to $Cl(\phi_j(B_x))$ is definable in $\mathbb{R}_{an}$. But then, $\phi_i(B_x) = \phi_{ij}(\phi_j(B_x))$ is definable as well, as required.

Claim Given $i, j \in I$, assume that $x \in V_i$, $y \in V_j$ and $B_x \cap B_y \neq \emptyset$. Then $Cl(B_x \cup B_y) \subseteq U_i \cap U_j$ and $\phi_i(B_x), \phi_i(B_y), \phi_j(B_x), \phi_j(B_y)$ are all definable in $\mathbb{R}_{an}$.

Indeed, since $B_x \cap B_y \neq \emptyset$ and $B_x \subseteq V_i$ we have $B_y \cap V_i \neq \emptyset$ and therefore $y \in Cl(V_i)$ (for otherwise, by the choice of B_y, we would have $B_y \subseteq Cl(V_i)^c$, a contradiction). By our choice the V_i's, it follows that $y \in U_i$ and therefore $Cl(B_y) \subseteq U_i$. We also have $Cl(B_x) \subseteq Cl(V_i) \subseteq U_i$ and therefore

$Cl(B_x \cup B_y) \subseteq U_i$. Similarly, we have $Cl(B_x \cup B_y) \subseteq U_j$. By our definition of B_x, B_y we have $\phi_i(B_x), \phi_i(B_y), \phi_j(B_x), \phi_j(B_y)$ all definable, proving the claim.

By compactness, there is a finite set $X \subseteq M$, such that $M = \bigcup_{x \in X} B_x$. For each $x \in X$ we choose $i(x) \in I$ such that $x \in V_{i(x)}$. By the claim, if $B_x \cap B_y \neq \emptyset$ then $\phi_{i(x)}(B_x \cap B_y) = \phi_{i(x)}(B_x) \cap \phi_{i(x)}(B_y)$ is definable in $\mathbb{R}_{an}$ and furthermore, the closure of $\phi_{i(x)}(B_x \cap B_y)$ is contained in $\phi_{i(x)}(U_{i(x)} \cap U_{i(y)})$. It follows that the restriction of $\phi_{i(y)}\phi_{i(x)}^{-1}$ to this set is definable. □

7.2. The proof of Theorem 5.7

Proof. Let $Fr(\phi(U)) = A \smallsetminus \phi(U)$ be the frontier of $\phi(U)$. We first prove that $\dim_{\mathbb{R}}(Fr(\phi(U))) \leq 2n - 2$.

Consider $\mathbb{C}^n$ as a subset $\mathbb{P}^n(\mathbb{C})$, namely we write $\mathbb{P}^n(\mathbb{C}) = \mathbb{C}^n \cup H$ for H a hyperplane at ∞. Let G be the closure in $\mathbb{P}^n(\mathbb{C}) \times N$ of the graph of ϕ and let $\pi : \mathbb{P}^n(\mathbb{C}) \times N \to N$ be the projection onto the second coordinate. We claim that $Fr(\phi(U)) = \pi(G \cap (H \times N))$. The right-to-left inclusion is immediate. For the converse, if $y \in Fr(\phi(U))$ then there is a sequence $x_n \in U$ such that $\phi(x_n)$ tends to y. Since $\phi(U) = \phi(D)$ we may assume that $x_n \in D$. Because D is closed in $\mathbb{C}^n$ and $y \notin \phi(U)$, the sequence x_n does not have any converging subsequence in $\mathbb{C}^n$ and therefore it is unbounded in $\mathbb{C}^n$. But then, viewed in $\mathbb{P}^n(\mathbb{C})$, the sequence has a converging subsequence to an element $z \in H$, and then $(z, y) \in G \cap (H \times N)$, hence $y \in \pi(G \cap (H \times N))$

Next, consider the set $B_{\inf}$ of all $(z, y) \in G \cap (H \times N)$ such that there are infinitely many $y' \in N$ with $(z, y') \in G$ and let $B_{fin} = G \cap (H \times N) \smallsetminus B_{\inf}$. By [29], Lemma 6.7 (ii), $\dim_{\mathbb{R}}(B_{\inf}) \leq 2n - 2$, and since $\dim_{\mathbb{R}} H = 2n - 2$ we also have $\dim_{\mathbb{R}} B_{fin} \leq 2n - 2$. It follows that $\dim_{\mathbb{R}}(G \cap (H \times N)) \leq 2n - 2$. We now have,

$$\dim_{\mathbb{R}}(Fr(\phi(U))) = \dim_{\mathbb{R}}(\pi(G \cap (H \times N))) \leq 2n - 2,$$

as claimed.

By Theorem 5.6 (1), there is a definable closed set $E \subseteq N$, with $\dim_{\mathbb{R}}(E) \leq \dim_{\mathbb{R}} \phi(U) - 2$, such that $\phi(U) \smallsetminus E$ is locally analytic in N. Because $A = (\phi(U) \smallsetminus E) \cup E \cup Fr(\phi(U))$, we have

$$Sing_{\mathbb{C}}(A) \subseteq E \cup Fr(\phi(U)) \cup Sing_{\mathbb{C}}(\phi(U) \smallsetminus E).$$

Since ϕ is finite-to-one, the real dimension of $\phi(U)$ is $2n$ everywhere and therefore the real dimension of $Sing_{\mathbb{C}}(\phi(U) \smallsetminus E)$ is at most $2n - 2$ everywhere (see 5.2(3)). We thus have for all open $W \subseteq N$,

$$\dim_{\mathbb{R}}(Sing_{\mathbb{C}}(A \cap W)) \leq 2n - 2 = \dim_{\mathbb{R}}(W \cap A) - 2.$$

We can now apply Theorem 5.2(3) once more and conclude that A is an analytic subset of N. □

References

[1] W. L. Baily *On the theory of θ-functions, the moduli of abelian varieties, and the moduli of curves*, Ann. Math., Second Series, **75**, No. 2 (Mar., 1962), 342–381.

[2] W. L. Baily and A. Borel, *Compactification of arithmetic quotients of bounded symmetric domains*, Ann. Math., Second series, **84** No. 3 (Nov., 1966), 442–528.

[3] R. Bianconi, *Undefinability results in o-minimal expansions of the real numbers*, Ann. Pure Appl. Logic **134**, (2005), 43–51.

[4] E. M. Chirka, *Complex Analytic Sets*, , translated from the Russian by R. A. M. Hoksbergen, Math. Appl. (Soviet Ser.) **46**, Kluwer Academic Publisher Group, Dordrecht 1989.

[5] O. Debarre, *Complex Tori and Abelian Varieties*, translated from the French by P. Mazaud, SMF/AMS texts and monographs **11**, 2005.

[6] I. Denef and L. van den Dries, *p-Adic and real subanalytic sets*, Ann. Math. **128** (1988), 79–138.

[7] L. van den Dries, *Tame topology and o-minimal structures*, London Math. Soc. Lect. Notes Ser. **248**, Cambridge University Press, Campbridge 1998.

[8] L. van den Dries, A. Macintyre and D. Marker, *The elementary theory of restricted analytic fields with exponentiation*, Ann. Math. **140** (1994), 183-205.

[9] L. van den Dries and C. Miller, *On the real exponential field with restricted analytic functions*, Israel J. Math. **85** (1994), 19-56.

[10] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J. **84** (1996), no. 2, 497–540.

[11] E. Fortuna and S. Łojasiewicz, *Sur l'algébricité des ensembles analytiques complexes*, J. Reine Angew. Math, **329**, 1981, 215–220.

[12] E. Fortuna, S. Łojasiewicz and M. Raimondo, *Algébricité de germes analytiques*, J. Reine Angew. Math, **374**, 1987, 208–213.

[13] E. Hrushovski, *Geometric model theory*, Documenta Mathematica extra volume ICM 1 (1998), 281-302.

[14] E. Hrushovski and Y. Peterzil, *A question of van den Dries and a theorem of Lipshitz and Robinson, not everything is standard*, J. Sym. Logic **72** 1 (2007), 119–122.

[15] R. Huber and M. Knebusch, *A glimpse at isoalgebraic spaces*, Note Mat. **10** (1990), suppl. 2, 315–336.

[16] J. Igusa, *Theta functions*, Springer-Verlag, **1972**.

[17] T. Kaiser, *Definability results for the Poisson equation*, Adv. Geom. **6** (2006), 627–644.

[18] T. Kaiser, *The Riemann mapping theorem for semianalytic domains and o-minimality*, Proc. London Math. Soc. (3) **98** (2009) 427–444.

[19] T. Kaiser, J-P. Rolin, P. Speissegger, *Transition maps at non-resonant hyperbolic singularities are o-minimal*. J. Reine Angew. Math. **636** (2009) 1–45.

[20] L. Lipshitz and Z. Robinson, *Overconvergent real closed quantifier elimination*, Bull. of London Math. Soc. 2006 **38** (6), 897–906.

[21] R. Moosa, *The model theory of compact complex spaces*, Logic Colloquium '01, Lecture Notes in Logic , **20** (M. Baaz, S. Friedman, and J. Krajicek, eds.), Assoc. for Symb. Logic, 2005, 317–349.

[22] R. Moosa, *A nonstandard Riemann existence theorem*, Trans. of AMS, **356** (2004), no. 5, 1781–1797.

[23] Y. Peterzil and S. Starchenko, *"Complex-like" analysis in o-minimal structures*, in *Proceedings of the RAAG Summer school Lisbon 2003, o-minimal structures* (M. Edmundo, D. Richardson and A.J. Wilkiem eds.) , Network RAAG 2005, 77–103.

[24] Y. Peterzil and S. Starchenko, *Expansions of algebraically closed fields in o-minimal structures*, Selecta Math. (N.S.) **7** (2001), no. 3, 409–445.

[25] Y. Peterzil and S. Starchenko, *Expansions of algebraically closed fields in o-minimal structures II. Functions of several variables*, J. Math. Log. **3** (2003), no. 1, 1–35.

[26] Y. Peterzil and S. Starchenko, *Uniform defianability of the Weierstrass $\wp$-function and generalized tori of dimension one*, Selecta Math. (N.S.) **10** (2004), 525–550.

[27] Y. Peterzil and S. Starchenko, *Mild manifolds and a non-standard Riemann existence theorem*, Selecta Math., (N.S) **14** (2009), 275–298.

[28] Y. Peterzil and S. Starchenko, *Complex analytic geometry in a nonstandard setting*, in *Model Theory with Applications to Algebra and Analysis I*, (Z. Chadzidakis, D. Macpherson, A. Pillay and Alex Wilkie, eds.), London Math. Soc. Lect. Notes Ser. **349**, Cambridge University Press, 2008, 117–166.

[29] Y. Peterzil and S. Starchenko, *Complex analytic geometry and analytic-geometric categories*, J. Reine Angew. Math. **626** (2009), 39–74.

[30] J. Pila, *Rational points of definable sets and results of Andre–Oort-Manin–Mumford type*. Int. Math. Res. Notices, (no. 13), (2009), 2476–2507.

[31] J. Pila, *Rational points of definable sets and finiteness results for special subvarieties*, preprint (2009).

[32] J. Pila and U. Zannier, *Rational points in periodic analytic sets and the Manin-Mumford conjecture*, Rend. Lincei Mat. Appl. 19 (2008), 149–162.

[33] J. Pila and A. Wilkie *The rational points of a definable set*, Duke Math. J. **133**, No. 3 (2006), 591–616.

[34] A. Pillay, *Some model theory of compact complex spaces*, Workshop on Hilbert's 10-th problem. Relations with Arithmetic and Algebraic Geometry (Ghent), Contemporary Math. 2000, 323–338.

[35] J. H. Silverman, *The arithmetic of elliptic curves*, Springer-verlag, New York, 1994, Corrected reprint of the 1986 original.

[36] L. Schneps and P. Lochak, *Geometric Galois Actions: I. Around Grothendieck's Esquisse d'un Programme*, Cambridge Univ. Press, 1997

[37] G. T. Whyburn, *Topological Analysis*, Princeton University Press, Princeton, 1964.

[38] A. J. Wilkie, *Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function*, J. Amer. Math. Soc. **9** (1996), 1051–1094.

[39] B. Zil'ber, *Model theory and algebraic geometry*, in: Proc. 10th Easter Conference on Model Thoery (wendisch Rietz, 1993), Seminarberichte 93, Humboldt Univ. Berlin, 93–117.

Section 2

Algebra

Paul Balmer
Tensor Triangular Geometry 85

David J. Benson
Modules for Elementary Abelian p-groups 113

Sergey Fomin
Total Positivity and Cluster Algebras 125

Nikita A. Karpenko
Canonical Dimension 146

Zinovy Reichstein
Essential Dimension 162

V. Suresh
Quadratic Forms, Galois Cohomology and Function Fields of p-adic Curves 189

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Tensor Triangular Geometry

Paul Balmer*

Abstract

We survey tensor triangular geometry: Its examples, early theory and first applications. We also discuss perspectives and suggest some problems.

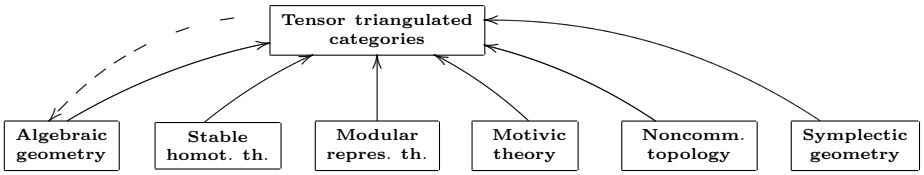
Mathematics Subject Classification (2010). Primary 18E30; Secondary 14F05, 19G12, 19K35, 20C20, 53D37, 55P42.

Keywords. Tensor triangulated categories, spectra.

Introduction

Tensor triangular geometry is the study of tensor triangulated categories by algebro-geometric methods. We invite the reader to discover this relatively new subject.

A great charm of this theory is the profusion of examples to be found throughout pure mathematics, be it in algebraic geometry, stable homotopy theory, modular representation theory, motivic theory, noncommutative topology, or symplectic geometry, to mention some of the most popular. We review them in Section 1. Here is an early photograph of tensor triangular geometry, in the crib:



Before climbing into vertiginous abstraction, it is legitimate to enquire about the presence of oxygen in the higher spheres. For instance, some readers might

*Research supported by NSF grant 0654397.
UCLA Mathematics Department, Los Angeles, CA 90095-1555.
E-mail: balmer@math.ucla.edu.

wonder whether tensor triangulated categories do not lose too much information about the more concrete mathematical objects to which they are associated. Our first answer is Theorem 54 below, which asserts that a scheme can be reconstructed from the associated tensor triangulated category, whereas a well-known result of Mukai excludes such reconstruction from the triangular structure alone. Informally speaking, *algebraic geometry embeds into tensor triangular geometry.*

The main tool for this result is the construction of a locally ringed space $\mathrm{Spec}(\mathcal{K}) = \big(\mathrm{Spc}(\mathcal{K})\,,\,\mathcal{O}_{\mathcal{K}}\big)$ for any tensor triangulated category $\mathcal{K}$, which gives back the scheme in the above geometric example. Interestingly, this construction also gives the projective support variety, $\mathcal{V}_G(k)$, in modular representation theory. This unification is one of the first achievements of tensor triangular geometry.

The most interesting part of our $\mathrm{Spec}(\mathcal{K})$ is the underlying space $\mathrm{Spc}(\mathcal{K})$, called the *spectrum* of $\mathcal{K}$. We shall see that determining $\mathrm{Spc}(\mathcal{K})$ is equivalent to the classification of thick triangulated tensor-ideals of $\mathcal{K}$. Indeed, in almost all examples, the classification of all objects of $\mathcal{K}$ is a wild problem. Nevertheless, using subsets of $\mathrm{Spc}(\mathcal{K})$, one can *always* classify objects of $\mathcal{K}$ modulo the basic operations available in $\mathcal{K}$: cones, direct summands and tensor products (Theorem 14). This marks the beginning of tensor triangular geometry, *per se*. See Section 2.

A general goal of this theory is to transpose ideas and techniques between the various areas of the above picture, via the abstract platform of tensor triangulated categories. For instance, from algebraic geometry, we shall abstract the technique of gluing and the concept of being *local*. From modular representation theory, we shall abstract Carlson's Theorem [18] and Rickard's idempotents. And of course many techniques used in triangulated categories have been borrowed from homotopy theory, not the least being the above idea of classifying thick tensor-ideals.

Finally, we also want applications, especially *strict* applications, i. e. results without tensor triangulated categories in the statement but only in the proof. Such applications already exist in algebraic geometry (for K-theory and Witt groups) and in modular representation theory (for endotrivial modules). And applications start to emerge in other areas as well. We discuss this in Section 3.

Let us illustrate our philosophy with a concrete abstraction. Take the notion of $\otimes$-*invertible* object $u \in \mathcal{K}$ (i. e. $u \otimes v \simeq \mathbb{1}$ for some $v \in \mathcal{K}$). This perfectly $\otimes$-triangular concept covers line bundles in algebraic geometry and endotrivial modules in modular representation theory. Now, in algebraic geometry, a line bundle is locally isomorphic to $\mathbb{1}$. Hence, the $\otimes$-triangular geometer asks:

(a) Can one make sense of "locally" in any $\otimes$-triangulated category?

(b) Are all $\otimes$-invertible objects "locally" isomorphic to $\mathbb{1}$, say, up to suspension?

(c) Can one use these ideas to relate line bundles and endotrivial modules?

We shall see that the respective answers are: yes, no (!) and, nonetheless, yes.

Acknowledgements. I'm indebted to many friends and colleagues, that I would like to thank, collectively but very sincerely, for their help and support.

1. Tensor Triangulated Categories in Nature

1.1. Basic definitions. Let us remind the reader of the notion of triangulated category, introduced by Grothendieck-Verdier [50] forty years ago. See Neeman [41] for a modern reference.

Definition 1. A *triangulated category* is an additive category $\mathcal{K}$ (we can add objects $a \oplus b$ and morphisms $f+g$) with a *suspension* $\Sigma : \mathcal{K} \xrightarrow{\sim} \mathcal{K}$ (treated here as an isomorphism of categories) and a class of so-called *distinguished triangles*

$$\Delta = \left(a \xrightarrow{f} b \xrightarrow{g} c \xrightarrow{h} \Sigma a \right)$$

which are like exact sequences in spirit and are subject to a list of simple axioms:

(TC 1) *Bookkeeping axiom*: Isomorphic triangles are simultaneously distinguished; Δ as above is distinguished if and only if its *rotated* $b \xrightarrow{g} c \xrightarrow{h} \Sigma a \xrightarrow{-\Sigma f} \Sigma b$ is distinguished; $a \xrightarrow{1} a \to 0 \to \Sigma a$ is distinguished for every object a.

(TC 2) *Existence axiom*: Every morphism $f : a \to b$ fits in some distinguished triangle Δ.

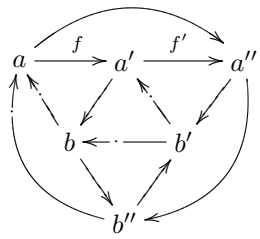
(TC 3) *Morphism axiom*: For every pair of distinguished triangles Δ and Δ'

$$\begin{array}{rcl} \Delta & = & \left(a \xrightarrow{f} b \xrightarrow{g} c \xrightarrow{h} \Sigma(a) \right) \\ & & \quad k\downarrow \qquad \ell\downarrow \qquad \exists\, m\downarrow \qquad \downarrow\Sigma k \\ \Delta' & = & \left(a' \xrightarrow[f']{} b' \xrightarrow[g']{} c' \xrightarrow[h']{} \Sigma(a') \right), \end{array}$$

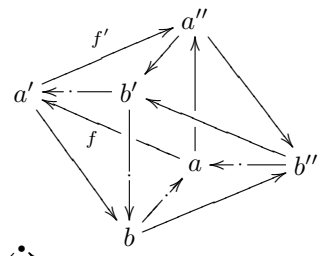
every commutative square (on the left) fits in a morphism of triangles.

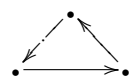
This was also proposed by Puppe in topology but Verdier's notorious addition is:

(TC 4) *Octahedron axiom*: Any two composable morphisms $a \xrightarrow{f} a' \xrightarrow{f'} a''$ fit in a commutative diagram (marked arrows $c \dashrightarrow c'$ mean $c \longrightarrow \Sigma(c')$)



or equivalently



in which the four triangles of the form  are distinguished.

A functor between triangulated categories is *exact* if it commutes with suspension (up to isomorphism) and preserves distinguished triangles in the obvious way.

Remark 2. Assuming (TC 1)-(TC 3), the third object c in a distinguished triangle Δ over a given $f : a \to b$ is unique up to (non-unique) isomorphism and is called the *cone* of f, denoted $\mathrm{cone}(f)$. The octahedron axiom simply says that there is a nice distinguished triangle relating $\mathrm{cone}(f)$, $\mathrm{cone}(f')$ and $\mathrm{cone}(f' \circ f)$.

The power of this axiomatic comes from its remarkable flexibility, compared for instance to the concepts of abelian or exact categories, which are somewhat too "algebraic". As we shall recall below, triangulated categories appear in a priori non-additive frameworks. In fact, the homotopy category of any stable Quillen model category is triangulated, see Hovey [27, Chap. 7].

Definition 3. A *tensor triangulated category* $(\mathcal{K}, \otimes, \mathbb{1})$ is a triangulated category $\mathcal{K}$ equipped with a monoidal structure (see Mac Lane [33, Chap. VII])

$$\mathcal{K} \times \mathcal{K} \xrightarrow{\otimes} \mathcal{K}$$

with unit object $\mathbb{1} \in \mathcal{K}$. We assume $- \otimes -$ exact in each variable, i. e. both functors $a \otimes - : \mathcal{K} \to \mathcal{K}$ and $- \otimes a : \mathcal{K} \to \mathcal{K}$ are exact, for every $a \in \mathcal{K}$. This involves natural isomorphisms $(\Sigma a) \otimes b \simeq \Sigma(a \otimes b)$ and $a \otimes (\Sigma b) \simeq \Sigma(a \otimes b)$ that we assume compatible, in that the two ways from $(\Sigma a) \otimes (\Sigma b)$ to $\Sigma^2(a \otimes b)$ only differ by a sign. Although some of the theory holds without further assumption, we are going to assume moreover that $\otimes$ is *symmetric monoidal*: $a \otimes b \cong b \otimes a$, see [33, § VII.7].

An exact functor F between tensor triangulated categories is $\otimes$-*exact* if it preserves the tensor structure, including the $\mathbb{1}$, up to isomorphisms which are compatible with the isomorphism $F\Sigma \simeq \Sigma F$, in the hopefully obvious way.

Remark 4. This is the most elementary axiomatic for "tensor triangulated"; see details in Hovey-Palmieri-Strickland [28, App. A]. May [34] proposed further compatibility axioms between tensor and octahedra, later extended by Keller-Neeman [30]. However, the elementary Definition 3 suffices for our purpose.

Such structures abound throughout pure mathematics, as we now review. See also [28, 1.2.3] for examples. We cannot provide background, motivation and explanations on all the following subjects and we assume some familiarity with at least some of the examples below, depending on the reader's own interests.

1.2. Examples from algebraic geometry. Let X be a scheme, here always assumed quasi-compact and quasi-separated (i. e. X admits a basis of quasi-compact open subsets); e.g. X affine, or X noetherian, like a variety over a field. Then $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(X)$, the derived category of perfect complexes over X, is a tensor triangulated category. See SGA 6 [14] or Thomason [49]. It sits $\mathcal{K} \subset \mathcal{T}$ inside the tensor triangulated category $\mathcal{T} = \mathrm{D}_{\mathrm{Qcoh}(X)}(X)$ of complexes of $\mathcal{O}_X$-modules with quasi-coherent homology. Such a complex is *perfect* if it is locally quasi-isomorphic to a bounded complex of finitely generated projective modules. When X is a quasi-projective variety over a field, $\mathrm{D}^{\mathrm{perf}}(X)$ is simply $\mathrm{D}^{\mathrm{b}}(\mathrm{VB}_X)$ the bounded derived category of vector bundles. The conceptual way of thinking of perfect complexes is as the compact objects in $\mathcal{T}$ (Def. 44). See Neeman [40] or Bondal-van den Bergh [15, Thm. 3.1.1]. The tensor $\otimes = \otimes^{\mathrm{L}}_{\mathcal{O}_X}$ is the left derived tensor product and the unit $\mathbb{1}$ is $\mathcal{O}_X$ (as a complex concentrated in degree 0).

When $X = \mathrm{Spec}(A)$ is affine, these categories are $\mathcal{T} = \mathrm{D}(A\text{–Mod})$, the derived category of A-modules, and $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(A) \cong \mathrm{K}^{\mathrm{b}}(A\text{–proj})$, the homotopy category of bounded complexes of finitely generated projective A-modules.

1.3. Examples from stable homotopy theory. Let $\mathcal{K} = \mathrm{SH}^{\mathrm{fin}}$ be the Spanier-Whitehead stable homotopy category of finite pointed CW-complexes. It sits $\mathcal{K} \subset \mathcal{T}$ as a tensor triangulated subcategory inside $\mathcal{T} = \mathrm{SH}$, the stable homotopy category of topological spectra. The tensor $\otimes = \wedge$ is the smash product and the unit $\mathbb{1} = S^0$ is the sphere spectrum. See Vogt [53]. One can also replace these by equivariant versions, use modules over a ring spectrum, or treat everything over a fixed base space.

1.4. Examples from modular representation theory. Let k be a field of positive characteristic and let G be a finite group, or a finite group scheme over k. (The adjective *modular* refers to kG not being semi-simple, i. e. to the existence of non-projective kG-modules.) Then $\mathcal{K} = \mathrm{stab}(kG)$, the stable module category of finitely generated kG-modules, modulo the projectives, is a tensor triangulated category. It sits $\mathcal{K} \subset \mathcal{T}$ inside the bigger tensor triangulated category $\mathcal{T} = \mathrm{Stab}(kG)$, the stable category of arbitrary kG-modules. Objects of $\mathrm{Stab}(kG)$ are k-representations of G and morphisms are equivalence classes of kG-linear maps under the relation $f \sim 0$ when f factors via a projective

(which is the same as an injective). The tensor is $\otimes_k$ with diagonal G-action and the unit is the trivial representation $\mathbb{1} = k$. See Happel [23], Carlson [19] or Benson [11]. One can alternatively consider $\mathrm{D}^{\mathrm{b}}(kG\text{–mod})$, inside $\mathrm{D}(kG\text{–Mod})$, with tensor product as above. Rickard [45] proved that the obvious functor $kG\text{–mod} \to \mathrm{D}^{\mathrm{b}}(kG\text{–mod})$ induces an equivalence of $\otimes$-triangulated categories

$$\mathrm{stab}(kG) \cong \mathrm{D}^{\mathrm{b}}(kG\text{–mod}) / \mathrm{K}^{\mathrm{b}}(kG\text{–proj})\,. \tag{5}$$

1.5. Examples from motivic theory. Let S be the spectrum of a perfect field (or some general base scheme). Then $\mathcal{K} = \mathrm{DM}_{\mathrm{gm}}(S)$, Voevodsky's derived category of geometric motives over S, is a tensor triangulated category. It sits $\mathcal{K} \subset \mathcal{T} = \mathrm{DM}(S)$ inside the derived category of motives over S. The tensor product extends the fiber product $X \times_S Y$. See [52]. The unit $\mathbb{1}$ is simply the motive of the base S (in degree zero).

1.6. Examples from $\mathbb{A}^1$-homotopy theory. Denote by $\mathcal{K} = \mathrm{SH}^{\mathbb{A}^1}_{\mathrm{gm}}(S)$ the triangulated subcategory generated by smooth S-schemes in the stable $\mathbb{A}^1$-homotopy category $\mathcal{T} = \mathrm{SH}^{\mathbb{A}^1}(S)$ of Morel-Voevodsky; see [51] or [36]. Again, the tensor $\otimes$ is essentially characterized as extending the fiber product $\times_S$ of S-schemes; and again $\mathbb{1}$ is the base S. In some sense, § 1.6 is to § 1.5 what § 1.3 is to § 1.2.

1.7. Examples from noncommutative topology. It is customary to think of C^*-algebras as noncommutative topological spaces. Let G be a second countable locally compact Hausdorff group – even G trivial is interesting. Then KK^G, the G-equivariant Kasparov category of separable G-C^*-algebras, is a tensor triangulated category, with $\otimes$ given by the minimal tensor product with diagonal G-action. See Meyer [35, § 4] for instance.

As the full category KK^G might be a little too overwhelming at first, we can follow Dell'Ambrogio [21] and consider the triangulated subcategory $\mathcal{K} = \mathcal{K}^G$ generated by the unit $\mathbb{1} = \mathbb{C}$. It actually sits inside the *Bootstrap category* $\mathcal{T} = \mathcal{T}^G$, which is the localizing subcategory of KK^G generated by the unit.

1.8. Further examples. There are examples in other areas of mathematics. For instance, triangulated categories famously appear in symplectic geometry, where Kontsevich's homological mirror symmetry conjecture [31] predicts an equivalence between the homotopy category of the Fukaya category of Calabi-Yau manifolds and the derived category of their mirror variety. Here, the tensor is a very interesting problem, which has seen recent progress in the work of Subotic [47].

As yet another example, Bühler recently proposed a triangulated category approach to bounded cohomology in [17]. Actually, examples of triangulated categories flourish in many directions, be it in connection to cluster algebras, knot theory, or theoretical physics, to mention a few less traditional examples.

In this luxuriant production of triangulated categories, we focus on *tensor* triangulated ones. And even if we "only" have the examples presented so far, the theory already calls for a unified treatment. Well, precisely, here comes one.

2. Abstract tensor triangular geometry

2.1. The spectrum. The basic idea of tensor triangular geometry, formulated in [1], is the construction of a topological space for every $\otimes$-triangulated category $\mathcal{K}$, called the *spectrum* of $\mathcal{K}$, in which every object b of $\mathcal{K}$ would have a *support*. This support should be understood as the non-zero locus of b. Since this idea admits no obvious formalization a priori, we follow the Grothendieckian philosophy of looking for the *best* such space, in a universal sense. To do this, we have to decide which properties this support should satisfy.

Theorem 6 ([1, Thm. 3.2]). *Let $\mathcal{K}$ be an essentially small $\otimes$-triangulated category. There exists a topological space* $\mathrm{Spc}(\mathcal{K})$ *and closed subsets* $\mathrm{supp}(a) \subset \mathrm{Spc}(\mathcal{K})$ *for all objects $a \in \mathcal{K}$, which form a* support datum *on $\mathcal{K}$, i. e. such that*

(SD 1) $\mathrm{supp}(0) = \varnothing$ *and* $\mathrm{supp}(\mathbb{1}) = \mathrm{Spc}(\mathcal{K})$,

(SD 2) $\mathrm{supp}(a \oplus b) = \mathrm{supp}(a) \cup \mathrm{supp}(b)$ *for every $a, b \in \mathcal{K}$,*

(SD 3) $\mathrm{supp}(\Sigma a) = \mathrm{supp}(a)$ *for every $a \in \mathcal{K}$,*

(SD 4) $\mathrm{supp}(c) \subset \mathrm{supp}(a) \cup \mathrm{supp}(b)$ *for every distinguished $a \to b \to c \to \Sigma a$,*

(SD 5) $\mathrm{supp}(a \otimes b) = \mathrm{supp}(a) \cap \mathrm{supp}(b)$ *for every $a, b \in \mathcal{K}$*

and such that $(\mathrm{Spc}(\mathcal{K}), \mathrm{supp})$ *is the final support datum on $\mathcal{K}$ in the sense that for every support datum (X, σ) on $\mathcal{K}$ (i. e. X a space with closed subsets $\sigma(a) \subset X$ for all $a \in \mathcal{K}$ satisfying (SD 1-5) above), there exists a unique continuous map $\varphi : X \to \mathrm{Spc}(\mathcal{K})$ such that $\sigma(a) = \varphi^{-1}(\mathrm{supp}(a))$ for every object $a \in \mathcal{K}$.*

Before explicitly constructing $\mathrm{Spc}(\mathcal{K})$, let us recall some standard terminology:

Definition 7. A non-empty full subcategory $\mathcal{J} \subset \mathcal{K}$ is a *triangulated subcategory* if for every distinguished triangle $a \to b \to c \to \Sigma a$ in $\mathcal{K}$, when two out of a, b, c belong to $\mathcal{J}$, so does the third; here, we call $\mathcal{J}$ *thick* if it is stable by direct summands: $a \oplus b \in \mathcal{J} \Rightarrow a, b \in \mathcal{J}$ (usual definition of thick) *and* triangulated; we say that $\mathcal{J}$ is $\otimes$-*ideal* if $\mathcal{K} \otimes \mathcal{J} \subset \mathcal{J}$; it is *radical* if $\sqrt[\otimes]{\mathcal{J}} = \mathcal{J}$, that is, $a^{\otimes n} \in \mathcal{J} \Rightarrow a \in \mathcal{J}$.

Construction 8. We baptize the universal support datum $(\mathrm{Spc}(\mathcal{K}), \mathrm{supp})$ of Theorem 6 the *spectrum* of $\mathcal{K}$. The content of the proof is the explicit construction of $\mathrm{Spc}(\mathcal{K})$. A thick $\otimes$-ideal $\mathcal{P} \subsetneq \mathcal{K}$ is called *prime* if it is proper ($\mathbb{1} \notin \mathcal{P}$) and if $a \otimes b \in \mathcal{P}$ implies $a \in \mathcal{P}$ or $b \in \mathcal{P}$. The spectrum of $\mathcal{K}$ is the set of primes:

$$\mathrm{Spc}(\mathcal{K}) := \left\{ \mathcal{P} \subsetneq \mathcal{K} \,\middle|\, \mathcal{P} \text{ is prime} \right\}.$$

(This is where we use $\mathcal{K}$ essentially small.) The *support* of an object $a \in \mathcal{K}$ is

$$\operatorname{supp}(a) := \left\{ \mathcal{P} \in \operatorname{Spc}(\mathcal{K}) \,\middle|\, a \notin \mathcal{P} \right\}.$$

The complements $U(a) := \left\{ \mathcal{P} \in \operatorname{Spc}(\mathcal{K}) \,\middle|\, a \in \mathcal{P} \right\}$, for all $a \in \mathcal{K}$, define an open basis of the topology of $\operatorname{Spc}(\mathcal{K})$. Examples of $\operatorname{Spc}(\mathcal{K})$ are given in § 3.1 below.

Remark 9. Of course, the above notion of *prime* reminds us of commutative algebra. Yet, this analogy is not a good reason for considering primes $\mathcal{P} \subset \mathcal{K}$. On the contrary, $\otimes$-triangular geometers should refrain from believing that everything works in all areas covered by $\otimes$-triangular geometry as simply as in their favorite toy area. The justification for the definition of $\operatorname{Spc}(\mathcal{K})$ is given by the universal property of Theorem 6 and by the Classification Theorem 14 below.

Remark 10. An important question is: Why do we ask $\operatorname{supp}(a)$ to be *closed*? After all, several notions of support involve non-closed subsets, if we deal with "big" objects. For instance, in $\mathrm{D}(\mathbb{Z}\text{–Mod})$, the object $\mathbb{Q}$ should certainly be supported only at (0), which is not closed in $\operatorname{Spec}(\mathbb{Z})$. This is a first indication that our theory is actually well suited for so-called *compact* objects (Def. 44). In fact, the assumption that $\mathcal{K}$ is essentially small points in the same direction: For instance, $\mathrm{D}(\mathbb{Z}\text{–Mod})$ is not essentially small but $\mathrm{D}^{\mathrm{perf}}(\mathbb{Z})$ is. We shall return to this discussion in a few places below, culminating in § 2.6.

Let us now collect some basic facts about the space $\operatorname{Spc}(\mathcal{K})$, all proven in [1].

Proposition 11. *Let $\mathcal{K}$ be an essentially small $\otimes$-triangulated category.*

(a) If $\mathcal{K}$ is non-zero then $\operatorname{Spc}(\mathcal{K})$ is non-empty.

(b) The space $\operatorname{Spc}(\mathcal{K})$ is spectral *in the sense of Hochster [24], that is, it is quasi-compact and quasi-separated (has a basis of quasi-compact open subsets) and every non-empty closed irreducible subset has a unique generic point (hence $\operatorname{Spc}(\mathcal{K})$ is T_0).*

(c) For every $\otimes$-exact functor $F : \mathcal{K} \to \mathcal{L}$, the assignment $\mathcal{Q} \mapsto F^{-1}(\mathcal{Q})$ defines a map $\varphi = \operatorname{Spc}(F) : \operatorname{Spc}(\mathcal{L}) \to \operatorname{Spc}(\mathcal{K})$ which is continuous and spectral (the preimage of a quasi-compact open subset is quasi-compact). So, $\operatorname{Spc}(-)$ is a contravariant functor. For every $a \in \mathcal{K}$, we have $\operatorname{supp}(F(a)) = \varphi^{-1}(\operatorname{supp}(a))$.

Remark 12. Hochster [24] observed that a spectral space X has a dual topology with dual-open subsets $Y \subset X$ being the arbitrary unions

$$Y = \cup_{i \in I} Y_i \quad \text{with each complement } X \smallsetminus Y_i \text{ open and quasi-compact.} \tag{13}$$

We call such a dual-open Y a *Thomason subset* of X, in honor of Thomason's insightful result [48, Thm. 4.1], which transposes remarkably well beyond algebraic geometry. When the space X is noetherian (every open is quasi-compact), a subset Y is Thomason if and only if it is specialization closed ($y \in Y \Rightarrow \overline{\{y\}} \subset Y$).

The next two results show that the computation of $\mathrm{Spc}(\mathcal{K})$ is equivalent to the classification of thick $\otimes$-ideals (see Definition 7 for terminology about ideals).

Theorem 14 (Classification of thick tensor-ideals [1, Thm. 4.10]). *Let $\mathcal{K}$ be an essentially small $\otimes$-triangulated category. Then the assignment*

$$Y \longmapsto \mathcal{K}_Y := \left\{ a \in \mathcal{K} \,\middle|\, \mathrm{supp}(a) \subset Y \right\}, \tag{15}$$

induces a bijection between Thomason subsets Y of the spectrum, see (13)*, and radical thick $\otimes$-ideals $\mathcal{J}$ of $\mathcal{K}$. Its inverse is $\mathcal{J} \mapsto \mathrm{supp}(\mathcal{J}) := \bigcup_{a \in \mathcal{J}} \mathrm{supp}(a)$.*

Being radical is a mild condition, as we shall see in Remark 23. Theorem 14 admits the following converse:

Theorem 16. *If the radical thick $\otimes$-ideals of $\mathcal{K}$ are classified as in* (15)*, by the Thomason subsets of a support datum (X, σ) with X spectral in the sense of Hochster, then the map $\varphi : X \to \mathrm{Spc}(\mathcal{K})$ of Theorem 6 is a homeomorphism.*

Theorem 16 was originally proven in [1, Thm. 5.2] under the assumption that X be a noetherian space. The ideal proof is due to Buan-Krause-Solberg [16, Cor. 5.2], who also extended our spectrum to lattices of ideals.

Remark 17. In categories like $\mathcal{K} = \mathrm{SH}^{\mathrm{fin}}$ or $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(A)$, which are generated by the unit $\mathbb{1}$, any thick subcategory is automatically $\otimes$-ideal. Similarly, $\mathcal{K} = \mathrm{stab}(kG)$ is generated by the unit $\mathbb{1} = k$ for G a p-group. However, the global study requires the tensor, see Remark 53.

We now indicate what happens to the spectrum under the few general constructions which are available for arbitrary $\otimes$-triangulated categories.

Theorem 18. *Let $\mathcal{K}$ be an essentially small $\otimes$-triangulated category.*

(a) Let $\mathcal{J} \subset \mathcal{K}$ be a thick $\otimes$-ideal. Then Verdier localization $\mathcal{K} \xrightarrow{q} \mathcal{K}/\mathcal{J}$ (Remark 19) induces a homeomorphism from $\mathrm{Spc}(\mathcal{K}/\mathcal{J})$ onto the subspace $\left\{ \mathcal{P} \,\middle|\, \mathcal{P} \supset \mathcal{J} \right\}$ of $\mathrm{Spc}(\mathcal{K})$. For instance, if $\mathcal{J} = \langle a \rangle = \mathcal{K}_{\mathrm{supp}(a)}$ is the thick $\otimes$-ideal generated by one object $a \in \mathcal{K}$, then $\mathrm{Spc}(\mathcal{K}/\langle a \rangle) \simeq U(a)$ is open in $\mathrm{Spc}(\mathcal{K})$.

(b) Idempotent completion $\iota : \mathcal{K} \to \mathcal{K}^\natural$ (see [10] or Remark 22 below) induces a homeomorphism $\mathrm{Spc}(\iota) : \mathrm{Spc}(\mathcal{K}^\natural) \xrightarrow{\sim} \mathrm{Spc}(\mathcal{K})$.

(c) Let $u \in \mathcal{K}$ be an object such that the cyclic permutation $(123) : u^{\otimes 3} \xrightarrow{\sim} u^{\otimes 3}$ is the identity and consider $F : \mathcal{K} \to \mathcal{K}[u^{\otimes -1}]$. Then $\mathrm{Spc}(F)$ yields a homeomorphism from $\mathrm{Spc}(\mathcal{K}[u^{\otimes -1}])$ onto the closed subspace $\mathrm{supp}(u)$ of $\mathrm{Spc}(\mathcal{K})$.

Proof. (a) and (b) are [1, Prop. 3.11 and Cor. 3.14]. For (c), recall that $\mathcal{K}[u^{\otimes -1}]$ has objects (a,m) with $a \in \mathcal{K}$ and $m \in \mathbb{Z}$ (the formal $a \otimes u^{\otimes m}$) and morphisms $\mathrm{Hom}\big((a,m),(b,n)\big) = \mathrm{colim}_{k\to+\infty} \mathrm{Hom}_{\mathcal{K}}(a \otimes u^{\otimes m+k}, b \otimes u^{\otimes n+k})$. This category inherits from $\mathcal{K}$ a unique $\otimes$-triangulation and the functor $F : a \mapsto (a,0)$ is $\otimes$-exact. The assumption on (123) ensures that the tensor structure on $\mathcal{K}[u^{\otimes -1}]$ is well-defined on morphisms. Then, the inverse of $\mathrm{Spc}(F)$ is defined by $\mathcal{P} \mapsto \mathcal{P}[u^{\otimes -1}]$ for every prime $\mathcal{P} \subset \mathcal{K}$ such that $\mathcal{P} \in \mathrm{supp}(u)$, that is, $u \notin \mathcal{P}$. Indeed, the latter condition implies that $\mathcal{P}[u^{\otimes -1}]$ is both proper and prime in $\mathcal{K}[u^{\otimes -1}]$. $\square$

Remark 19. Recall that the Verdier quotient $q : \mathcal{K} \longrightarrow \mathcal{K}/\mathcal{J}$ is the universal functor out of $\mathcal{K}$ such that $q(\mathcal{J}) = 0$. It is the localization of $\mathcal{K}$ with respect to the morphisms s in $\mathcal{K}$ such that $\mathrm{cone}(s) \in \mathcal{J}$. It can be constructed by keeping the same objects as $\mathcal{K}$ and defining morphisms as equivalence classes of fractions $\cdot \xleftarrow{s} \cdot \to \cdot$ with $\mathrm{cone}(s) \in \mathcal{J}$, under amplification.

We now introduce a very useful condition on $\mathcal{K}$:

Definition 20. A $\otimes$-triangulated category $\mathcal{K}$ is *rigid* if there exists an exact functor $D : \mathcal{K}^{\mathrm{op}} \to \mathcal{K}$ and a natural isomorphism $\mathrm{Hom}_{\mathcal{K}}(a \otimes b, c) \cong \mathrm{Hom}_{\mathcal{K}}(b, Da \otimes c)$ for every $a, b, c \in \mathcal{K}$. One calls Da the *dual* of a. In the terminology of [33] and [28], $(\mathcal{K}, \otimes)$ is *closed* symmetric monoidal and every object is *strongly dualizable*.

Hypothesis 21. From now on, we assume our $\otimes$-triangulated category $\mathcal{K}$ to be essentially small, rigid and idempotent complete.

Remark 22. Following up on Remark 10, the assumption that $\mathcal{K}$ is rigid is another indication that our input category $\mathcal{K}$ cannot be chosen too big. Much milder is the assumption that $\mathcal{K}$ is *idempotent-complete*, i. e. every idempotent $e = e^2 : a \to a$ in $\mathcal{K}$ yields a decomposition $a = \mathrm{im}(e) \oplus \ker(e)$, since $\mathcal{K}$ can always be idempotent completed $\mathcal{K} \overset{\iota}{\hookrightarrow} \mathcal{K}^{\natural}$ (see [10]) without changing the spectrum (Thm. 18 (b)).

Remark 23. Under Hypothesis 21, some natural properties become true in $\mathcal{K}$. For instance, $\mathrm{supp}(a) = \varnothing$ forces $a = 0$ (not only $\otimes$-nilpotent) by [3, Cor. 2.5]. Moreover, if $\mathrm{supp}(a) \cap \mathrm{supp}(b) = \varnothing$ then $\mathrm{Hom}_{\mathcal{K}}(a,b) = 0$, see [3, Cor. 2.8]. Finally, every thick $\otimes$-ideal $\mathcal{J} \subset \mathcal{K}$ is automatically radical $\sqrt[\otimes]{\mathcal{J}} = \mathcal{J}$ by [3, Prop. 2.4].

2.2. Localization.

Let us introduce the most important basic construction of $\otimes$-triangular geometry, which gives a meaning to "the category $\mathcal{K}$ *over some open* U of its spectrum".

Construction 24. For every quasi-compact open $U \subset \mathrm{Spc}(\mathcal{K})$, with closed complement $Z := \mathrm{Spc}(\mathcal{K}) \smallsetminus U$, we define the tensor triangulated category $\mathcal{K}(U)$ as

$$\mathcal{K}(U) := \big(\mathcal{K}/\mathcal{K}_Z\big)^{\natural}.$$

It is the idempotent completion of the Verdier quotient (Rem. 19) $\mathcal{K}/\mathcal{K}_Z$ of $\mathcal{K}$ by the thick $\otimes$-ideal $\mathcal{K}_Z = \{ a \in \mathcal{K} \mid \operatorname{supp}(a) \subset Z \}$ of those objects supported outside U. We have a natural functor $\operatorname{res}_U : \mathcal{K} \to \mathcal{K}(U)$. One can prove that $\operatorname{Spc}(\operatorname{res}_U)$ induces a conceptually pleasant homeomorphism, see [9, Prop. 1.11],

$$\operatorname{Spc}(\mathcal{K}(U)) \cong U\,.$$

Hence quasi-compactness of U is necessary since $\operatorname{Spc}(\mathcal{K})$ is always quasi-compact, see Prop. 11 (b). Informally, the category $\mathcal{K}(U)$ is the piece of $\mathcal{K}$ living above the open U. For every $a, b \in \mathcal{K}$, we abbreviate $\operatorname{Hom}_{\mathcal{K}(U)}(\operatorname{res}_U(a), \operatorname{res}_U(b))$ by $\operatorname{Hom}_U(a, b)$. In the same spirit, we say that something about $\mathcal{K}$ happens "over U", when it happens in the category $\mathcal{K}(U)$ after applying the restriction functor res_U.

Theorem 25 ([5, § 4]). *Let $\mathcal{K}$ be a $\otimes$-triangulated category as in Hypothesis 21.*

(a) The topological space $\operatorname{Spc}(\mathcal{K})$ *is local (i. e. every open cover contains the whole space) if and only if $a \otimes b = 0$ implies $a = 0$ or $b = 0$. Then $\{0\}$ is the unique closed point of* $\operatorname{Spc}(\mathcal{K})$ *and we call $\mathcal{K}$ a* local *$\otimes$-triangulated category.*

(b) For every $\mathcal{P} \in \operatorname{Spc}(\mathcal{K})$, the category $\mathcal{K}/\mathcal{P}$ is local in the above sense. Its idempotent completion $(\mathcal{K}/\mathcal{P})^\natural$ is the colimit of the $\mathcal{K}(U)$ over those quasi-compact open $U \subset \operatorname{Spc}(\mathcal{K})$ containing the point $\mathcal{P} \in \operatorname{Spc}(\mathcal{K})$.

Remark 26. Roughly speaking, $\mathcal{K}/\mathcal{P}$ (or rather $(\mathcal{K}/\mathcal{P})^\natural$) is the *stalk* of $\mathcal{K}$ at the point $\mathcal{P} \in \operatorname{Spc}(\mathcal{K})$. The support $\operatorname{supp}(a) = \{ \mathcal{P} \mid a \notin \mathcal{P} \} = \{ \mathcal{P} \mid a \neq 0 \text{ in } \mathcal{K}/\mathcal{P} \}$ of an object $a \in \mathcal{K}$ can now be understood as the points of $\operatorname{Spc}(\mathcal{K})$ where a does not vanish in the stalk. This expresses the non-zero locus of a, as initially wanted.

Remark 27. Amusingly, a *local* $\otimes$-triangulated category $\mathcal{K}$ (i. e. $a \otimes b = 0 \Rightarrow a$ or $b = 0$) could hastily be baptized "integral" if one was to follow algebraic gut feeling. Extending standard terminology to $\otimes$-triangular geometry requires some care. Indeed, "local" is correct because of the conceptual characterization of Theorem 25 (a). And comfortingly, for X a scheme, the $\otimes$-triangulated category $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(X)$ is local if and only if $X \cong \operatorname{Spec}(A)$ with A a local ring.

Remark 28. When $\mathcal{K}$ is local, $\operatorname{Spc}(\mathcal{K})$ has a unique closed point by Thm. 25 (a). Then, the smallest possible support for a non-zero object is exactly that closed point $*$. We define $\operatorname{FL}(\mathcal{K}) := \{ a \in \mathcal{K} \mid \operatorname{supp}(a) \subset * \}$ and call such objects the *finite length* objects, by analogy with commutative algebra. (This somewhat improper terminology might need improvement; see the comments in Remark 27.)

We now use $\mathcal{K}(U)$ to create a structure sheaf on $\operatorname{Spc}(\mathcal{K})$.

Construction 29. For every quasi-compact open $U \subset \mathrm{Spc}(\mathcal{K})$, we can consider the commutative ring $\mathrm{End}_{\mathcal{K}(U)}(\mathbb{1})$. Since the unit $\mathbb{1}$ of $\mathcal{K}(U)$ is simply the restriction of the unit $\mathbb{1}$ of $\mathcal{K}$, and since $(\mathcal{K}(U))(V) \cong \mathcal{K}(V)$ for every $V \subset U \cong \mathrm{Spc}(\mathcal{K}(U))$, we obtain a presheaf of commutative rings $\mathrm{p}\mathcal{O}_{\mathcal{K}}$, at least on the open basis consisting of quasi-compact open subsets. This presheaf $\mathrm{p}\mathcal{O}_{\mathcal{K}}(U) = \mathrm{End}_U(\mathbb{1})$ is already useful in itself but can also be sheafified into a sheaf $\mathcal{O}_{\mathcal{K}}$ of commutative rings on $\mathrm{Spc}(\mathcal{K})$. We denote by

$$\mathrm{Spec}(\mathcal{K}) := \big(\mathrm{Spc}(\mathcal{K})\,,\, \mathcal{O}_{\mathcal{K}}\big)$$

the corresponding ringed space. It is a locally ringed space by [5, Cor. 6.6].

Remark 30. The above construction has an obvious algebro-geometric bias and one should not expect too much from this sheaf of rings $\mathcal{O}_{\mathcal{K}}$ in general. Still, it will be important in Theorems 54 and 57 below. Our preferred presheaf on $\mathrm{Spc}(\mathcal{K})$ is not $\mathcal{O}_{\mathcal{K}}$ but the more fundamental "presheaf" of $\otimes$-triangulated categories: $U \mapsto \mathcal{K}(U)$ of Construction 24.

2.3. Support and decomposition.

Here comes the first $\otimes$-triangular result which really opens the door to geometry. It extends a famous result of Carlson [18] in representation theory.

Theorem 31 ([3, Thm. 2.11]). *Let $\mathcal{K}$ be a $\otimes$-triangulated category as in Hypothesis 21 and let $a \in \mathcal{K}$ be an object. Suppose that its support is disconnected, i. e.* $\mathrm{supp}(a) = Y_1 \sqcup Y_2$ *with each Y_i closed and $Y_1 \cap Y_2 = \varnothing$. Then the object decomposes accordingly, that is, $a \simeq a_1 \oplus a_2$ with* $\mathrm{supp}(a_1) = Y_1$ *and* $\mathrm{supp}(a_2) = Y_2$.

It is easy to build counter-examples to the above statement if we remove the assumption that $\mathcal{K}$ is idempotent complete, see [3, Ex. 2.13]. This explains why we insist on idempotent-completion, for instance in the construction of $\mathcal{K}(U)$ above. Theorem 31 has the following application.

Theorem 32 ([3, Thm. 3.24]). *Let $\mathcal{K}$ be a $\otimes$-triangulated category as in Hypothesis 21 and assume that* $\mathrm{Spc}(\mathcal{K})$ *is a noetherian topological space (every open is quasi-compact). Let* $\dim : \mathrm{Spc}(\mathcal{K}) \to \mathbb{Z} \cup \{\pm\infty\}$ *be a dimension function, i. e.* $\mathcal{Q} \subsetneqq \mathcal{P} \Rightarrow \dim(\mathcal{Q}) + 1 \leq \dim(\mathcal{P})$. *Consider the filtration of $\mathcal{K}$ by the $\otimes$-ideals* $\mathcal{K}_{(d)} := \big\{\, a \in \mathcal{K} \,\big|\, \dim(\mathcal{P}) \leq d \text{ for all } \mathcal{P} \in \mathrm{supp}(a) \,\big\}$. *Then for every finite $d \in \mathbb{Z}$, the corresponding subquotient $\mathcal{K}_{(d)}/\mathcal{K}_{(d-1)}$ decomposes into a coproduct of local parts. More precisely, after idempotent completion, we have an equivalence*

$$\big(\mathcal{K}_{(d)}/\mathcal{K}_{(d-1)}\big)^{\natural} \xrightarrow{\sim} \coprod_{\mathcal{P}\in\mathrm{Spc}(\mathcal{K}),\ \dim(\mathcal{P})=d} \big(\mathrm{FL}(\mathcal{K}/\mathcal{P})\big)^{\natural}$$

where the subcategories of finite-length objects $\mathrm{FL}(\mathcal{K}/\mathcal{P})$ *are as in Remark 28.*

Examples of dimension functions, $\dim(\mathcal{P})$, include the Krull dimension of the irreducible closed $\overline{\{\mathcal{P}\}}$, or the opposite of its Krull codimension, in $\mathrm{Spc}(\mathcal{K})$.

2.4. Gluing and Picard groups.

The true power of Theorem 31 appears in the following gluing method.

Theorem 33 (B.-Favi [9, Cor. 5.8 and 5.10]). *Let $\mathcal{K}$ be a $\otimes$-triangulated category as in Hypothesis 21 and let $\mathrm{Spc}(\mathcal{K}) = U_1 \cup U_2$ be a cover with both U_i quasi-compact open. Set $U_{12} := U_1 \cap U_2$ and consider the commutative square of $\otimes$-triangulated categories and restriction functors*

$$\begin{tikzcd} \mathcal{K} \arrow[r] \arrow[d] & \mathcal{K}(U_1) \arrow[d] \\ \mathcal{K}(U_2) \arrow[r] & \mathcal{K}(U_{12}) \,. \end{tikzcd}$$

(a) Gluing of morphisms: For every pair of objects $a, b \in \mathcal{K}$, we have a Mayer-Vietoris long exact sequence of abelian groups

$$\cdots \xrightarrow{\partial} \mathrm{Hom}_{\mathcal{K}}(a,b) \longrightarrow \begin{matrix} \mathrm{Hom}_{U_1}(a,b) \\ \oplus \\ \mathrm{Hom}_{U_2}(a,b) \end{matrix} \longrightarrow \mathrm{Hom}_{U_{12}}(a,b) \xrightarrow{\partial} \mathrm{Hom}_{\mathcal{K}}(a, \Sigma b) \longrightarrow \cdots$$

(b) Gluing of objects: Given two objects $a_i \in \mathcal{K}(U_i)$, $i = 1, 2$, and an isomorphism $\sigma : a_1 \xrightarrow{\sim} a_2$ over U_{12}, there exists a triple (a, f_1, f_2) where a is an object of $\mathcal{K}$ and $f_i : a \xrightarrow{\sim} a_i$ is an isomorphism over U_i such that $\sigma \circ f_1 = f_2$ over U_{12}. This gluing is unique up to possibly non-unique isomorphism of triples in $\mathcal{K}$.

Remark 34. The apparently anodyne non-uniqueness of the isomorphism in (b) has a cost. Namely, gluing of three objects over three open subsets is still possible but without uniqueness [9, Cor. 5.11]. And gluing of more than three pieces might simply not exist unless some connectivity conditions are imposed [9, Thm. 5.13].

Here is an application of the gluing technique to Picard groups.

Definition 35. The *Picard group*, $\mathrm{Pic}(\mathcal{K})$, is the group of isomorphism classes of $\otimes$-invertible objects of $\mathcal{K}$, that is, those $u \in \mathcal{K}$ for which there exists $v \in \mathcal{K}$ with $u \otimes v \simeq \mathbb{1}$. (As $\mathcal{K}$ is rigid, $v \simeq Du$.) This does not use the triangulation.

We can now construct $\otimes$-invertible objects by gluing copies of the $\otimes$-unit $\mathbb{1}$.

Definition 36. For every quasi-compact open $U \subset \mathrm{Spc}(\mathcal{K})$, denote by $\mathbb{G}_{\mathrm{m}}(U) := \mathrm{Aut}_U(\mathbb{1})$ the group of automorphisms of $\mathbb{1}$ in $\mathcal{K}(U)$.

Theorem 37 (B.-Favi [9, Thm. 6.7]). *Under Hypothesis 21, if* $\mathrm{Spc}(\mathcal{K}) = U_1 \cup U_2$ *with each* U_i *quasi-compact, then gluing induces a well-defined group homomorphism* $\delta : \mathbb{G}_{\mathrm{m}}(U_{12}) \to \mathrm{Pic}(\mathcal{K})$*, where* $U_{12} := U_1 \cap U_2$*. We have an exact sequence*

$$\begin{tikzcd}
\cdots \mathrm{Hom}_{U_{12}}(\Sigma \mathbb{1}, \mathbb{1}) \arrow[r, "1+\partial"] & \mathbb{G}_{\mathrm{m}}(\mathrm{Spc}(\mathcal{K})) \arrow[r] & \mathbb{G}_{\mathrm{m}}(U_1) \oplus \mathbb{G}_{\mathrm{m}}(U_2) \arrow[r] & \mathbb{G}_{\mathrm{m}}(U_{12}) \arrow[dll, "\delta"] \\
& \mathrm{Pic}(\mathcal{K}) \arrow[r] & \mathrm{Pic}(\mathcal{K}(U_1)) \oplus \mathrm{Pic}(\mathcal{K}(U_2)) \arrow[r] & \mathrm{Pic}(\mathcal{K}(U_{12}))\,,
\end{tikzcd}$$

which continues on the left as in Theorem 33 (where ∂ *also comes from).*

It remains an open problem how to extend this sequence on the right, say, with Brauer groups. The other natural thing one might want to do is to glue any $\mathbb{G}_{\mathrm{m}}$-cocycle on $\mathrm{Spc}(\mathcal{K})$ into an invertible object of $\mathcal{K}$. Then the difficulty of gluing more than three pieces (Remark 34) becomes an obstacle. It can be circumvented in positive characteristic p, at the price of inverting p on the Picard group:

Theorem 38 ([6, Thm. 3.9]). *Let* p *be a prime and* $\mathcal{K}$ *a* $\otimes$*-triangulated* $\mathbb{Z}/p$*-category satisfying Hypothesis 21. Let* $\check{H}^1(\mathrm{Spc}(\mathcal{K}), \mathbb{G}_{\mathrm{m}})$ *be the first Čech cohomology group with coefficients in the above presheaf of units* $\mathbb{G}_{\mathrm{m}}$*. Let* $\mathrm{Pic}_{loc.tr.}(\mathcal{K}) := \big\{ [u] \,\big|\, u \simeq \mathbb{1} \text{ in } \mathcal{K}/\mathcal{P} \text{ for all } \mathcal{P} \in \mathrm{Spc}(\mathcal{K}) \big\} \subset \mathrm{Pic}(\mathcal{K})$ *be the subgroup of locally (very) trivial invertibles. Then, gluing induces a well-defined isomorphism* β

$$\check{H}^1(\mathrm{Spc}(\mathcal{K}), \mathbb{G}_{\mathrm{m}}) \underset{\mathbb{Z}}{\otimes} \mathbb{Z}[1/p] \xrightarrow[\simeq]{\beta} \mathrm{Pic}_{loc.tr.}(\mathcal{K}) \underset{\mathbb{Z}}{\otimes} \mathbb{Z}[1/p] \;\subset\; \mathrm{Pic}(\mathcal{K}) \underset{\mathbb{Z}}{\otimes} \mathbb{Z}[1/p]\,.$$

We call $\mathbb{1}$ the *very* trivial $\otimes$-invertible because the right notion of a *trivial* $\otimes$-invertible is probably one of the form $\Sigma^n \mathbb{1}$ for some $n \in \mathbb{Z}$. See more in § 4.5.

Remark 39. In algebraic geometry, invertible objects are (shifted) line bundles. Hence they are locally trivial for the Zariski topology, which explains why the Picard group, $\mathrm{Pic}(X)$, is the first Zariski cohomology group of $\mathbb{G}_{\mathrm{m}}$. However, there are local $\otimes$-triangulated categories with non-trivial Picard group. See Remark 71 for an example in modular representation theory. The following result shows that the Picard group can be as large as one wants with given (even local) spectrum.

Proposition 40 (B. - Rahbar Virk). *Let* $\mathcal{K}$ *be a local* $\otimes$*-triangulated category (*$\mathrm{Spc}(\mathcal{K})$ *connected is enough). Let* G *be an abelian group. Define a tensor on the triangulated category* $\mathcal{L} := \coprod_G \mathcal{K}$ *by* $a_g \otimes b_h := (a \otimes b)_{g+h}$*, where* $a_g \in \mathcal{L}$ *is the object corresponding to* $a \in \mathcal{K}$ *in the copy indexed by* $g \in G$*. Then* $\mathrm{Spc}(\mathcal{L}) \cong \mathrm{Spc}(\mathcal{K})$ *whereas* $\mathrm{Pic}(\mathcal{L}) \cong \mathrm{Pic}(\mathcal{K}) \times G$*.*

Proof. Easy exercise using the $\otimes$-invertible objects $\mathbb{1}_g \in \mathcal{L}$ for all $g \in G$ and the fact that every object of $\mathcal{L}$ is a finite direct sum $\bigoplus_{g\in G} a(g)_0 \otimes \mathbb{1}_g$ for objects $a(g) \in \mathcal{K}$. $\square$

2.5. Comparing triangular spectra and algebraic spectra.

Remark 41. It should be clear by now that the main key to the geometry of a given $\otimes$-triangulated category $\mathcal{K}$, is the determination of its spectrum, $\mathrm{Spc}(\mathcal{K})$. We have seen in Theorem 16 that this problem amounts to the classification of thick $\otimes$-ideals of $\mathcal{K}$. This is very nice when the latter classification has been kindly performed by our predecessors but in most new areas such a classification is not yet under roof and actually constitutes a very interesting challenge. See § 4.1 below. To study $\mathrm{Spc}(\mathcal{K})$ without classification, we need some comparison with other spaces that might appear in examples. This is the purpose of [5], where we relate $\mathrm{Spc}(\mathcal{K})$ to the spectrum of the endomorphism ring $\mathrm{R}_{\mathcal{K}} = \mathrm{End}_{\mathcal{K}}(\mathbb{1})$ of the $\otimes$-unit $\mathbb{1}$, and to the homogeneous spectrum of the graded ring $\mathrm{R}^{\bullet}_{\mathcal{K}} = \mathrm{Hom}_{\mathcal{K}}(\mathbb{1}, \Sigma^{\bullet}\mathbb{1})$.

Theorem 42 ([5, Thm. 5.3]). *There exist two natural continuous maps*

$$\rho^{\bullet}_{\mathcal{K}} : \mathrm{Spc}(\mathcal{K}) \longrightarrow \mathrm{Spec}^{\mathrm{h}}(\mathrm{R}^{\bullet}_{\mathcal{K}}) \qquad \textit{and} \qquad \rho_{\mathcal{K}} : \mathrm{Spc}(\mathcal{K}) \longrightarrow \mathrm{Spec}(\mathrm{R}_{\mathcal{K}})$$

defined by $\rho^{\bullet}_{\mathcal{K}}(\mathcal{P}) = \bigoplus_{d\in\mathbb{Z}} \left\{ f \in \mathrm{R}^d_{\mathcal{K}} \;\middle|\; \mathrm{cone}(f) \notin \mathcal{P} \right\}$ *and* $\rho_{\mathcal{K}}(\mathcal{P}) = \rho^0_{\mathcal{K}}(\mathcal{P})$.

In fact, these maps are often surjective (yet, not always, see [5, Ex. 8.3]):

Theorem 43 ([5, § 7]). *With the notation of Theorem 42, we have:*

(a) Suppose that $\mathcal{K}$ is connective, i. e. that $\mathrm{Hom}(\Sigma^i\mathbb{1}, \mathbb{1}) = 0$ *for* $i < 0$ *(which reads* $\mathrm{R}^d_{\mathcal{K}} = 0$ *for* $d > 0$*). Then* $\rho_{\mathcal{K}} : \mathrm{Spc}(\mathcal{K}) \to \mathrm{Spec}(\mathrm{R}_{\mathcal{K}})$ *is a surjective map.*

(b) Suppose that $\mathrm{R}^{\bullet}_{\mathcal{K}}$ *is coherent (e.g. noetherian) in the graded sense. Then both* $\rho^{\bullet}_{\mathcal{K}} : \mathrm{Spc}(\mathcal{K}) \to \mathrm{Spec}^{\mathrm{h}}(\mathrm{R}^{\bullet}_{\mathcal{K}})$ *and* $\rho_{\mathcal{K}}$ *are surjective maps.*

Injectivity is more delicate, see Theorem 51. However, in "algebraic" examples, these maps are (local) homeomorphisms, see Remark 56 and Theorem 57.

2.6. Non-compact objects.

As indicated a couple of times above, the natural input $\mathcal{K}$ to our $\otimes$-triangular geometry machine consists of small enough categories. Let us now be more precise.

Definition 44. Let $\mathcal{T}$ be a triangulated category admitting arbitrary small coproducts $\coprod_{i\in I} t_i$. An object $c \in \mathcal{T}$ is called *compact* if for every set of objects $\{t_i\}_{i\in I}$ in $\mathcal{T}$, the natural map $\coprod_{i\in I} \mathrm{Hom}_{\mathcal{T}}(c, t_i) \to \mathrm{Hom}_{\mathcal{T}}(c, \coprod_{i\in I} t_i)$ is an isomorphism. The subcategory $\mathcal{T}^c$ of compact objects is triangulated but

not closed under coproducts. We say that $\mathcal{T}$ is a *compactly generated tensor triangulated category* if

(i) $\mathcal{T}^c$ generates $\mathcal{T}$, that is, $\mathcal{T} = \mathrm{Loc}(\mathcal{T}^c)$ is the smallest localizing (i. e. closed under small coproducts) triangulated subcategory of $\mathcal{T}$ which contains $\mathcal{T}^c$.

(ii) $\mathcal{T}^c$ is essentially small, $\mathcal{T}^c$ is rigid and $\mathbb{1}$ is compact.

In that case, an object is compact if and only if it is rigid (i. e. strongly dualizable) and the $\otimes$-triangulated category $\mathcal{K} := \mathcal{T}^c$ of rigid-compact objects satisfies our Hypothesis 21. We can then apply the above $\otimes$-triangular geometry to $\mathcal{K} = \mathcal{T}^c$.

Examples 45. Examples § 1.2-1.4 fit in this picture with the $\mathcal{T}$ provided each time. (Examples § 1.5–1.7 require some care.) In [28], $\otimes$-triangulated categories $\mathcal{T}$ as above are studied under the name *unital algebraic stable homotopy categories.*

Remark 46. Our spectrum $\mathrm{Spc}(\mathcal{K})$ is the right space for the compact part but $\mathrm{Spc}(\mathcal{T})$ is not an appropriate invariant of $\mathcal{T}$ for it might not even be a set. Moreover, we do not need supports of non-compact objects to be closed and we would like $\mathrm{supp}(\coprod_{i\in I} t_i) = \cup_{i\in I}\,\mathrm{supp}(t_i)$. The question of $\mathrm{supp}(s\otimes t)$ is not entirely clear. One expects $\mathrm{supp}(s\otimes t) \subset \mathrm{supp}(s)\cap\mathrm{supp}(t)$ with equality when s is compact. Putting all this together, one can actually define a "big spectrum" of $\mathcal{T}$ as the universal space with supports, satisfying (SD' 1)-(SD' 7) below. Since it is not clear yet how useful this big spectrum can be, we do not make a theory out of this. The following result, due independently to Pevtsova-Smith [43] and Dell'Ambrogio, indicates that such a big spectrum might often coincide with $\mathrm{Spc}(\mathcal{K})$ anyway.

Theorem 47 ([21, Thm. 3.1]). *Let $\mathcal{T}$ be a compactly generated $\otimes$-triangulated category as in Definition 44. Let X be a topological space with a choice of a subset $\sigma(t) \subset X$ for every object $t \in \mathcal{T}$ satisfying the following conditions:*

(SD' 1) $\sigma(0) = \varnothing$ *and* $\sigma(\mathbb{1}) = X$,

(SD' 2) $\sigma(s \oplus t) = \sigma(s) \cup \sigma(t)$ *for every* $s, t \in \mathcal{T}$,

(SD' 3) $\sigma(\Sigma t) = \sigma(t)$ *for every* $t \in \mathcal{T}$,

(SD' 4) $\sigma(u) \subset \sigma(s) \cup \sigma(t)$ *for every distinguished triangle* $s \to t \to u \to \Sigma s$,

(SD' 5) $\sigma(s \otimes t) \subset \sigma(s) \cap \sigma(t)$ *for every* $s, t \in \mathcal{T}$, *with equality if* s *or* t *is compact,*

(SD' 6) $\sigma(\coprod_{i\in I} t_i) = \cup_{i\in I}\sigma(t_i)$ *for every set* $\{t_i\}_{i\in I}$ *of objects of* $\mathcal{T}$,

(SD' 7) $\sigma(c)$ *is closed for every compact object* $c \in \mathcal{T}^c$.

In particular (X, σ) is a support datum on $\mathcal{K} = \mathcal{T}^c$. Suppose moreover:

(i) X is spectral in the sense of Hochster [24], see Proposition 11 (b).

(ii) An open $U \subset X$ is quasi-compact if and only if $U = X \smallsetminus \sigma(c)$ for $c \in \mathcal{T}^c$.

(iii) For $t \in \mathcal{T}$, if $\sigma(t) = \varnothing$ then $t = 0$.

Then the canonical map $X \to \operatorname{Spc}(\mathcal{T}^c)$ of Theorem 6 is a homeomorphism.

In examples where $\mathcal{T}$ is given with such supports, Theorem 47 might be used to compute $\operatorname{Spc}(\mathcal{K})$. Conversely, $\operatorname{Spc}(\mathcal{K})$, for $\mathcal{K} = \mathcal{T}^c$, yields information about the big category $\mathcal{T}$, via the following inflating technique, see [41, Chap. 4]:

Remark 48. For $U \subset \operatorname{Spc}(\mathcal{K})$ quasi-compact open with closed complement Z, set $\mathcal{T}_Z = \operatorname{Loc}(\mathcal{K}_Z)$ the localizing subcategory of $\mathcal{T}$ generated by $\mathcal{K}_Z \subset \mathcal{K}$. In [8], we define the category "$\mathcal{T}$ over U" as the localization $\mathcal{T}(U) := \mathcal{T}/\mathcal{T}_Z$. The $\otimes$-triangulated category $\mathcal{T}(U)$ remains compactly generated and Neeman's generalization [41, Thm. 4.4.9] of Thomason's result (Rem. 55) reads: $\big(\mathcal{T}(U)\big)^c = \mathcal{K}(U)$. This also justifies the idempotent completion in the definition of $\mathcal{K}(U)$.

Transposing Rickard's idempotents [46] to $\otimes$-triangular geometry gives:

Theorem 49 (B.-Favi [8]). *Let $\mathcal{T}$ be a compactly generated $\otimes$-triangulated category (Def. 44) and $\mathcal{K} = \mathcal{T}^c$ its compact objects. For every Thomason subset $Y \subset \operatorname{Spc}(\mathcal{K})$, there exists a distinguished triangle $e(Y) \to \mathbb{1} \to f(Y) \to \Sigma(e(Y))$ in $\mathcal{T}$ such that $e(Y) \otimes f(Y) = 0$ (hence $e(Y)^{\otimes 2} \simeq e(Y)$ and $f(Y)^{\otimes 2} \simeq f(Y)$ are $\otimes$-idempotents) and such that $f(Y) \otimes - : \mathcal{T} \longrightarrow \mathcal{T}$ realizes Bousfield localization with respect to $\mathcal{T}_Y := \operatorname{Loc}(\mathcal{K}_Y) = e(Y) \otimes \mathcal{T}$, the localizing subcategory of $\mathcal{T}$ generated by the compact objects $\mathcal{K}_Y = \big\{\, a \in \mathcal{K} \,\big|\, \operatorname{supp}(a) \subset Y \,\big\}$. Moreover, for every pair of Thomason subsets $Y_1, Y_2 \subset \operatorname{Spc}(\mathcal{K})$, we have isomorphisms $e(Y_1 \cap Y_2) \cong e(Y_1) \otimes e(Y_2)$ and $f(Y_1 \cup Y_2) \cong f(Y_1) \otimes f(Y_2)$ and two Mayer-Vietoris distinguished triangles in $\mathcal{T}$:*

$$e(Y_1 \cap Y_2) \longrightarrow e(Y_1) \oplus e(Y_2) \longrightarrow e(Y_1 \cup Y_2) \longrightarrow \Sigma e(Y_1 \cap Y_2)$$

$$f(Y_1 \cap Y_2) \longrightarrow f(Y_1) \oplus f(Y_2) \longrightarrow f(Y_1 \cup Y_2) \longrightarrow \Sigma f(Y_1 \cap Y_2)\,.$$

Using these $\otimes$-idempotents, we get the announced definition of a support inside $\operatorname{Spc}(\mathcal{K})$, for all objects of $\mathcal{T}$ (compare Benson-Iyengar-Krause [13]):

Theorem 50 (B.-Favi [8, § 7]). *Let $\mathcal{T}$ and $\mathcal{K} = \mathcal{T}^c$ be as above and suppose that $\operatorname{Spc}(\mathcal{K})$ is noetherian. Define $\kappa(\mathcal{P}) = e(\overline{\{\mathcal{P}\}}) \otimes f(\operatorname{supp}(\mathcal{P})) \in \mathcal{T}$, for all $\mathcal{P} \in \operatorname{Spc}(\mathcal{K})$ (here $\operatorname{supp}(\mathcal{P})$ is the Thomason subset corresponding to $\mathcal{P}$ in the Classification Theorem 14). Then, the support admits the following extension to all objects $t \in \mathcal{T}$:*

$$\operatorname{supp}(t) := \big\{\, \mathcal{P} \in \operatorname{Spc}(\mathcal{K}) \,\big|\, t \otimes \kappa(\mathcal{P}) \neq 0 \,\big\}\,.$$

This support satisfies all properties (SD' 1)-(SD' 7) of Theorem 47.

Note that (i) and (ii) of Theorem 47 are trivial here. It is not clear when this support detects vanishing, i. e. when $t \otimes \kappa(\mathcal{P}) = 0$ for all $\mathcal{P} \in \operatorname{Spc}(\mathcal{K})$ implies $t = 0$.

3. Examples and Applications

We now apply the theory of Part 2 to the examples of Part 1.

3.1. Classification of thick $\otimes$-ideals, after Hopkins. Such classifications began in stable homotopy theory, see § 1.3, long before the start of $\otimes$-triangular geometry. Via Theorem 16, this becomes:

Theorem 51 (Hopkins-Smith [26], see [5, Cor. 9.5]). *The spectrum of* $\mathrm{SH}^{\mathrm{fin}}$ *is*

$$\begin{tikzcd}
\mathcal{P}_{2,\infty} \arrow[d, dash] & \mathcal{P}_{3,\infty} \arrow[d, dash] & \cdots & \mathcal{P}_{p,\infty} \arrow[d, dash] & \cdots \\
\vdots \arrow[d, dash] & \vdots \arrow[d, dash] & & \vdots \arrow[d, dash] & \\
\mathcal{P}_{2,n+1} \arrow[d, dash] & \mathcal{P}_{3,n+1} \arrow[d, dash] & \cdots & \mathcal{P}_{p,n+1} \arrow[d, dash] & \cdots \\
\mathcal{P}_{2,n} \arrow[d, dash] & \mathcal{P}_{3,n} \arrow[d, dash] & \cdots & \mathcal{P}_{p,n} \arrow[d, dash] & \cdots \\
\vdots \arrow[d, dash] & \vdots \arrow[d, dash] & & \vdots \arrow[d, dash] & \\
\mathcal{P}_{2,1} \arrow[drr, dash] & \mathcal{P}_{3,1} \arrow[dr, dash] & \cdots & \mathcal{P}_{p,1} \arrow[dl, dash] & \cdots \\
& & \mathrm{SH}^{\mathrm{fin}}_{\mathrm{tor}} & &
\end{tikzcd}$$

The lines $\mathcal{P} - \mathcal{P}'$ indicate that the higher prime is in the closure of the lower one. For every prime number p and every $n \geq 1$, the prime $\mathcal{P}_{p,n}$ of $\mathrm{SH}^{\mathrm{fin}}$ is the kernel of the n-th Morava K-theory (composed with localization at p) and $\mathcal{P}_{p,\infty} = \cap_{n\geq 1}\mathcal{P}_{p,n}$ is the kernel of localization at p. Finally, $\mathrm{SH}^{\mathrm{fin}}_{\mathrm{tor}} := \operatorname{Ker}(\mathrm{H}(-,\mathbb{Q}))$ is the subcategory of torsion spectra. The surjective continuous map $\rho = \rho_{\mathrm{SH}^{\mathrm{fin}}} : \operatorname{Spc}(\mathrm{SH}^{\mathrm{fin}}) \longrightarrow \operatorname{Spec}(\mathbb{Z})$ of Theorem 42 is given by $\rho(\mathrm{SH}^{\mathrm{fin}}_{\mathrm{tor}}) = (0)$ and $\rho(\mathcal{P}_{p,n}) = p\mathbb{Z}$ for all $1 \leq n \leq \infty$.

Remark 52. This example yields many observations. First, $\operatorname{Spc}(\mathrm{SH}^{\mathrm{fin}})$ is not noetherian and the closed subsets $\{\mathcal{P}_{n,\infty}\}$ are not the support of any object. In particular, in the local category $\mathrm{SH}^{\mathrm{fin}}_p$ at p, we have $\mathrm{FL}(\mathrm{SH}^{\mathrm{fin}}_p) = 0$. Finally, $\operatorname{Spec}(\mathrm{SH}^{\mathrm{fin}})$ is a locally ringed space but is not a scheme. See more in [5].

Remark 53. Hopkins [25] also understood that this classification could be transposed to algebra and indicated that (15) should provide the classification for $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(A)$, with the subsets $Y \subset \operatorname{Spec}(A)$ being all specialization closed subsets. The actual proof of this statement requires A to be noetherian and

was given by Neeman [39]. But it is Thomason who nailed down the dual-open subsets (our Thomason subsets) in [48, Thm. 3.15]. His result settles the non-noetherian affine case and, most interestingly, works for any quasi-separated scheme if one insists on $\otimes$-*ideal* thick subcategories. Via Theorem 16 and Construction 29, this yields:

Theorem 54 (Reconstruction [1, Thm. 6.3]). *Let X be a quasi-separated scheme; see Section 1.2. We have an isomorphism* $\operatorname{Spec}(\mathrm{D}^{\mathrm{perf}}(X)) \simeq X$ *of ringed spaces.*

Remark 55. Under the underlying homemorphism $\operatorname{Spc}(\mathrm{D}^{\mathrm{perf}}(X)) \simeq X$, we can reformulate another famous result of Thomason's [49, § 5]: *For every quasi-compact $U \subset X$, we have* $\mathcal{K}(U) \cong \mathrm{D}^{\mathrm{perf}}(U)$, where $\mathcal{K}(U)$ is as in Construction 24.

Remark 56. The map $\varphi : X \to \operatorname{Spc}(\mathrm{D}^{\mathrm{perf}}(X))$ of Theorem 16 sends $x \in X$ to $\operatorname{Ker}\big(\mathrm{D}^{\mathrm{perf}}(X) \to \mathrm{D}^{\mathrm{perf}}(\mathcal{O}_{X,x})\big)$. For $X = \operatorname{Spec}(A)$ affine and $\mathfrak{p} \in \operatorname{Spec}(A)$, the quotient $\mathcal{K}/\varphi(\mathfrak{p}) \cong \mathrm{D}^{\mathrm{perf}}(A_{\mathfrak{p}})$ is indeed the expected local category. Let us make two further observations. First, φ reverses inclusions, i. e. if $\mathfrak{p} \subset \mathfrak{q}$ in A then $\varphi(\mathfrak{p}) \supset \varphi(\mathfrak{q})$ in $\mathcal{K}$. This phenomenon is in line with other mildly surprising facts, to an algebraist's eye, like $\overline{\{\mathcal{P}\}} = \{\, \mathcal{Q} \mid \mathcal{Q} \subset \mathcal{P} \,\}$ for every $\mathcal{P} \in \operatorname{Spc}(\mathcal{K})$.

Secondly, an inverse to φ is given by the map $\rho_{\mathcal{K}} : \operatorname{Spc}(\mathcal{K}) \to \operatorname{Spec}(\mathrm{R}_{\mathcal{K}}) = \operatorname{Spec}(A)$ of Theorem 42. Hence $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(A)$ provides an example where $\rho_{\mathcal{K}}$ is not only surjective, as follows from Theorem 43, but also injective. Interestingly, one can actually give a direct proof of the injectivity of $\rho_{\mathcal{K}}$ in this case and obtain the Hopkins-Neeman-Thomason classification for $\mathrm{D}^{\mathrm{perf}}(A)$ by Theorem 14. See details in [5, Rem. 8.4].

Walking in Hopkins's steps, Benson-Carlson-Rickard [12] and later Friedlander-Pevtsova [22] performed the classification in modular representation theory for finite groups and finite group schemes. Combined with Theorem 16, this reads:

Theorem 57 ([1, Thm. 6.3] and [5, Cor. 9.5]). *Let k be a field of positive characteristic and G be a finite group (scheme over k). See Section 1.4. Consider the graded-commutative cohomology ring* $\mathrm{H}^{\bullet}(G,k)$. *Then, for the derived category* $\mathcal{K} = \mathrm{D}^{\mathrm{b}}(kG\text{–mod})$, *the map $\rho^{\bullet}_{\mathcal{K}}$ of Theorem 42 induces an isomorphism*

$$\operatorname{Spec}(\mathrm{D}^{\mathrm{b}}(kG\text{–mod})) \simeq \operatorname{Spec}^{\mathrm{h}}(\mathrm{H}^{\bullet}(G,k))$$

between the triangular spectrum of $\mathcal{K}$ and the homogeneous spectrum of the cohomology. Via (5), *it restricts to an isomorphism* $\operatorname{Spec}(\operatorname{stab}(kG)) \simeq \operatorname{Proj}(\mathrm{H}^{\bullet}(G,k))$, *where the latter is the so-called* projective support variety $\mathcal{V}_G(k)$.

Indeed, Friedlander and Pevtsova were able to reconstruct the structure sheaf of $\mathcal{V}_G$ by computing the triangular structure sheaf $\mathcal{O}_{\mathcal{K}}$ of our Construction 29.

Recently, Krishna [32, Thm. 7.10] proved that the spectrum of the category of perfect complexes over a (reasonable) stack is the associated moduli space.

3.2. Further computations. It is now natural to turn to other, newer areas, where the classification of thick $\otimes$-ideals is not yet known, to see whether the spectrum can be computed by some other means. Here are some first results in motivic theory and noncommutative topology. In both cases, the spectrum is only known in the simplest $\otimes$-triangulated category that one can produce. But these should be considered as bridgeheads in two unknown (but friendly) territories.

Let us start with motivic theory, see § 1.5-§ 1.6. Here, the simplest category is probably that of mixed Tate motives with rational coefficients, i. e. the triangulated subcategory of $\mathrm{DM}(k)_{\mathbb{Q}}$ generated by the Tate objects $\mathbb{Q}(i)$, for all $i \in \mathbb{Z}$.

Theorem 58 (Peter [42]). *Let k be a number field and $\mathrm{DMT}(k)_{\mathbb{Q}}$ be the triangulated category of mixed Tate motives. Then $\mathrm{Spc}(\mathrm{DMT}(k)_{\mathbb{Q}})$ is just a point.*

At the other end of the motivic game, the computation of the spectrum of $\mathrm{SH}^{\mathbb{A}^1}_{\mathrm{gm}}(S)$ as in § 1.6 is probably a difficult long-term challenge. Using Theorem 43 and Morel's computation [37] of $\mathrm{End}_{\mathrm{SH}^{\mathbb{A}^1}}(\mathbb{1})$, we can still get:

Theorem 59 ([5, Cor. 10.1]). *Let $\mathcal{K} = \mathrm{SH}^{\mathbb{A}^1}_{\mathrm{gm}}(k)$ for a perfect field k of characteristic different from 2 as in Section 1.6. Then the continuous map $\rho_{\mathcal{K}}$ of Theorem 42 defines a surjection from the triangular spectrum $\mathrm{Spc}(\mathcal{K})$ onto the Zariski spectrum $\mathrm{Spec}(\mathrm{GW}(k))$ of the Grothendieck-Witt ring of quadratic forms over k.*

The second area we want to discuss is noncommutative topology, see § 1.7. In that case, the baby $\otimes$-triangulated category is the thick subcategory $\mathcal{K}^G$ of KK^G generated by the unit. The ring of endomorphisms of the unit $R(G) = \mathrm{End}_{KK^G}(\mathbb{1})$ is the Grothendieck group of continuous complex representations of G.

Theorem 60 (Dell'Ambrogio [21]). *Let G be a finite group. Then the map $\rho_{\mathcal{K}^G}$ of Theorem 42 is split surjective. It is a homeomorphism for G trivial, i. e. $\mathrm{Spc}(\mathcal{K}) \simeq \mathrm{Spec}(\mathbb{Z})$ where $\mathcal{K} \subset KK$ is the triangulated subcategory generated by $\mathbb{1} = \mathbb{C}$.*

Dell'Ambrogio also conjectured [21, Conj. 1.3] that $\rho_{\mathcal{K}^G}$ is injective for every finite group G. Again, our surjectivity Theorem 43 applies in big generality:

Theorem 61 ([5, Cor. 8.8]). *Let G be a compact Lie group. Then the continuous map $\rho_{\mathcal{K}^G} : \mathrm{Spc}(\mathcal{K}^G) \to \mathrm{Spec}(R(G))$ of Theorem 42 is surjective.*

Remark 62. A famous result of Quillen in modular representation theory of a finite group G asserts that $\mathcal{V}_G$ is covered by the images of the $\mathcal{V}_E$ under the

maps $\mathrm{Spc}(\mathrm{res}^G_E) : \mathcal{V}_E \to \mathcal{V}_G$, where $E < G$ runs through the elementary abelian p-subgroups. Dell'Ambrogio explains in [21] how the celebrated Baum-Connes conjecture with coefficients would follow from an analogous property in KK-theory, namely that the spectrum of KK^G (G as in §1.7) be covered by the images of the various spectra of KK^H, where $H < G$ runs through compact subgroups.

3.3. Applications to algebraic geometry.

The following result is an immediate corollary of Theorem 54:

Corollary 63. *Let X and Y be two quasi-separated (e. g. noetherian) schemes. If their derived categories of perfect complexes are equivalent $\mathrm{D}^{\mathrm{perf}}(X) \simeq \mathrm{D}^{\mathrm{perf}}(Y)$ as* tensor *triangulated categories then the schemes $X \simeq Y$ are isomorphic.*

Remark 64. A $\otimes$-triangular equivalence $\mathrm{D}_{\mathrm{Qcoh}(X)}(X) \simeq \mathrm{D}_{\mathrm{Qcoh}(Y)}(Y)$ restricts to a $\otimes$-triangular equivalence on the compact parts, $\mathrm{D}^{\mathrm{perf}}(X) \simeq \mathrm{D}^{\mathrm{perf}}(Y)$, hence implies $X \simeq Y$ as well. This reconstruction result is known to fail without the tensor: There exist non-isomorphic schemes, even abelian varieties, with triangular equivalent derived categories. See Mukai [38].

Remark 65. In homological mirror symmetry, or more generally each time that one expects a given triangulated category $\mathcal{K}$ to be equivalent to $\mathrm{D}^{\mathrm{perf}}(X)$ for some (maybe conjectural) scheme X, it becomes interesting to construct the tensor product on $\mathcal{K}$ which should correspond to that of $\mathrm{D}^{\mathrm{perf}}(X)$. See [47]. In this situation, the scheme X *must* be $\mathrm{Spec}(\mathcal{K})$ by Theorem 54. This does not guarantee that $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(X)$ but it tells us what X must be.

The abstract results of $\otimes$-triangular geometry apply in particular to $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(X)$. For instance, the filtration by (co)dimension of support in Theorem 32 yields a spectral sequence in any cohomology theory "defined" on derived categories, like K-theory or Witt theory, for instance. In particular, we get the following generalization of Quillen's famous spectral sequence for regular schemes [44]:

Theorem 66 ([4, Thm. 1]). *Let X be a (topologically) noetherian scheme of finite Krull dimension. Then there is a cohomologically indexed and converging spectral sequence in Thomason non-connective K-theory [49], of local-global nature:*

$$E_1^{p,q} = \bigoplus_{x \in X^{(p)}} K_{-p-q}(\mathcal{O}_{X,x} \text{ on } \{x\}) \qquad \xRightarrow[p,q,n \in \mathbb{Z}]{p+q=n} \qquad K_{-n}(X)\,.$$

Remark 67. This theorem is a first *strict* application of $\otimes$-triangular geometry, since the statement does not involve $\otimes$-triangulated categories. Yet, the deeper result is Theorem 32 which says that the quotient $\mathrm{D}^{\mathrm{perf}}(X)_{(d)} / \mathrm{D}^{\mathrm{perf}}(X)_{(d-1)}$

decomposes, up to idempotent completion, as the coproduct of the categories $\mathrm{FL}(\mathrm{D}^{\mathrm{perf}}(\mathcal{O}_{X,x})) = \{\, a \in \mathrm{D}^{\mathrm{perf}}(\mathcal{O}_{X,x}) \mid \mathrm{supp}(a) \subset \{x\} \,\}$ over all $x \in X_{(d)}$.

This illustrates the "boomerang effect" of abstraction: Inspired by Quillen [44], we started from the well-known fact that for a *regular* scheme, the above quotient is exactly equivalent to $\coprod_{x \in X_{(d)}} \mathrm{FL}(\mathrm{D}^{\mathrm{perf}}(\mathcal{O}_{X,x}))$, without idempotent completion, and we tried to extend it to $\otimes$-triangular geometry. This simply fails! But it works if one adds the idempotent completion to the picture. Then, Theorem 32 holds in *all* areas of $\otimes$-triangular geometry. Now, this yields a gain even in algebraic geometry where we started, for we understand that the regularity assumption was not that important after all. In K-theory, the idempotent completion explains the presence of negative K-theory in Theorem 66. Of course, all this has its origin in Thomason's description of $\mathrm{D}^{\mathrm{perf}}(U)$ (Remark 55) and it is fair to say that he had everything in [49] to prove Theorem 66. It is nonetheless remarkable that these ideas extend so far beyond algebraic geometry.

3.4. Applications to modular representation theory. In modular representation theory, see § 1.4, the filtration Theorem 32 applied to $\mathcal{K} = \mathrm{stab}(kG)$ recovers, and slightly improves, a result of Carlson-Donovan-Wheeler [20, Thm. 3.5]. Let us rather comment on the Picard group, $\mathrm{Pic}(\mathrm{stab}(kG))$, which is a classical invariant, known as the group $T(G) = T_k(G)$ of endotrivial kG-modules up to isomorphism. A kG-module M is *endotrivial* if $\mathrm{End}_k(M) \simeq k \oplus (\mathrm{proj})$ which simply means that $M^* \otimes M \simeq \mathbb{1}$ in $\mathrm{stab}(kG)$. We proved:

Theorem 68 (B.-Benson-Carlson [7]). *The endotrivial modules obtained by the gluing technique of Theorem 37 generate a finite-index subgroup of $T(G)$.*

Remark 69. Recall the $\otimes$-triangulated category $\mathcal{K}(U)$ of Construction 24 for every quasi-compact open $U \subset \mathrm{Spc}(\mathcal{K})$. In algebraic geometry, for X a scheme and $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(X)$, Thomason proved $\mathcal{K}(U) \simeq \mathrm{D}^{\mathrm{perf}}(U)$, see Remark 55. In other words, the construction $(\mathcal{K}, U) \longmapsto \mathcal{K}(U)$ "stays inside algebraic geometry".

On the other hand, for $\mathcal{K} = \mathrm{stab}(kG)$ and $U \subset \mathcal{V}_G(k)$ non-trivial, $\mathcal{K}(U)$ is never equivalent to a stable category $\mathrm{stab}(kH)$, no matter what finite group H one tries. See [6, Prop. 4.2]. Hence, although Thomason's result *does* work abstractly and transposes to modular representation theory via the $\otimes$-triangular construction $\mathcal{K}(U)$, the resulting construction takes us out of basic modular representation theory. Here is a nice *strict* application of Theorem 38 (without $\otimes$-triangulated categories in the statement):

Theorem 70 ([6, Thm. 4.7]). *Let G be a finite group and $\mathcal{V}_G = \mathrm{Proj}(\mathrm{H}^\bullet(G,k))$ its projective support variety over a field k of characteristic dividing the order of G. Then gluing induces an injection $\beta : \mathrm{Pic}(\mathcal{V}_G) \underset{\mathbb{Z}}{\otimes} \mathbb{Z}[1/p] \hookrightarrow T(G) \underset{\mathbb{Z}}{\otimes} \mathbb{Z}[1/p]$.*

Combining with Theorem 68, we obtain a rational isomorphism

$$\mathrm{Pic}(\mathcal{V}_G) \underset{\mathbb{Z}}{\otimes} \mathbb{Q} \simeq T(G) \underset{\mathbb{Z}}{\otimes} \mathbb{Q}\,.$$

Remark 71. The above result fails integrally. For instance, for $G = Q_8$ the quaternion group and k containing a cubic root of unity, the group of endotrivials is $T(Q_8) = \mathbb{Z}/4 \oplus \mathbb{Z}/2$ although $\mathrm{Spc}(\mathrm{stab}(kQ_8)) = \mathcal{V}_{Q_8}(k) = *$ is just a point, hence $\mathrm{Pic}(\mathcal{V}_{Q_8}) = 0$. Note also that $\mathrm{stab}(kQ_8)$ is a local $\otimes$-triangulated category.

3.5. Intra-utero applications. While $\otimes$-triangular geometry was still in the making, $\otimes$-triangulated categories showed useful in the theory of Witt groups of quadratic forms over schemes. This abstract theory, of so-called *triangular Witt groups*, has been quite useful. It led to the proof of the Gersten conjecture for Witt groups, among many other (strict) applications, including the computation of several classical Witt groups. For a survey, the interested reader is referred to [2]. In retrospect, many of these triangular Witt groups results fit very well in the language of $\otimes$-triangular geometry.

4. Problems

We have already mentioned a few open questions in the above text. In conclusion, we briefly suggest some additional directions of possible interest. We refrain from insisting on the wildest dreams (as in Remarks 62 and 65 for instance) and favor of a few problems reasonably close to the current stage of the theory.

4.1. Computing the spectrum in more examples. As discussed in § 2.5, the most basic question is to compute $\mathrm{Spc}(\mathcal{K})$ for more $\otimes$-triangulated categories $\mathcal{K}$, preferably without using the classification of thick $\otimes$-ideals, in order to deduce the latter via Theorem 14 and show off a little. Theorem 47 offers an angle of attack. Still, we need more results telling us how to compare $\mathrm{Spc}(\mathcal{K})$ to other spaces. Such a comparison is provided by the maps $\rho_{\mathcal{K}}$ and $\rho^{\bullet}_{\mathcal{K}}$ of Theorem 42. We have seen that these maps are often surjective (Thm. 43). It then becomes interesting to decide when they are injective and more generally to study their fibers.

In algebraic examples like $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(A)$ or $\mathcal{K} = \mathrm{D}^{\mathrm{b}}(kG\text{–mod})$, the map $\rho^{\bullet}_{\mathcal{K}}$ is injective (see § 3.1) but we have seen in the very first example (Thm. 51) that injectivity fails completely outside algebra. The tempting guess would be:

Conjecture 72. The map $\rho^{\bullet}_{\mathcal{K}}$ is (locally) injective when $\mathcal{K}$ is "algebraic enough".

Here "algebraic enough" could mean those triangulated categories $\mathcal{K}$ which arise as stable categories of Frobenius exact categories, or, alternatively, those

$\mathcal{K}$ which are the derived category of some dg-category, see Keller [29]. It might also be necessary to add some hypothesis like $\mathcal{K}$ being locally generated by $\mathbb{1}$.

Remark 73. By Hochster [24], any spectral space, like our $\mathrm{Spc}(\mathcal{K})$, is the spectrum of *some* commutative ring. It would be pleasant to construct such a ring explicitly in terms of $\mathcal{K}$. The above use of $\mathrm{R}_{\mathcal{K}}$ and $\mathrm{R}^{\bullet}_{\mathcal{K}}$ was a first attempt to do this.

4.2. Image of algebraic geometry in $\otimes$-triangular geometry. We have seen in Theorem 54 that a scheme X can be reconstructed from the $\otimes$-triangulated category $\mathrm{D}^{\mathrm{perf}}(X)$. An important question is to decide which $\otimes$-triangulated categories $\mathcal{K}$ are $\otimes$-equivalent to $\mathrm{D}^{\mathrm{perf}}(\mathrm{Spec}(\mathcal{K}))$. Actually, it would also be interesting to know when the locally ringed space $\mathrm{Spec}(\mathcal{K})$ is a scheme. As already mentioned in Remark 65, this could have consequences beyond algebraic geometry, as for instance in homological mirror symmetry.

Also interesting would probably be the tensor-triangular characterization of some properties of morphisms of schemes, like being smooth or étale.

4.3. Residue fields. In examples, triangular primes $\mathcal{P} \subset \mathcal{K}$ are often the kernel of a tensor functor $\mathcal{K} \to \mathcal{F}$ with $\mathcal{F} = \mathrm{VB}_k$ being the category of k-vector spaces over a field k (in algebraic geometry), or $\mathcal{F}$ being the category of graded modules over a graded field $k[t, t^{-1}]$ (in homotopy theory), or $\mathcal{F} = \mathrm{stab}(kC_p)$ being the stable category of kC_p-modules, for C_p the cyclic group of order $p = \mathrm{char}(k)$ (in modular representation theory, although this case is still unclear). This observation calls for two things:

(a) The definition of *$\otimes$-triangular fields* $\mathcal{F}$, which would imply in particular that $\mathrm{Spc}(\mathcal{F}) = \{*\}$ is reduced to a point.

(b) The construction, for every local category $\mathcal{K}$ (Thm. 25), of a conservative $\otimes$-exact functor $\pi : \mathcal{K} \to \mathcal{F}$ into some $\otimes$-triangular field, that would be a "residue field". Conservative means that $\mathrm{Ker}(\pi) = 0$, i. e. that the image of $\mathrm{Spc}(\pi) : \{*\} = \mathrm{Spc}(\mathcal{F}) \longrightarrow \mathrm{Spc}(\mathcal{K})$ would be the unique closed point of $\mathrm{Spc}(\mathcal{K})$.

Note that there might be several such residue field functors, as seems to be the case in modular representation theory. It is not at all clear whether such functors can be constructed from the $\otimes$-triangular structure alone but they should certainly be looked for in examples where one tries to determine $\mathrm{Spc}(\mathcal{K})$.

Regarding the definition of $\otimes$-triangular fields, the naive idea of requesting the category $\mathcal{F}$ to be semi-simple does not cover $\mathrm{stab}(kC_p)$ for instance. Indeed, $\mathrm{Spc}(\mathrm{stab}(kC_p))$ is a point but there is no non-zero $\otimes$-exact functor from $\mathrm{stab}(kC_p)$ into a semi-simple $\otimes$-category as soon as $p \geq 3$. (For $p = 2$, $\mathrm{stab}(kC_2) \cong \mathrm{VB}_k$.) Currently, my favorite guess is to define $\mathcal{F}$ to be a triangular field if every non-zero object $x \in \mathcal{F}$ is faithful (i. e. $x \otimes f = 0$ forces $x = 0$ or $f = 0$). This covers all three examples above and still forces $\mathrm{Spc}(\mathcal{F}) = \{*\}$ but

there is no solid conceptual motivation for this definition at this stage, beyond unification of examples.

4.4. Nilpotence. A clear understanding of nilpotence phenomena in triangulated categories still eludes us, even in the presence of a tensor. First, we do not know how to define *reduced* $\otimes$-triangulated categories. Nor do we know how to construct $\mathrm{D}^{\mathrm{perf}}(X_{\mathrm{red}})$ out of the $\otimes$-triangulated category $\mathcal{K} = \mathrm{D}^{\mathrm{perf}}(X)$, except via the odious cheat: $\mathrm{D}^{\mathrm{perf}}\big((\mathrm{Spec}(\mathcal{K}))_{\mathrm{red}}\big)$. For instance, even when $\mathrm{Spc}(\mathcal{K}) = \{*\}$ is a point, that is, when $\mathcal{K}$ is something like an "artinian local" $\otimes$-triangulated category, it is not clear how to obtain a residue field (§ 4.3) by reduction modulo nilpotents.

Also, there seems to be no obvious way to construct a $\otimes$-triangulated category "$\mathcal{K}$ over Z", for a closed subset $Z \subset \mathrm{Spc}(\mathcal{K})$ of the spectrum, say, with what should be the "reduced structure". Neither do I know which closed subsets $Z \subset \mathrm{Spc}(\mathcal{K})$ are the support of an object $u \in \mathcal{K}$ as in Theorem 18 (c). Again, this relates to the residue field of § 4.3 when $\mathcal{K}$ is local and $Z = \{*\}$ is the closed point.

4.5. Torsion in the Picard group. This is a follow-up on Remarks 39 and 71. First, let us note that the isomorphism $\mathrm{Pic}(\mathcal{V}_G) \otimes \mathbb{Q} \simeq T(G) \otimes \mathbb{Q}$ of Theorem 70 is still unknown for G a finite group *scheme*, because we do not know whether the Picard group is locally torsion in that case. We have seen in Proposition 40 that the Picard group can be locally wild. Yet, the example $\coprod_G \mathcal{K}$ can be ruled out if we further require $\mathcal{K}$ to be generated by $\mathbb{1}$, as a thick triangulated subcategory. Hence the following hope survives:

Conjecture 74. Let $\mathcal{K}$ be a $\otimes$-triangulated category as in Hypothesis 21. Assume that $\mathcal{K}$ is local (Thm. 25) and that $\mathcal{K}$ is generated by $\mathbb{1}$. Let $u \in \mathcal{K}$ be $\otimes$-invertible. Then there exists $m > 0$ such that $u^{\otimes m}$ is trivial in the sense that $u^{\otimes m} \simeq \Sigma^n \mathbb{1}$ for some $n \in \mathbb{Z}$. That is, $\mathrm{Pic}(\mathcal{K})$ is rationally trivial: $\mathrm{Pic}(\mathcal{K}) \otimes_{\mathbb{Z}} \mathbb{Q} = \mathbb{Q} \cdot [\Sigma \mathbb{1}]$.

References

[1] P. Balmer. The spectrum of prime ideals in tensor triangulated categories. *J. Reine Angew. Math.*, 588:149–168, 2005.

[2] P. Balmer. Witt groups. In *Handbook of K-theory. Vol. 2*, pages 539–576. Springer, Berlin, 2005.

[3] P. Balmer. Supports and filtrations in algebraic geometry and modular representation theory. *Amer. J. Math.*, 129(5):1227–1250, 2007.

[4] P. Balmer. Niveau spectral sequences on singular schemes and failure of generalized Gersten conjecture. *Proc. Amer. Math. Soc.*, 137(1):99–106, 2009.

[5] P. Balmer. Spectra, spectra, spectra - Tensor triangular spectra versus Zariski spectra of endomorphism rings. Preprint, 2009.

[6] P. Balmer. Picard groups in triangular geometry and applications to modular representation theory. *Trans. Amer. Math. Soc.*, 362(7), 2010.

[7] P. Balmer, D. J. Benson, and J. F. Carlson. Gluing representations via idempotent modules and constructing endotrivial modules. *J. Pure Appl. Algebra*, 213(2):173–193, 2009.

[8] P. Balmer and G. Favi. Generalized tensor idempotents and the telescope conjecture. Preprint, 24 pages, available at `www.math.ucla.edu/~balmer`, 2009.

[9] P. Balmer and G. Favi. Gluing techniques in triangular geometry. *Q. J. Math.*, 51(4):415–441, 2007.

[10] P. Balmer and M. Schlichting. Idempotent completion of triangulated categories. *J. Algebra*, 236(2):819–834, 2001.

[11] D. J. Benson. *Representations and cohomology I & II*, volume 30 & 31 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1998.

[12] D. J. Benson, J. F. Carlson, and J. Rickard. Thick subcategories of the stable module category. *Fund. Math.*, 153(1):59–80, 1997.

[13] D. J. Benson, S. B. Iyengar, and H. Krause. Local cohomology and support for triangulated categories. *Ann. Sci. Éc. Norm. Supér. (4)*, 41(4):573–619, 2008.

[14] P. Berthelot, A. Grothendieck, and L. Illusie, editors. *SGA 6: Théorie des intersections et théorème de Riemann-Roch*. Springer LNM 225. 1971.

[15] A. Bondal and M. van den Bergh. Generators and representability of functors in commutative and noncommutative geometry. *Mosc. Math. J.*, 3(1):1–36, 258, 2003.

[16] A. B. Buan, H. Krause, and Ø. Solberg. Support varieties: an ideal approach. *Homology, Homotopy Appl.*, 9(1):45–74, 2007.

[17] T. Bühler. On the algebraic foundations of bounded cohomology. ETH Thesis 2008. To appear in *Mem. Amer. Math. Soc.*

[18] J. F. Carlson. The variety of an indecomposable module is connected. *Invent. Math.*, 77(2):291–299, 1984.

[19] J. F. Carlson. *Modules and group algebras*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1996.

[20] J. F. Carlson, P. W. Donovan, and W. W. Wheeler. Complexity and quotient categories for group algebras. *J. Pure Appl. Algebra*, 93(2):147–167, 1994.

[21] I. Dell'Ambrogio. Tensor triangular geometry and KK-theory. Preprint, 33 pages, available online, 2009.

[22] E. M. Friedlander and J. Pevtsova. Π-supports for modules for finite group schemes. *Duke Math. J.*, 139(2):317–368, 2007.

[23] D. Happel. *Triangulated categories in the representation theory of finite-dimensional algebras*, volume 119 of *LMS Lecture Note*. Cambr. Univ. Press, Cambridge, 1988.

[24] M. Hochster. Prime ideal structure in commutative rings. *Trans. Amer. Math. Soc.*, 142:43–60, 1969.

[25] M. J. Hopkins. Global methods in homotopy theory. In *Homotopy theory (Durham, 1985)*, volume 117 of *LMS Lect. Note*, pages 73–96. Cambridge Univ. Press, 1987.

[26] M. J. Hopkins and J. H. Smith. Nilpotence and stable homotopy theory. II. *Ann. of Math. (2)*, 148(1):1–49, 1998.

[27] M. Hovey. *Model categories*, volume 63 of *Mathematical Surveys and Monographs*. American Mathematical Society, 1999.

[28] M. Hovey, J. H. Palmieri, and N. P. Strickland. Axiomatic stable homotopy theory. *Mem. Amer. Math. Soc.*, 128(610), 1997.

[29] B. Keller. Deriving DG categories. *Ann. Sci. École Norm. Sup. (4)*, 27(1):63–102, 1994.

[30] B. Keller and A. Neeman. The connection between May's axioms for a triangulated tensor product and Happel's description of the derived category of the quiver D_4. *Doc. Math.*, 7:535–560, 2002.

[31] M. Kontsevich. Homological algebra of mirror symmetry. In *Proc. of the Intern. Congress of Mathematicians (Zürich, 1994)*, pages 120–139. Birkhäuser, 1995.

[32] A. Krishna. Perfect complexes on Deligne-Mumford stacks and applications. *J. K-Theory*, 4(3):559–603, 2009.

[33] S. Mac Lane. *Categories for the working mathematician*, volume 5 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

[34] J. P. May. The additivity of traces in triangulated categories. *Adv. Math.*, 163(1):34–73, 2001.

[35] R. Meyer. Categorical aspects of bivariant K-theory. In *K-theory and noncommutative geometry*, EMS Ser. Congr. Rep., pages 1–39. Eur. Math. Soc., 2008.

[36] F. Morel. An introduction to $\mathbb{A}^1$-homotopy theory. In *Contemporary developments in algebraic K-theory*, ICTP Lect. Notes, XV, pages 357–441. Trieste, 2004.

[37] F. Morel. On the motivic π_0 of the sphere spectrum. In *Axiomatic, enriched and motivic homotopy theory*, volume 131 of *NATO Sci. Ser. II*, pages 219–260. Kluwer, Dordrecht, 2004.

[38] S. Mukai. Duality between $D(X)$ and $D(\hat{X})$ with its application to Picard sheaves. *Nagoya Math. J.*, 81:153–175, 1981.

[39] A. Neeman. The chromatic tower for $D(R)$. *Topology*, 31(3):519–532, 1992.

[40] A. Neeman. The Grothendieck duality theorem via Bousfield's techniques and Brown representability. *J. Amer. Math. Soc.*, 9(1):205–236, 1996.

[41] A. Neeman. *Triangulated categories*, volume 148 of *Annals of Mathematics Studies*. Princeton University Press, 2001.

[42] T. Peter. Prime ideals of mixed Tate motives. Preprint, 2010.

[43] J. Pevtsova. Spectra of tensor triangulated categories. MSRI talk, 2008 April 4, available online at `www.msri.org`.

[44] D. Quillen. Higher algebraic K-theory. I. In *Algebraic K-theory, I: Higher K-theories (Proc. Conf., Battelle Memorial Inst., Seattle, Wash., 1972)*, pages 85–147. Lecture Notes in Math., Vol. 341. Springer, Berlin, 1973.

[45] J. Rickard. Derived categories and stable equivalence. *J. Pure Appl. Algebra*, 61(3):303–317, 1989.

[46] J. Rickard. Idempotent modules in the stable category. *J. London Math. Soc. (2)*, 56(1):149–170, 1997.

[47] A. Subotic. Monoidal structures on the Fukaya category. PhD, Harward (2010).

[48] R. W. Thomason. The classification of triangulated subcategories. *Compositio Math.*, 105(1):1–27, 1997.

[49] R. W. Thomason and T. Trobaugh. Higher algebraic K-theory of schemes and of derived categories. In *The Grothendieck Festschrift, Vol. III*, volume 88 of *Progr. Math.*, pages 247–435. Birkhäuser, Boston, MA, 1990.

[50] J.-L. Verdier. Des catégories dérivées des catégories abéliennes. *Astérisque*, 239, 1996.

[51] V. Voevodsky. $\mathbf{A}^1$-homotopy theory. In *Proceedings of the International Congress of Mathematicians, Vol. I (Berlin, 1998)*, pages 579–604, 1998.

[52] V. Voevodsky, A. Suslin, and E. M. Friedlander. *Cycles, transfers, and motivic homology theories*, volume 143 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 2000.

[53] R. Vogt. *Boardman's stable homotopy category*. Lecture Notes Series, No. 21. Matematisk Institut, Aarhus Universitet, Aarhus, 1970.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Modules for Elementary Abelian p-groups

David J. Benson*

Abstract

Let $E \cong (\mathbb{Z}/p)^r$ $(r \geq 2)$ be an elementary abelian p-group and let k be an algebraically closed field of characteristic p. A finite dimensional kE-module M is said to have constant Jordan type if the restriction of M to every cyclic shifted subgroup of kE has the same Jordan canonical form. I shall begin by discussing theorems and conjectures which restrict the possible Jordan canonical form. Then I shall indicate methods of producing algebraic vector bundles on projective space from modules of constant Jordan type. I shall describe realisability and non-realisability theorems for such vector bundles, in terms of Chern classes and Frobenius twists. Finally, I shall discuss the closely related question: can a module of small dimension have interesting rank variety? The case p odd behaves throughout these discussions somewhat differently to the case $p = 2$.

Mathematics Subject Classification (2010). Primary: 20C20; Secondary: 14F05

Keywords. Modular representations, elementary abelian groups, constant Jordan type, vector bundles.

Dramatis Personæ: B=Benson, C=Carlson, F=Friedlander, P=Pevtsova, S=Suslin.

1. Introduction

Many questions in modular representation theory of finite groups reduce to questions about elementary abelian subgroups. A prototype for such a reduction is Chouinard's Theorem [11], which states that a module is projective if and only if its restriction to every elementary abelian subgroup is projective. Quillen [19, 20] described the spectrum of the cohomology ring in terms of the elementary abelian subgroups, and this was generalised to the theory of varieties for modules by Carlson [7, 8], Alperin and Evens [1], Avrunin and Scott [2]. The classification of localising subcategories of the stable module category also

*David Benson, Aberdeen. E-mail: bensondj@maths.abdn.ac.uk

reduces to elementary abelian subgroups, see Benson, Iyengar and Krause [5, 6]. These theorems and others motivate the study of modules for an elementary abelian p-group

$$E = \langle g_1, \dots, g_r \rangle \cong (\mathbb{Z}/p)^r$$

over an algebraically closed field k of characteristic p. These are generally unclassifiable, so we often restrict our attention to particular classes of modules that we might hope to understand better.

We shall only be interested in finite dimensional kE-modules in this talk. Dade's Lemma [12] states that a finite dimensional kE-module M is projective (or equivalently, free) if and only if its restriction to every cyclic shifted subgroup is free. A cyclic shifted subgroup is a subgroup of the group algebra kE of order p generated by an element of the form $1 + X_\alpha$ where

$$X_\alpha = \lambda_1 X_1 + \cdots + \lambda_r X_r, \tag{1.1}$$

$X_i = g_i - 1 \in J(kE)$ and $\alpha = (\lambda_1, \dots, \lambda_r) \in \mathbb{A}^r(k) \setminus \{0\}$. This motivates Carlson's definition of the *rank variety* of M:

Definition 1.2. The rank variety of a kE-module M is defined to be the closed homogeneous subset of $\mathbb{A}^r = \mathbb{A}^r(k)$ given by

$$V_E^r(M) = \{\alpha \in \mathbb{A}^r \setminus \{0\} \mid M{\downarrow_{\langle 1+X_\alpha \rangle}} \text{ is not free}\} \cup \{0\}.$$

Example 1.3. Let $r = 4$, let $p = 2$, and let M be the four dimensional kE-module defined by

$$g_1 \mapsto \begin{pmatrix} 1&0&0&0\\0&1&0&0\\1&0&1&0\\0&0&0&1 \end{pmatrix} \quad g_2 \mapsto \begin{pmatrix} 1&0&0&0\\0&1&0&0\\0&1&1&0\\0&0&0&1 \end{pmatrix} \quad g_3 \mapsto \begin{pmatrix} 1&0&0&0\\0&1&0&0\\0&0&1&0\\1&0&0&1 \end{pmatrix} \quad g_4 \mapsto \begin{pmatrix} 1&0&0&0\\0&1&0&0\\0&0&1&0\\0&1&0&1 \end{pmatrix}.$$

Then

$$X_\alpha \mapsto \begin{pmatrix} 0&0&0&0\\0&0&0&0\\\lambda_1&\lambda_2&0&0\\\lambda_3&\lambda_4&0&0 \end{pmatrix}$$

and $V_E^r(M)$ is the variety defined by the vanishing of the minor $\lambda_1\lambda_4 - \lambda_2\lambda_3$.

We write $k[Y_1, \dots, Y_r]$ for the coordinate ring of affine space $k[\mathbb{A}^r]$. With this notation, in Example 1.3, $V_E^r(M)$ is the irreducible quadric with equation given by $Y_1Y_4 - Y_2Y_3 = 0$. This, when projectivised, gives the Segre embedding of $\mathbb{P}^1 \times \mathbb{P}^1$ in $\mathbb{P}^3$.

2. Jordan Type

Let k, E and M be as described in the introduction. The elements $X_1, \dots, X_r$ are generators for $J(kE)$, and the elements X_α of equation (1.1) form a set of coset representatives of $J^2(kE)$ in $J(kE)$. Since $X_\alpha^p = 0$, the action of X_α on

M breaks up into Jordan blocks of length between 1 and p with eigenvalue 0. We write

$$[p]^{a_p} \dots [1]^{a_1}$$

for the Jordan type.

Warning 2.1. If $x, y \in J(kE)$, $x - y \in J^2(kE)$, it can happen that x and y have different Jordan types on M.

Definition 2.2. Nilpotent Jordan types are partially ordered: if A and B are nilpotent square matrices of the same size, $A \geq B$ if and only if for all $s > 0$ we have $\operatorname{rank}(A^s) \geq \operatorname{rank}(B^s)$. If we think of Jordan types as represented by partitions, then this is the well known dominance ordering.

We say that $x \in J(kE) \setminus J^2(kE)$ has *maximal* Jordan type if it is maximal with respect to this partial order.

Maximal Jordan type was examined in depth by Friedlander, Pevtsova and Suslin [14], and the following theorem relating it to generic Jordan type forms the beginning of their investigation.

Theorem 2.3 (FPS). *If $x, y \in J(kE)$ and $x - y \in J^2(kE)$ then x has maximal Jordan type if and only if y does, so that the elements of maximal Jordan type determine a well defined subset of $J(kE)/J^2(kE)$. This is a dense open subset. This Jordan type is the same as that of the element X_α for a generic point α defined over a large enough transcendental extension of k.*

Because of this theorem, we talk of the *generic* Jordan type of M, which is the Jordan type of a generic element X_α of M.

3. Modules of Constant Jordan Type

Definition 3.1 (CFP [10]). We say that M has *constant Jordan type* $[p]^{a_p} \dots [1]^{a_1}$ if every element of $J(kE) \setminus J^2(kE)$ has the same Jordan canonical form on M.

Rather than working in the module category, we often wish to work in the stable module category where we ignore projective summands of a module. So if we forget the term $[p]^{a_p}$ in the Jordan type of a module, as it can be recovered from the remaining terms and the dimension, we talk of a module M of *stable constant Jordan type* $[p-1]^{a_{p-1}} \dots [1]^{a_1}$. Dade's lemma shows that a module M has empty stable constant Jordan type if and only if M is projective.

It is obvious that a direct sum of modules of constant Jordan type is again such. It is not quite so obvious, but it follows from Theorem 2.3, that a direct summand of a module of constant Jordan type is again such. After all, if there's a closed subset where one summand has smaller Jordan type, the other would have to have larger Jordan type on the same closed subset, in order for the sum to be constant.

Warning 3.2. If M and N are kE-modules, it is not necessarily true that if $\alpha \in \mathbb{A}^r \setminus \{0\}$ then $M \otimes_k N$ (with diagonal group action) restricted to X_α is isomorphic to $M{\downarrow}_{X_\alpha} \otimes_k N{\downarrow}_{X_\alpha}$.

Nonetheless, we have the following theorem.

Theorem 3.3 (CFP [10]). *If M and N have constant Jordan type then so do M^* and $M \otimes_k N$.*

The fundamental question then arises:

Question 3.4. Let us assume that $r \geq 2$, so that the situation is not trivial. Then what stable constant Jordan types can occur?

4. Endotrivial Modules

What modules have stable constant Jordan type [1]? Suppose that M is such a module. Then $M \otimes_k M^* \cong \mathrm{Hom}_k(M, M)$ also has stable constant Jordan type [1]. The canonical maps $k \to M \otimes_k M^* \to k$, defined using the inclusion of the identity map and the trace map on matrices, compose to give $\dim M$ times the identity map. But $\dim M \equiv 1 \pmod p$, so $M \otimes_k M^*$ decomposes as a direct sum of a trivial module and another summand; this other summand is projective, by Dade's lemma.

Definition 4.1. A module M is said to be *endotrivial* if $M \otimes_k M^*$ is a direct sum of a trivial module and a projective module.

Theorem 4.2 (Dade [12]). *If M is an endotrivial module for an elementary abelian p-group then $M \cong \Omega^n(k) \oplus$ (projective) $(n \in \mathbb{Z})$.*

The notation here is as follows. If $n \geq 0$ then $\Omega^n(k)$ is the nth kernel in a minimal projective resolution of k while $\Omega^{-n}(k)$ is the nth cokernel in a minimal injective resolution of k. The module $\Omega^n(k)$ has stable constant Jordan type [1] if $p = 2$ or n is even, while if both p and n are odd it has stable constant Jordan type $[p-1]$.

So we have seen that the indecomposable modules of stable constant Jordan type [1] are precisely the modules $\Omega^n(k)$, where n is even if p is odd.

We can deal with indecomposable modules of stable constant Jordan type $[p-1]$ (p odd) in the same way. If M is such a module, then again $M \otimes_k M^*$ has stable constant Jordan type [1], and we deduce using the same arguments that M is isomorphic to $\Omega^n(k)$ with n odd.

So what about modules of stable constant Jordan type $[a]$ with $1 < a < p-1$? It was conjectured by CFP [10] that there are no modules of stable constant Jordan type [2] if $p \geq 5$. I proved this conjecture while visiting MSRI in 2008 [4].

Theorem 4.3 (B, MSRI 2008). *If $r \geq 2$ then there are no kE-modules of stable constant Jordan type $[a]$ with $1 < a < p-1$.*

The technique of proof was to take exterior and symmetric powers, and use Dade's lemma to get incompatible congruences on $\dim_k M$.

There are two conjectures, each of which imply the theorem above. The first was formulated by Rickard at MSRI in 2008, based on a computer printout of data from a large number of modules of constant Jordan type.

Conjecture 4.4 (Rickard, MSRI 2008). Suppose that $r \geq 2$ and M is a module of constant Jordan type. If there are no Jordan blocks of length i then the total number of Jordan blocks of length greater than i is divisible by p.

The second was formulated by Suslin and recorded in CFP [10].

Conjecture 4.5 (S). Suppose that $r \geq 2$ and M is a module of constant Jordan type. If for some i with $2 \leq i \leq p-1$, the module M has a Jordan block of length i, then it also has a Jordan block either of length $i-1$ or of length $i+1$.

The smallest cases which remain unresolved, and which would follow from either of these conjectures, are the existence of modules of stable constant Jordan type $[3][1]$ and $[2]^2$ for $p \geq 5$. But each conjecture disallows types allowed by the other, and we have no reasonable conjecture in general as to exactly what types occur.

5. Vector Bundles on Projective Space

We use the phrase *vector bundle* on $\mathbb{P}^{r-1}$ in the algebraic sense, so that it is equivalent to the phrase *locally free sheaf of $\mathcal{O}$-modules*, where $\mathcal{O}$ is the structure sheaf of $\mathbb{P}^{r-1}$.

Remarks 5.1. The only line bundles on $\mathbb{P}^{r-1}$ are $\mathcal{O}(n)$ for $n \in \mathbb{Z}$.

$r = 2$: every vector bundle on $\mathbb{P}^1$ is a direct sum of line bundles; the decomposition is essentially unique (Grothendieck).

$r \geq 3$: It is moderately easy to construct indecomposable vector bundles on $\mathbb{P}^{r-1}$ of every rank $\geq r-2$.

The only known indecomposable vector bundles with rank bigger than 1 and less than $r-2$ are:

On $\mathbb{P}^4$, the *Horrocks–Mumford bundle* [16] $\mathcal{F}_{HM}$ of rank 2 with $15,000$ symmetries (the group is $5^{1+2}SL(2,5)$),

On $\mathbb{P}^5$, *Horrocks' Parent bundle* [15] of rank 3,

On $\mathbb{P}^5$ in characteristic 2, the *Tango bundle* [21] of rank 2,

> Some further rank 2 bundles on $\mathbb{P}^4$ and rank 3 bundles on $\mathbb{P}^5$ constructed by Kumar [17], and by Kumar, Peterson and Rao [18], in positive characteristic only...
>
> ...and bundles obtained from these by twisting and pulling back through self-maps of projective space.

In particular, it remains unknown whether there are indecomposable vector bundles of rank two on $\mathbb{P}^6$ in any characteristic and on $\mathbb{P}^5$ in characteristic other than 2.

We construct vector bundles $\mathcal{F}_i(M)$ $(1 \leq i \leq p)$ from a module M of constant Jordan type as follows. We let

$$\mathbb{P}^{r-1} = \mathsf{Proj}\, k[Y_1, \dots, Y_r]$$

where the Y_i are the functions on $\mathbb{A}^r$ defined by $Y_i(X_j) = \delta_{ij}$ (Kronecker delta). Given a kE-module M, we set $\tilde{M} = M \otimes_k \mathcal{O}$, a trivial bundle over $\mathbb{P}^{r-1}$ whose rank is equal to the dimension of M.

Definition 5.2 (FP [13]). We define a map of vector bundles

$$\theta\colon \tilde{M}(j) \to \tilde{M}(j+1) \qquad (j \in \mathbb{Z})$$

by the formula

$$\theta(m \otimes f) = \sum_i X_i m \otimes Y_i f.$$

Definition 5.3 (BP, MSRI 2008).

$$\mathcal{F}_i(M) = \frac{\operatorname{Ker}\theta \cap \operatorname{Im}\theta^{i-1}}{\operatorname{Ker}\theta \cap \operatorname{Im}\theta^{i}}$$

as a subquotient of $\tilde{M}$, for $1 \leq i \leq p$.

Remark 5.4. This definition has the feature that $\mathcal{F}_i(M)$ is a vector bundle for $1 \leq i \leq p$ if and only if M has constant Jordan type. The rank of $\mathcal{F}_i(M)$ is equal to the number of Jordan blocks of length i on a cyclic shifted subgroup.

Example 5.5. Let $E = (\mathbb{Z}/p)^2 = \langle g_1, g_2 \rangle$, $kE = k[X_1, X_2]/(X_1^p, X_2^p)$, $X_i = g_i - 1$. Let M be the kE-module given by

$$g_1 \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad g_2 \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Then

$$\theta = \begin{pmatrix} 0 & 0 & 0 \\ Y_1 & 0 & 0 \\ Y_2 & 0 & 0 \end{pmatrix}$$

has kernel of rank two and image of rank one; $\mathcal{F}_1(M)$ and $\mathcal{F}_2(M)$ are both rank one bundles.

Example 5.6. More generally, for E of any rank, the module $M = kE/J^2(kE)$ has constant Jordan type $[2][1]^{r-1}$. We have $\mathcal{F}_1(M) \cong \mathcal{T}(-1)$ and $\mathcal{F}_2(M) \cong \mathcal{O}(-1)$, where $\mathcal{T}$ is the tangent bundle.

Example 5.7. The kE-module $M = \mathrm{Soc}^2(kE)$ also has constant Jordan type $[2][1]^{r-1}$, but $\mathcal{F}_1(M) \cong \Omega(1)$, $\mathcal{F}_2(M) \cong \mathcal{O}$ where $\Omega = \mathcal{T}^*$ denotes the cotangent bundle.

Example 5.8 (B, MSRI 2008). If $p \geq 7$ and $r = 5$ then there exists a kE-module M of dimension $30p^5$ of stable constant Jordan type $[p-1]^{30}[2]^2[1]^{26}$ such that $\mathcal{F}_2(M) \cong \mathcal{F}_{\mathrm{HM}}(-2)$, a twist of the Horrocks–Mumford bundle. See the preprint [3] for details.

Example 5.9 (B, Summer 2008). Let $E = (\mathbb{Z}/2)^6$, and M be the module for which θ is the following 30×30 matrix:

$$\begin{pmatrix}
a & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
b & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
c & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
d & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
e & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
f & & & & & & & & & & & & & & & & & & & & & & & & & & & & & \\
 & a & & & & & & & f & e & & & & & & & & & & & & & & & & & & & \\
 & b & & & & & & f & & d & & & & & & & & & & & & & & & & & & & \\
 & c & & & & & & e & d & & & & & & & & & & & & & & & & & & & & \\
 & d & & & c & b & & & & & & & & & & & & & & & & & & & & & & & \\
 & e & & c & & a & & & & & & & & & & & & & & & & & & & & & & & \\
 & f & & b & a & & & & & & & & & & & & & & & & & & & & & & & & \\
 & & & d & & f & a & & & c & & & & & & & & & & & & & & & & & & & \\
 & & & e & & & & & a & & & & & & & & & & & & & & & & & & & & \\
 & & & f & & & & & & & a & & & & & & & & & & & & & & & & & & \\
 & & & & d & & & b & & & & & & & & & & & & & & & & & & & & & \\
 & & & & e & f & & & & b & c & & & & & & & & & & & & & & & & & & \\
 & & & & f & & & & & & b & & & & & & & & & & & & & & & & & & \\
 & & & & & d & c & & & & & & & & & & & & & & & & & & & & & & & \\
 & & & & & e & & c & & & & & & & & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & f & e & & b & c & & a & & & & & & & & & & & \\
 & & & & & & & & & & & f & & d & b & & & a & & c & & & & & & & & & & \\
 & & & & & & & & & & & e & d & & c & & & & c & & a & b & & & & & & & & \\
 & & & & & & & & c & b & & & & & & & & e & d & & & f & & & & & & & & \\
 & & & & & & c & & & a & & & & & & e & d & & & & & & f & & & & & & & \\
 & & & & & & b & a & & & & & & & & f & & d & & f & e & & & & & & & & & \\
 & & & & & & d & e & f & a & b & c & & & & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & & & & & & & & & & d & e & f & a & b & c
\end{pmatrix}$$

Then M has constant Jordan type $[2]^{14}[1]^2$ and $\mathcal{F}_1(M)$ is the Tango bundle of rank two on $\mathbb{P}^5$.

6. Properties of $\mathcal{F}_i(M)$

The following properties of the vector bundles $\mathcal{F}_i(M)$ were found by BP, MSRI 2008.

- $\mathcal{F}_{p-i}(\Omega M) \cong \mathcal{F}_i(M)(-p+i)$ $(1 \leq i \leq p-1))$
- $\mathcal{F}_i(M^*) \cong \mathcal{F}_i(M)^\vee(-i+1)$ $(1 \leq i \leq p)$
- $\mathcal{F}_1(M \otimes_k N) \cong \bigoplus_{i=1}^{p-1} \mathcal{F}_i(M) \otimes_{\mathcal{O}} \mathcal{F}_i(N)(i-1)$

- The short exact sequence $0 \to \Omega M \to P_M \to M \to 0$ induces a complex

$$0 \to \mathcal{F}_p(\Omega M) \to \mathcal{F}_p(P_M) \to \mathcal{F}_p(M) \to 0$$

which is not exact, but has homology only in the middle, where it is

$$\bigoplus_{i=1}^{p-1} \mathcal{F}_i(M)(-p+i).$$

Realisability of vector bundles by modules of constant Jordan type is addressed in the following theorem.

Theorem 6.1 (BP, MSRI 2008). *Given a vector bundle $\mathcal{F}$ of rank s on $\mathbb{P}^{r-1}$, there exists a kE-module of stable constant Jordan type $[1]^s$ such that*

1. *if $p = 2$ then $\mathcal{F}_1(M) \cong \mathcal{F}$,*
2. *if p is odd then $\mathcal{F}_1(M) \cong F^*(\mathcal{F})$*

where $F\colon \mathbb{P}^{r-1} \to \mathbb{P}^{r-1}$ is the Frobenius map.

The method of proof of the theorem is to take a resolution of the vector bundle by sums of twists of the structure sheaf. The maps in the resolution are polynomials in $Y_1, \dots, Y_r$. If $p = 2$, a polynomial of degree d is realised by a map from $\Omega^d k$ to k. Then there is a construction in the stable module category of kE which produces a single module from these maps.

If p is odd then the pth power of the polynomial is realised by a map from $\Omega^{2d} k$ to k, and so we only get to realise bundles whose resolutions involve only pth powers of polynomials. This is where the Frobenius twist comes in.

When p is odd, there is an obstruction to improving the theorem to $\mathcal{F}_1(M) \cong \mathcal{F}$ coming from Chern classes.

7. Chern Classes

The *Chow group* $A^*(\mathbb{P}^{r-1})$ is isomorphic to $\mathbb{Z}[h](h^r)$. Given a vector bundle $\mathcal{F}$ on $\mathbb{P}^{r-1}$, there is a *Chern polynomial*

$$c(\mathcal{F}) = 1 + c_1(\mathcal{F})h + \cdots + c_{r-1}(\mathcal{F})h^{r-1} \in A^*(\mathbb{P}^{r-1})$$

whose coefficients $c_i(\mathcal{F})$ are the *Chern numbers* of $\mathcal{F}$.

Theorem 7.1 (B, Summer 2008). *Suppose that $r \geq 2$, and let M be a kE-module of stable constant Jordan type $[1]^s$. Then $p \mid c_i(\mathcal{F}_1(M))$ for $1 \leq i \leq p-2$.*

If $p = 2$ this gives no information, but for p odd it gives a genuine restriction on the vector bundles that can occur this way.

Example 7.2. The rank two Horrocks–Mumford bundle $\mathcal{F}_{\mathrm{HM}}$ on $\mathbb{P}^4$ has Chern numbers $c_1(\mathcal{F}_{\mathrm{HM}}(i)) = 2i + 5$ and $c_2(\mathcal{F}_{\mathrm{HM}}(i)) = i^2 + 5i + 10$. So no twist of $\mathcal{F}_{\mathrm{HM}}$ can occur as $\mathcal{F}_1(M)$ for a module of stable constant Jordan type $[1]^2$. This explains why, in Example 5.8, it was necessary to have Jordan blocks of lengths other than one and p.

8. Small Modules with Interesting Varieties

We now move away from constant Jordan type, and look at some more speculative questions. Looking back at Example 1.3, can we mimic this construction if p is odd? The most obvious attempts fail, so we are left with the general question: are there small modules with interesting varieties, when p is odd? Can we find good bounds and good constructions to address this question?

A construction of Carlson produces modules with any desired closed homogeneous subvariety of $\mathbb{A}^r$ as its rank variety, but the dimension of the module produced this way is large. For example, if we do this with the ruled quadric $Y_1Y_4 - Y_2Y_3 = 0$ for $(\mathbb{Z}/2)^4$, then we are required to interpret this as an element of $H^2(E, k) \cong \mathrm{Ext}^1_{kE}(\Omega k, k)$ and take the corresponding extension of k by Ωk, of dimension 16. Carlson's method in general is to realise hypersurfaces in this way and then tensor modules for a general variety written as an intersection of hypersurfaces.

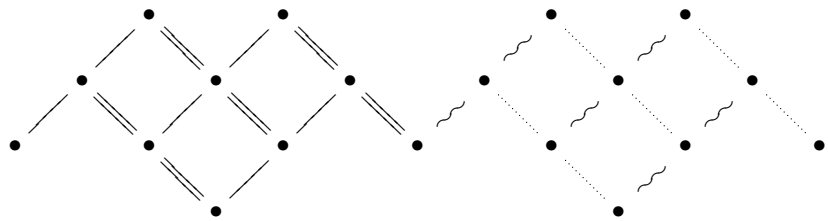
A modified version of Carlson's method produces Example 1.3. Namely, instead of using a single element of cohomology, we use a 2×2 matrix of elements $\left(\begin{smallmatrix} Y_1 & Y_2 \\ Y_3 & Y_4 \end{smallmatrix}\right)$ and interpret it as an element of $\mathrm{Ext}^1_{kE}(k \oplus k, k \oplus k)$. The corresponding four dimensional module gives the 4×4 matrices of the example.

When p is odd, the single element method is even worse. The generators in degree one of cohomology are nilpotent, and cannot be used in Carlson's construction. One is forced to use the generators of degree two, so $Y_1Y_4 - Y_2Y_3$ corresponds to an element of $H^4(E, k) \cong \mathrm{Ext}^1_{kE}(\Omega^3 k, k)$. The dimension of the module obtained in this way is $13p^4$.

If we go with the 2×2 matrix method for p odd, we still end up with an element of $\mathrm{Ext}^1_{kE}(\Omega k \oplus \Omega k, k \oplus k)$, giving a module of dimension $2p^4$.

But we can do better than this. If a row of the matrix does not use all of the variables, we can use a relative syzygy of k instead of the absolute syzygy. The condition for making this work is that the variety of the subgroup used for the relative syzygy should be contained in the variety defined by the polynomials appearing in the row. So for the 2×2 matrix we're considering, we can use relative syzygies for two subgroups of order p^2 to obtain a module of dimension

$2p^2$ with the required variety. Here is a diagram for this module when $p = 3$:



In this diagram the actions of the four elements $X_i = g_i - 1$ $(1 \leq i \leq 4)$ are represented by the single, double, wavy and dotted edges respectively. The leftmost and rightmost vertices are identified to make a module of dimension 18.

This module of dimension $2p^2$ is still far from the lower bound given by Bézout's theorem, as described in Carlson [9]:

Theorem 8.1 (C, 1993). *Let E be an elementary abelian p-group of rank r. If M is a kE-module whose rank variety has dimension m and degree d then $\dim_k M/p^{r-m}$ is an integer at least as big as d.*

In our case, the Bézout bound is $2p$, but it is not hard to prove that there is no module of dimension $2p$ with the required variety for p odd. I suspect that the smallest module with the ruled quadric as an irreducible component of its variety is the above one of dimension $2p^2$.

Question 8.2. Let E be an elementary abelian p-group of rank r. Given a closed homogeneous subvariety $V \subseteq \mathbb{A}^r$, what is the smallest dimension of a kE-module M with $V_E(M) = V$?

9. Modules for $(\mathbb{Z}/p)^2$

For an elementary abelian group of rank two, Question 8.2 is not interesting, because the projective line does not have interesting subvarieties. But when we mix conditions on the variety with conditions on the Jordan type, we do get some very interesting questions. I shall describe just one theorem in this direction.

Let $E = (\mathbb{Z}/p)^2$ with p odd. For an indecomposable kE-module the possible varieties are the whole of $\mathbb{A}^2$ and a single line through the origin. In the latter case, namely the case of periodic modules, we can look at the Jordan type of the module on the cyclic shifted subgroup corresponding to that line. The following theorem gives rather surprising information about the dimensions of such modules.

Theorem 9.1 (B, 2010). *If an indecomposable periodic kE-module M has Jordan blocks of length p on all cyclic shifted subgroups then the dimension of*

M is at least $p^{\frac{3}{2}}$. Furthermore, there exists such a module M of dimension at most $p^{\frac{3}{2}} + \sqrt{2}\,p^{\frac{5}{4}}$.

The constant $\sqrt{2}$ in the theorem is probably not best possible. There must be further theorems of a similar type waiting to be discovered, but at the moment this seems to be the only one known.

References

[1] J. L. Alperin and L. Evens, *Varieties and elementary abelian subgroups*, J. Pure & Applied Algebra **26** (1982), 221–227.

[2] G. S. Avrunin and L. L. Scott, *Quillen stratification for modules*, Invent. Math. **66** (1982), 277–286.

[3] D. J. Benson, *Modules of constant Jordan type and the Horrocks–Mumford bundle*, preprint, 2008.

[4] ———, *Modules of constant Jordan type with one non-projective block*, Algebras and Representation Theory **13** (2010).

[5] D. J. Benson, S. B. Iyengar, and H. Krause, *Local cohomology and support for triangulated categories*, Ann. Scient. Éc. Norm. Sup. (4) **41** (2008), 1–47.

[6] ———, *Stratifying modular representations of finite groups*, Preprint, 2008.

[7] J. F. Carlson, *The complexity and varieties of modules*, Integral representations and their applications, Oberwolfach, 1980, Lecture Notes in Mathematics, vol. 882, Springer-Verlag, Berlin/New York, 1981, pp. 415–422.

[8] ———, *The varieties and cohomology ring of a module*, J. Algebra **85** (1983), 104–143.

[9] ———, *Varieties and modules of small dimension*, Arch. Math. (Basel) **60** (1993), 425–430.

[10] J. F. Carlson, E. M. Friedlander, and J. Pevtsova, *Modules of constant Jordan type*, J. Reine & Angew. Math. **614** (2008), 191–234.

[11] L. Chouinard, *Projectivity and relative projectivity over group rings*, J. Pure & Applied Algebra **7** (1976), 278–302.

[12] E. C. Dade, *Endo-permutation modules over p-groups, II*, Ann. of Math. **108** (1978), 317–346.

[13] E. M. Friedlander and J. Pevtsova, *Constructions for infinitesimal group schemes*, Preprint, 2008.

[14] E. M. Friedlander, J. Pevtsova, and A. Suslin, *Generic and maximal Jordan types*, Invent. Math. **168** (2007), 485–522.

[15] G. Horrocks, *Examples of rank three vector bundles on five dimensional projective space*, J. London Math. Soc. **18** (1978), 15–27.

[16] G. Horrocks and D. Mumford, *A rank* 2 *vector bundle on* $\mathbb{P}^4$ *with* $15,000$ *symmetries*, Topology **12** (1973), 63–81.

[17] N. M. Kumar, *Construction of rank two vector bundles on $\mathbb{P}^4$ in positive characteristic*, Invent. Math. **130** (1997), 277–286.

[18] N. M. Kumar, C. Peterson, and A. P. Rao, *Construction of low rank vector bundles on $\mathbb{P}^4$ and $\mathbb{P}^5$*, J. Alg. Geometry **11** (2002), 203–217.

[19] D. G. Quillen, *The spectrum of an equivariant cohomology ring, I*, Ann. of Math. **94** (1971), 549–572.

[20] ———, *The spectrum of an equivariant cohomology ring, II*, Ann. of Math. **94** (1971), 573–602.

[21] H. Tango, *On morphisms from projective space $\mathbb{P}^n$ to the Grassmann variety* $\mathrm{Gr}(n, d)$, J. Math. Kyoto Univ. **16** (1976), 201–207.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Total Positivity and Cluster Algebras

Sergey Fomin*

Abstract

This is a brief and informal introduction to cluster algebras. It roughly follows the historical path of their discovery, made jointly with A. Zelevinsky. Total positivity serves as the main motivation.

Mathematics Subject Classification (2010). Primary 13F60, Secondary 05E10, 05E15, 14M15, 15A23, 15B48, 20F55, 22E46.

Keywords. Total positivity, cluster algebra, chamber minors, quiver mutation.

Introduction

Cluster algebras are encountered in many algebraic and geometric contexts, with combinatorics providing a unifying framework. This short paper reviews the origins of cluster algebras, their deep connections with total positivity phenomena, and some of their recent manifestations in Teichmüller theory.

The introduction of cluster algebras, made in joint work with A. Zelevinsky [26], was rooted in the desire to understand, in a concrete and combinatorial way, G. Lusztig's theory of total positivity and canonical bases in quantum groups (see, e.g., [44, 47]). Although this goal remains largely elusive (cf. [43]), the concept proved valuable due to its surprising ubiquity, and to the connections it helped uncover between diverse and seemingly unrelated areas of mathematics.

This paper gives a popular and quick introduction to the subjects in the title, aimed at an uninitiated reader, and roughly following the historical order of modern developments in the two related fields. Cumbersome technicalities involved in the usual definition of cluster algebras are largely omitted, giving

*Partially supported by NSF grant DMS-0555880.
Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA.
E-mail: fomin@umich.edu.

way to prototypical examples from which the reader is invited to generalize, to discussions of underlying motivations, and to hints concerning further applications and extensions of the basic theory. Many important aspects are left out due to space limitations.

The style is rather informal, owing to the desire to see the forest through the trees, and to make the paper accessible to a general mathematical audience. There are no numbered formulas or theorems: results are stated as part of the general narrative. Some attributions are missing; they can be found in the sources quoted. The goal is to give the reader an intuitive feel for what cluster algebras are, and motivate her/him to read the more formal expositions elsewhere.

Several survey/introductory papers dedicated to the subjects in the title, approached from various perspectives, have already appeared in the literature; see in particular [1, 5, 19, 25, 29, 34, 39, 43, 54, 55, 56]. An excellent introduction to applications of cluster algebras in representation theory is given in B. Leclerc's contribution [43] to these proceedings. Besides consulting these sources and references therein, the reader is invited to visit the *Cluster Algebras Portal* [18], which provides numerous links to publications, conferences, seminars, thematic programs, software packages, etc.

Our presentation is loosely based on the papers [2, 3, 20, 21, 23, 25, 26, 27, 29], joint with A. Berenstein, M. Shapiro, D. Thurston, and A. Zelevinsky. Section 1 introduces total positivity and the idea of a positive/nonnegative part of an algebraic variety. Section 2 presents the basic notions of cluster algebra theory, emphasizing its roots in total positivity. Section 3 discusses the occurence of cluster algebras in combinatorial topology of triangulated surfaces, and connections with Teichmüller spaces.

The format of this brief survey does not allow us to discuss several important directions of current research on cluster algebras and related fields. In particular, not covered here are the theory of *cluster categories* and the various facets of *categorification* [39, 40, 41, 50]; the connections between cluster algebras and *Poisson geometry* [32, 33]; closely related work on cluster varieties arising in *higher Teichmüller theory* [16, 17]; the polyhedral combinatorics of *cluster fans* and *Cambrian lattices* [52]; applications to *discrete integrable systems* [13, 28, 37, 41]; the machinery of *quivers with potentials* [11, 12]; connections with *Donaldson-Thomas invariants* [42, 49]; and other exciting topics.

Acknowledgments. The discovery of cluster algebras, the main work leading to it, and the development of fundamentals of the general theory were all done jointly with my longtime collaborator Andrei Zelevinsky. I am indebted to him, and to my co-authors Arkady Berenstein, Michael Shapiro, and Dylan Thurston for their invaluable contributions to our joint work discussed below. Catharina Stroppel persuaded me to give a talk in Bonn whose design this presentation follows. Bernhard Keller, George Lusztig, and Kelli Talaska made valuable editorial suggestions.

1. Total Positivity

A matrix x with real entries is called *totally positive* (resp., *totally nonnegative*) if all its minors—that is, determinants of square submatrices—are positive (resp., nonnegative). Following the pioneering work of I. Schoenberg, the systematic study of these classes of matrices was initiated in the 1930s by F. Gantmacher and M. Krein [31] who in particular showed that the eigenvalues of an $n \times n$ totally positive matrix are real, positive, and distinct.

Total positivity is a remarkably widespread phenomenon: matrices with positive/nonnegative minors play an important role in classical mechanics (theory of small oscillations), probability (one-dimensional diffusion processes), discrete potential theory (planar resistor networks), asymptotic representation theory (the Edrei-Thoma theorem), algebraic and enumerative combinatorics (immanants, lattice paths), and of course in linear algebra and its applications. See [1, 25, 31, 35, 38] for a plethora of examples and results, and for additional references.

A key technical fact from the classical theory of total positivity is C. Cryer's "splitting lemma" [8, 9]: an invertible square matrix x (say of determinant 1) is totally nonnegative if and only if it has a *Gaussian decomposition*

$$x = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ * & 1 & 0 & \cdots & 0 \\ * & * & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & \cdots & 1 \end{bmatrix} \begin{bmatrix} * & 0 & 0 & \cdots & 0 \\ 0 & * & 0 & \cdots & 0 \\ 0 & 0 & * & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & * \end{bmatrix} \begin{bmatrix} 1 & * & * & \cdots & * \\ 0 & 1 & * & \cdots & * \\ 0 & 0 & 1 & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

in which all three factors (lower-triangular unipotent, diagonal, and upper-triangular unipotent) are totally nonnegative. There is also a counterpart of this statement for totally positive matrices.

The Binet-Cauchy theorem implies that totally positive (resp., nonnegative) matrices in $G = \mathrm{SL}_n$ form a multiplicative semigroup, denoted by $G_{\geq 0}$. In view of Cryer's lemma, the study of $G_{\geq 0}$ can be reduced to the investigation of its subsemigroup $N_{\geq 0} \subset G_{\geq 0}$ of upper-triangular unipotent totally nonnegative matrices.

The celebrated Loewner-Whitney Theorem [45, 53] identifies the infinitesimal generators of $N_{\geq 0}$ as the *Chevalley generators* of the corresponding Lie algebra. In pedestrian terms, each upper-triangular unipotent totally nonnegative $n \times n$ matrix can be written as a product of (totally nonnegative) matrices

of the form

$$x_i(t) = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & t & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix};$$

here the matrix $x_i(t)$ differs from the identity matrix by a single entry $t \geq 0$ in row i and column $i+1$. This led G. Lusztig [46] to the idea of extending the notion of total positivity to other semisimple groups G, by defining the set $G_{\geq 0}$ of totally nonnegative elements in G as the semigroup generated by the Chevalley generators. Lusztig has shown that $G_{\geq 0}$ can be described by inequalities of the form $\Delta(x) \geq 0$ where Δ ranges over the appropriate *dual canonical basis* (at $q = 1$). This set is infinite, and very hard to understand; fortunately, it can be replaced [24] by a much simpler (and finite) set of *generalized minors* [22].

A yet more general (if informal) concept is one of a *totally positive/nonnegative variety*. Vaguely, the idea is this: take a complex variety X together with a family $\mathbf{\Delta}$ of "important" regular functions on X. The corresponding totally positive (resp., totally nonnegative) variety $X_{>0}$ (resp., $X_{\geq 0}$) is the set of points at which all of these functions take positive (resp., nonnegative) values:

$$X_{>0} = \{x \in X : \Delta(x) > 0 \text{ for all } \Delta \in \mathbf{\Delta}\}.$$

If X is the affine space of matrices of a given size (or $\mathrm{GL}_n(\mathbb{C})$ or $\mathrm{SL}_n(\mathbb{C})$), and $\mathbf{\Delta}$ is the set of all minors, then we recover the classical notion. One can restrict this construction to matrices lying in a given stratum of a Bruhat decomposition, or in a given *double Bruhat cell* [22, 46]. Another important example is the *totally positive (resp., nonnegative) Grassmannian* consisting of the points in a usual Grassmann manifold where all Plücker coordinates can be chosen to be positive (resp., nonnegative).

In each of these examples, the notion of positivity depends on a particular choice of a coordinate system: a basis in a vector space allows us to view linear transformations as matrices; a choice of reference flag determines a system of Plücker coordinates; and so on.

Why study totally nonnegative varieties? Besides the connections to Lie theory alluded to above, there are at least three more reasons.

First, some totally nonnegative varieties are interesting in their own right as they can be identified with important spaces, e.g. some of those arising in Teichmüller theory; cf. Section 3. One can hope to gain additional insight into the structure of such spaces and their compactifications by "upgrading" them to complex varieties, studying associated quantizations, etc. The nascent "higher Teichmüller theory" [7, 17] is one prominent expression of this paradigm.

Second, passing from a complex variety to its positive part can be viewed as a step towards its *tropicalization*. The deep connections between total positivity,

tropical geometry, and cluster theory lie outside the scope of this short paper; see [17, 21, 30] for some aspects of this emerging research area.

Yet another reason to study totally nonnegative varieties lies in the fact that their structure as semialgebraic sets reveals important features of related complex varieties. We illustrate this phenomenon using the example first studied in [46] (cf. also [2, 23]). Consider $N \subset \mathrm{SL}_n(\mathbb{C})$, the subgroup of $n \times n$ unipotent upper-triangular matrices. The corresponding totally nonnegative variety is the semigroup $N_{\geq 0}$ of totally nonnegative matrices in N. Take $n = 3$; then

$$N_{\geq 0} = \left\{ \begin{bmatrix} 1 & x & y \\ 0 & 1 & z \\ 0 & 0 & 1 \end{bmatrix} : \begin{array}{l} x \geq 0 \\ y \geq 0 \\ z \geq 0 \end{array} \text{ and } xz - y \geq 0 \right\}.$$

The inequalities defining $N_{\geq 0}$ are homogeneous in the following sense: replacing (x, y, z) by (ax, a^2y, az), with $a > 0$, does not change them. Consequently, the space $N_{\geq 0}$ is topologically a cone with the apex $x = y = z = 0$ (the identity matrix) over the base $M_{\geq 0} \subset N_{\geq 0}$ cut out by the plane $x + z = 1$. Thus

$$M_{\geq 0} \cong \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1 \text{ and } \ y \leq x(1 - x)\}$$

is the subset of the coordinate plane $\mathbb{R}^2$ bounded by the x axis and the parabola $y = x(1 - x)$, as shown in Figure 1(a).

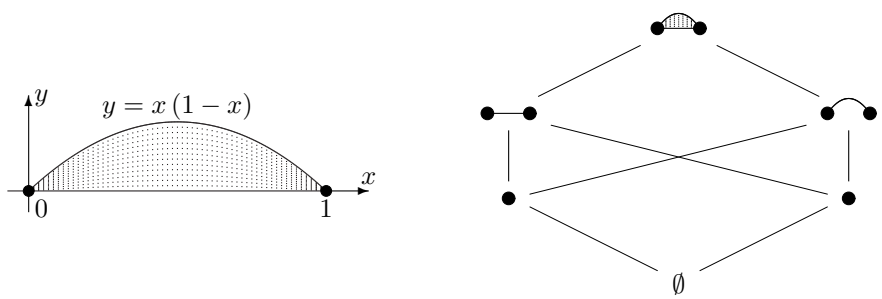


Figure 1. (a) The base $M_{\geq 0}$ of the cone $N_{\geq 0}$. (b) The attachment of algebraic strata.

The semialgebraic set $M_{\geq 0}$ naturally decomposes into 5 algebraic strata: two of dimension 0, two of dimension 1, and one of dimension 2. Accordingly, the cone $N_{\geq 0}$ decomposes into 6 algebraic strata of dimension 1 higher; the apex of $N_{\geq 0}$ corresponds to the "empty face" of $M_{\geq 0}$. See Figure 1(b).

The adjacency of these strata is described by a partial order isomorphic to the *Bruhat order* on the symmetric group $\mathcal{S}_3$. This happens in general, for any n: the decomposition of $N_{\geq 0}$ into algebraic strata produces a CW-complex with cell attachments described by the Bruhat order on $\mathcal{S}_n$. Recall that the same partial order describes the attachment of Schubert cells in the manifold of complete flags in $\mathbb{C}^n$. The latter has rich topology, and is a central object of study in modern Schubert Calculus. By contrast, $N_{\geq 0}$ and $M_{\geq 0}$ have no topology to speak of (in fact, $M_{\geq 0}$ is expected to be homeomorphic to a ball [23, 36])

but has a cell decomposition with exactly the same cell attachments. The big difference of course is that the complex Schubert cells have twice the dimensions of their real (more precisely, positive real) counterparts. Still, the stratification of $M_{\geq 0}$ resulting from its semialgebraic structure somehow "remembers"the Bruhat order—which is all one needs to know in order to reconstruct the topology of the flag manifold and its Schubert cells/varieties—including Schubert and Kazhdan-Lusztig polynomials, etc.

2. Cluster Algebras

The discussion in Section 1 prompts one to ask: Which algebraic varieties X have a natural notion of positivity? Which families $\mathbf{\Delta}$ of regular functions should one consider in defining this notion? The concept of a cluster algebra can be viewed as an attempt to provide a general answer to these questions. Since the definition is fairly technical, we start with an example and then generalize.

Our prototypical example of a cluster algebra $\mathcal{A}$ is the coordinate ring of the *base affine space* for the special linear group $G = \mathrm{SL}_n(\mathbb{C})$, defined as follows. The subgroup $N \subset G$ of unipotent upper-triangular matrices acts on G by right multiplication. The algebra $\mathcal{A} = \mathbb{C}[G/N]$ consists of regular functions on G which are invariant under this action of N. Thus elements of $\mathcal{A}$ can be viewed as polynomials in the entries x_{ij} of a matrix $x = (x_{ij}) \in \mathrm{SL}_n(\mathbb{C})$ which are invariant under column operations that add to a column of x a linear combination of preceding columns. Classical invariant theory tells us that $\mathcal{A}$ is generated by the *flag minors*

$$\Delta_I : x \mapsto \det(x_{ij} | i \in I, j \leq |I|)$$

where I ranges over nonempty proper subsets of $\{1, \ldots, n\}$. That is, Δ_I is a minor occupying the rows in I and the first several columns. The generators Δ_I satisfy certain well known homogeneous quadratic identities sometimes called *generalized Plücker relations*.

A point in G/N represented by a matrix x is, by definition, totally positive/nonnegative if all flag minors Δ_I take positive/nonnegative values at x. Total positivity in G/N is closely related to the classical notion of total positivity in G: it is not hard to deduce from Cryer's lemma that a matrix x is totally positive if and only if both x and its transpose represent totally positive elements in G/N.

There are $2^n - 2$ flag minors; do we really have to test all of them to verify that a point $x \in G/N$ is totally positive? The answer is no: it suffices to test positivity of $\dim(G/N) = \frac{(n-1)(n+2)}{2}$ minors; one could hardly hope for a more efficient test.

To design such tests, we will need the notion of a *pseudoline arrangement*. The latter is a collection of n "pseudolines" each of which is a graph of a continuous function on $[0, 1]$; each pair of pseudolines must have exactly one crossing

point in common. (See Figure 2.) The resulting arrangement is considered up to isotopy.

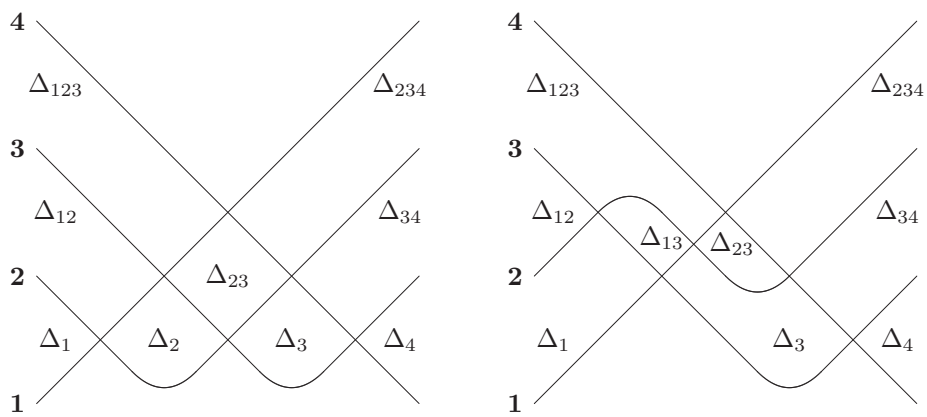


Figure 2. Two pseudoline arrangements, and associated chamber minors

We label the pseudolines **1** through **n** by numbering their left endpoints from the bottom up. To each *region* R of a pseudoline arrangement, with the exception of the top and the bottom regions, we associate the *chamber minor* $\Delta_{I(R)}$ (cf. [2]) defined as the flag minor indexed by the set $I(R)$ of labels of the pseudolines passing *below* R. The $\frac{(n-1)(n+2)}{2}$ chamber minors associated with a given pseudoline arrangement form an *extended cluster*; we shall see that the positivity of these minors implies that *all* flag minors of a given matrix are positive.

There are two types of regions: the *bounded* regions entirely surrounded by pseudolines, and the *unbounded* ones, adjacent to the left and right borders. The $2(n-1)$ chamber minors associated with unbounded regions are called *frozen*: these minors are present in every arrangement. For $n = 4$, the frozen minors are Δ_1, Δ_{12}, Δ_{123}, Δ_4, Δ_{34}, and Δ_{234} (cf. Figure 2).

The chamber minors corresponding to the bounded regions form the *cluster* associated with the given pseudoline arrangement. (Thus an extended cluster is a cluster plus the frozen minors.) Each cluster contains $\binom{n-1}{2}$ chamber minors. The two pseudoline arrangements shown in Figure 2 have clusters $\{\Delta_2, \Delta_3, \Delta_{23}\}$ and $\{\Delta_{13}, \Delta_3, \Delta_{23}\}$, respectively.

These two clusters differ in one element only. This is because the corresponding two arrangements are related to each other by a *local move* consisting in dragging one of the pseudolines through an intersection of two others; see Figure 3. As a result of such a move, one chamber minor (namely e in Figure 3, and Δ_2 in Figure 2) disappears (we say that this minor is *flipped*), and a new one (namely f in Figure 3, and Δ_{13} in Figure 2) is introduced.

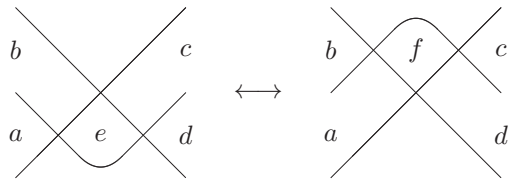


Figure 3. A local move in a pseudoline arrangement

It can be shown that for a local move as in Figure 3, the chamber minors associated with the regions where the action takes place satisfy the identity

$$ef = ac + bd.$$

This identity is one of the generalized Plücker relations alluded to above. We call it an *exchange relation*, as the chamber minors e and f are exchanged by the local move. For the local move shown in Figure 2, the exchange relation is

$$\Delta_2\Delta_{13} = \Delta_{12}\,\Delta_3 + \Delta_1\,\Delta_{23}\,.$$

The new chamber minor f produced by a local move is given by a simple rational expression $f = \frac{ac+bd}{e}$ in the chamber minors of the original arrangement. Note that this expression is *subtraction-free* (no minus signs). One can now start with a particular pseudoline arrangement, label its regions by indeterminates, then use iterated local moves (combined with the corresponding birational transformations) to generate all possible arrangements, and in doing so write *all* flag minors as rational expressions in the initial extended cluster. All these expressions are clearly subtraction-free, and the claim follows: if the elements of the initial extended cluster evaluate positively at a given point in G/N, then so do all flag minors.

Let $\mathcal{F}$ denote the field of rational functions in the formal variables making up the initial extended cluster. Inside $\mathcal{F}$, the rational expressions discussed in the previous paragraph generate the subalgebra $\mathcal{A}$ canonically isomorphic to $\mathbb{C}[G/N]$. Notice that our construction does not explicitly involve the group G: we can pretend to be unaware that we are dealing with matrices, their minors, etc. Yet the construction produces, by design, an algebra $\mathcal{A}$ equipped with a distinguished set of generators $\mathbf{\Delta}$ (the rational expressions corresponding to the flag minors), and thus endowed with a notion of (total) positivity.

The example of a base affine space treated above displays, in a rudimentary form, the main features of a general cluster algebra set-up. We next proceed to describing the latter on an informal level, with details to be filled in later on.

Fix a field $\mathcal{F}$ of rational functions in several variables, some of which are designated as "frozen." Imagine a (potentially infinite) family of equinumerous finite collections ("clusters") of elements in $\mathcal{F}$. (These elements, called *cluster variables*, can be thought of as regular functions on some "cluster variety" X.)

Each cluster can be "extended" by adjoining the frozen variables. The (extended) clusters are the vertices of a connected regular graph in which adjacent clusters are related by birational transformations of the most simple kind, replacing an arbitrary element of a cluster by a sum of two monomials divided by the element being removed. (By a monomial we mean a product of elements of a given extended cluster.) These transformations are subtraction-free, so positivity of the elements of a cluster at a point $x \in X$ does not depend on the choice of a cluster. The birational maps between adjacent clusters are encoded by appropriate combinatorial data, and the construction is made rigid by mandating that these data are transformed (as one moves to an adjacent cluster) according to certain canonical rules. These combinatorial rules define a discrete dynamics that drives the algebraic dynamics of cluster transformations. Consequently, the choice of initial combinatorial data (the pseudoline arrangement in the example of G/N) determines, in a recursive fashion, the entire structure of clusters and exchanges. The corresponding cluster algebra is then defined as the subring of the ambient field $\mathcal{F}$ generated by the elements of all extended clusters.

In the example of the base affine space, one key feature of the set-up described above is lacking: we do not always know how to exchange an element of a cluster. If a region in a pseudoline arrangement is bounded by more than three pseudolines, then the corresponding chamber minor cannot be readily flipped by a local move. For instance, how do we exchange the chamber minor Δ_{23} in Figure 2 on the left? There is in fact a "hidden" exchange relation of the form $\Delta_{23} \circledast = \circledast + \circledast$ —but how do we guess what those $\circledast$'s are?

The answer to this question will fall into our lap once we replace the language of pseudoline arrangements, too specialized for a general theory, by a more universal combinatorial language of quivers. (Using quivers somewhat restricts the generality of the cluster theory, but is general enough for the purposes of this paper.) Developing this language will take a little time—but will pay off quickly.

A *quiver* is a finite oriented graph. We allow multiple edges, but not loops (i.e., edges connecting a vertex to itself) or oriented 2-cycles (i.e., edges of opposite orientation connecting the same pair of vertices). We will need a slightly richer notion, with some vertices in a quiver designated as *frozen*. The remaining vertices are called *mutable*. We assume that no edges connect frozen vertices to each other. (Such edges would make no difference in what follows.)

Quivers play the role of the aforementioned combinatorial data accompanying the clusters. We think of the vertices of a quiver as labeled by the elements of an extended cluster, so that the frozen vertices are labeled by the frozen variables, and the mutable vertices by the cluster variables.

We next describe the quiver analogue of a local move. Let z be a mutable vertex in a quiver Q. The *quiver mutation* μ_z transforms Q into a new quiver $Q' = \mu_z(Q)$ via a sequence of three steps. At the first step, for each pair of directed edges $x \to z \to y$ passing through z, we introduce a new edge

$x \to y$ (unless both x and y are frozen, in which case do nothing). At the second step, we reverse the direction of all edges incident to z. At the third step, we repeatedly remove oriented 2-cycles until unable to do so. See Figure 4. It is easy to check that mutating Q' at z' recovers Q.

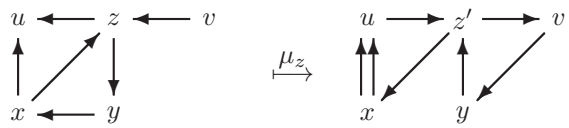


Figure 4. A quiver mutation. Vertices u and v are frozen.

Quiver mutation can be viewed as a generalization of the notion of a local move: there is a combinatorial rule associating a quiver with an arbitrary pseudoline arrangement so that local moves translate into quiver mutations. Rather than stating this rule precisely, we refer to Figure 5, and let the reader guess.

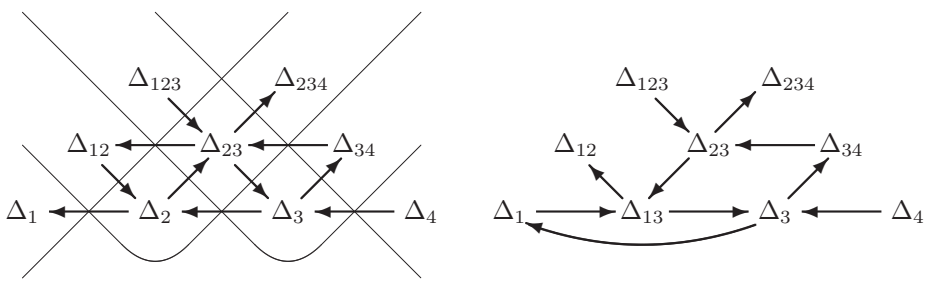


Figure 5. The quivers corresponding to the pseudoline arrangements shown in Figure 2. The chambers of an arrangement correspond to the vertices of the associated quiver.

Let us now define cluster exchanges using the language of quivers. This turns out to be very simple. Consider a quiver Q accompanied by an extended cluster $\mathbf{z}$, a finite collection of algebraically independent elements in our ambient field of rational functions $\mathcal{F}$. (Such a pair $(Q, \mathbf{z})$ is called a *seed*.) Pick a mutable vertex labeled by a cluster variable z. A *seed mutation* at z replaces $(Q, \mathbf{z})$ by the seed $(Q', \mathbf{z}')$ whose quiver is $Q' = \mu_z(Q)$ and whose extended cluster is $\mathbf{z}' = \mathbf{z} \cup \{z'\} \setminus \{z\}$; here the new cluster variable z' is determined by the *exchange relation*

$$z\, z' = \prod_{z \leftarrow y} y + \prod_{z \to y} y\,.$$

(The products are over the edges directed at/from z, respectively.) For example, the exchange relation associated with the quiver mutation shown in Figure 4 is $zz' = vx + uy$; applying mutation μ_x to the quiver on the right would invoke the exchange relation $xx' = z' + u^2$.

Following the blueprint outlined earlier, we now define a *cluster algebra* $\mathcal{A}(Q)$ associated to an arbitrary quiver Q. Assign a formal variable to each

vertex of Q; these variables form the initial extended cluster $\mathbf{z}$, and generate the ambient field $\mathcal{F}$. Starting with the initial seed $(Q, \mathbf{z})$, repeatedly apply seed mutations in all possible directions. The cluster algebra $\mathcal{A}(Q)$ is defined as the subring of $\mathcal{F}$ generated by all the elements of all extended clusters obtained by this recursive process.

Returning to our running example, we illustrate this definition by describing the cluster algebra structure in $\mathbb{C}[\mathrm{SL}_4/N]$. Let us start with the quiver shown on the left in Figure 5. We view the 9 variables Δ_I labeling the vertices of this quiver as formal indeterminates (secretly, they are chamber minors). We declare the variables Δ_2, Δ_3, and Δ_{23} mutable; the remaining six variables are frozen. There are three possible mutations out of this seed; we use the quiver to write the corresponding exchange relations:

$$\begin{aligned}
\Delta_2\Delta_{13} &= \Delta_{12}\,\Delta_3 + \Delta_1\,\Delta_{23}\,,\\
\Delta_3\,\Delta_{24} &= \Delta_4\,\Delta_{23} + \Delta_{34}\,\Delta_2\,,\\
\Delta_{23}\;\Omega &= \Delta_{123}\,\Delta_{34}\,\Delta_2 + \Delta_{12}\,\Delta_{234}\,\Delta_3\,.
\end{aligned}$$

At this point, these relations merely *define* Δ_{13}, Δ_{24}, and Ω as rational functions in the original extended cluster. The first two relations look familiar: they correspond to the two local moves that can be applied to the given pseudoline arrangement. The third relation is new: it enables us to flip the chamber minor Δ_{23}, something we could not do before. Although the resulting cluster does not correspond to a pseudoline arrangement, we can still determine its associated quiver using the definition of quiver mutation. Continuing this process recursively *ad infinitum* yields more and more extended clusters; taken together, they generate a cluster algebra.

If one interprets the elements of the initial cluster as actual flag minors, then the generators produced by this process become rational functions on the base affine space. Remarkably, all these generators are regular functions, and generate the ring of all such functions. This holds for any n, resulting in a cluster algebra structure in $\mathbb{C}[\mathrm{SL}_n/N]$; see, e.g., [3, 34, 43].

In the special case $n = 4$, this recursive process produces a *finite* number of distinct extended clusters, 14 of them to be exact. Altogether they contain 15 generators: in addition to the $2^4 - 2 = 14$ flag minors Δ_I, there is a single new cluster variable

$$\Omega = -\Delta_1\Delta_{234} + \Delta_2\Delta_{134}$$

that already appeared in the third exchange relation above.

Figure 6 shows the 14 clusters for $\mathbb{C}[\mathrm{SL}_4/N]$ as vertices of a planar graph; note that there is one additional vertex at infinity, so that the graph should be viewed as drawn on a sphere rather than a plane. The regions are labeled by cluster variables. Each cluster consists of the three elements labeling the regions adjacent to the corresponding vertex. The edges of the graph correspond to seed mutations. The 6 frozen variables are not shown.

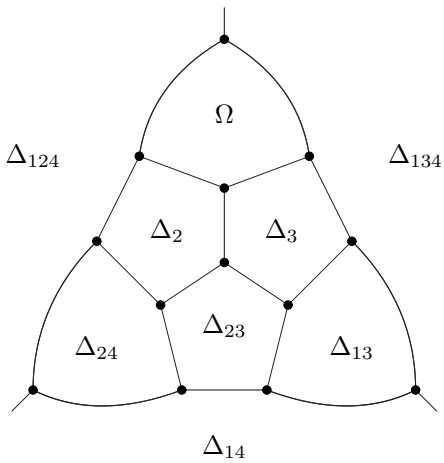


Figure 6. Clusters in $\mathbb{C}[\mathrm{SL}_4/N]$

What do we gain by introducing a cluster algebra structure into a commutative ring that already appears well understood? One reason has been given earlier: such a structure gives rise to a well-defined notion of the (totally) positive part of the associated algebraic variety. Another reason has to do with defining a "canonical basis" in the algebra at hand; the next paragraph hints at a possible approach.

Let us call two generators of a cluster algebra *compatible* if they appear together in some extended cluster. A *cluster monomial* is a product of pairwise compatible (not necessarily distinct) generators. It is not too hard to show that in the cluster algebra $\mathcal{A} = \mathbb{C}[\mathrm{SL}_4/N]$, the cluster monomials form a linear basis. This is a particular instance of the *dual canonical basis* of G. Lusztig (called the "upper global basis" by M. Kashiwara).

Unfortunately, the general picture (for arbitrary SL_n) is much more complicated: the cluster monomials seem to form just a part of the dual canonical (or dual semicanonical) basis; see [43]. The challenge of describing the rest of the dual canonical basis in concrete terms remains unmet.

Many other algebraic varieties of representation-theoretic importance turn out to possess a natural structure of a cluster algebra (hence the notions of positivity, cluster monomials, perhaps canonical bases, etc.). The list includes Grassmannians, flag manifolds, Schubert varieties, and double Bruhat cells in arbitrary semisimple Lie groups. See [22, 29, 34, 39, 43, 54, 55].

We conclude this section by mentioning some of the most basic structural results in the general theory of cluster algebras. The first such result is the *Laurent phenomenon*: the cluster variables are not merely rational functions in the elements of the initial extended cluster—all of them are in fact Laurent polynomials! We conjectured [26] that these Laurent polynomials always have

positive coefficients; many instances of this conjecture have been proved (see in particular [4, 14, 48, 50]) but the general case seems out of reach at the moment.

Another basic structural result is the classification [27] of the cluster algebras of *finite type*, i.e., those with finitely many seeds (equivalently, finitely many generators). In the generality presented here, the classification theorem states that a cluster algebra has finite type if and only if one of its seeds has a quiver whose subquiver formed by the mutable vertices is an orientation of a disjoint union of simply-laced Dynkin diagrams. (The full-blown version of the cluster theory leads to a complete analogue of the Cartan-Killing classification.)

The combinatorial scaffolding for a cluster algebra is provided by its *cluster complex*, a simplicial complex whose vertices are the cluster variables, and whose maximal simplices are the clusters. In the finite type case, this simplicial complex can be identified as the dual complex of a *generalized associahedron*, a remarkable convex polytope [6, 28] associated with the corresponding root system. In particular, the cluster complex of finite type is homeomorphic to a sphere. This can be observed in our running example of $\mathbb{C}[\mathrm{SL}_4/N]$: the cluster complex is the dual simplicial complex of the spherical cell complex shown in Figure 6.

3. Triangulations and Laminations

Cluster algebras owe much of their appeal to the ubiquity of the combinatorial and algebraic dynamics that underlies them. *A priori*, one might not expect the fairly rigid axioms governing quiver mutations and exchange relations to be satisfied in a large variety of contexts. Yet this is exactly what happens. Moreover, in each instance the framework of clusters and mutations seems to arise organically rather than artificially. A case in point is discussed in this section: the classical (by now) machinery of triangulations and laminations on bordered Riemann surfaces, which goes back to W. Thurston, can be naturally recast in the language of quiver mutations. The resulting connection between combinatorial topology and cluster theory is bound to benefit both.

This section is based on the papers [20, 21], which were in turn inspired by the work of V. Fock and A. Goncharov [16, 17], M. Gekhtman, M. Shapiro, and A. Vainshtein [32, 33], and R. Penner[51].

Let $\mathbf{S}$ be a connected oriented surface with boundary. (A few simple cases must be ruled out.) Fix a finite nonempty set $\mathbf{M}$ of *marked points* in the closure of $\mathbf{S}$. An *arc* in $(\mathbf{S}, \mathbf{M})$ is a non-selfintersecting curve in $\mathbf{S}$, considered up to isotopy, which connects two points in $\mathbf{M}$, does not pass through $\mathbf{M}$, and does not cut out an unpunctured monogon or digon. Arcs are *compatible* if they have non-intersecting realizations. Collections of pairwise compatible arcs are the simplices of the *arc complex* of $\mathbf{S}$. The facets of this simplicial complex

correspond to (ideal) *triangulations*. Note that these triangulations may contain *self-folded triangles*. See Figure 7.

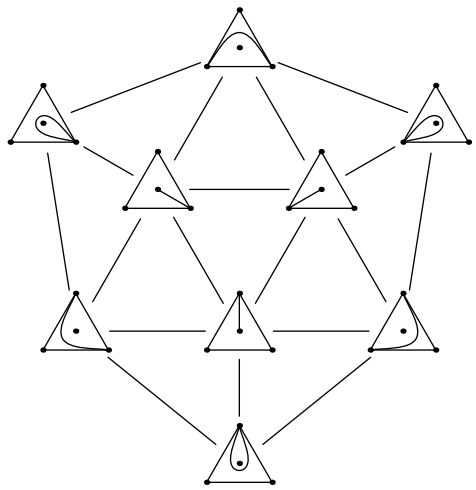


Figure 7. The arc complex of a once-punctured triangle. Its 10 two-dimensional simplices correspond to ideal triangulations. Among them, 6 contain self-folded triangles.

The vertices of the dual graph of the arc complex correspond to the triangulations; the edges in this graph correspond to *flips*. A flip replaces an arc in a triangulation by another (uniquely defined) arc. Note that an edge inside a self-folded triangle cannot be flipped. The situation is akin to pseudoline arrangements, which are likewise related to each other by flips (of a different kind).

This analogy goes much deeper than it might appear at first. To see that, we translate the setting into the *lingua franca* of quivers. Let us define the quiver $Q(T)$ associated to a triangulation T. The vertices of $Q(T)$ are labeled by the arcs in T. If two arcs belong to the same triangle, we connect the corresponding vertices of the quiver $Q(T)$ by an edge whose orientation is determined by the clockwise orientation of the boundary of the triangle. See Figure 8. For triangulations containing self-folded triangles, the definition is more complicated but is nevertheless completely elementary and explicit.

As the reader may have guessed by now, flips in ideal triangulations translate into mutations of the associated quivers. Furthermore, the quiver language suggests what we should do about the "forbidden" flips (of interior edges in self-folded triangles): forget about triangulations and just mutate the corresponding quivers.

It is easy to check that a quiver mutation corresponding to an edge inside a self-folded triangle transforms any quiver into an isomorphic one. Another simple observation is that the number of different (up to isomorphism) quivers $Q(T)$ associated to triangulations T of a given surface is *finite* (because the action of

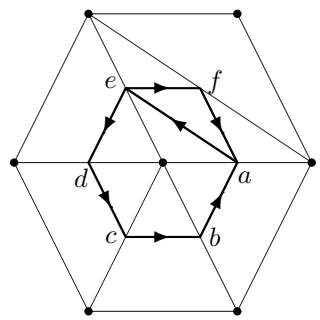


Figure 8. A triangulation T of a once-punctured hexagon and the associated quiver $Q(T)$.

the mapping class group on triangulations has finitely many orbits). Combining these two observations, one concludes that any quiver $Q(T)$ associated to a triangulated surface is of *finite mutation type*: its iterated mutations produce finitely many distinct (non-isomorphic) quivers. In fact, as shown in [15], all connected quivers of finite mutation type, with a few exceptions, are of the form $Q(T)$, for some triangulation T of some marked bordered surface $(\mathbf{S}, \mathbf{M})$. (We assume that there are no frozen vertices.) The complete list of exceptions consists of (a) quivers with two vertices and more than one edge, and (b) 11 quivers listed in [10].

The construction of quivers $Q(T)$ can be generalized by involving W. Thurston's machinery of laminations on Riemann surfaces. An integral (unbounded measured) *lamination* on $(\mathbf{S}, \mathbf{M})$ is a finite collection of non-selfintersecting and pairwise non-intersecting curves in $\mathbf{S}$, considered modulo isotopy. The curves in a lamination must satisfy certain constraints. In particular, each of them is either closed, or runs from boundary to boundary, or spirals into an interior marked point (a puncture). See Figure 9.

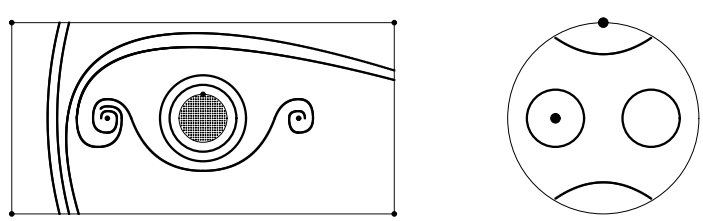


Figure 9. (a) A lamination; (b) curves not allowed in a lamination.

Let L be an integral lamination, and T a triangulation without self-folded triangles. For an arc γ in T, the *shear coordinate* $b_\gamma(T, L)$ is the signed number of curves in L which intersect γ and in doing so, connect the opposite sides of

the quadrilateral surrounding γ. The sign depends on which pair of opposite sides the curves connect; see Figure 10.

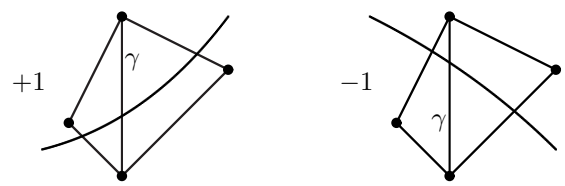


Figure 10. A (signed) contribution of a curve in L to the shear coordinate $b_\gamma(T, L)$.

By a theorem of W. Thurston, the shear coordinates *coordinatize* integral laminations in the following sense: for a fixed triangulation T, the map

$$L \mapsto (b_\gamma(T, L))_{\gamma \in T}$$

is a bijection between integral laminations and $\mathbb{Z}^n$.

A *multi-lamination* $\mathbf{L}$ on $(\mathbf{S}, \mathbf{M})$ is an arbitrary finite family of laminations. Given such $\mathbf{L}$ and a triangulation T of the surface $(\mathbf{S}, \mathbf{M})$, we construct the "extended" quiver $Q(T, \mathbf{L})$ by adding vertices and oriented edges to $Q(T)$ as follows. For each lamination L in $\mathbf{L}$, we introduce a new vertex labeled by L. We then connect this vertex to each vertex in Q, say labeled by an arc γ, by $|b_\gamma(T, L)|$ edges whose direction is determined by the sign of $b_\gamma(T, L)$. See Figure 11.

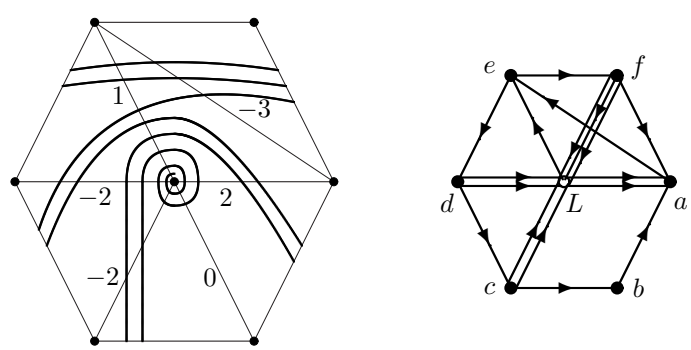


Figure 11. (a) Shear coordinates of a lamination L; (b) the quiver $Q(T, \{L\})$.

Amazingly, the same property as before holds: for a fixed multi-lamination $\mathbf{L}$, a flip in a triangulation T translates into the corresponding mutation in the quiver $Q(T, \mathbf{L})$. (The definition of the latter can be generalized to allow for self-folded triangles.) This strongly suggests the existence of a cluster algebra structure associated with any given marked surface $(\mathbf{S}, \mathbf{M})$ and any multi-lamination $\mathbf{L}$ on it.

This class of cluster algebras can be understood on several levels. On the combinatorial level, the cluster complex of such an algebra can be explicitly described in terms of *tagged arcs*, which are ordinary arcs adorned with very simple combinatorial decorations. This description represents the cluster complex as a finite covering space for the arc complex. The cluster complex turns out to be either contractible or homotopy equivalent to a sphere. Unlike the generalized associahedra mentioned above, these cluster complexes are usually not compact; moreover, with a few exceptions, they exhibit exponential growth. See [20].

The coordinatization theorem implies that any quiver Q whose mutable part can be interpreted as a quiver $Q(T)$ corresponding to a triangulation T of some marked surface $(\mathbf{S}, \mathbf{M})$, there exists a (unique) multi-lamination $\mathbf{L}$ on $(\mathbf{S}, \mathbf{M})$ such that $Q = Q(T, \mathbf{L})$. In view of the discussion above, the cluster algebra $\mathcal{A}(Q)$ associated with such a quiver Q depends only on $(\mathbf{S}, \mathbf{M})$ and $\mathbf{L}$ but not on the triangulation T. Consequently, one should be able to understand this cluster algebra in terms of the topology of the surface $(\mathbf{S}, \mathbf{M})$ and the multi-lamination $\mathbf{L}$.

We illustrate this construction by returning, once again, to the example of the cluster algebra $\mathcal{A} = \mathbb{C}[\mathrm{SL}_4 / N]$. The mutable part of any quiver Q defining this algebra (see, e.g., Figure 5) has 3 vertices, and is isomorphic to a quiver $Q(T)$ associated to a triangulation of a hexagon, i.e., a disk with 6 marked points on the boundary. Thus, we can let $(\mathbf{S}, \mathbf{M})$ be a hexagon. Due to the absence of marked points in the interior of $\mathbf{S}$, the construction simplifies considerably: there are no self-folded triangles, and the cluster complex coincides with the arc complex. The underlying combinatorics of $\mathcal{A}$ is thus modeled as follows: cluster variables correspond to arcs (that is, the diagonals of the hexagon), clusters correspond to triangulations, and exchanges correspond to flips. It remains to determine the appropriate multi-lamination $\mathbf{L}$. This is done by interpreting the multiplicities of edges connecting the frozen vertices in Q to the mutable ones as shear coordinates of laminations, and then constructing the unique laminations having those shear coordinates. The result is shown in Figure 12.

It is natural to ask whether cluster variables in the cluster algebra associated with a multi-lamination on a bordered surface can be given an intrinsic geometric interpretation. The answer is yes: each cluster variable can be viewed as a suitably renormalized *lambda length* [51] (a.k.a. Penner coordinate) of the corresponding (tagged) arc. For a given arc, such a lambda length is a real function on (an appropriate generalization of) the *decorated Teichmüller space* for $(\mathbf{S}, \mathbf{M})$; see [21] for further details. Thus in this geometric realization, the decorated Teichmüller space plays the role of the corresponding totally positive variety. This brings us back full circle to the problems discussed at the end of Section 1, namely to the challenges of understanding the stratification of a totally nonnegative variety (in this case, a compactified decorated Teichmüller space) and the singularities of its boundary.

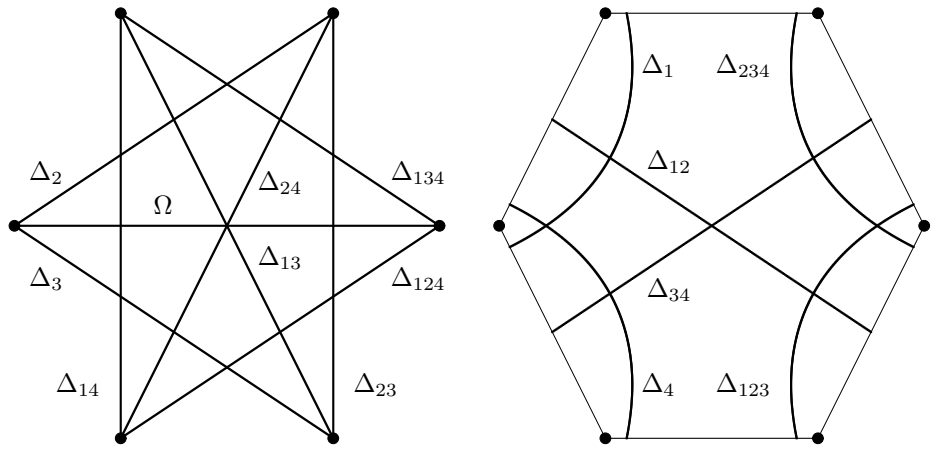


Figure 12. (a) Labeling the cluster variables in $\mathbb{C}[\mathrm{SL}_4/N]$ by the diagonals of a hexagon. (b) Labeling the frozen variables by laminations, each consisting of a single curve.

References

[1] T. Ando, Totally positive matrices, *Linear Algebra Appl.* **90** (1987), 165–219.

[2] A. Berenstein, S. Fomin, and A. Zelevinsky, Parametrizations of canonical bases and totally positive matrices, *Adv. Math.* **122** (1996), 49–149.

[3] A. Berenstein, S. Fomin and A. Zelevinsky, Cluster algebras III: Upper bounds and double Bruhat cells, *Duke Math. J.* **126** (2005), 1–52.

[4] G. Carroll and D. Speyer, The cube recurrence, *Electron. J. Combin.* **11** (2004), no. 1, Research Paper 73, 31 pp.

[5] R. W. Carter, *Cluster algebras*. Textos de Matemática, Série B, 37. Universidade de Coimbra, 2006.

[6] F. Chapoton, S. Fomin, and A. Zelevinsky, Polytopal realizations of generalized associahedra, *Canad. Math. Bull.* **45** (2002), 537–566.

[7] L. O. Chekhov, Orbifold Riemann surfaces and geodesic algebras, *J. Phys. A* **42** (2009), no. 30, 304007, 32 pp.

[8] C. Cryer, The *LU*-factorization of totally positive matrices, *Linear Algebra Appl.* **7** (1973), 83–92.

[9] C. Cryer, Some properties of totally positive matrices, *Linear Algebra Appl.* **15** (1976), 1–25.

[10] H. Derksen and T. Owen, New graphs of finite mutation type, *Electron. J. Combin.* **15** (2008), no. 1, Research Paper 139, 15 pp.

[11] H. Derksen, J. Weyman, and A. Zelevinsky, Quivers with potentials and their representations I: Mutations, *Selecta Math. (N.S.)* **14** (2008), 59–119.

[12] H. Derksen, J. Weyman, and A. Zelevinsky, Quivers with potentials and their representations II: Applications to cluster algebras, to appear in *J. Amer. Math. Soc.*, `arXiv:0904.0676`.

[13] P. Di Francesco and R. Kedem, Q-systems as cluster algebras. II. Cartan matrix of finite type and the polynomial property, *Lett. Math. Phys.* **89** (2009), 183–216.

[14] P. Di Francesco and R. Kedem, Q-systems, heaps, paths and cluster positivity, *Comm. Math. Phys.* **293** (2010), 727–802.

[15] A. Felikson, M. Shapiro, and P. Tumarkin, Skew-symmetric cluster algebras of finite mutation type, `arXiv:0811.1703`.

[16] V. V. Fock and A. B. Goncharov, Moduli spaces of local systems and higher Teichmüller theory, *Publ. Math. Inst. Hautes Études Sci.* (2006), no. 103, 1–211.

[17] V. V. Fock and A. B. Goncharov, Dual Teichmuller and lamination spaces, *Handbook of Teichmüller theory, vol. I,* 647–684, Eur. Math. Soc., Zürich, 2007.

[18] S. Fomin, Cluster Algebras Portal, `http://www.math.lsa.umich.edu/~fomin/cluster.html`.

[19] S. Fomin and N. Reading, Root systems and generalized associahedra, *Geometric Combinatorics (Park City, UT, 2003),* 63–131, IAS/Park City Math. Ser., 14, Amer. Math. Soc., Providence, RI, 2007.

[20] S. Fomin, M. Shapiro, and D. Thurston, Cluster algebras and triangulated surfaces. Part I: Cluster complexes, *Acta Math.* **201** (2008), 83–146.

[21] S. Fomin and D. Thurston, Cluster algebras and triangulated surfaces. Part II: Lambda lengths, preprint.

[22] S. Fomin and A. Zelevinsky, Double Bruhat cells and total positivity, *J. Amer. Math. Soc.* **12** (1999), 335–380.

[23] S. Fomin and M. Z. Shapiro, Stratified spaces formed by totally positive varieties, *Michigan Math. J.* **48** (2000), 253–270.

[24] S. Fomin and A. Zelevinsky, Totally nonnegative and oscillatory elements in semisimple groups, *Proc. Amer. Math. Soc.* **128** (2000), 3749–3759.

[25] S. Fomin and A. Zelevinsky, Total positivity: tests and parametrizations, *Math. Intelligencer* **22** (2000), 23–33.

[26] S. Fomin and A. Zelevinsky, Cluster algebras I: Foundations, *J. Amer. Math. Soc.* **15** (2002), 497–529.

[27] S. Fomin and A. Zelevinsky, Cluster algebras II: Finite type classification, *Invent. Math.* **154** (2003), 63–121.

[28] S. Fomin and A. Zelevinsky, Y-systems and generalized associahedra, *Ann. of Math.* **158** (2003), 977–1018.

[29] S. Fomin and A. Zelevinsky, Cluster algebras: Notes for the CDM-03 conference, *Current Developments in Mathematics, 2003*, 1–34, Int. Press, 2004.

[30] S. Fomin and A. Zelevinsky, Cluster algebras IV: Coefficients, *Compos. Math.* **143** (2007), 112–164.

[31] F. R. Gantmacher and M. G. Krein, *Oscillation matrices and kernels and small vibrations of mechanical systems*, AMS Chelsea Publishing, Providence, RI, 2002. (Original Russian edition, 1941.)

[32] M. Gekhtman, M. Shapiro, and A. Vainshtein, Cluster algebras and Poisson geometry, *Mosc. Math. J.* **3** (2003), 899–934.

[33] M. Gekhtman, M. Shapiro, and A. Vainshtein, Cluster algebras and Weil-Petersson forms, *Duke Math. J.* **127** (2005), 291–311.

[34] C. Geiss, B. Leclerc, and J. Schröer, Preprojective algebras and cluster algebras, *in:* Trends in representation theory of algebras and related topics, 253–283, *EMS Ser. Congr. Rep.*, Eur. Math. Soc., Zrich, 2008.

[35] *Total positivity and its applications*, M. Gasca and C. A. Micchelli (Eds.), *Mathematics and its Applications* **359**, Kluwer Academic Publishers, Dordrecht, 1996.

[36] P. Hersh, Regular cell complexes in total positivity, `arXiv:0711.1348`.

[37] R. Inoue, O. Iyama, A. Kuniba, T. Nakanishi, and J. Suzuki, Periodicities of T-systems and Y-systems, to appear in *Nagoya Math. J.*, `arXiv:0812.0667`.

[38] S. Karlin, *Total positivity*, Stanford University Press, 1968.

[39] B. Keller, Algèbres amassées et applications, *Séminaire Bourbaki, 2009/2010, exposé 1014*, `arXiv:0911.2903`.

[40] B. Keller, Categorification of acyclic cluster algebras: an introduction, to appear in: *Proceedings of the conference "Higher structures in Geometry and Physics 2007,"* Birkhäuser.

[41] B. Keller, Cluster algebras, quiver representations and triangulated categories, to appear in: *Triangulated categories,* London Math. Soc., 2010.

[42] M. Kontsevich and Y. Soibelman, Stability structures, motivic Donaldson-Thomas invariants and cluster transformations, `arXiv:0811.2435`, November 2008.

[43] B. Leclerc, Cluster algebras and representation theory, *Proceedings of the International Congress of Mathematicians*, Hyderabad, India, 2010.

[44] P. Littelmann, Bases canoniques et applications, *Séminaire Bourbaki, 1997/1998, exposé 847*; *Astérisque* **252** (1998), 287–306.

[45] C. Loewner, On totally positive matrices, *Math. Z.* **63** (1955), 338–340.

[46] G. Lusztig, Total positivity in reductive groups, in: *Lie theory and geometry: in honor of Bertram Kostant, Progress in Mathematics* **123**, Birkhäuser, 1994, 531–568.

[47] G. Lusztig, Introduction to total positivity, in: *Positivity in Lie theory: open problems, de Gruyter Exp. Math.* **26**, de Gruyter, Berlin, 1998, 133–145.

[48] G. Musiker, R. Schiffler, and L. Williams, Positivity for cluster algebras from surfaces, `arXiv:0906.0748`.

[49] K. Nagao, Donaldson-Thomas theory and cluster algebras, `arXiv:1002.4884`, February 2010.

[50] H. Nakajima, Quiver varieties and cluster algebras, `arXiv:0905.0002`, May 2009.

[51] R. C. Penner, Lambda lengths, University of Aarhus, lecture notes, August 2006, `http://www.ctqm.au.dk/research/MCS/lambdalengths.pdf`.

[52] N. Reading and D. Speyer, Cambrian fans, *J. Eur. Math. Soc.* **11** (2009), 407–447.

[53] A. M. Whitney, A reduction theorem for totally positive matrices, *J. d'Analyse Math.* **2** (1952), 88–92.

[54] A. Zelevinsky, Cluster algebras: origins, results and conjectures. *Advances in algebra towards millennium problems*, 85–105, SAS Int. Publ., Delhi, 2005.

[55] A. Zelevinsky, From Littlewood-Richardson coefficients to cluster algebras in three lectures, in: *Symmetric functions 2001: surveys of developments and perspectives*, edited by S. Fomin, 253–273, NATO Sci. Ser. II Math. Phys. Chem., 74, Kluwer Acad. Publ., Dordrecht, 2002.

[56] A. Zelevinsky, What is ... a cluster algebra? *Notices Amer. Math. Soc.* **54** (2007), no. 11, 1494–1495.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Canonical Dimension

Nikita A. Karpenko*

Abstract

Canonical dimension is an integral-valued invariant of algebraic structures. We are mostly interested in understanding the canonical dimension of projective homogeneous varieties under semisimple affine algebraic groups over arbitrary fields. Known methods, results, applications, and open problems are reviewed, some new ones are provided.

Mathematics Subject Classification (2010). Primary 14L17; Secondary 14C25.

Keywords. Algebraic groups, projective homogeneous varieties, Chow groups and motives.

0. Introduction

A smooth projective variety X is *incompressible*, if any rational map $X \dashrightarrow X$ is dominant. *Canonical dimension* $\operatorname{cdim} X$, an invariant measuring the level of compressibility of X, is the minimum of the dimension of the image of a rational map $X \dashrightarrow X$. Formally introduced by G. Berhuy and Z. Reichstein only in 2005, [3], this invariant has been implicitly studied for a long time before. For instance, an old question of M. Knebusch, [19, Question 4.13], answered in Example 1.5, was about the canonical dimension of a quadric. Also the incompressibility of the Severi-Brauer variety of a primary division algebra – see Example 2.3 – has been known and intensively applied since 1995.

In this talk we look at the canonical dimension of a projective homogeneous variety X, mainly, through the motive of X. This approach is justified by Theorem 5.1.

*Supported by the Max-Planck-Institut für Mathematik in Bonn
UPMC Univ Paris 06, Institut de Mathématiques de Jussieu, F-75252 Paris, France, `www.math.jussieu.fr/~karpenko`. E-mail: karpenko *at* math.jussieu.fr.

1. Definitions of Canonical Dimension

By *variety* we mean an *integral* separated scheme of finite type over a field.

Since we are mainly interested in canonical dimension of projective homogeneous varieties, we define it for smooth projective varieties only. We refer to [25] for the case of a more general variety.

Let X be a smooth projective variety over a field F.

Definition 1.1. *Canonical dimension* $\operatorname{cdim} X$ of X is the minimum of $\dim Y$, where Y runs over the closed subvarieties of X admitting a rational map $X \dashrightarrow Y$. Equivalently, Y runs over the closed subvarieties of X such that the scheme $Y_{F(X)}$ has a rational point.

Of course, $\operatorname{cdim} X = 0$ if X has a rational point. We are basically interested in varieties without rational points.

In general, $\operatorname{cdim} X$ is an integer satisfying

$$0 \leq \operatorname{cdim} X \leq \dim X.$$

Let p be a positive prime integer. We write Ch for the Chow group [7, §57] with coefficients in $\mathbb{F}_p$, the finite field of p elements. By a *correspondence* $X \rightsquigarrow Y$ we mean an element of the Chow group $\operatorname{Ch}_{\dim X}(X \times Y)$. The *multiplicity* $\operatorname{mult} \alpha \in \mathbb{F}_p$ of a correspondence $\alpha : X \rightsquigarrow Y$ (also called *degree* in the literature) is its image under the push-forward homomorphism

$$\operatorname{Ch}_{\dim X}(X \times Y) \to \operatorname{Ch}_{\dim X}(X) = \mathbb{F}_p$$

with respect to the projection $X \times Y \to X$. Finally, a 0-*cycle class* is an element of $\operatorname{Ch}_0(X)$, its *degree* is therefore an element of $\operatorname{Ch}_0(\operatorname{Spec} F) = \mathbb{F}_p$.

Our actual subject of study is the *canonical p-dimension*, a p-local version of the above notion, defined as follows:

Definition 1.2. *Canonical p-dimension* $\operatorname{cdim}_p X$ of X is the minimum of $\dim Y$, where Y runs over the closed subvarieties of X admitting a multiplicity 1 correspondence $X \rightsquigarrow Y$. Equivalently, Y runs over the closed subvarieties of X such that the scheme $Y_{F(X)}$ has a 0-cycle class of degree 1.

Of course, $\operatorname{cdim}_p X = 0$ if X has a 0-cycle class of degree 1. We are basically interested in varieties without 0-cycle classes of degree 1, that is, varieties where the degree of each closed point is divisible by p.

In general, $\operatorname{cdim}_p X$ is an integer satisfying

$$0 \leq \operatorname{cdim}_p X \leq \operatorname{cdim} X.$$

There are at least two more definitions of the canonical (p-)dimension looking quite differently. We refer to [25] for a proof that they are equivalent to the initial one. We start by the definition via the *essential dimension*. We refer to [25, §1.1] for the definition of the essential (p-)dimension of an arbitrary functor $\mathbf{Fields}_F \to \mathbf{Sets}$ of the category of the field extensions of F to the category of sets.

Definition 1.3. Let $\mathcal{F}_X : \mathbf{Fields}_F \to \mathbf{Sets}$ be the functor defined by the formulas $\mathcal{F}_X(L) = \emptyset$ if $X(L) = \emptyset$ and $\mathcal{F}_X(L) = \{L\}$ (a singleton) otherwise. We define $\operatorname{cdim} X$ as the essential dimension of the functor $\mathcal{F}_X$, and we define $\operatorname{cdim}_p X$ as its essential p-dimension.

We come to the last definition. It makes use of the notion of a *generic splitting field* of a variety. We say that a field L/F is a *splitting field* (or *isotropy field*) of X is $X(L) \neq \emptyset$. A splitting field E/F is *generic*, if for each splitting field L/F of X there exists an F-place $E \dashrightarrow L$. A splitting field E/F is *p-generic*, if for each splitting field L/F of X there exist a finite field extension L'/L of a p-prime degree and an F-place $E \dashrightarrow L'$. Of course, any generic splitting field is also p-generic (for any p); the function field $F(X)$ is a generic splitting field.

Definition 1.4. We define the canonical (p-)dimension of X as the minimum of the transcendence degree of a (p-)generic splitting field of X.

The last definition (as well as the previous one) naturally generalizes to the case of an arbitrary "algebraic structure" A in place of X as soon as we have a notion of a *splitting field* for A. We consider two examples of such a generalization. (However, one easily comes back to varieties in both examples.)

Example 1.5. Let φ be a finite-dimensional non-degenerate quadratic form over F. A field L/F is a splitting field (or isotropy field) of φ if the quadratic form φ_L has a non-trivial zero. This way we get the notion of the canonical (p-)dimension of φ. Let X be the projective quadric of φ. We have $\operatorname{cdim} \varphi = \operatorname{cdim} X$ and $\operatorname{cdim}_p \varphi = \operatorname{cdim}_p X$, because a splitting field of φ is the same as a splitting field of X. These invariants are computed. If $X(F) = \emptyset$, i.e., if the quadric X is *anisotropic*, then we have $\operatorname{cdim} \varphi = \operatorname{cdim}_2 \varphi = \dim X - \mathfrak{i}_1 + 1$, where $\mathfrak{i}_1$ is the *first Witt index* of φ, [7, Theorem 90.2]. (Of course, $\operatorname{cdim}_p \varphi = 0$ for $p \neq 2$.)

Example 1.6. Let A be a finite p-subgroup of the Brauer group $\operatorname{Br} F$ of F. A field L/F is a splitting field of A if $A_L = 0$, i.e., if A vanishes under the change of field homomorphism $\operatorname{Br} F \to \operatorname{Br} L$. We get the notion of the canonical (p-) dimension of A. Let $A_1, \dots, A_n$ be central simple F-algebras such that their classes are in A and generate A; let X be the direct product of the corresponding Severi-Brauer varieties. We have $\operatorname{cdim} A = \operatorname{cdim} X$ and $\operatorname{cdim}_p A = \operatorname{cdim}_p X$ (for any X obtained this way), because a splitting field of A is the same as a splitting field of X. These invariants are computed as $\operatorname{cdim} A = \operatorname{cdim}_p A = \min \dim X$, [18, §2]. (Of course, $\operatorname{cdim}_{p'} A = 0$ for $p' \neq p$.)

The result of Example 1.6 has numerous applications. Many of them compute the essential dimension of algebraic groups as the following series of papers on finite p-groups. It was initiated by M. Florence who used the case of *cyclic* A to compute the essential dimension of a finite *cyclic* p-group in [8]. Arbitrary finite constant p-groups have been treated later on in [18]. Finally, the case of an arbitrary finite p-group (as well as the case of an algebraic tori), still essentially

using Example 1.6, has been recently done by R. Lötscher, M. Macdonald, A. Meyer, and Z. Reichstein, [22].

Here is an example of a class of projective homogeneous varieties for which the canonical p-dimension is computed in terms of their Chow groups. These are the *generically split* projective homogeneous varieties. A projective homogeneous variety X is *generically split*, if the $F(X)$-variety $X_{F(X)}$ is cellular.

Example 1.7 ([17, Theorem 5.8]). Let X be a generically split projective homogeneous variety and let $\bar{X} := X_{\bar{F}}$ with an algebraic closure $\bar{F}$ of F. The canonical p-dimension $\operatorname{cdim}_p X$ coincides with the minimal integer i such that the change of field homomorphism $\operatorname{Ch}_i(X) \to \operatorname{Ch}_i(\bar{X})$ is non-zero.

Let G be a *split* simple affine algebraic group, T a *generic* G-torsor, B a Borel subgroup of G. Using the result of Example 1.7, the canonical dimension of the (generically split) projective homogeneous variety T/B is determined: the case of a classical G is done in [17], the case of an exceptional G in [28].

Example 1.8. Let n be a positive integer and X be the variety of n-dimensional totally isotropic subspaces of a $2n+1$-dimensional non-degenerate quadratic form φ. The variety X is homogeneous and generically split. Its canonical (2-) dimension is the canonical (2-)dimension of φ if defining the splitting fields of φ we require that φ becomes completely split (i.e., almost hyperbolic). The canonical 2-dimension of X is known, [7, Theorem 90.3]; $\operatorname{cdim} X$, however, is not known in general. It is conjectured in [27, Conjecture 6.6] that $\operatorname{cdim} X = \operatorname{cdim}_2 X$.

2. Incompressible Varieties

A smooth projective variety X is *incompressible*, if $\operatorname{cdim} X = \dim X$; X is *p-incompressible*, if $\operatorname{cdim}_p X = \dim X$. Equivalently, X is incompressible if any rational map $X \dashrightarrow X$ is dominant, that is, no proper closed subset $Y \subset X$ admits a rational map $X \dashrightarrow Y$; X is p-incompressible, if no proper closed subset $Y \subset X$ admits a degree 1 correspondence $X \rightsquigarrow Y$.

Of course, any p-incompressible variety (for some p) is incompressible. An example of an incompressible and p-compressible (for any p) projective homogeneous variety is obtained in [21] with a help of the birational classification of geometrically rational surfaces:

Example 2.1. Let X_1 be the Severi-Brauer variety of a *quaternion* (i.e., *degree* 2 central) division algebra and let X_2 be the Severi-Brauer variety of a *degree* 3 central division algebra. The (projective homogeneous, 3-dimensional) variety $X := X_1 \times X_2$ is incompressible. However, $\operatorname{cdim}_2 X = \operatorname{cdim}_2 X_1 = \dim X_1 = 1$ and $\operatorname{cdim}_3 X = \operatorname{cdim}_2 X_2 = \dim X_2 = 2$ (and $\operatorname{cdim}_p X = 0$ for any other p).

An important source of p-incompressible varieties is Proposition 2.2 below which is a consequence of the A. Merkurjev *degree formula* [24, Theorem 6.4], a generalization of the M. Rost degree formula.

For any sequence $R = (r_1, r_2, \dots)$ of non-negative almost all zero integers r_i, a homogeneous integral polynomial $T_R \in \mathbb{Z}[\sigma_1, \sigma_2, \dots]$ in variables $\sigma_1, \sigma_2, \dots$ of degree $|R| := \sum_{i\geq 1} r_i(p^i - 1)$ is defined in [24, §4], where for any $i \geq 1$ the degree of the variable σ_i is defined as i. (The polynomial T_R also depends on the prime p which we have fixed before.) For any smooth projective variety X of dimension $|R|$, the characteristic number $R(X)$ is defined as $R(X) := \deg c_R(-T_X) \in \mathbb{Z}$, where c_R is the characteristic class $c_R := T_R(c_1, c_2, \dots)$ (the polynomial T_R evaluated on the Chern classes $c_1, c_2, \dots$) and T_X (which has nothing to do with T_R) is the tangent bundle of X.

For any integer n, we write $v_p(n)$ for the value on n of the p-adic valuation. For any F-scheme X, we write $v_p(X)$ for the value of the p-adic valuation on the greatest common divisor of the degrees of the closed points on X.

Clearly, $v_p(R(X)) \geq v_p(X)$ for any R. A smooth projective variety X is *p-rigid*, if $v_p(R(X)) = v_p(X)$ for some R.

A smooth projective variety X is *strongly p-incompressible*, if for any projective variety Y with $v_p(Y) \geq v_p(X)$, $\dim Y \leq \dim X$, and a multiplicity 1 correspondence $X \rightsquigarrow Y$, one has: $\dim Y = \dim X$ (in particular, any strongly p-incompressible variety is p-incompressible) and there also exists a multiplicity 1 correspondence $Y \rightsquigarrow X$.

Proposition 2.2 ([24, Theorem 7.2]). *Assume that* $\operatorname{char} F \neq p$. *Then any p-rigid F-variety is strongly p-incompressible.*

For any projective scheme X and any positive integer $l \leq v_p(X)$, we define a homomorphism $\deg/p^l : \mathrm{Ch}_0(X) \to \mathbb{F}_p$ associating to the class $[x] \in \mathrm{Ch}_0(X)$ of a closed points $x \in X$ the class in $\mathbb{F}_p$ of the integer $(\deg x)/p^l$. Of course, $\deg/p^l = 0$ for $l < v_p(X)$. For any morphism $f : X \to Y$ of projective schemes X and Y and any $l \leq \min\{v_p(X), v_p(Y)\}$, the push-forward homomorphism $f_* : \mathrm{Ch}_0(X) \to \mathrm{Ch}_0(Y)$ satisfies $(\deg/p^l) \circ f_* = \deg/p^l$.

Since $\operatorname{char} F \neq p$, any sequence R as above determines certain degree $|R|$ homological operation S_R on the (modulo p) Chow group Ch, [24, §5]. This means that for any projective (not necessarily smooth) F-scheme Z, we are given a degree $-|R|$ homogeneous group homomorphism

$$S_R^Z : \mathrm{Ch}_*(Z) \to \mathrm{Ch}_{*-|R|}(Z)$$

commuting with the push-forward homomorphisms and such that

$$S_R^Z([Z]) = c_R(-T_Z) \mod p$$

if Z is smooth.

Proof of Proposition 2.2. Let X be a p-rigid variety and let R be a sequence such that $v_p(R(X)) = v_p(X)$. For checking the strong p-incompressibility of X, let us take a projective variety Y with $v_p(Y) \geq v_p(X)$, $\dim Y \leq \dim X$, and a multiplicity 1 correspondence $X \rightsquigarrow Y$. Then there exists a closed subvariety

$Z \subset X \times Y$ such that the degree $\deg pr_X \in \mathbb{F}_p$ of the projection $pr_X : Z \to X$ is non-zero. The proof plays with the following commutative diagram:

$$\begin{tikzcd}
 & \mathrm{Ch}_{|R|}(Z) \arrow[dl, "(pr_X)_*"'] \arrow[dr, "(pr_Y)_*"] \arrow[dd, "S_R^Z"] & \\
\mathrm{Ch}_{|R|}(X) \arrow[dd, "S_R^X"'] & & \mathrm{Ch}_{|R|}(Y) \arrow[dd, "S_R^Y"] \\
 & \mathrm{Ch}_0(Z) \arrow[dl, "(pr_X)_*"'] \arrow[dr, "(pr_Y)_*"] \arrow[dd, "\deg/p^l"] & \\
\mathrm{Ch}_0(X) \arrow[dr, "\deg/p^l"'] & & \mathrm{Ch}_0(Y) \arrow[dl, "\deg/p^l"] \\
 & \mathbb{F}_p &
\end{tikzcd}$$

Since the operation S_R commutes with the push-forward $(pr_X)_*$ and $(pr_X)_*([Z]) = (\deg pr_X) \cdot [Z]$, we have $(\deg/p^l)\big(S_R^Z([Z])\big) = (\deg pr_X) \cdot (\deg/p^l)\big(S_R^X([X])\big) \neq 0$, where $l = v_p(X)$. Since the operation S_R also commutes with the push-forward with respect to the projection $pr_Y : Z \to Y$, we have $(\deg/p^l)\big(S_R^Z([Z])\big) = (\deg/p^l)\big(S_R^Y \circ (pr_Y)_*([Z])\big)$. It follows that $(pr_Y)_*([Z]) \neq 0$, that is, $\dim Y = \dim Z$ $(= \dim X)$ and $\deg pr_Y \neq 0$. Therefore the the class in $\mathrm{Ch}_{\dim Y}(Y \times X)$ of the transposition of Z is a required correspondence $Y \rightsquigarrow X$. $\square$

Certainly, the strength of the above approach to the p-incompressibility is in the fact that it gives a stronger property – the strong p-incompressibility. Moreover, if X is a p-rigid variety, then for any field extension L/F, any twisted form X'/L of X with $v_p(X') = v_p(X)$ is also p-rigid. Therefore we get the p-incompressibility not only for X, but also for any such X'. Sometimes, however, this is too much, becoming a weakness of the approach: it cannot possibly succeed for a variety possessing a p-compressible twisted form with the same v_p. Besides that, the approach does not exist in characteristic p at all because a construction of the operations on the Chow group modulo p is not available in characteristic p.

Example 2.3. Let n be a positive integer and let D be a central division F-algebra of degree p^n. The Severi-Brauer variety X of D is p-rigid, [24, §7.2]. Therefore, if $\operatorname{char} F \neq p$, the variety X is strongly p-incompressible. Consequently, X is p-incompressible. (This is the particular case of Example 1.6 with cyclic A and $\operatorname{char} F \neq p$.) For F with $\operatorname{char} F = p$, it is not known whether X is strongly p-incompressible. The general case of Example 1.6 (even with the characteristic p excluded) cannot be done by the degree formula method. For instance, the product of two non-isomorphic anisotropic

conics possessing a common quadratic splitting field is 2-incompressible but not 2-rigid (and even not strongly 2-incompressible). In general, a product $X = X_1 \times \cdots \times X_n$ of arbitrary smooth projective varieties $X_1, \ldots, X_n$ can be p-rigid only if $v_p(X) = v_p(X_1) + \cdots + v_p(X_n)$.

Example 2.4 ([13]). Here is another proof of the p-incompressibility for the variety X of Example 2.3, which works for F of arbitrary characteristic and which also works in the general case of Example 1.6. Using the computation of $K_0(X)$ and the relationship between $K_0(X)$ and $\mathrm{Ch}(X)$, one shows that the image of $\mathrm{Ch}(X) \to \mathrm{Ch}(\bar{X})$ is generated by the class of X. Since the variety X is projective homogeneous and generically split, it follows by Example 1.7 that X is p-incompressible.

Example 2.5. An immediate consequence of the above result concerns an orthogonal involution σ on a central division F-algebra D. An F-linear involution – a self-inverse anti-automorphism of the algebra D – is *orthogonal*, if the induced involution on the split algebra $D_{F(X)} \simeq \mathrm{End}(V)$ is adjoint to a non-alternating bilinear form b on the vector space V. Possessing an involution, D has to be 2-primary, so that we have the incompressibility statement which implies that b is anisotropic, or, equivalently, that $\sigma_{F(X)}$ is anisotropic. Indeed, otherwise the proper closed subvariety $Y \subset X$ of the *isotropic* ideals in D (i.e., ideals $I \subset D$ with $\sigma(I) \cdot I = 0$) would have an $F(X)$-point. Note that in contrast to the original paper [15], containing this observation, we do not exclude the case of characteristic 2 here. Moreover, we can replace the involution by a *quadratic pair*, [20, Definition 5.4]; the conclusion obtained this way differs from the previous one in characteristic 2 (and coincides with it in characteristic $\neq 2$).

Example 2.6 (cf. Example 1.5). Let X be an anisotropic smooth projective quadric of the first Witt index 1. Then X is strongly 2-incompressible, [7, Theorem 76.1]. The degree formula approach works only if $\dim X + 1$ is a 2-power: otherwise, X has a 2-compressible twisted form X' (another quadric) with $v_2(X') = v_2(X) = 1$ so that the degree formula approach cannot possibly work.

We terminate this Section by a criterion of p-incompressibility in terms of the correspondence multiplicities:

Lemma 2.7. *A projective homogeneous variety X is p-incompressible if and only if* $\mathrm{mult}\,\alpha = \mathrm{mult}\,\alpha^t$ *for any correspondence $\alpha : X \rightsquigarrow X$, where α^t is the transposition of α.*

Proof. If X is p-compressible, there exists a multiplicity 1 correspondence $\alpha : X \rightsquigarrow Y$ to a proper closed subvariety $Y \subset X$. Considering α as a correspondence $X \rightsquigarrow X$, we have $\mathrm{mult}\,\alpha = 1$ and $\mathrm{mult}\,\alpha^t = 0$. Therefore the "only if" part of Lemma 2.7 holds for an arbitrary X, not only for a homogeneous one.

The other way round, suppose that we are given a correspondence $\alpha : X \rightsquigarrow X$ with $\operatorname{mult}\alpha \neq \operatorname{mult}\alpha^t$. Adding a multiple of the diagonal class and multiplying by an element of $\mathbb{F}_p$, we may achieve that $\operatorname{mult}\alpha = 1$ and $\operatorname{mult}\alpha^t = 0$. In this case the pull-back of α with respect to the morphism $X_{F(X)} \to X \times X$ induced by the generic point of the second factor of the product $X \times X$, is a 0-cycle class of degree 0. Since X is homogeneous, the degree homomorphism $\operatorname{Ch}_0(X_{F(X)}) \to \mathbb{F}_p$ is an isomorphism. Therefore the pull-back of α is 0. By the continuity property of Chow groups [7, Proposition 52.9], there exists a non-empty open subset $U \subset X$ such that the pull-back of α to $X \times U$ is already 0. By the localization sequence [7, Proposition 57.9], it follows that α is the push-forward of some correspondence $\beta : X \rightsquigarrow Y \in \operatorname{Ch}_{\dim X}(X \times Y)$, where Y is the proper closed subset $Y = X \setminus U$ of X. Since $\operatorname{mult}\beta = \operatorname{mult}\alpha = 1$, the variety X is p-compressible. $\square$

3. Motives

The classical Grothendieck Chow motives [7, Chapter XII] we are going to use are simply a convenient language to work with the correspondences. Since our correspondences live in the Chow groups with coefficients in $\mathbb{F}_p$, our motives also have coefficients in $\mathbb{F}_p$. Thus, a motive is a direct sum of triples (X, π, i), where X is a smooth projective variety, $\pi : X \rightsquigarrow X$ a projector, and i an integer. Given two such triples (X_1, π_1, i_1) and (X_2, π_2, i_2), one defines

$$\operatorname{Hom}\big((X_1, \pi_1, i_1), (X_2, \pi_2, i_2)\big) := \pi_2 \circ \operatorname{Ch}_{\dim X_1 + i_1 - i_2}(X_1 \times X_2) \circ \pi_1.$$

For any smooth projective X, the motive $M(X)$ of X is the triple $(X, \mathrm{id}_X, 0)$. For any integer j, the shift functor $M \mapsto M(j)$ is identity on the homomorphisms, additive, and takes (X, π, i) to $(X, \pi, i + j)$. The motive $M(\operatorname{Spec} F)$ is denoted by $\mathbb{F}_p$; any its shift $\mathbb{F}_p(j)$ is called a *Tate* motive.

The *Krull-Schmidt principle* holds for the motives of projective homogeneous varieties: any direct summand of the motive of a projective homogeneous variety decomposes – and in a unique way – into a direct sum of indecomposable motives, see [6] or [12].

The *nilpotence principle*, initially discovered in the case of quadrics by M. Rost, holds for the motives of projective homogeneous varieties, [5, Theorem 8.2]. In particular, a motivic summand of a projective homogeneous variety becoming 0 over an extension of F is 0. However, in contrast to the Krull-Schmidt principle, the nilpotence principle is not really required for our purposes. It allows us to work with the usual Chow motives with coefficients in $\mathbb{F}_p$ (which is probably more interesting from the view point of the theory of motives itself). Alternatively, we could have constructed our motives out of the *reduced* Chow groups $\bar{\operatorname{Ch}}$ which are defined as Ch modulo everything vanishing over an extension of the base field. In this "simplified" motivic category, the nilpotence principle vanishes as well.

Let X be a projective homogeneous variety. The motive $\bar{M}(X)$ (which is $M(X)$ over an algebraic closure of F) is a sum of Tate motives $\mathbb{F}_p(j)$, with j varying between 0 and $\dim X$; moreover, there is precisely one summand with $j = 0$ (as well as with $j = \dim X$). Therefore, there is one and unique (up to an isomorphism) indecomposable summand $U(X)$ of $M(X)$ such that the Tate motive $\mathbb{F}_p$ is a summand of $\bar{U}(X)$. We call this $U(X)$ the *upper indecomposable motivic summand* of X or simply the *upper motive* of X. (The *lower* motive of X is defined in the same way by taking the Tate motive $\mathbb{F}_p(\dim X)$ in place of $\mathbb{F}_p$.)

Upper motives are easy to handle. For instance, $U(X) \simeq U(Y)$ for two projective homogeneous varieties X and Y if and only if $v_p(Y_{F(X)}) = 0 = v_p(X_{F(Y)})$, [12].

Upper motives are important: any indecomposable summand of the motive of a projective homogeneous variety under an algebraic group of inner type is the upper motive of some (other) projective homogeneous variety. A more precise statement is given in [12]. A generalization including the outer type case is given in [16, Theorem 1.1].

A projector $\pi : X \rightsquigarrow X$ determines an upper summand of $M(X)$ if and only if $\operatorname{mult} \pi = 1$; π determines a lower summand if and only if $\operatorname{mult} \pi^t = 1$ (see [12]). Since moreover, an appropriate power of any correspondence $X \rightsquigarrow X$ is a projector (see [12]), Lemma 2.7 can be reformulated as follows:

Lemma 3.1. *A projective homogeneous variety X is p-incompressible if and only if its upper motive is lower.* □

A simple but extremely useful tool for proving p-incompressibility is the following lemma. For any direct summand M of the motive of a projective homogeneous variety X, the *rank* $\operatorname{rk} M$ of M is the number of summands in the complete decomposition of $\bar{M}$.

Lemma 3.2 ([12]). $v_p(\operatorname{rk} M) \geq v_p(X)$.

Proof. Let π be a lifting of the projector on X defining M to the Chow group with coefficients in $\mathbb{Z}/p^l\mathbb{Z}$, where $l = v_p(X)$. Some power of the correspondence π is a projector and its pull-back with respect to the diagonal morphism $X \to X \times X$ is a (modulo p^l) 0-cycle class on X of degree $\operatorname{rk} M \mod p^l$. □

Example 3.3. Let X be the Severi-Brauer variety of a p-primary central division F-algebra D. Lemma 3.2 shows that the motive of X is indecomposable. Indeed, if $\deg D = p^n$, where $\deg D := \sqrt{\dim_F D} \in \mathbb{Z}$, then $v_p(X) = n$ and it follows that the rank of any non-zero summand of $M(X)$ is at least $p^n = \operatorname{rk} M(X)$. After the proofs of Examples 2.3 and 2.4, this is the third proof of the p-incompressibility of X.

Let A be a central simple F-algebra. For any integer i with $0 \leq i \leq \deg A$ we write $\mathrm{SB}_i(A)$ for the following generalized Severi-Brauer variety of A: the

variety of the right ideals in A of the *reduced dimension* i (that is, of the F-dimension $i \cdot \deg A$). For instance, $\mathrm{SB}_1(A)$ is the usual Severi-Brauer variety $\mathrm{SB}(A)$.

For $p = 2$, the opposite to the Severi-Brauer case has been considered by B. Mathews:

Example 3.4 ([23]). Let D be a non-trivial 2-primary central division F-algebra. Then the variety $X := \mathrm{SB}_{(\deg D)/2}(D)$ is 2-incompressible. Indeed, according to [4] or [5] or [14], the motive $M(X)_{F(X)}$ is a sum of one $\mathbb{F}_2$, one $\mathbb{F}_2(\dim X)$, and of shifts of $M(Y)$, where Y runs over some projective homogeneous $F(X)$-varieties with $v_2(Y) > 0$. It follows that $U(X)_{F(X)}$ contains the summand $\mathbb{F}_2(\dim X)$. Therefore X is 2-incompressible by Lemma 3.1. (This proof differs from the original one.) In contrast to Example 3.3, the motive of X is *decomposable* as far as $v_2(\deg D) > 2$: this is a special case of motivic decompositions found by M. Zhykhovich in [29].

Although the rank of $U(X)$ in Example 3.4 is not determined, one can show that $v_2 \operatorname{rk} U(X) = 1$, [12]. Together with the incompressibility of X, this is a basement for the following result concerning isotropy of an orthogonal involution on an arbitrary (not necessarily division) central simple algebra:

Theorem 3.5 ([11]). *Assume that* $\operatorname{char} F \neq 2$. *Any orthogonal involution* σ *on a central simple* F*-algebra* A *becoming isotropic over the function field of* $\mathrm{SB}(A)$*, also becomes isotropic over a finite odd degree field extension of* F.

An F-linear involution on a central simple F-algebra A is *hyperbolic*, if A possesses a σ-isotropic ideal of the reduced dimension $(\deg A)/2$.

The following non-hyperbolicity result is an immediate consequence of Theorem 3.5 and [1, Proposition 1.2]:

Theorem 3.6 ([9]). *Assume that* $\operatorname{char} F \neq 2$. *Any non-hyperbolic orthogonal involution* σ *on a central simple* F*-algebra* A *remains non-hyperbolic over the function field of* $\mathrm{SB}(A)$.

The symplectic version of Theorem 3.6 has been obtained by J.-P. Tignol:

Theorem 3.7 ([26]). *Assume that* $\operatorname{char} F \neq 2$. *Any non-hyperbolic* symplectic *(i.e., non-orthogonal) involution* σ *on a central simple* F*-algebra* A *remains non-hyperbolic over the function field of* $\mathrm{SB}_2(A)$.

Tensor products of F-linear involutions on quaternion F-algebras are called *Pfister involutions*. This is a generalization of the classical *Pfister forms*. Any isotropic Pfister form is hyperbolic. An over 30 years old conjecture saying that any isotropic Pfister involution on a central simple algebra A is hyperbolic, has been proved for algebras A of index ≤ 2 by K. Becher 3 years ago, [2]. Theorems 3.6 and 3.7 give the general case:

Theorem 3.8. *Any isotropic Pfister involution (over a field of characteristic* $\neq 2$*) is hyperbolic.*

Proof. Let σ be an isotropic Pfister involution on a central simple F-algebra A. If σ is orthogonal, $\sigma_{F(X)}$ with $X := \mathrm{SB}(A)$ is hyperbolic by [2, Theorem 1]; therefore σ is hyperbolic by Theorem 3.6. If σ is symplectic, $\sigma_{F(X)}$ with $X := \mathrm{SB}_2(A)$ is hyperbolic by [2, Corollary]; therefore σ is hyperbolic by Theorem 3.7. □

4. General Generalized Severi-Brauer Varieties

The following result generalizes Examples 3.3 and 3.4:

Theorem 4.1 ([12]). *Let n be a positive integer and let D be a central division F-algebra of degree p^n. For any integer i with $0 \le i < n$, the generalized Severi-Brauer variety $\mathrm{SB}_{p^i}(D)$ is p-incompressible.*

The proof is based on the properties of upper motives formulated in Section 3. It makes use of a double induction on n and i with a simultaneous computation of the p-adic valuation of the rank of the upper motive of $\mathrm{SB}_{p^i}(D)$ which turns out to be

$$v_p \operatorname{rk} U\big(\mathrm{SB}_{p^i}(D)\big) = v_p \operatorname{rk} M\big(\mathrm{SB}_{p^i}(D)\big) = n - i.$$

Theorem 4.1 actually computes the canonical p-dimension of an arbitrary generalized Severi-Brauer variety:

Corollary 4.2 ([12]). *Let A be a central simple F-algebra, i any integer with $0 \le i \le \deg A$. Then*

$$\operatorname{cdim}_p \mathrm{SB}_i(A) = \dim \mathrm{SB}_{p^{v_p(i)}}(D_p) = p^{v_p(i)}(p^{v_p(\operatorname{ind} A)} - p^{v_p(i)}),$$

where D_p is the p-primary part of a central division algebra Brauer-equivalent to A.

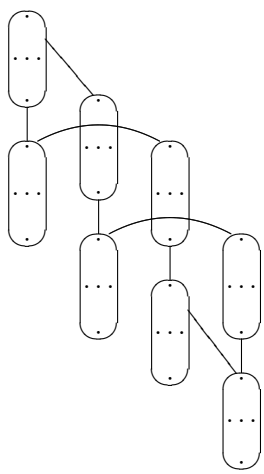
Example 4.3 (J.-P. Tignol, [26]). The particular case of Theorem 4.1 with $p = 2$ and $i = 1$ has the following application to a symplectic involution σ on a central division F-algebra D: $\sigma_{F(X)}$ is anisotropic, where $X = \mathrm{SB}_2(D)$. Indeed, otherwise the proper closed subvariety $Y \subset X$ of the *isotropic* ideals in D would have an $F(X)$-point. (This proof differs from the original one.) Note that the characteristic 2 case is included here. We do not get the same result for $X = \mathrm{SB}_1(D)$ because $Y = X$ for such X.

We have already spoken in Example 1.6 about the incompressibility of some products of Severi-Brauer varieties. There is one more related class of incompressible projective homogeneous varieties. It is useful in study of *unitary* involutions.

Given a finite separable field extension L/F, we write $\mathcal{R}_{L/F}X$ for the *Weil transfer* of an L-variety X.

Theorem 4.4 ([10]). *Let F be a field, L/F a quadratic separable field extension, n a non-negative integer, and D a central division L-algebra of degree 2^n such that the norm algebra $N_{L/F}D$ is trivial. For any integer $i \in [0,\ n]$, the variety $X := \mathcal{R}_{L/F}\,\mathrm{SB}_{2^i}(D)$ is 2-incompressible.*

The proof, using induction on n, considers some indecomposable motivic summands – the upper one, the lower one, and some of their shifts – of the variety X_{EL}, where $E = F\big(\mathcal{R}_{L/F}\,\mathrm{SB}_{2^{n-1}}(D)\big)$. "Connections" between these summands existing over E (known by induction) and over L are represented in the diagram below, where the ovals represent the summands. Since the upper and the lower summand are connected (by a chain of connections), the variety X is 2-incompressible.



Example 4.5 (J.-P. Tignol, [26]). Theorem 4.4 with $i = 0$ has the following application to a *unitary* involution σ on a 2-primary central division L-algebra D (an F-linear involution σ on D is unitary if it acts on L by the non-trivial F-automorphism): $\sigma_{F(X)}$ is anisotropic, where $X = \mathcal{R}_{L/F}\,\mathrm{SB}_1(D)$. Indeed, otherwise the proper closed subvariety $Y \subset X$ of the *isotropic* ideals in D would have an $F(X)$-point. (This proof differs from the original one.) Characteristic 2 case is included here. Unlike [26], we do not need to assume that the exponent of D is 2.

5. Dimension of Upper Motive

Let X be a projective homogeneous variety. In this final section we will show that $\mathrm{cdim}_p(X)$ is determined by the upper motive $U(X)$. Since $\mathrm{cdim}_p(X)$ is not changed under field extensions of p-prime degrees, [25, Proposition 1.5], we may assume that the semisimple affine algebraic group G acting on X has the

following property: G becomes of inner type over some p-primary field extension of F.

Dimension $\dim U(X)$ of $U(X)$ is the biggest integer i such that the Tate motive $\mathbb{F}_p(i)$ is a summand of $\bar{U}(X)$. More generally, *dimension* of a summand M of the motive of a projective homogeneous variety is the maximum of $i-j$, where i and j run over the integers such that $\mathbb{F}_p(i)$ and $\mathbb{F}_p(j)$ are summands of $\bar{M}$.

Theorem 5.1. $\dim U(X) = \operatorname{cdim}_p X$.

For a motive M, M^* is its dual. The cofunctor $M \mapsto M^*$ transposes the homomorphisms, is additive, and takes (Y, π, i) to $(Y, \pi^t, -\dim Y - i)$ for any smooth projective variety Y, where π^t is the transposition of the projector π. In particular, $M(Y)^* = M(Y)(-\dim Y)$.

Proposition 5.2. $U(X)^* \simeq U(X)\big(-\dim U(X)\big)$. *In other words, the lower indecomposable motivic summand of X, that is, $U(X)^*(\dim X)$, is isomorphic to*

$$U(X)\big(\dim X - \dim U(X)\big).$$

Remark 5.3. Note that $\operatorname{Ch}_i \bar{U}(X) = 0 = \operatorname{Ch}^i \bar{U}(X)$ for any integer $i > \dim U(X)$ by the very definition of $\dim U(X)$. Proposition 5.2 shows that actually

$$\operatorname{Ch}_i U(X) = 0 = \operatorname{Ch}^i U(X)$$

for i as above. Indeed, for $d := \dim U(X)$, we have:

$$\operatorname{Ch}^i U(X) = \operatorname{Ch}_{-i} U(X)^* \simeq \operatorname{Ch}_{-i} U(X)(-d) = \operatorname{Ch}_{d-i} U(X) \subset \operatorname{Ch}_{d-i} X = 0$$

and $\operatorname{Ch}_i U(X) = \operatorname{Ch}^{-i} U(X)^* \simeq \operatorname{Ch}^{d-i} U(X) \subset \operatorname{Ch}^{d-i} X = 0$. (Of course, since $U(X)$ is a summand of the motive of a variety, we also have $\operatorname{Ch}_i(U) = 0 = \operatorname{Ch}^i(U)$ for any $i < 0$.)

Proof of Proposition 5.2. For G as above, let $r = r(X)$ be the rank of the semisimple anisotropic kernel of $G_{F(X)}$. We induct on r.

The motive $U(X)^*(d)$, where now $d := \dim X$, is an indecomposable summand of $M(X)$. Therefore, by [16, Theorem 1.1] and according to the assumption on G made in the beginning of this Section, there exists a finite separable field extension L/F, a projective G_L-homogeneous L-variety Y, and an integer n such that $U(X)^*(d) \simeq U(Y)(n)$ and the Tits index of $G_{L(Y)}$ contains the Tits index of $G_{F(X)}$. Here we consider the upper motive of Y, which originally lives over L, as a motive over F (strictly speaking, we apply to the L-motive $U(Y)$ the functor $\operatorname{cor}_{L/F}$ of [16, §3]).

Since $\operatorname{Ch}_d \bar{U}(X)^*(d) = \operatorname{Ch}_0 \bar{U}(X)^* = \operatorname{Ch}^0 \bar{U}(X) = \mathbb{F}_p$ and $\dim_{\mathbb{F}_p} \operatorname{Ch}_d \bar{U}(Y)(n)$ is a multiple of $[L:F]$, it follows that $L = F$. Besides,

$$n = \min\{i \mid \operatorname{Ch}^i \bar{U}(Y)(n) \neq 0\}$$

and $\min\{i \mid \operatorname{Ch}^i \bar{U}(X)^*(d) \neq 0\} = d - \dim U(X)$, therefore $n = d - \dim U(X)$, and we have $U(X)^* \simeq U(Y)\big(-\dim U(X)\big)$.

If the Tits index of $G_{F(Y)}$ coincides with the Tits index of $G_{F(X)}$, the motives $U(X)$ and $U(Y)$ are isomorphic, and we are done in this case. Otherwise, the rank of the semisimple anisotropic kernel of $G_{F(Y)}$ is smaller than r, and, by the induction hypothesis, we have $U(Y)^* \simeq U(Y)\big(-\dim U(Y)\big)$. Dualizing and substituting, we see that

$$U(X) \simeq U(Y)\big(\dim U(X) - \dim U(Y)\big).$$

It follows that $\dim U(X) = \dim U(Y)$ and $U(X) \simeq U(Y)$. □

Proof of Theorem 5.1. We start by proving the easier inequality

$$\dim U(X) \leq \operatorname{cdim}_p X.$$

We can find a closed subvariety $Y \subset X$ with $\dim Y = \operatorname{cdim}_p X$ and with a multiplicity 1 correspondence $\pi : X \rightsquigarrow Y$. Considering π as a correspondence $X \rightsquigarrow X$, we can find an integer $m \geq 1$ such that $\pi^{\circ m}$ is a projector. Let $M = (X, \pi^{\circ m})$. Since $\operatorname{mult} \pi^{\circ m} = \operatorname{mult} \pi = 1$, the motivic summand M of X is upper and so, $\dim U(X) \leq \dim M$. Since $\operatorname{Ch}_i \bar{M} \subset \operatorname{Im}(\operatorname{Ch}_i \bar{Y} \to \operatorname{Ch}_i \bar{X})$ for any integer i, and $\operatorname{Ch}_i \bar{Y} = 0$ for $i > \dim Y$, we get the inequality $\dim M \leq \dim Y$ proving that $\dim U(X) \leq \operatorname{cdim}_p X$.

The opposite inequality $\dim U(X) \geq \operatorname{cdim}_p X$ requires Proposition 5.2. We set $n := \dim X - \dim U(X)$. Since $U(X)(n)$ is a motivic summand of X, shifting, we have morphisms

$$U(X) \xrightarrow{\ f\ } M(X)(-n) \xrightarrow{\ g\ } U(X)$$

with $g \circ f = \mathrm{id}$. Since $U(X)$ is an upper summand of $M(X)$, the subgroup $\operatorname{Ch}^0 U(X)$ of $\operatorname{Ch}^0 X$ coincides with $\operatorname{Ch}^0 X$ and, in particular, the class $[X] \in \operatorname{Ch}^0 X$ belongs to $\operatorname{Ch}^0 U(X)$. Applying $f_* : \operatorname{Ch}^0 U(X) \to \operatorname{Ch}^0 M(X)(-n) = \operatorname{Ch}^n X$, we get an element $\alpha := f_*([X]) \in \operatorname{Ch}^n X$ such that $g_*(\alpha) = [X]$. Therefore, there exists a closed subvariety $Y \subset X$ of codimension n such that $g_*([Y]) \neq 0$. We claim that $Y_{F(X)}$ has a closed point of a p-prime degree, and this claim proves Theorem 5.1.

To prove the claim, it suffices to notice that the relation $g_*([Y]) \neq 0 \in \operatorname{Ch}^0(X)$ implies that $\xi^* g_*([Y]) \neq 0 \in \operatorname{Ch}^0 \operatorname{Spec} F(X) = \mathbb{F}_p$, where $\xi : \operatorname{Spec} F(X) \to X$ is the generic point. In the same time, the modulo p integer $\xi^* g_*([Y]) \in \mathbb{F}_p$ is the degree of the 0-cycle class $[Y_{F(X)}] \cdot (\mathrm{id}_X \times \xi)^*(g)$ which is represented by a 0-cycle on $Y_{F(X)}$. □

References

[1] Bayer-Fluckiger, E., and Lenstra, Jr., H. W. Forms in odd degree extensions and self-dual normal bases. *Amer. J. Math. 112*, 3 (1990), 359–373.

[2] Becher, K. J. A proof of the Pfister factor conjecture. *Invent. Math. 173*, 1 (2008), 1–6.

[3] Berhuy, G., and Reichstein, Z. On the notion of canonical dimension for algebraic groups. *Adv. Math. 198*, 1 (2005), 128–171.

[4] Brosnan, P. On motivic decompositions arising from the method of Białynicki-Birula. *Invent. Math. 161*, 1 (2005), 91–111.

[5] Chernousov, V., Gille, S., and Merkurjev, A. Motivic decomposition of isotropic projective homogeneous varieties. *Duke Math. J. 126*, 1 (2005), 137–159.

[6] Chernousov, V., and Merkurjev, A. Motivic decomposition of projective homogeneous varieties and the Krull-Schmidt theorem. *Transform. Groups 11*, 3 (2006), 371–386.

[7] Elman, R., Karpenko, N., and Merkurjev, A. *The algebraic and geometric theory of quadratic forms*, vol. 56 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2008.

[8] Florence, M. On the essential dimension of cyclic p-groups. *Invent. Math. 171*, 1 (2008), 175–189.

[9] Karpenko, N. A. Hyperbolicity of orthogonal involutions. *Doc. Math.,* to appear.

[10] Karpenko, N. A. Incompressibility of quadratic Weil transfer of generalized Severi-Brauer varieties. Linear Algebraic Groups and Related Structures (preprint server) 362 (2009, Oct 28), 12 pages.

[11] Karpenko, N. A. Isotropy of orthogonal involutions. Linear Algebraic Groups and Related Structures (preprint server) 371 (2009, Nov 21), 9 pages.

[12] Karpenko, N. A. Upper motives of algebraic groups and incompressibility of Severi-Brauer varieties. Linear Algebraic Groups and Related Structures (preprint server) 333 (2009, Apr 3, revised: 2009, Apr 24), 18 pages.

[13] Karpenko, N. A. Grothendieck Chow motives of Severi-Brauer varieties. *Algebra i Analiz 7*, 4 (1995), 196–213.

[14] Karpenko, N. A. Cohomology of relative cellular spaces and of isotropic flag varieties. *Algebra i Analiz 12*, 1 (2000), 3–69.

[15] Karpenko, N. A. On anisotropy of orthogonal involutions. *J. Ramanujan Math. Soc. 15*, 1 (2000), 1–22.

[16] Karpenko, N. A. Upper motives of outer algebraic groups. In *Quadratic forms, linear algebraic groups, and cohomology*, vol. (to appear) of *Dev. Math.* Springer, New York, 2010.

[17] Karpenko, N. A., and Merkurjev, A. S. Canonical p-dimension of algebraic groups. *Adv. Math. 205*, 2 (2006), 410–433.

[18] Karpenko, N. A., and Merkurjev, A. S. Essential dimension of finite p-groups. *Invent. Math. 172*, 3 (2008), 491–508.

[19] Knebusch, M. Generic splitting of quadratic forms. I. *Proc. London Math. Soc. (3) 33*, 1 (1976), 65–93.

[20] Knus, M.-A., Merkurjev, A., Rost, M., and Tignol, J.-P. *The book of involutions*, vol. 44 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1998. With a preface in French by J. Tits.

[21] Kol′o-Telen, Z.-L., Karpenko, N. A., and Merkur′ev, A. S. Rational surfaces and the canonical dimension of the group PGL_6. *Algebra i Analiz 19*, 5 (2007), 159–178.

[22] Lötscher, R., Macdonald, M., Meyer, A., and Reichstein, Z. Essential p-dimension of algebraic tori. Linear Algebraic Groups and Related Structures (preprint server) 363 (2009, Oct 28), 30 pages.

[23] Mathews, B. G. Canonical dimension of projective $\mathrm{PGL}_1(A)$-homogeneous varieties. Linear Algebraic Groups and Related Structures (preprint server) 332 (2009, Mar 30), 7 pages.

[24] Merkurjev, A. S. Steenrod operations and degree formulas. *J. Reine Angew. Math. 565* (2003), 13–26.

[25] Merkurjev, A. S. Essential dimension. In *Quadratic Forms – Algebra, Arithmetic, and Geometry*, vol. 493 of *Contemp. Math.* Amer. Math. Soc., Providence, RI, 2009, pp. 299–326.

[26] Tignol, J.-P. Hyperbolicity of symplectic and unitary involutions. Appendix to a paper of N. Karpenko. *Doc. Math.,* to appear.

[27] Vishik, A. On the Chow groups of quadratic Grassmannians. *Doc. Math. 10* (2005), 111–130 (electronic).

[28] Zainoulline, K. Canonical p-dimensions of algebraic groups and degrees of basic polynomial invariants. *Bull. Lond. Math. Soc. 39*, 2 (2007), 301–304.

[29] Zhykhovich, M. Motivic decomposability of generalized Severi-Brauer varieties. Linear Algebraic Groups and Related Structures (preprint server) 361 (2009, Oct 26), 6 pages.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Essential Dimension

Zinovy Reichstein*

Abstract

Informally speaking, the essential dimension of an algebraic object is the minimal number of independent parameters one needs to define it. This notion was initially introduced in the context where the objects in question are finite field extensions [BuR97]. Essential dimension has since been investigated in several broader contexts, by a range of techniques, and has been found to have interesting and surprising connections to many problems in algebra and algebraic geometry.

The goal of this paper is to survey some of this research. I have tried to explain the underlying ideas informally through motivational remarks, examples and proof outlines (often in special cases, where the argument is more transparent), referring an interested reader to the literature for a more detailed treatment. The sections are arranged in rough chronological order, from the definition of essential dimension to open problems.

Mathematics Subject Classification (2010). Primary 14L30, 20G10, 11E72.

Keywords. Essential dimension, linear algebraic group, Galois cohomology, cohomological invariant, quadratic form, central simple algebra, algebraic torus, canonical dimension

1. Definition of Essential Dimension

Informally speaking, the essential dimension of an algebraic object is the minimal number of parameters one needs to define it. To motivate this notion, let us consider an example, where the object in question is a quadratic form.

Let k be a base field, K/k be a field extension and q be an n-dimensional quadratic form over K. Assume that $\mathrm{char}(k) \neq 2$ and denote the symmetric

*The author is grateful to S. Cernele, A. Duncan, S. Garibaldi, R. Lötscher, M. Macdonald, A. Merkurjev and A. Meyer for helpful comments and to the National Science and Engineering Council of Canada for financial support through its Discovery and Accelerator Supplement grants.

Department of Mathematics, University of British Columbia, Vancouver, BC, Canada. E-mail: reichst@math.ubc.ca.

bilinear form associated to q by b. We would now like to see if q can be defined over a smaller field $k \subset K_0 \subset K$. This means that there is a K-basis $e_1, \ldots, e_n$ of K^n such that $b(e_i, e_j) \in K_0$ for every $i, j = 1, \ldots, n$. If we can find such a basis, we will say that q *descends* to K_0 or that K_0 is a *field of definition* of q. It is natural to ask if there is a minimal field K_{min}/k (with respect to inclusion) to which q descends. The answer to this question is usually "no". For example, it is not difficult to see that the "generic" form $q(x_1, \ldots, x_n) = a_1x_1^2 + \cdots + a_nx_n^2$ over the field $K = k(a_1, \ldots, a_n)$, where $a_1, \ldots, a_n$ are independent variables, has no minimal field of definition. We will thus modify our question: instead of asking for a minimal field of definition K_0 for q, we will ask for the minimal value of the transcendence degree $\operatorname{tr}\deg_k(K_0)$. [1] This number is called the *essential dimension* of q and is denoted by $\operatorname{ed}(q)$.

Note that the above definition of $\operatorname{ed}(q)$ is in no way particular to quadratic forms. In a similar manner one can consider fields of definition of any polynomial in $K[x_1, \ldots, x_n]$, any finite-dimensional K-algebra, any algebraic variety defined over K, etc. In each case the minimal transcendence degree of a field of definition is an interesting numerical invariant which gives us some insight into the "complexity" of the object in question.

We will now state these observations more formally. Let k be a base field, Fields_k be the category of field extensions K/k, Sets be the category of sets, and $\mathcal{F}\colon \mathrm{Fields}_k \to \mathrm{Sets}$ be a covariant functor. In the sequel the word "functor" will always refer to a functor of this type. If $\alpha \in \mathcal{F}(K)$ and L/K is a field extension, we will denote the image of α in $\mathcal{F}(L)$ by α_L.

For example, $\mathcal{F}(K)$ could be the set of K-isomorphism classes of quadratic forms on K^n, or of n-dimensional K-algebras, for a fixed integer n, or of elliptic curves defined over K. In general we think of $\mathcal{F}$ as specifying the type of algebraic object we want to work with, and elements of $\mathcal{F}(K)$ as the of algebraic objects of this type defined over K.

Given a field extension K/k, we will say that $a \in \mathcal{F}(K)$ *descends* to an intermediate field $k \subseteq K_0 \subseteq K$ if a is in the image of the induced map $\mathcal{F}(K_0) \to \mathcal{F}(K)$. The *essential dimension* $\operatorname{ed}(a)$ of $a \in \mathcal{F}(K)$ is the minimum of the transcendence degrees $\operatorname{tr}\deg_k(K_0)$ taken over all fields $k \subseteq K_0 \subseteq K$ such that a descends to K_0. The essential dimension $\operatorname{ed}(\mathcal{F})$ of the functor $\mathcal{F}$ is the supremum of $\operatorname{ed}(a)$ taken over all $a \in \mathcal{F}(K)$ with K in Fields_k.

These notions are relative to the base field k; we will sometimes write ed_k in place of ed to emphasize the dependence on k. If $\mathcal{F}\colon \mathrm{Fields}_k \to \mathrm{Sets}$ be a covariant functor and $k \subset k'$ is a field extension, we will write $\operatorname{ed}_{k'}(\mathcal{F})$ for $\operatorname{ed}(\mathcal{F}_{k'})$, where $\mathcal{F}_{k'}$ denotes the restriction of $\mathcal{F}$ to $\mathrm{Fields}_{k'}$. Is easy to see that in this situation

$$\operatorname{ed}_k(\mathcal{F}) \geq \operatorname{ed}_{k'}(\mathcal{F})\,; \tag{1.1}$$

[1] One may also ask which quadratic forms have a minimal field of definition. To the best of my knowledge, this is an open question; see Section 7.1.

cf. [BF03, Proposition 1.5]. In particular, taking k' to be an algebraic closure of k, we see that for the purpose of proving a lower bound of the form $\mathrm{ed}_k(\mathcal{F}) \geq d$, where d does not depend on k, we may assume that k is algebraically closed.

Let $\boldsymbol{\mu}_n$ denote the group of nth roots of unity, defined over k. Whenever we consider this group, we will assume that it is smooth, i.e., that $\mathrm{char}(k)$ does not divide n.

Example 1.1. Let $\mathcal{F}(K) := H^r(K, \boldsymbol{\mu}_n)$ be the Galois cohomology functor. Assume k is algebraically closed. If $\alpha \in H^r(K, \boldsymbol{\mu}_n)$ is non-trivial then by the Serre vanishing theorem (see, e.g., [Se02, II.4.2, Prop. 11, p. 83]) $\mathrm{ed}(\alpha) \geq r$.

Example 1.2. Once again, assume that k is algebraically closed. Let $\mathbf{Forms}_{n,d}(K)$ be the set of homogeneous polynomials of degree d in n variables. If $\alpha \in \mathbf{Forms}_{n,d}(K)$ is anisotropic over K then by the Tsen-Lang theorem (see, e.g., [Pf95]), $n \leq d^{\mathrm{ed}(\alpha)}$ or equivalently, $\mathrm{ed}(\alpha) \geq \log_d(n)$.

Of particular interest to us will be the functors $\mathcal{F}_G$ given by $K \to H^1(K, G)$, where G is an algebraic group over k. Here, as usual, $H^1(K, G)$ denotes the set of isomorphism classes of G-torsors over $\mathrm{Spec}(K)$. The essential dimension of this functor is a numerical invariant of G, which, roughly speaking, measures the complexity of G-torsors over fields. We write $\mathrm{ed}\, G$ for $\mathrm{ed}\, \mathcal{F}_G$. Essential dimension was originally introduced in this context (and only in characteristic 0); see [BuR97, Rei00, RY00]. The above definition of essential dimension for a general functor $\mathcal{F}$ is due to A. Merkurjev; see [BF03].

In special cases this notion was investigated much earlier. To the best of my knowledge, the first non-trivial result related to essential dimension is due to F. Klein [Kl1884]. In our terminology, Klein showed that the essential dimension of the symmetric group S_5 over $k = \mathbb{C}$, is 2. (Klein referred to this result as "Kroeneker's theorem", so it may in fact go back even further.) The essential dimension of the projective linear group $\mathbf{PGL}_n$ first came up in C. Procesi's pioneering work on universal division algebras in the 1960s; see [Pr67, Section 2]. The problems of computing the essential dimension of the symmetric group S_n and the projective linear group $\mathbf{PGL}_n$ remain largely open; see Section 7.

If k is an algebraically closed field then groups of essential dimension zero are precisely the *special groups*, introduced by J.-P. Serre [Se58]. Recall that an algebraic group G over k is called special if $H^1(K, G) = 0$ for every field extension K/k. Over an algebraically closed field of characteristic zero these groups were classified by A. Grothendieck [Gro58] in the 1950s. The problem of computing the essential dimension of an algebraic group may be viewed as a natural extension of the problem of classifying special groups.

2. First Examples

Recall that an action of an algebraic group G on an algebraic k-variety X is called *generically free* if X has a dense G-invariant open subset U such that

the stabilizer $\mathrm{Stab}_G(x) = \{1\}$ for every $x \in U(\overline{k})$ and *primitive* if G permutes the irreducible components of X. Here $\overline{k}$ denotes an algebraic closure of k. Equivalently, X is primitive if $k(X)^G$ is a field.

If K/k is a finitely generated field extension then elements of $H^1(K, G)$ can be interpreted as birational isomorphism classes of generically free primitive G-varieties (i.e., k-varieties with a generically free primitive G-action) equipped with a k-isomorphism of fields $k(X)^G \simeq K$; cf. [BF03, Section 4]. If X is a generically free primitive G-variety, and $[X]$ is its class in $H^1(K, G)$ then

$$\mathrm{ed}([X]) = \min \dim(Y) - \dim(G)\,, \tag{2.1}$$

where the minimum is taken over all dominant rational G-equivariant maps $X \dashrightarrow Y$ such that the G-action on Y is generically free.

An important feature of the functor $H^1(\,*\,, G)$ is the existence of so-called *versal objects*; see [GMS03, Section I.5]. If $\alpha \in H^1(K, G)$ is a versal torsor then it is easy to see that $\mathrm{ed}(\alpha) \geq \mathrm{ed}(\beta)$ for any field extension L/k and any $\beta \in H^1(L, G)$. In other words, $\mathrm{ed}(\alpha) = \mathrm{ed}(G)$. If $G \to \mathbf{GL}(V)$ is a generically free k-linear representation of G then the class $[V]$ of V in $H^1(k(V)^G, G)$ is versal. By (2.1), we see that

$$\mathrm{ed}(G) = \min \dim(Y) - \dim(G)\,, \tag{2.2}$$

where the minimum is taken over all dominant rational G-equivariant maps $V \dashrightarrow Y$, such that G-action on Y is generically free. In particular,

$$\mathrm{ed}(G) \leq \dim(V) - \dim(G)\,. \tag{2.3}$$

Moreover, unless k is a finite field and G is special, we only need to consider closed G-invariant subvarieties Y of V. That is,

$$\mathrm{ed}(G) = \min\{\dim \mathrm{Im}(f)\} - \dim(G)\,, \tag{2.4}$$

where the minimum is taken over all G-equivariant rational maps $f\colon V \dashrightarrow V$ such that the G-action on $\mathrm{Im}(f)$ is generically free; see [Me09, Theorem 4.5].

Example 2.1. Let G be a connected adjoint semisimple group over k. Then $\mathrm{ed}(G) \leq \dim(G)$. To prove this inequality, apply (2.3) to the generically free representation $V = \mathcal{G} \times \mathcal{G}$, where $\mathcal{G}$ is the adjoint representation of G on its Lie algebra.

Note that the inequality $\mathrm{ed}(G) \leq \dim(G)$ can fail dramatically if G is not adjoint; see Corollary 4.3.

We now turn to lower bounds on $\mathrm{ed}(G)$ for various algebraic groups G and more generally, on $\mathrm{ed}(\mathcal{F})$ for various functors $\mathcal{F}\colon \mathrm{Fields}_k \to \mathrm{Sets}$. The simplest approach to such bounds is to relate the functor $H^1(\,*\,, G)$ (and more generally, $\mathcal{F}$) to the functors in Examples 1.1 or 1.2, using the following lemma, whose proof is immediate from the definition; cf. [BF03, Lemma 1.9].

Lemma 2.2. *Suppose a morphism of functors $\phi\colon \mathcal{F} \to \mathcal{F}'$ takes α to β. Then* $\mathrm{ed}(\alpha) \geq \mathrm{ed}(\beta)$. *In particular, if ϕ is surjective then* $\mathrm{ed}(\mathcal{F}) \geq \mathrm{ed}(\mathcal{F}')$.

A morphism of functors $\mathcal{F} \to H^d(\,*\,, \boldsymbol{\mu}_n)$ is called a *cohomological invariant* of degree d; it is said to be nontrivial if $\mathcal{F}(K)$ contains a non-zero element of $H^d(K, \boldsymbol{\mu}_n)$ for some K/k. Using Lemma 2.2 and Example 1.1 we recover the following observation, due to Serre.

Lemma 2.3. *Suppose k is algebraically closed. If there exists a non-trivial cohomological invariant $\mathcal{F} \to H^d(\,*\,, \boldsymbol{\mu}_n)$ then* $\mathrm{ed}(\mathcal{F}) \geq d$.

In the examples below I will, as usual, write $\langle a_1, \dots, a_n\rangle$ for the quadratic form $(x_1, \dots, x_n) \mapsto a_1x_1^2 + \dots + a_nx_n^2$ and $\ll a_1, \dots, a_r \gg$ for the r-fold Pfister form $\langle 1, -a_1\rangle \otimes_K \dots \otimes \langle 1, -a_r\rangle$.

Example 2.4. Suppose $\mathrm{char}(k) \neq 2$. Let Pf_r be the functor that assigns to a field K/k the set of K-isomorphism classes of r-fold Pfister forms, $q = \ll a_1, \dots, a_r \gg$. Then $\mathrm{ed}_k(\mathrm{Pf}_r) = r$.

Indeed, since q is defined over $k(a_1, \dots, a_r)$, we have $\mathrm{ed}_k(\mathrm{Pf}_r) \leq r$. To prove the opposite inequality, we may assume that k is algebraically closed. Let $a_1, \dots, a_r$ be independent variables and $K = k(a_1, \dots, a_r)$. Then the tautological map $\mathrm{Pf}_r(K) \to \mathbf{Forms}_{2^r,2}(K)$ takes $q = \ll a_1, \dots, a_r \gg$ to an anisotropic form in 2^r variables; see, e.g., [Pf95, p. 111]. Combining Lemma 2.2 and Example 1.2 we conclude that $\mathrm{ed}(\mathrm{Pf}_r) \geq r$, as desired.

Alternatively, the inequality $\mathrm{ed}(\mathrm{Pf}_r) \geq r$ also follows from Lemma 2.3, applied to the cohomological invariant $\mathrm{Pf}_r \to H^r(\,*\,, \boldsymbol{\mu}_2)$, which takes $\ll a_1, \dots, a_r \gg$ to the cup product $(a_1) \cup \dots \cup (a_r)$.

Since $H^1(\,*\,, \mathbf{G}_2)$ is naturally isomorphic to Pf_3, we conclude that $\mathrm{ed}(\mathbf{G}_2) = 3$. Here $\mathbf{G}_2$ stands for the split exceptional group of type $\mathbf{G}_2$ over k.

Example 2.5. If $\mathrm{char}(k) \neq 2$ then $\mathrm{ed}_k(\mathbf{O}_n) = n$.

Indeed, since every quadratic form over K/k can be diagonalized, we see that $\mathrm{ed}_k(\mathbf{O}_n) \leq n$. To prove the opposite inequality, we may assume that k is algebraically closed. Define the functor $\phi\colon H^1(\,*\,, \mathbf{O}_n) \to \mathrm{Pf}_n$ as follows. Let b be the bilinear form on $V = K^n$, associated to $q = \langle a_1, \dots, a_n\rangle \in H^1(K, \mathbf{O}_n)$. Then b naturally induces a non-degenerate bilinear form on the 2^n-dimensional K-vector space $\bigwedge(V)$. We now set $\phi(q)$ to be the 2^r-dimensional quadratic form associated to $\wedge(b)$. One easily checks that $\phi(q)$ is the n-fold Pfister form $\phi(q) = \ll a_1, \dots, a_n \gg$. Since ϕ is clearly surjective, Lemma 2.2 and Example 2.4 tell us that $\mathrm{ed}(\mathbf{O}_n) \geq \mathrm{ed}(\mathrm{Pf}_n) = n$.

We remark that $\phi(q)$ is closely related to the nth Stiefel-Whitney class $\mathrm{sw}_n(q)$ (see [GMS03, p. 41]), and the inequality $\mathrm{ed}(\mathbf{O}_n) \geq n$ can also be deduced by applying Lemma 2.3 to the cohomological invariant $\mathrm{sw}_n\colon H^1(K, \mathbf{O}_n) \to H^n(K, \boldsymbol{\mu}_2)$.

Example 2.6. If k contains a primitive nth root of unity then $\mathrm{ed}_k(\boldsymbol{\mu}_n^r) = r$.

Indeed, the upper bound, $\mathrm{ed}(\boldsymbol{\mu}_n^r) \le r$, follows from (2.3). Alternatively, note that any $(\overline{a_1}, \dots, \overline{a_r}) \in H^1(K, \boldsymbol{\mu}_n^r)$ is defined over the subfield $k(a_1, \dots, a_r)$ of K, of transcendence degree $\le r$.

To prove the opposite inequality we may assume that k is algebraically closed. Now apply Lemma 2.3 to the cohomological invariant

$$H^1(K, \boldsymbol{\mu}_n^r) = K^*/(K^*)^n \times \cdots \times K^*/(K^*)^n \to H^r(K, \boldsymbol{\mu}_n)$$

given by $(\overline{a_1}, \dots, \overline{a_r}) \to (a_1) \cup \cdots \cup (a_r)$. Here $\overline{a}$ denotes the class of $a \in K^*$ in $K^*/(K^*)^n$.

Remark 2.7. Suppose H is a closed subgroup of G and $G \to \mathbf{GL}(V)$ is a generically free linear representation. Since every rational G-equivariant map $V \dashrightarrow Y$ is also H-equivariant, (2.2) tells us that

$$\mathrm{ed}(G) \ge \mathrm{ed}(H) + \dim(H) - \dim(G)\,. \tag{2.5}$$

In particular, if a finite group G contains a subgroup $H \simeq (\mathbb{Z}/p\mathbb{Z})^r$ for some prime p and if $\mathrm{char}(k) \ne p$ (so that we can identify $(\mathbb{Z}/p\mathbb{Z})^r$ with $\boldsymbol{\mu}_p^r$ over $\overline{k}$) then

$$\mathrm{ed}_k(G) \ge \mathrm{ed}_{\overline{k}}(G) \ge \mathrm{ed}_{\overline{k}}(H) = r\,.$$

In the case where G is the symmetric group S_n and $H \simeq (\mathbb{Z}/2\mathbb{Z})^{[n/2]}$ is the subgroup generated by the commuting 2-cycles (12), (34), (56), etc., this yields $\mathrm{ed}(S_n) \ge [n/2]$; cf. [BuR97].

Example 2.8. Recall that elements of $H^1(K, \mathbf{PGL}_n)$ are in a natural bijective correspondence with isomorphism classes of central simple algebras of degree n over K. Suppose $n = p^s$ is a prime power, and k contains a primitive pth root of unity. Consider the morphism of functors $\phi\colon H^1(K, \mathbf{PGL}_n) \to \mathbf{Forms}_{n^2,p}$ given by sending a central simple K-algebra A to the degree p trace form $x \to \mathrm{Tr}_{A/K}(x^p)$.

If $a_1, \dots, a_{2s}$ are independent variables over k, $K = k(a_1, \dots, a_{2s})$, and

$$A = (a_1, a_2)_p \otimes_K \otimes \cdots \otimes (a_{2s-1}, a_{2s})_p$$

is a tensor product of s symbol algebras of degree p then one can write out $\phi(A)$ explicitly and show that it is anisotropic over K; see [Rei99]. Lemma 2.2 and Example 1.2 now tell us that $\mathrm{ed}_k(A) \ge \mathrm{ed}_{\overline{k}}(A) \ge 2s$. Since $\mathrm{tr\,deg}_k(K) = 2s$, we conclude that, in fact

$$\mathrm{ed}_k(A) = 2s \text{ and consequently, } \mathrm{ed}_k(\mathbf{PGL}_{p^s}) \ge 2s. \tag{2.6}$$

The following alternative approach to proving (2.6) was brought to my attention by P. Brosnan. Consider the cohomological invariant given by the composition of the natural map $H^1(K, \mathbf{PGL}_n) \to H^2(K, \boldsymbol{\mu}_n)$, which sends a central simple algebra to its Brauer class, and the divided power map

$H^2(*, \boldsymbol{\mu}_n) \to H^{2s}(*, \boldsymbol{\mu}_n)$; see [Kahn00, Appendix]. The image of A under the resulting cohomological invariant

$$H^1(*, \mathbf{PGL}_n) \to H^{2s}(*, \boldsymbol{\mu}_n)$$

is $(a_1) \cup (a_2) \cup \cdots \cup (a_{2s}) \neq 0$ in $H^{2s}(K, \boldsymbol{\mu}_n)$. Lemma 2.3 now tells us that $\mathrm{ed}_k(A) \geq 2s$, and (2.6) follows. The advantage of this approach is that it shows that the essential dimension of the Brauer class of A is also $2s$.

3. The Fixed Point Method

The following lower bound on $\mathrm{ed}(G)$ was conjectured by Serre and proved in [GR07]. Earlier versions of this theorem have appeared in [RY00] and [CS06].

Theorem 3.1. *If G is connected, A is a finite abelian subgroup of G and* $\mathrm{char}(k)$ *does not divide $|A|$, then* $\mathrm{ed}_k(G) \geq \mathrm{rank}(A) - \mathrm{rank}\ C_G^0(A)$.

Here $\mathrm{rank}(A)$ stands for the minimal number of generators of A and $\mathrm{rank}\ C_G^0(A)$ for the dimension of the maximal torus of the connected group $C_G^0(A)$. Note that if A is contained in a torus $T \subset G$ then $\mathrm{rank}(C_G^0(A)) \geq \mathrm{rank}(T) \geq \mathrm{rank}(A)$, and the inequality of Theorem 3.1 becomes vacuous. Thus we are primarily interested in non-toral finite abelian subgroups A of G. These subgroups have come up in many different contexts, starting with the work of Borel in the 1950s. For details and further references, see [RY00].

The proof of Theorem 3.1 relies on the following two simple results.

Theorem 3.2 (Going Down Theorem). *Suppose k is an algebraically closed base field and A is an abelian group such that* $\mathrm{char}(k)$ *does not divide $|A|$. Suppose A acts on k-varieties X and Y and $f\colon X \dashrightarrow Y$ is an A-equivariant rational map. If X has a smooth A-fixed point and Y is complete then Y has an A-fixed point.*

A short proof of Theorem 3.2, due to J. Kollár and E. Szabó, can be found in [RY00, Appendix].

Lemma 3.3. *Let A be a finite abelian subgroup, acting faithfully on an irreducible k-variety X. Suppose* $\mathrm{char}(k)$ *does not divide $|A|$. If X has a smooth A-fixed point then* $\dim(X) \geq \mathrm{rank}(A)$.

The lemma follows from the fact that the A-action on the tangent space $T_x(X)$ at the fixed point x has to be faithful; see [GR07, Lemma 4.1].

For the purpose of proving Theorem 3.1 we may assume that k is algebraically closed. To convey the flavor of the proof I will make the following additional assumptions: (i) $C_G(A)$ is finite and (ii) $\mathrm{char}(k) = 0$. The conclusion then reduces to

$$\mathrm{ed}(G) \geq \mathrm{rank}(A)\,. \tag{3.1}$$

This special case of Theorem 3.1 is proved in [RY00] but I will give a much simplified argument here, based on [GR07, Section 4].

Let $G \to \mathbf{GL}(V)$ be a generically free representation. By (2.2) we need to show that if $V \dashrightarrow Y$ is a G-equivariant dominant rational map and the G-action on Y is generically free, then

$$\dim(Y) - \dim(G) \geq \operatorname{rank}(A)\,. \tag{3.2}$$

To see how to proceed, let us first consider the "toy" case, where G is finite. Here (3.1) follows from (2.5), but I will opt for a different argument below, with the view of using a variant of it in greater generality.

After birationally modifying Y, we may assume that it is smooth and projective. (Note that this step relies on G-equivariant resolution of singularities and thus uses the characteristic 0 assumption.) Since V has a smooth A-fixed point (namely, the origin), the Going Down Theorem 3.2 tells us that so does Y. By Lemma 3.3, $\dim(Y) \geq \operatorname{rank}(A)$, which proves (3.2) in the case where G is finite.

If G is infinite, we can no longer hope to prove (3.2) by applying Lemma 3.3 to the A-action on Y. Instead, we will apply Lemma 3.3 to a suitable A-invariant subvariety $Z \subset Y$. This subvariety Z will be a cross-section for the G-action on Y, in the sense that a G-orbit in general position will intersect Z in a finite number of points. Hence, $\dim(Z) = \dim(Y) - \dim(G)$, and (3.2) reduces to $\dim(Z) \geq \operatorname{rank}(A)$. We will then proceed as in the previous paragraph: we will use Theorem 3.2 to find an A-fixed point on a smooth complete model of Z, then use Lemma 3.3 to show that $\dim(Z) \geq \operatorname{rank}(A)$.

Let me now fill in the details. By [CGR06] Y is birationally isomorphic to $G \times^S Z$, where S is a finite subgroup of G and Z is an algebraic variety equipped with a faithful S-action. (A priori Z does not carry an A-action; however, we will show below that some conjugate A' of A lies in S and consequently, acts on Z. We will then replace A by A' and argue as above.) We also note that we are free to replace Z by an (S-equivariantly) birationally isomorphic variety, so we may (and will) take it to be smooth and projective.

Here, as usual, if S acts on normal quasi-projective varieties X and Z then $X \times^S Z$ denotes the geometric quotient of $X \times Z$ by the natural (diagonal) action of S. Since S is finite, there is no difficulty in forming the quotient map $\pi\colon X \times Z \to X \times^S Z$; cf. [GR07, Lemma 3.1]. Moreover, if the S-action on X extends to a $G \times S$-action, then by the universal property of geometric quotients $X \times^S Z$ inherits a G-action from $X \times Z$, where G acts on the first factor. I will write $[x, z] \in X \times^S Z$ for the image of (x, z) under π.

We now compactify $Y = G \times^S Z$ by viewing it as a G-invariant open subset of the projective variety $\overline{Y} := \overline{G} \times^S Z$, where $\overline{G}$ is a so-called "wonderful" (or "regular") compactification of G. Recall that $G \times G$ acts on $\overline{G}$, extending the right and left multiplication action of G on itself, The complement $\overline{G} \smallsetminus G$ is a normal crossing divisor $D_1 \cup \cdots \cup D_r$, where each D_i is irreducible, and the intersection of any number of D_i is the closure of a single $G \times G$-orbit in $\overline{G}$.

The compactification $\overline{G}$ has many wonderful properties; the only one we will need is Lemma 3.4 below. For a proof, see [Br98, Proposition A1].

Lemma 3.4. *For every $x \in \overline{G}$, $P = \mathrm{pr}_1(\mathrm{Stab}_{G\times G}(x))$ is a parabolic subgroup of G. Here pr_1 is projection to the first factor. Moreover, $P = G$ if and only if $x \in G$.*

We are now ready to complete the proof of the inequality (3.2) (and thus of (3.1)) by showing that S contains a conjugate A' of A, and A' has a fixed point in Z. In other words, our goal is to show that some conjugate A' of A lies in $\mathrm{Stab}_S(z)$.

By the Going Down Theorem 3.2, $\overline{Y}$ has an A-fixed point. Denote this point by $[x, z]$ for some $x \in \overline{G}$ and $z \in Z$. That is, for every $a \in A$, $[ax, z] = [x, z]$ in $\overline{Y}$. Equivalently,

$$\begin{cases} ax = xs^{-1} \\ sz = z \end{cases} \tag{3.3}$$

for some $s \in S$. In other words, for every $a \in A$, there exists an $s \in \mathrm{Stab}_S(z)$ such that $(a^{-1}, s) \in \mathrm{Stab}_{G\times G}(x)$. Equivalently, the image of the natural projection $\mathrm{pr}_1\colon \mathrm{Stab}_{G\times G}(x) \to G$ contains A. Since we are assuming that $C_G^0(A)$ is finite, A cannot be contained in any proper parabolic subgroup of G. Thus $x \in G$; see Lemma 3.4. Now the first equation in (3.3) tells us that $A' := x^{-1}Ax \subset \mathrm{Stab}_S(z)$, as desired. □

Remark 3.5. The above argument proves Theorem 3.1 under two simplifying assumptions: (i) $C_G(A)$ is finite and (ii) $\mathrm{char}(k) = 0$. If assumption (i) is removed, a variant of the same argument can still be used to prove Theorem 3.1 in characteristic 0; see [GR07, Section 4]. Assumption (ii) is more serious, because our argument heavily relies on resolution of singularities. Consequently, the proof of Theorem 3.1 in prime characteristic is considerably more complicated; see [GR07].

Corollary 3.6. *(a)* $\mathrm{ed}(\mathbf{SO}_n) \geq n-1$ *for any* $n \geq 3$, *(b)* $\mathrm{ed}(\mathbf{PGL}_{p^s}) \geq 2s$,

(c) $\mathrm{ed}(\mathbf{Spin}_n) \geq \begin{cases} [n/2] \text{ for any } n \geq 11, \\ [n/2]+1 \text{ if } n \equiv -1,\ 0 \text{ or } 1 \text{ modulo } 8, \end{cases}$

(d) $\mathrm{ed}(\mathbf{G}_2) \geq 3$, *(e)* $\mathrm{ed}(\mathbf{F}_4) \geq 5$, *(f)* $\mathrm{ed}(\mathbf{E}_6^{sc}) \geq 4$.

(g) $\mathrm{ed}(\mathbf{E}_7^{sc}) \geq 7$, *(h)* $\mathrm{ed}(\mathbf{E}_7^{ad}) \geq 8$, *(i)* $\mathrm{ed}(\mathbf{E}_8) \geq 9$.

Here the superscript ^{sc} stands for "simply connected" and ^{ad} for "adjoint".

Each of these inequalities is proved by exhibiting a non-toral abelian subgroup $A \subset G$ whose centralizer is finite. For example, in part (a) we can take $A \simeq (\mathbb{Z}/2\mathbb{Z})^{n-1}$ to be the subgroup of diagonal matrices of the form

$$\mathrm{diag}(\epsilon_1, \ldots, \epsilon_n), \text{ where each } \epsilon_i = \pm 1 \text{ and } \epsilon_1 \cdot \ldots \cdot \epsilon_n = 1. \tag{3.4}$$

The details are worked out in [RY00], with the exception of the first line in part (c), which was first proved by V. Chernousov and J.-P. Serre [CS06], by a

different method. I later noticed that it can be deduced from Theorem 3.1 as well; the finite abelian subgroups one uses here can be found in [Woo89].

Remark 3.7. The inequalities in parts (a), (b), (d), (e) and (f) can be recovered by applying Lemma 2.3 to suitable cohomological invariants. For parts (b) and (d), this is done in Examples 2.8 and 2.4, respectively; for parts (a), (e) and (f), see [Rei00, Example 12.7], [Rei00, Example 12.10] and [Gar01, Remark 2.12].

It is not known whether or not parts (g), (h) and (i) can be proved in a similar manner, i.e., whether or not there exist cohomological invariants of E_7^{sc}, E_7^{ad} and E_8 of dimensions 7, 8, and 9, respectively.

4. Central Extensions

In this section we will discuss another more recent method of proving lower bounds on $\mathrm{ed}(G)$. This method does not apply as broadly as those described in the previous two sections, but in some cases it leads to much stronger bounds. Let

$$1 \to C \to G \to \overline{G} \to 1 \tag{4.1}$$

be an exact sequence of algebraic groups over k such that C is central in G and isomorphic to $\boldsymbol{\mu}_p^r$ for some $r \geq 1$. Given a character $\chi\colon C \to \boldsymbol{\mu}_p$, we will, following [KM07], denote by Rep^χ the set of irreducible representations $\phi\colon G \to \mathbf{GL}(V)$, defined over k, such that $\phi(c) = \chi(c)\,\mathrm{Id}_V$ for every $c \in C$.

Theorem 4.1. *Assume that k is a field of characteristic $\neq p$ containing a primitive pth root of unity. Then*

$$\mathrm{ed}_k(G) \geq \min_{\langle \chi_1,\ldots,\chi_r\rangle = C^*} \left(\sum_{i=1}^{r} \gcd_{\rho_i \in \mathrm{Rep}^{\chi_i}} \dim(\rho_i) \right) - \dim G\,. \tag{4.2}$$

Here gcd *stands for the greatest common divisor and the minimum is taken over all minimal generating sets $\chi_1, \ldots, \chi_r$ of $C^* \simeq (\mathbb{Z}/p\mathbb{Z})^r$.*

Theorem 4.1 has two remarkable corollaries.

Corollary 4.2. (N. Karpenko – A. Merkurjev [KM07]) *Let G be a finite p-group and k be a field containing a primitive pth root of unity. Then*

$$\mathrm{ed}_k(G) = \min \dim(\phi)\,, \tag{4.3}$$

where the minimum is taken over all faithful k-representations ϕ of G.

Proof. We apply Theorem 4.1 to the exact sequence $1 \to C \to G \to G/C \to 1$, where C be the socle of G, i.e., $C := \{g \in Z(G) \mid g^p = 1\}$. Since $\dim(\rho)$ is a power of p for every irreducible representation ρ of G, we may replace gcd by min in (4.2). Choosing a minimal set of generators $\chi_1, \ldots, \chi_r$ of C^* so that the sum

on the right hand side of (4.2) has minimal value, and $\rho_i \in \mathrm{Rep}^{\chi_i}$ of minimal dimension, we see that (4.2) reduces to $\mathrm{ed}_k(G) \geq \dim(\rho_1) + \cdots + \dim(\rho_r)$. Equivalently, $\mathrm{ed}_k(G) \geq \dim(\rho)$, where $\rho := \rho_1 \oplus \cdots \oplus \rho_r$ is faithful by elementary p-group theory. This shows that $\mathrm{ed}_k(G) \geq \min \dim(\phi)$ in (4.3). The opposite inequality follows from (2.3). □

Corollary 4.3. *Let* $\mathbf{Spin}_n$ *be the split spinor group over a a field* k *of characteristic* 0*. Assume* $n \geq 15$*. Then*

(a) $\mathrm{ed}(\mathbf{Spin}_n) = 2^{(n-1)/2} - \frac{n(n-1)}{2}$*, if* n *is odd,*

(b) $\mathrm{ed}(\mathbf{Spin}_n) = 2^{(n-2)/2} - \frac{n(n-1)}{2}$*, if* $n \equiv 2 \pmod 4$*, and*

(c) $2^{(n-2)/2} - \frac{n(n-1)}{2} + 2^m \leq \mathrm{ed}(\mathbf{Spin}_n) \leq 2^{(n-2)/2} - \frac{n(n-1)}{2} + n$*, if* $n \equiv 0 \pmod 4$*. Here* 2^m *is the largest power of* 2 *dividing* n*.*

We remark that M. Rost and S. Garibaldi have computed the essential dimension of $\mathbf{Spin}_n$ for every $n \leq 14$; see [Rost06] and [Gar09].

Proof outline. The lower bounds (e.g., $\mathrm{ed}(\mathbf{Spin}_n) \geq 2^{(n-1)/2} - \frac{n(n-1)}{2}$, in part (a)) are valid whenever $\mathrm{char}(k) \neq 2$; they can be deduced either directly from Theorem 4.1 or by applying the inequality (2.5) to the finite 2-subgroup H of $G = \mathbf{Spin}_n$, where H is the inverse image of the diagonal subgroup $\boldsymbol{\mu}_2^{n-1} \subset \mathbf{SO}_n$, as in (3.4), under the natural projection $\pi\colon \mathbf{Spin}_n \to \mathbf{SO}_n$. Here $\mathrm{ed}(H)$ is given by Corollary 4.2.

The upper bounds (e.g., $\mathrm{ed}(\mathbf{Spin}_n) \leq 2^{(n-1)/2} - \frac{n(n-1)}{2}$, in part (a)) follow from the inequality (2.3), where V is spin representation V_{spin} in part (a), the half-spin representation $V_{\mathrm{half-spin}}$ in part (b), and to $V_{\mathrm{half-spin}} \oplus V_{\mathrm{natural}}$ in part (c), where V_{natural} is the natural n-dimensional representation of $\mathbf{SO}_n$, viewed as a representation of $\mathbf{Spin}_n$ via π. The delicate point here is to check that these representations are generically free. In characteristic 0 this is due to E. Andreev and V. Popov [AP71] for $n \geq 29$ and to A. Popov [Po85] in the remaining cases.

For details, see [BRV10a] and (for the lower bound in part (c)) [Me09, Theorem 4.9]. □

To convey the flavor of the proof of Theorem 4.1, I will consider a special case, where G is finite and $r = 1$. That is, I will start with a sequence

$$1 \to \boldsymbol{\mu}_p \to G \to \overline{G} \to 1 \tag{4.4}$$

of finite groups and will aim to show that

$$\mathrm{ed}_k(G) \geq \gcd_{\rho \in \mathrm{Rep}'} \dim(\rho)\,, \tag{4.5}$$

where k contains a primitive pth root and Rep' denotes the set of irreducible representations of G whose restriction to $\boldsymbol{\mu}_p$ is non-trivial. The proof relies on the following two results, which are of independent interest.

Theorem 4.4. (Karpenko's Incompressibility Theorem; [Kar00, Theorem 2.1]) *Let X be a Brauer-Severi variety of prime power index p^m, over a field K and let $f\colon X \dashrightarrow X$ be a rational map defined over K. Then* $\dim_K \operatorname{Im}(f) \geq p^m - 1$.

Theorem 4.5. (Merkurjev's Index Theorem [KM07, Theorem 4.4]; cf. also [Me96]) *Let K/k be a field extension, and $\partial_K\colon H^1(K,\overline{G}) \to H^2(K,\boldsymbol{\mu}_p)$ be the connecting map induced by the short exact sequence* (4.4)*. Then the maximal value of the index of $\partial_K(a)$, as K ranges over all field extension of k and a ranges over $H^1(K,\overline{G})$, equals* $\gcd_{\rho\in\mathrm{Rep}'} \dim(\rho)$.

Recall that $H^2(K,\boldsymbol{\mu}_p)$ is naturally isomorphic to the p-torsion subgroup of the Brauer group $\operatorname{Br}(K)$, so that it makes sense to talk about the index.

I will now outline an argument, due to M. Florence [Fl07], which deduces the inequality (4.5) from these two theorems. To begin with, let us choose a faithful representation V of G, where C acts by scalar multiplication. In particular, we can induce V from a faithful 1-dimensional representation $\chi\colon C = \boldsymbol{\mu}_p \to k^*$. We remark that χ exists because we assume that k contains a primitive pth root of unity and that we do not require V to be irreducible.

By (2.4), there exists a non-zero G-equivariant rational map $f\colon V \dashrightarrow V$ defined over k (or a *rational covariant*, for short) whose image has dimension $\operatorname{ed}(G)$. We will now replace f by a non-zero *homogeneous* rational covariant $f_{\mathrm{hom}}\colon V \dashrightarrow V$. Here "homogeneous" means that $f_{\mathrm{hom}}(tv) = t^d f_{\mathrm{hom}}(v)$ for some $d \geq 1$. Roughly speaking, f_{hom} is the "leading term" of f, relative to some basis of V, and it can be chosen so that

$$\dim \operatorname{Im}(f_{\mathrm{hom}}) \leq \dim \operatorname{Im}(f) = \operatorname{ed}(G)\,;$$

see [KLS09, Lemma 2.1]. Since we no longer need the original covariant f, we will replace f by f_{hom} and thus assume that f is homogeneous. Note that G may not act faithfully on the image of f but this will not matter to us in the sequel. Since f is homogeneous and non-zero, it descends to an $\overline{G}$-equivariant rational map $\overline{f}\colon \mathbb{P}(V) \dashrightarrow \mathbb{P}(V)$ defined over k, whose image has dimension $\leq \operatorname{ed}_k(G) - 1$.

Now, given a field extension K/k and a $\overline{G}$-torsor $T \to \operatorname{Spec}(K)$ in $H^1(K,\overline{G})$, we can twist $\mathbb{P}(V)$ by T. The resulting K-variety ${}^T\mathbb{P}(V)$ is defined as the quotient of $\mathbb{P}(V) \times_K T$ by the natural (diagonal) $\overline{G}$-action. One can show, using the theory of descent, that this action is in fact free, i.e., the natural projection map $\mathbb{P}(V) \times_K T \to^T \mathbb{P}(V)$ is a $\overline{G}$-torsor; see [Fl07, Proposition 2.12 and Remark 2.13]. Note that we have encountered a variant of this construction in the previous section, where we wrote $\mathbb{P}(V) \times^{\overline{G}} T$ in place of ${}^T\mathbb{P}(V)$.

We also remark that ${}^T\mathbb{P}(V)$ is a K-form of $\mathbb{P}(V)$, i.e., is a Brauer-Severi variety defined over K. Indeed, if a field extension L/K splits T then it is easy to see that ${}^T\mathbb{P}(V)$ is isomorphic to $\mathbb{P}(V)$ over L. One can now show that the index of this Brauer-Severi variety equals the index of $\partial_K(T) \in H^2(K,\boldsymbol{\mu}_p)$; in

particular, it is a power of p. By Theorem 4.5 we can choose K and T so that

$$\operatorname{ind}({}^T\mathbb{P}(V)) = \gcd_{\rho\in\mathrm{Rep}'} \dim(\rho)\,.$$

The $\overline{G}$-equivariant rational map $\overline{f}\colon \mathbb{P}(V) \dashrightarrow \mathbb{P}(V)$ induces a $\overline{G}$-equivariant rational map $\overline{f}\times\mathrm{id}\colon \mathbb{P}(V)\times T \dashrightarrow \mathbb{P}(V)\times T$, which, in turn, descends to a K-rational map ${}^T\overline{f}\colon {}^T\mathbb{P}(V) \dashrightarrow^T \mathbb{P}(V)$. Since the dimension of the image of $\overline{f}$ is $\le \mathrm{ed}_k(G)-1$, the dimension of the image of $\overline{f}\times\mathrm{id}$ is $\le \mathrm{ed}_k(G)-1+\dim(G)$, and thus the dimension of the image of ${}^T\overline{f}$ is $\le \mathrm{ed}_k(G)-1$. By Theorem 4.4,

$$\mathrm{ed}_k(G)-1 \ge \dim_K(\mathrm{Im}({}^T\overline{f})) \ge \operatorname{ind}({}^T\mathbb{P}(V))-1 = \gcd_{\rho\in\mathrm{Rep}'} \dim(\rho)-1\,,$$

and (4.5) follows. □

Remark 4.6. Now suppose G is finite but $r \ge 1$ is arbitrary. The above argument has been modified by R. Lötscher [Lö08] to prove Theorem 4.1 in this more general setting. The proof relies on Theorem 4.5 and a generalization of Theorem 4.4 to the case where X is the direct product of Brauer-Severi varieties $X_1\times\cdots\times X_r$, such that $\operatorname{ind}(X_i)$ is a power of p for each i; see [KM07, Theorem 2.1].

Here we choose our faithful k-representation V so that $V = V_1\times\cdots\times V_r$, where C acts on V_i by scalar multiplication via a multiplicative character $\chi_i\in C^*$, and $\chi_1,\dots,\chi_r$ generate C^*. Once again, there exists a G-equivariant rational map $f\colon V \dashrightarrow V$ whose image has dimension $\mathrm{ed}(G)$. To make the rest of the argument go through in this setting one needs to show that f can be chosen to be multi-homogeneous, so that it will descend to a $\overline{G}$-equivarint rational map

$$\overline{f}\colon \mathbb{P}(V_1)\times\dots\mathbb{P}(V_r) \dashrightarrow\colon \mathbb{P}(V_1)\times\dots\mathbb{P}(V_r)\,.$$

If G is finite, this is done in [Lö08]. The rest of the argument goes through unchanged.

In his (still unpublished) Ph. D. thesis Lötscher has extended this proof of Theorem 4.1 to the case where G is no longer assumed to be finite. His only requirement on G is that it should have a completely reducible faithful k-representation. The only known proof of Theorem 4.1 in full generality uses the notion of essential dimension for an algebraic stack, introduced in [BRV07]; cf. also [BRV10b]. For details, see [Me09, Theorem 4.8 and Example 3.7], in combination with [KM07, Theorem 4.4 and Remark 4.5].

5. Essential Dimension at p and Two Types of Problems

Let p be a prime integer. I will say that a field extension L/K is *prime-to-p* if $[L:K]$ is finite and not divisible by p.

This section is mostly "metamathematical"; the main point I would like to convey is that some problems in Galois cohomology and related areas are sensitive to prime-to-p extensions and some aren't. Loosely speaking, I will refer to such problems as being of "Type 2" and "Type 1", respectively.

More precisely, suppose we are given a functor $\mathcal{F}\colon \mathrm{Fields}_k \to \mathrm{Sets}$ and we would like to show that some (or every) $\alpha \in \mathcal{F}(K)$ has a certain property. For example, this property may be that $\mathrm{ed}(\alpha) \le d$ for a given d. If our functor is $\mathcal{F}(K) = H^1(K, \mathbf{O}_n)$, we may want to show that the quadratic form representing α is isotropic over K. If our functor is $\mathcal{F}(K) = H^1(K, \mathbf{PGL}_n)$, we may ask if the central simple algebra representing α is a crossed product. Note that in many interesting examples, including the three examples above, the property in question is functorial, i.e., if $\alpha \in \mathcal{F}(K)$ has it then so does α_L for every field extension L/K.

The problem of whether or not $\alpha \in \mathcal{F}(K)$ has a property we are interested in can be broken into two steps. For the first step we choose a prime p and ask whether or not α_L has the desired property for some prime-to-p extension L/K. This is what I call a *Type 1 problem*. If the answer is "no" for some p then we are done: we have solved the original problem in the negative. If the answer is "yes" for every prime p, then the remaining problem is to determine whether or not α itself has the desired property. I refer to problems of this type as *Type 2 problems*. Let me now explain what this means in the context of essential dimension.

Let $\mathcal{F}\colon \mathrm{Fields}_k \to \mathrm{Sets}$ be a functor and $a \in \mathcal{F}(K)$ for some field K/k. The essential dimension $\mathrm{ed}(a;p)$ of a at a prime integer p is defined as the minimal value of $\mathrm{ed}(a_L)$, as L ranges over all finite field extensions L/K such that p does not divide the degree $[L : K]$. The essential dimension $\mathrm{ed}(\mathcal{F};p)$ is then defined as the maximal value of $\mathrm{ed}(a;p)$, as K ranges over all field extensions of k and a ranges over $\mathcal{F}(K)$.

As usual, in the case where $\mathcal{F}(K) = H^1(K, G)$ for some algebraic group G defined over k, we will write $\mathrm{ed}(G;p)$ in place of $\mathrm{ed}(\mathcal{F};p)$. Clearly, $\mathrm{ed}(a;p) \le \mathrm{ed}(a)$, $\mathrm{ed}(\mathcal{F};p) \le \mathrm{ed}(\mathcal{F})$, and $\mathrm{ed}(G;p) \le \mathrm{ed}(G)$ for every prime p.

In the previous three sections we proved a number of lower bounds of the form $\mathrm{ed}(G) \ge d$, where G is an algebraic group and d is a positive integer. A closer look reveals that in every single case the argument can be modified to show that $\mathrm{ed}(G;p) \ge d$, for a suitable prime p. (Usually p is a so-called "exceptional prime" for G; see, e.g., [St75] or [Me09]. Sometimes there is more than one such prime.) In particular, the arguments we used in Examples 2.4, 2.5, 2.6 and 2.8 show that $\mathrm{ed}(\mathbf{G}_2;2) = 3$, $\mathrm{ed}(\mathbf{O}_n;2) = n$, $\mathrm{ed}(\boldsymbol{\mu}_p^r;p) = r$ and $\mathrm{ed}(\mathbf{PGL}_{p^s};p) \ge 2s$, respectively. In Theorem 3.1 we may replace $\mathrm{ed}(G)$ by $\mathrm{ed}(G;p)$, as long as A is a p-group; see [GR07, Theorem 1.2(b)]. Consequently, in Corollary 3.6 $\mathrm{ed}(G)$ can be replaced by $\mathrm{ed}(G;p)$, where $p = 2$ in parts (a), (c), (d), (e), (g), (h), (i) and $p = 3$ in part (f). Theorem 4.1 remains valid

with $\mathrm{ed}(G)$ replaced by $\mathrm{ed}(G;p)$.[2] Consequently, Corollary 4.2 remains valid if $\mathrm{ed}(G)$ is replaced by $\mathrm{ed}(G;p)$ (see [KM07]), and Corollary 4.3 remains valid if $\mathrm{ed}(\mathbf{Spin}_n)$ is replaced by $\mathrm{ed}(\mathbf{Spin}_n;2)$ (see [BRV10a]).

The same is true of virtually all existing methods for proving lower bounds on $\mathrm{ed}(\mathcal{F})$ and, in particular, on $\mathrm{ed}(G)$: they are well suited to address Type 1 problems and poorly suited for Type 2 problems. In this context a Type 1 problem is the problem of computing $\mathrm{ed}(\mathcal{F};p)$ for various primes p and a Type 2 problem is the problem of computing $\mathrm{ed}(\mathcal{F})$, assuming $\mathrm{ed}(\mathcal{F};p)$ is known for all p.

I will now make an (admittedly vague) claim that this phenomenon can be observed in a broader context and illustrate it with three examples not directy related to essential dimension.

Observation 5.1. *Most existing methods in Galois cohomology and related areas apply to Type 1 problems only. On the other hand, many long-standing open problems are of Type 2.*

Example 5.2. The crossed product problem. Recall that a central simple algebra A/K of degree n is a crossed product if it contains a commutative Galois subalgebra L/K of degree n. We will restrict our attention to the case where $n = p^r$ is a prime power; the general case reduces to this one by the primary decomposition theorem. In 1972 Amitsur [Am72] showed that for $r \geq 3$ a generic division algebra $U(p^r)$ of degree p^r is not a crossed product, solving a long-standing open problem. L. H. Rowen and D. J. Saltman [RS92, Theorem 2.2] modified Amitsur's argument to show that, in fact, $UD(p^r)_L$ is a non-crossed product for any prime-to-p extension L of the center of $\mathrm{UD}(p^r)$.

For $r = 1, 2$ it is not known whether or not every central smple algebra A of degree p^r is a crossed product. It is, however, known that every such algebra becomes a crossed product after a prime-to-p extension of the center; see [RS92, Section 1]. In other words, the Type 1 part of the crossed product problem has been completely solved, and the remaining open questions, for algebras of degree p and p^2, are of Type 2.

Example 5.3. The torsion index. Let G be an algebraic group defined over k and K/k be a field extension. The torsion index n_α of $\alpha \in H^1(K, G)$ was defined by Grothendieck as the greatest common divisor of the degrees $[L : K]$, where L ranges over all finite splitting fields L/K. The torsion index n_G of $\mathcal{G}$ is then the least common multiple of n_α taken over all K/k and all $\alpha \in H^1(K, G)$. One can show that $n_G = n_{\alpha_{\mathrm{ver}}}$, where $\alpha_{\mathrm{ver}} \in H^1(K_{versal}, G)$ is a versal G-torsor. One can also show, using a theorem of J. Tits [Se95], that the prime divisors of n_G are precisely the exceptional primes of G.

[2] At the moment the only known proof of this relies on the stack-theoretic approach; see the references at the end of Remark 4.6. The more elementary "homogenization" argument I discussed in the previous section has not (yet?) yielded a lower bound on $\mathrm{ed}(G;p)$.

The problem of computing n_G and more generally, of n_α for $\alpha \in H^1(K, G)$ can thus be rephrased as follows. Given an exceptional prime p for G, find the highest exponent d_p such that p^{d_p} divides $[L : K]$ for every splitting extension L/K. It is easy to see that this is a Type 1 problem; d does not change if we replace α by $\alpha_{K'}$, where K'/K is a prime-to-p extension. This torsion index n_G has been computed Tits and B. Totaro, for all simple groups G that are either simply connected or adjoint; for details and further references, see [Ti92, To05].

The remaining Type 2 problem consists of finding the possible values of $e_1, \ldots, e_r$ such that α_{ver} is split by a field extension L/K of degree $p_1^{e_1} \ldots p_r^{e_r}$, where $p_1, \ldots, p_r$ are the exceptional primes for G. This problem is open for many groups G. It is particularly natural for those G with only one exceptional prime, e.g., $G = \mathbf{Spin}_n$.

Example 5.4. Canonical dimension. Let G be a connected linear algebraic group defined over k, K/k be a field extension, and X be a G-torsor over K. Recall that the *canonical dimension* $\mathrm{cdim}(X)$ of X is the minimal value of $\dim_K(\mathrm{Im}(f))$, where the minimum is taken over all rational maps $f\colon X \dashrightarrow X$ defined over K. In particular, X is split if and only if $\mathrm{cdim}(X) = 0$. The maximal possible value of $\mathrm{cdim}(X)$, as X ranges over all G-torsors over K and K ranges over all field extensions of k, is called the canonical dimension of G and is denoted by $\mathrm{cdim}(G)$. Clearly $0 \le \mathrm{cdim}(G) \le \dim(G)$ and $\mathrm{cdim}(G) = 0$ if and only if G is special. For a detailed discussion of the notion of canonical dimension, we refer the reader to [BerR05], [KM06] and [Me09].

Computing the canonical dimension $\mathrm{cdim}(G)$ of an algebraic group G is a largely open Type 2 problem. The associated Type 1 problem of computing the canonical p-dimension $\mathrm{cdim}(G; p)$ has been solved by Karpenko-Merkurjev [KM06] and K. Zainoulline [Zai07].

6. Finite Groups of Low Essential Dimension

Suppose we would like to determine the essential dimension of a finite group G. To keep things simple, we will assume throughout this section that, unless otherwise specified, the base field k is algebraically closed and of characteristic 0. Let us break up the problem of computing $\mathrm{ed}(G)$ into a Type 1 part and a Type 2 part, as we did in the previous section.

The Type 1 problem is to determine $\mathrm{ed}(G; p)$ for a prime p. It is not difficult to show that $\mathrm{ed}(G; p) = \mathrm{ed}(G_p; p)$, where p is a prime and G_p is a p-Sylow subgroup of G; see [MR09a, Lemma 4.1] or [Me96, Proposition 5.1]. The value of $\mathrm{ed}(G_p; p)$ is given by Corollary 4.2. So, to the extent that we are able to compute the dimension of the smallest faithful representation of G_p, our Type 1 problem has been completely solved, i.e., we know $\mathrm{ed}(G; p)$ for every prime p.

Now our best hope of computing $\mathrm{ed}(G)$ is to obtain a strong upper bound $\mathrm{ed}(G) \le n$, e.g., by constructing an explicit G-equivariant dominant rational map $V \dashrightarrow Y$, as in (2.2), with $\dim(Y) = n$. If $n = \mathrm{ed}(G; p)$ then we conclude

that $\mathrm{ed}(G) = \mathrm{ed}(G;p)$, i.e., the remaining Type 2 problem is trivial, and we are done. In particular, this is what happens if G is a p-group.

If the best upper bound we can prove is $\mathrm{ed}(G) \leq n$, where n is strictly greater than $\mathrm{ed}(G;p)$ for every p then we are entering rather murky waters. Example 6.2 below shows that it is indeed possible for $\mathrm{ed}(G)$ to be strictly greater than $\mathrm{ed}(G;p)$ for every prime p. On the other hand, there is no general method for computing $\mathrm{ed}(G)$ in such cases. The only ray of light in this situation is that it may be possible to prove a lower bound of the form $\mathrm{ed}(G) > d$, where $d = 1$ or (with more effort) 2 and sometimes even 3.

Let us start with the simplest case where $d = 1$.

Lemma 6.1 (cf. [BuR97, Theorem 6.2]). *Let G be a finite group. Then*

(a) $\mathrm{ed}(G) = 0$ *if and only if* $G = \{1\}$,

(b) $\mathrm{ed}(G) = 1$ *if and only if* $G \neq \{1\}$ *is either cyclic or odd dihedral.*

Proof. Let V be a faithful linear representation of G. By (2.4) there exists a dominant G-equivariant rational map $V \dashrightarrow X$, where G acts faithfully on X and $\dim(X) = \mathrm{ed}(G)$.

(a) If $\mathrm{ed}(G) = 0$ then X is a point. This forces G to be trivial.

(b) If $\mathrm{ed}(G) = 1$ then X is a rational curve, by a theorem of Lüroth. We may assume that X is smooth and complete, i.e., we may assume that $X = \mathbb{P}^1$. Consequently, G is isomorphic to a subgroup of $\mathbf{PGL}_2$. By a theorem of Klein [Kl1884], G is cyclic, dihedral or is isomorphic to S_4, A_4 or S_5. If G is an even dihedral group, S_4, A_4 or S_5 then G contains a copy of $\mathbb{Z}/2 \times \mathbb{Z}/2\mathbb{Z} \simeq \boldsymbol{\mu}_2 \times \boldsymbol{\mu}_2$. Hence,

$$\mathrm{ed}(G) \geq \mathrm{ed}(\boldsymbol{\mu}_2^2) = 2\,;$$

see Example 2.6. This means that if $\mathrm{ed}(G) = 1$ then G is cyclic or odd dihedral.

Conversely, if G is cyclic or odd dihedral then one can easily check that, under our assumption on k, $\mathrm{ed}_k(G) = 1$. □

Example 6.2. Suppose q and r are odd primes and q divides $r - 1$. Let $G = \mathbb{Z}/r\mathbb{Z} \rtimes \mathbb{Z}/q\mathbb{Z}$ be a non-abelian group of order rq. Clearly all Sylow subgroups of G are cyclic; hence, $\mathrm{ed}(G;p) \leq 1$ for every prime p. On the other hand, since G is neither cyclic nor odd dihedral, Lemma 6.1 tells us that $\mathrm{ed}(G) \geq 2$. □

Similar reasoning can sometimes be used to show that $\mathrm{ed}(G) > 2$. Indeed, assume that $\mathrm{ed}(G) = 2$. Then there is a faithful representation V of G and a dominant rational G-equivariant map

$$V \dashrightarrow X\,, \tag{6.1}$$

where G acts faithfully on X and $\dim(X) = 2$. By a theorem of G. Castelnuovo, X is a rational surface. Furthermore, we may assume that X is smooth, complete, and is minimal with these properties (i.e., does not allow any G-equivariant blow-downs $X \to X_0$, with X_0 smooth). Such surfaces (called minimal rational G-surfaces) have been classified by Yu. Manin and V. Iskovskikh,

following up on classical work of F. Enriques; for details and further references, see [Du09a]. This classification is significantly more complicated than Klein's classification of rational curves but one can use it to determine, at least in principle, which finite groups G can act on a rational surface and describe all such actions; cf. [DI06]. Once all minimal rational G-surfaces X are accounted for, one then needs to decide, for each X, whether or not the G-action is versal, i.e., whether or not a dominant rational G-equivariant map (6.1) can exist for some faithful linear representation $G \to \mathbf{GL}(V)$. Note that by the Going Down Theorem 3.2 if some abelian subgroup A of G acts on X without fixed points then the G-action on X cannot be versal. If every minimal rational G-surface X can be ruled out this way (i.e., is shown to be non-versal) then one can conclude that $\mathrm{ed}(G) > 2$.

This approach was used by Serre to show that $\mathrm{ed}(\mathrm{A}_6) > 2$; see [Se08, Proposition 3.6]. Since the upper bound $\mathrm{ed}(\mathrm{A}_6) \le \mathrm{ed}(\mathrm{S}_6) \le 3$ was previously known (cf. (7.2) and the references there) this implies $\mathrm{ed}(\mathrm{A}_6) = 3$. Note that

$$\mathrm{ed}(\mathrm{A}_6; p) = \begin{cases} 2, \text{ if } p = 2 \text{ or } 3, \\ 1, \text{ if } p = 5, \text{ and} \\ 0, \text{ otherwise;} \end{cases}$$

see (7.1). A. Duncan [Du09a] has recently refined this approach to give the following complete classification of groups of essential dimension ≤ 2.

Theorem 6.3. *Let k be an algebraically closed field of characteristic 0 and $T = \mathbb{G}_m^2$ be the 2-dimensional torus over k. A finite group G has essential dimension ≤ 2 if and only if it is isomorphic to a subgroup of one of the following groups:*

(1) The general linear group $\mathbf{GL}_2(k)$,

(2) $\mathrm{PSL}_2(\mathbb{F}_7)$, the simple group of order 168,

(3) S_5, the symmetric group on 5 letters,

(4) $T \rtimes G_1$, where $|G \cap T|$ is coprime to 2 and 3 and

$$G_1 = \left\langle \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\rangle \simeq D_{12},$$

(5) $T \rtimes G_2$, where $|G \cap T|$ is coprime to 2 and

$$G_2 = \left\langle \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\rangle \simeq D_8,$$

(6) $T \rtimes G_3$, where $|G \cap T|$ is coprime to 3 and

$$G_3 = \left\langle \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \right\rangle \simeq \mathrm{S}_3,$$

(7) $T \rtimes G_4$*, where* $|G \cap T|$ *is coprime to* 3 *and*

$$G_4 = \left\langle \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\rangle \simeq \mathrm{S}_3 .$$

If one would like to go one step further and show that $\mathrm{ed}(G) > 3$ by this method, for a particular finite group G, the analysis becomes considerably more complicated. First of all, while X in (6.1) is still unirational, if $\dim(X) \geq 3$, we can no longer conclude that it is rational. Secondly, there is no analogue of the Enriques-Manin-Iskovskikh classification of rational surfaces in higher dimensions. Nevertheless, in dimension 3 one can sometimes use Mori theory to get a handle on X. In particular, Yu. Prokhorov [Pr09] recently classified the finite simple groups with faithful actions on rationally connected threefolds. This classification was used by Duncan [Du09b] to prove the following theorem, which is out of the reach of all previously existing methods.

Theorem 6.4. *Let* k *be a field of characteristic* 0*. Then* $\mathrm{ed}_k(A_7) = \mathrm{ed}_k(S_7) = 4$*.*

Note that $\mathrm{ed}(\mathrm{A}_7; p) \leq \mathrm{ed}(\mathrm{S}_7; p) \leq 3$ for every prime p; cf. [MR09a, Corollary 4.2].

7. Open Problems

7.1. Strongly incompressible elements. Let $\mathcal{F}\colon \mathrm{Fields}_k \to \mathrm{Sets}$ be a covariant functor. We say that an object $\alpha \in \mathcal{F}(K)$ is *strongly incompressible* if α does not descend to any proper subfield of K.

Examples of strongly incompressible elements in the case where G is a finite group, K is the function field of an algebraic curve Y over k, and $\mathcal{F} = H^1(\,*\,, G)$, are given in [Rei04]. In these examples α is represented by a (possibly ramified) G-Galois cover $X \to Y$. I do not know any such examples in higher dimensions.

Problem. *Does there exist a finitely generated field extension* K/k *of transcendence degree* ≥ 2 *and a finite group* G *(or an algebraic group* G *defined over* k*) such that* $H^1(K, G)$ *has a strongly incompressible element?*

For $G = \mathrm{O}_n$ Problem 7.1 is closely related to the questions of existence of a minimal field of definition of a quadratic form posed at the beginning of Section 1.

It is easy to see that if an element of $H^1(K, \mathbf{PGL}_n)$ represented by a non-split central simple algebra A is strongly incompressible and $\mathrm{tr}\deg_k(K) \geq 2$ then A cannot be cyclic. In particular, if $n = p$ is a prime then the existence of a strongly incompressible element in $H^1(K, \mathbf{PGL}_n)$ would imply the existence of a non-cyclic algebra of degree p over K, thus solving (in the negative) the long-standing cyclicity conjecture of Albert.

7.2. Symmetric groups.

Problem. *What is the essential dimension of the symmetric group* S_n*? of the alternating group* A_n*?*

Let us assume that $\mathrm{char}(k)$ does not divide $n!$. Then in the language of Section 5, the above problem is of Type 2. The associated Type 1 problem has been solved: $\mathrm{ed}(\mathrm{S}_n; p) = [n/p]$ (see [MR09a, Corollary 4.2]) and similarly

$$\mathrm{ed}(\mathrm{A}_n; p) = \begin{cases} 2[\frac{n}{4}], \text{ if } p = 2, \text{ and} \\ [\frac{n}{p}], \text{ otherwise.} \end{cases} \tag{7.1}$$

It is shown in [BuR97] that $\mathrm{ed}(\mathrm{S}_{n+2}) \geq \mathrm{ed}(\mathrm{S}_n) + 1$, $\mathrm{ed}(\mathrm{A}_{n+4}) \geq \mathrm{ed}(\mathrm{A}_n) + 2$, and

$$\mathrm{ed}(\mathrm{A}_n) \leq \mathrm{ed}(\mathrm{S}_n) \leq n - 3\,. \tag{7.2}$$

I believe the true value of $\mathrm{ed}(\mathrm{S}_n)$ is closer to $n-3$ than to $[n/2]$; the only piece of evidence for this is Theorem 6.4.

7.3. Cyclic groups.

Problem. *What is the essential dimension* $\mathrm{ed}_k(\mathbb{Z}/n\mathbb{Z})$*?*

Let us first consider the case where $\mathrm{char}(k)$ is prime to n. Under further assumptions that $n = p^r$ is a prime power and k contains a primitive pth root of unity ζ_p, Problem 7.3 has been solved by Florence [Fl07]. It is now a special case of Corollary 4.2:

$$\mathrm{ed}_k(\mathbb{Z}/p^r\mathbb{Z}) = \mathrm{ed}_k(\mathbb{Z}/p^r\mathbb{Z}; p) = [k(\zeta_{p^r}) : k]\,; \tag{7.3}$$

see [KM07, Corollary 5.2]. This also settles the (Type 1) problem of computing $\mathrm{ed}_k(\mathbb{Z}/n\mathbb{Z}; p)$ for every integer $n \geq 1$ and every prime p. Indeed,

$$\mathrm{ed}_k(\mathbb{Z}/n\mathbb{Z}; p) = \mathrm{ed}_k(\mathbb{Z}/p^r\mathbb{Z}; p)\,,$$

where p^r is the largest power of p dividing n. Also, since $[k(\zeta_p) : k]$ is prime to p, for the purpose of computing $\mathrm{ed}_k(\mathbb{Z}/n\mathbb{Z}; p)$ we are allowed to replace k by $k(\zeta_p)$; then formula (7.3) applies.

If we do not assume that $\zeta_p \in k$ then the best currently known upper bound on $\mathrm{ed}_k(\mathbb{Z}/p^r\mathbb{Z})$, due to A. Ledet [Led02], is $\mathrm{ed}_k(\mathbb{Z}/p^r\mathbb{Z}) \leq \varphi(d)p^e$. Here $[k(\zeta_{p^r}) : k] = dp^e$, where d divides $p-1$, and φ is the Euler φ-function.

Now let us suppose $\mathrm{char}(k) = p > 0$. Here it is easy to see that $\mathrm{ed}_k(\mathbb{Z}/p^r\mathbb{Z}) \leq r$; Ledet [Led04] conjectured that equality holds. This seems to be out of reach at the moment, at least for $r \geq 5$. More generally, essential dimension of finite (but not necessarily smooth) group schemes over a field k of prime characteristic is poorly understood; some interesting results in this direction can be found in [TV10].

7.4. Quadratic forms. Let us assume that $\mathrm{char}(k) \neq 2$. The following question is due to J.-P. Serre (private communication, April 2003).

Problem. *If q is a quadratic form over K/k, is it true that* $\mathrm{ed}(q;2) = ed(q)$*?*

A similar question for central simple algebras A of prime power degree p^r is also open: is it true that $\mathrm{ed}(A;p) = \mathrm{ed}(A)$?

Here is another natural essential dimension question in the context of quadratic form theory.

Problem. *Assume* $\mathrm{char}(k) \neq 2$. *If q and q' are Witt equivalent quadratic forms over a field K/k, is it true that* $\mathrm{ed}_k(q) = \mathrm{ed}_k(q')$*?*

The analogous question for central simple algebras, with Witt equivalence replaced by Brauer equivalence, has a negative answer. Indeed, assume k contains a primitive 4th root of unity and $D = UD_k(4)$ is a universal division algebra of degree 4. Then $\mathrm{ed}(D) = 5$ (see [Me10a, Corollary 1.2], cf. also [Rost00]) while $\mathrm{ed}\ \mathrm{M}_2(D) = 4$ (see [LRRS03, Corollary 1.4]).

7.5. Canonical dimension of Brauer-Severi varieties. Let X be a smooth complete variety defined over a field K/k. The canonical dimension $\mathrm{cdim}(X)$ is the minimal dimension of the image of a K-rational map $X \dashrightarrow X$. For a detailed discussion of this notion, see [KM06] and [Me09].

Conjecture. (Colliot-Thélène, Karpenko, Merkurjev [CKM08]) *Suppose X is a Brauer-Severi variety of index n. If $n = p_1^{e_1} \dots p_r^{e_r}$ is the prime decomposition of n then* $\mathrm{cdim}(X) = p_1^{e_1} + \dots + p_r^{e_r} - r$.

This is a Type 2 problem. The associated Type 1 question is completely answered by Theorem 4.4: $\mathrm{cdim}(X;p_i) = p_i^{e_1} - 1$. Also, by Theorem 4.4 the conjecture is true if $r = 1$. The only other case where this conjecture has been proved is $n = 6$; see [CKM08]. The proof is similar in spirit to the results of Section 6; it relies on the classification of rational surfaces over a non-algebraically closed field. For other values of n the conjecture has not even been checked for one particular X.

Note that the maximal value of $\mathrm{cdim}(X)$, as X ranges over the Brauer-Severi varieties of index n, equals $\mathrm{cdim}(\mathbf{PGL}_n)$. As I mentioned in Example 5.4, computing the canonical dimension $\mathrm{cdim}(G)$ of a linear algebraic (and in particular, simple) group G is a largely open Type 2 problem. In particular, the exact value of $\mathrm{cdim}(\mathbf{PGL}_n)$ is only known if $n = 6$ or a prime power.

7.6. Essential dimension of $\mathbf{PGL}_n$.

Problem. *What is* $\mathrm{ed}(\mathbf{PGL}_n;p)$*?* $\mathrm{ed}(\mathbf{PGL}_n)$*?*

As I mentioned in Section 1, this problem originated in the work of Procesi [Pr67]; for a more detailed history, see [MR09a, MR09b]. The second question appears to be out of reach at the moment, except for a few small values of

n. However, there has been a great deal of progress on the first (Type 1) question in the past year. By primary decomposition $\mathrm{ed}(\mathbf{PGL}_n; p) = \mathrm{ed}(\mathbf{PGL}_{p^r}; p)$, where p^r is the highest power of p dividing n. Thus we may assume that $n = p^r$. As I mentioned in Example 5.2, every central simple algebra A of degree p becomes cyclic after a prime-to-p extension. Hence, $\mathrm{ed}(\mathbf{PGL}_p; p) = 2$; cf. [RY00, Lemma 8.5.7]. For $r \geq 2$ we have

$$(r-1)p^r + 1 \leq \mathrm{ed}(\mathbf{PGL}_{p^r}; p) \leq p^{2r-2} + 1\,.$$

The lower bound is due to Merkurjev [Me10b]; the upper bound is proved in a recent preprint of A. Ruozzi [Ru10]. (A weaker upper bound, $\mathrm{ed}(\mathbf{PGL}_n; p) \leq 2p^{2r-2} - p^r + 1$, is proved in [MR09b].) In particular, $\mathrm{ed}(\mathbf{PGL}_{p^2}; p) = p^2 + 1$; see [Me10a].

Note that the argument in [Me10b] shows that if A is a generic $(\mathbb{Z}/p\mathbb{Z})^r$-crossed product then $\mathrm{ed}(A; p) = (r-1)p^r + 1$. As mentioned in Example 5.2, for $r \geq 3$ a general division algebra A/K of degree p^r is not a crossed product and neither is $A_L = A \otimes_K L$ for any prime-to-p extension L/K. Thus for $r \geq 3$ it is reasonable to expect the true value of $\mathrm{ed}(\mathbf{PGL}_{p^r}; p)$ to be strictly greater than $(r-1)p^r + 1$.

7.7. Spinor groups.

Problem. *Does Corollary 4.3 remain valid over an algebraically closed field of characteristic $p > 2$?*

As I mentioned at the beginning of the proof of Corollary 4.3, the lower bound in each part remains valid over any field of characteristic > 2. Consequently, Problem 7.7 concerns only the upper bounds. It would, in fact, suffice to show that the spin representation V_{spin} and the half-spin representation $V_{\mathrm{half-spin}}$ of $\mathbf{Spin}_n$ are generically free, if n is odd or $n \equiv 2 \pmod 4$, respectively; see [BRV10a, Lemma 3-7 and Remark 3-8].

Problem. *What is* $\mathrm{ed}_k(\mathbf{Spin}_{4m}; 2)$*?* $\mathrm{ed}_k(\mathbf{Spin}_{4m})$*? Here $m \geq 5$ is an integer.*

Corollary 4.3 answers this question in the case where m is a power of 2. In the other cases there is a gap between the upper and the lower bound in that corollary, even for $k = \mathbb{C}$.

7.8. Exceptional groups.

Problem. *Let G be an exceptional simple group and p be an exceptional prime for G. What is* $\mathrm{ed}_k(G; p)$*?* $\mathrm{ed}_k(G)$*? Here we assume that k is an algebraically closed field of characteristic* 0 *(or at least,* $\mathrm{char}(k)$ *is not an exceptional prime for G).*

For the exceptional group $G = \mathbf{G}_2$ we know that $\mathrm{ed}(\mathbf{G}_2) = \mathrm{ed}(\mathbf{G}_2; 2) = 3$; see Example 2.4.

For $G = \mathbf{F}_4$, the Type 1 problem has been completely solved: $\mathrm{ed}(\mathbf{F}_4; 2) = 5$ (see [MacD08, Section 5]), $\mathrm{ed}(\mathbf{F}_4; 3) = 3$ (see [GR07, Example 9.3]), and $\mathrm{ed}(\mathbf{F}_4; p) = 0$ for all other primes. It is claimed in [Ko00] that $\mathrm{ed}(\mathbf{F}_4) = 5$. However, the argument there appears to be incomplete, so the (Type 2) problem of computing $\mathrm{ed}(\mathbf{F}_4)$ remains open.

The situation is similar for the simply connected group $\mathbf{E}_6^{sc}$. The Type 1 problem has been solved,

$$\mathrm{ed}(\mathbf{E}_6^{sc}; p) = \begin{cases} 3, \text{ if } p = 2 \text{ (see [GR07, Example 9.4])}, \\ 4, \text{ if } p = 3 \text{ (see [RY00, Theorem 8.19.4 and Remark 8.20])}, \\ 0, \text{ if } p \geq 5. \end{cases}$$

(For the upper bound on the second line, cf. also [Gar09, 11.1].) The Type 2 problem of computing $\mathrm{ed}(\mathbf{E}_6^{sc})$ remains open.

For the other exceptional groups, $\mathbf{E}_6^{ad}$, $\mathbf{E}_7^{ad}$ $\mathbf{E}_7^{sc}$ and $\mathbf{E}_8$, even the Type 1 problem of computing $\mathrm{ed}(G; p)$ is only partially solved. It is known that $\mathrm{ed}(\mathbf{E}_6^{ad}; 2) = 3$ (see [GR07, Remark 9.7]), $\mathrm{ed}(\mathbf{E}_7^{ad}; 3) = \mathrm{ed}(\mathbf{E}_7^{sc}; 3) = 3$ (see [GR07, Example 9.6 and Remark 9.7]; cf. also [Gar09, Lemma 13.1]) and $\mathrm{ed}(\mathbf{E}_8; 5) = 3$ (see [RY00, Theorem 18.19.9] and [Gar09, Proposition 14.7]). On the other hand, the values of $\mathrm{ed}(\mathbf{E}_6^{ad}; 3)$, $\mathrm{ed}(\mathbf{E}_7^{ad}; 2)$, $\mathrm{ed}(\mathbf{E}_7^{sc}; 2)$, $\mathrm{ed}(\mathbf{E}_8; 3)$ and $\mathrm{ed}(\mathbf{E}_8; 2)$ are wide open, even for $k = \mathbb{C}$. For example, the best known lower bound on $\mathrm{ed}_{\mathbb{C}}(\mathbf{E}_8; 2)$ is 9 (see Corollary 3.6(i)) but the best upper bound I know is $\mathrm{ed}_{\mathbb{C}}(\mathbf{E}_8; 2) \leq 120$. The essential dimension $\mathrm{ed}(G)$ for these groups is largely uncharted territory, beyond the upper bounds in [Lem04].

7.9. Groups whose connected component is a torus. Let G be an algebraic group over k and p be a prime. We say that a linear representation $\phi\colon G \to \mathbf{GL}(V)$ is *p-faithful* (respectively, *p-generically free*) if $\mathrm{Ker}(\phi)$ is a finite group of order prime to p and ϕ descends to a faithful (respectively, generically free) representation of $G/\mathrm{Ker}(\phi)$.

Suppose the connected component G^0 of G is a k-torus. One reason such groups are of interest is that the normalizer G of a maximal torus in a reductive k-group Γ is of this form and $\mathrm{ed}(G)$ (respectively $\mathrm{ed}(G; p)$) is an upper bound on $\mathrm{ed}(\Gamma)$ (respectively, $\mathrm{ed}(\Gamma; p)$). The last assertion follows from [Se02, III.4.3, Lemma 6], in combination with Lemma 2.2.

For the sake of computing $\mathrm{ed}(G; p)$ we may assume that G/G^0 is a p-group and k is p-closed, i.e., the degree of every finite field extension k'/k is a power of p; see [LMMR09, Lemma 3.3]. It is shown in [LMMR09] that

$$\min \dim \nu - \dim(G) \leq \mathrm{ed}(G; p) \leq \min \dim \rho - \dim G\,, \tag{7.4}$$

where the two minima are taken over all p-faithful representations ν, and p-generically free representations ρ, respectively. In the case where $G = T$ is a torus or $G = F$ is a finite p-group or, more generally, G is a direct product $T \times F$, a faithful representation is automatically generically free. Thus in these

cases the lower and upper bounds of (7.4) coincide, yielding the exact value of $\mathrm{ed}_k(G;p)$. If we only assume that G^0 is a torus, I do not know how to close the gap between the lower and the upper bound in (7.4). However, in every example I have been able to work out the upper bound in (7.4) is, in fact, sharp.

Conjecture. ([LMMR09]) *Let G be an extension of a p-group by a torus, defined over a p-closed field k of characteristic $\neq p$. Then* $\mathrm{ed}(G;p) = \min \dim \rho - \dim G$*, where the minimum is taken over all p-generically free k-representations ρ of G.*

References

[Am72] S. A. Amitsur, *On central division algebras*, Israel J. Math. **12** (1972), 408–420.

[AP71] E. M. Andreev, V. L. Popov, *The stationary subgroups of points in general position in a representation space of a semisimple Lie group* (in Russian), Funkcional. Anal. i Priložen. **5** (1971), no. 4, 1–8. English Translation in Functional Anal. Appl. **5** (1971), 265–271.

[BF03] G. Berhuy and G. Favi, *Essential dimension: a functorial point of view (after A. Merkurjev)*, Doc. Math. **8** (2003), 279–330.

[BerR05] G. Berhuy, Z. Reichstein, *On the notion of canonical dimension for algebraic groups*, Adv. Math. **198** (2005), no. 1, 128–171.

[Br98] M. Brion, *The behaviour at infinity of the Bruhat decomposition*, Comment. Math. Helv. **73** (1998), no. 1, 137–174.

[BRV07] P. Brosnan, Z. Reichstein, A. Vistoli, *Essential dimension and algebraic stacks*, 2007, arXiv:math/0701903.

[BRV10a] P. Brosnan, Z. Reichstein, A. Vistoli, *Essential dimension, spinor groups and quadratic forms*, Annals of Math., **171**, no. 1 (2010), 533–544.

[BRV10b] P. Brosnan, Z. Reichstein, A. Vistoli, *Essential dimension of moduli of curves and other algebraic stacks*, with an appendix by N. Fakhruddin, J. European Math. Soc., to appear.

[BuR97] J. Buhler and Z. Reichstein, *On the essential dimension of a finite group*, Compositio Math. **106** (1997), no. 2, 159–179.

[CGR06] V. Chernousov, Ph. Gille, Z. Reichstein, *Resolving G-torsors by abelian base extensions*, J. Algebra **296** (2006), 561–581.

[CS06] V. Chernousov and J.-P. Serre, *Lower bounds for essential dimensions via orthogonal representations*, J. Algebra **305** (2006), no. 2, 1055–1070.

[CKM08] J.-L. Colliot-Thélène, N. A. Karpenko, A. S. Merkurjev, *Rational surfaces and the canonical dimension of the group* $\mathbf{PGL}_6$ (in Russian), Algebra i Analiz **19** (2007), no. 5, 159–178. English translation in St. Petersburg Math. J. **19** (2008), no. 5, 793–804.

[DI06] I. V. Dolgachev, V. A. Iskovskikh, *Finite subgroups of the plane Cremona group*, In Algebra, Arithmetic and Geometry: In honor of Yu. I. Manin, 443–549, vol. 1, Progress in Math. **269**, Springer-Verlag, 2009.

[Du09a] A. Duncan, *Finite Groups of Essential Dimension 2*, arXiv:0912.1644.

[Du09b] A. Duncan, *Essential Dimensions of* A_7 *and* S_7, Math. Res. Lett. **17** (2010), no. 2, 265–268.

[Fl07] M. Florence, *On the essential dimension of cyclic p-groups*, Inventiones Math., **171** (2007), 175–189.

[Gar01] S. Garibaldi, *Structurable Algebras and Groups of Type* E_6 *and* E_7, J. Algebra, **236** (2001), no. 2, 651–691.

[Gar09] S. Garibaldi, *Cohomological invariants: exceptional groups and spin groups*, with an appendix by D. W. Hoffmann, Memoirs of the American Mathematical Society **200** (2009), no. 937.

[GMS03] S. Garibaldi, A. Merkurjev, and J.-P. Serre, *Cohomological invariants in Galois cohomology*, University Lecture Series, vol. 28, American Mathematical Society, Providence, RI, 2003.

[GR07] Ph. Gille, Z. Reichstein, *A lower bound on the essential dimension of a connected linear group*, Comment. Math. Helv. **84** (2009), no. 1, 189–212.

[Gro58] A. Grothendieck, *Torsion homologique et sections rationnelles*, in: *Anneaux de Chow et Applications*, Séminaire C. Chevalley, 1958, exposé 5.

[Kahn00] B. Kahn, *Comparison of some field invariants*, J. Algebra, **232** (2000), no. 2, 485–492.

[Kar00] N. A. Karpenko, *On anisotropy of orthogonal involutions*, J. Ramanujan Math. Soc. **15** (2000), no. 1, 1–22.

[KM06] N. A. Karpenko, and A. S. Merkurjev, *Canonical p-dimension of algebraic groups*, Adv. Math. **205** (2006), no. 2, 410–433.

[KM07] N. Karpenko, A. Merkurjev, *Essential dimension of finite p-groups*, Inventiones Math., **172**, (2008), no. 3, 491–508.

[Kl1884] F. Klein, *Vorlesungen über das Ikosaeder und die Auflösung der Gleichungen vom 5ten Grade*, 1884.

[Ko00] V. E. Kordonskiĭ, *On the essential dimension and Serre's conjecture II for exceptional groups*, Mat. Zametki, **68**, (2000), no. 4, 539–547.

[KLS09] H. Kraft, R. Lötscher, G. M. Schwarz, *Compression of finite group actions and covariant dimension. II*, J. Algebra **322** (2009), no. 1, 94–107.

[Led02] A. Ledet, *On the essential dimension of some semi-direct products*, Canad. Math. Bull. **45** (2002), no. 3, 422–427.

[Led04] A. Ledet, *On the essential dimension of p-groups*, Galois theory and modular forms, 159–172, Dev. Math., 11, Kluwer Acad. Publ., Boston, MA, 2004.

[Lem04] N. Lemire, *Essential dimension of algebraic groups and integral representations of Weyl groups*, Transform. Groups **9** (2004), no. 4, 337–379.

[LRRS03] M. Lorenz, Z. Reichstein, L. H. Rowen, D. J. Saltman, *Fields of definition for division algebras*, J. London Math. Soc. (2) **68** (2003), no. 3, 651–670.

[Lö08] R. Lötscher, *Application of multihomogeneous covariants to the essential dimension of finite groups*, arXiv:0811.3852.

[LMMR09] R. Lötscher, M. MacDonald, A. Meyer, Z. Reichstein, *Essential p-dimension of algebraic tori*, arXiv:0910.5574.

[MacD08] M. MacDonald, *Cohomological invariants of odd degree Jordan algebras*, Math. Proc. Cambridge Philos. Soc. **145** (2008), no. 2, 295–303.

[Me96] A. S. Merkurjev, *Maximal indexes of Tits algebras*, Doc. Math. **1** (1996), no. 12, 229–243

[Me09] A. S Merkurjev, *Essential dimension*, in Quadratic forms – algebra, arithmetic, and geometry (R. Baeza, W.K. Chan, D.W. Hoffmann, and R. Schulze-Pillot, eds.), Contemporary Mathematics **493** (2009), 299–326.

[Me10a] A. S. Merkurjev, *Essential p-dimension of* $\mathbf{PGL}(p^2)$, *J. Amer. Math. Soc.* **23** (2010), 693–712.

[Me10b] A. Merkurjev *A lower bound on the essential dimension of simple algebras*, `http://www.math.ucla.edu/~merkurev/publicat.htm`.

[MR09a] A. Meyer, Z. Reichstein, *The essential dimension of the normalizer of a maximal torus in the projective linear group*, Algebra Number Theory, **3**, no. 4 (2009), 467–487.

[MR09b] A. Meyer, Z. Reichstein, *An upper bound on the essential dimension of a central simple algebra*, to appear in Journal of Algebra, arXiv:0907.4496

[Pf95] A. Pfister, *Quadratic forms with applications to geometry and topology*, Cambridge University Press, 1995.

[Po85] A. M. Popov, *Finite stationary subgroups in general position of simple linear Lie groups* (Russian), Trudy Moskov. Mat. Obshch. **48** (1985), 7–59.

English translation in *Transactions of the Moscow Mathematical Society*, A translation of Trudy Moskov. Mat. Obshch. 48 (1985). Trans. Moscow Math. Soc. 1986. American Mathematical Society, Providence, RI, 1986, 3–63.

[Pr67] C. Procesi, *Non-commutative affine rings*, Atti Acc. Naz. Lincei, S. VIII, v. VIII, fo. 6 (1967), 239–255.

[Pr09] Yu. Prokhorov, *Simple finite subgroups of the Cremona group of rank* 3, arXiv:0908.0678

[Rei99] Z. Reichstein, *On a theorem of Hermite and Joubert*, Can. J. Math. **51**, No.1, 69–95 (1999).

[Rei00] Z. Reichstein, *On the notion of essential dimension for algebraic groups*, Transform. Groups **5** (2000), no. 3, 265–304.

[Rei04] Z. Reichstein, *Compressions of group actions*, Invariant theory in all characteristics, 199–202, CRM Proc. Lecture Notes, 35, Amer. Math. Soc., Providence, RI, 2004.

[RY00] Z. Reichstein and B. Youssin, *Essential dimensions of algebraic groups and a resolution theorem for G-varieties*, Canad. J. Math. **52** (2000), no. 5, 1018–1056, With an appendix by János Kollár and Endre Szabó.

[Rost00] M. Rost, *Computation of some essential dimensions*, 2000, `http://www.math.uni-bielefeld.de/~rost/ed.html`

[Rost06] M. Rost, *On the Galois cohomology of* **Spin**(14), `http://www.mathematik.uni-bielefeld.de/~rost/spin-14.html#ed`

[RS92] L. H. Rowen, D. J. Saltman, *Prime-to-p extensions of division algebras*, Israel J. Math. **78** (1992), no. 2-3, 197–207.

[Ru10] A. Ruozzi, *Essential p-dimension of* **PGL**, `http://www.mathematik.uni-bielefeld.de/lag/man/385.html`

[Se58] J.-P. Serre, *Espaces fibrés algébriques*, in: *Anneaux de Chow et Applications*, Séminaire C. Chevalley, 1958, exposé 1. Reprinted in J.-P. Serre, *Exposés de séminaires* 1950–1999, deuxième édition, augmentée, Documents mathématiques **1**, Société mathématique de France 2008, 107–140.

[Se95] J.-P. Serre, *Cohomologie galoisienne: progrès et problèmes*, Séminaire Bourbaki, Vol. 1993/94. Astérisque **227** (1995), Exp. No. 783, 4, 229–257.

[Se02] J.-P. Serre, *Galois cohomology*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2002.

[Se08] J.-P. Serre, *Le group de cremona et ses sous-groupes finis*, Séminaire Bourbaki, 2008/2009, Exp. No. 1000.

[St75] R. Steinberg, *Torsion in reductive groups*, Advances in Math. **15** (1975), 63–92.

[Ti92] J. Tits, *Sur les degrés des extensions de corps déployant les groupes algébriques simples*, C. R. Acad. Sci. Paris, t. 315, Série I (1992), 1131–1138.

[TV10] D. Tossici, A. Vistoli, *On the essential dimension of infinitesimal group schemes*, arXiv:1001.3988.

[To05] B. Totaro, *The torsion index of E_8 and other groups*, Duke Math. J. **129** (2005), no. 2, 219–248.

[Woo89] J. A. Wood, *Spinor groups and algebraic coding theory*, J. Combin. Theory Ser. A **51** (1989), no. 2, 277–313.

[Zai07] K. Zainoulline, *Canonical p-dimensions of algebraic groups and degrees of basic polynomial invariants*, Bull. Lond. Math. Soc. **39** (2007), no. 2, 301–304.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Quadratic Forms, Galois Cohomology and Function Fields of p-adic Curves

V. Suresh*

Dedicated to my parents V. Narasimha Rao and V. Lakshmi

Abstract

Let k be a p-adic field and K a function field of a curve over k. It was proved in ([PS3]) that if $p \neq 2$, then the u-invariant of K is 8. Let l be a prime number not equal to p. Suppose that K contains a primitive l^{th} root of unity. It was also proved that every element in $H^3(K, \mathbb{Z}/l\mathbb{Z})$ is a symbol ([PS3]) and that every element in $H^2(K, \mathbb{Z}/l\mathbb{Z})$ is a sum of two symbols ([Su]). In this article we discuss these results and explain how the Galois cohomology methods used in the proof lead to consequences beyond the u-invariant computation.

Mathematics Subject Classification (2010). Primary 11E04, 11R34; Secondary 11G35, 14C25.

Keywords. Quadratic forms, Galois cohomology, u-invariant, p-adic curves.

Let k be a field of characteristic not equal to 2. By a *quadratic form* q over k we mean a homogeneous polynomial $q(X_1, \cdots, X_n) = \sum a_{ij} X_i X_j$ of degree 2 with $a_{ij} \in k$. The number of variables n is called the *dimension* of q. We say that a quadratic form q over k is *isotropic* if there exist $\lambda_1, \cdots, \lambda_n \in k$, not all zero, such that $q(\lambda_1, \cdots, \lambda_n) = 0$. If q is not isotropic, then q is called *anisotropic*. Kaplansky attached an invariant to a field k, known as u-invariant of k, denoted by $u(k)$, defined as the supremum of the dimensions of anisotropic quadratic forms over k. A theorem of Chevalley asserts that every quadratic form of dimension at least 3 over a finite field is isotropic: the u-invariant of a finite field is 2. It follows from Hensel's lemma that the u-invariant of a p-adic field is 4. A theorem of Tsen-Lang on C_i-fields yields that $u(k(T_1, \cdots, T_n))$ is 2^n if k is an algebraically closed field and 2^{n+1} if k is a finite field.

Kaplansky conjectured that the u-invariant of a field is always a power of 2. For a given integer $n \geq 3$, Merkurjev ([M]) constructed a field with u-invariant

*Department of Mathematics and Statistics, University of Hyderabad, Hyderabad, India 500046. E-mail: vssm@uohyd.ernet.in.

$2n$, thereby disproving the conjecture of Kaplansky. Later on, Izhboldin ([I]) constructed a field with u-invariant 9. The more recent results of Vishik ([V]) assert that there exist fields of u-invariant $2^n + 1$ for each $n \geq 3$.

The fields constructed by Merkurjev/Izhboldin are very large, obtained by taking iterated function fields of quadrics.

Conjecture. Let K be a finitely generated field extension of either $\mathbb{Q}$ or $\mathbb{Q}_p$. If $u(K)$ is finite, then $u(K)$ is a power of 2.

Let k be a p-adic field and K the function field of a curve over k. Kaplansky conjectured that $u(K) = 8$. If $p \neq 2$, it was proved in ([PS3]) that the u-invariant of K is 8. We shall discuss the Galois cohomology methods used in ([PS3]) and show how they lead to further consequences.

Let k be a global field or a local field and l a prime not equal to the characteristic of k. Class field asserts that every element in $H^2(k, \mu_l)$ is a symbol. Let k be either a global field of positive characteristic p or a p-adic field and l a prime not equal to p. Let K be a function field of a curve over k. In ([PS2], [PS3], [PS4]), we proved that if K contains a primitive l^{th} root of unity, then every element in $H^3(K, \mu_l)$ is a symbol. This result, for $l = 2$ and k a non-dyadic p-adic field, was used in the proof of the above conjecture of Kaplansky for non-dyadic p-adic field case. The main ingredient is a certain local global principle for elements of $H^3(K, \mu_l)$ in terms of symbols in $H^2(K, \mu_l)$. Let X be a smooth projective surface over a finite field $\mathbb{F}$. Let Y be a smooth projective 3-fold with a surjective morphism $Y \to X$ whose generic fibre is a smooth conic. The above local-global principle also leads to the fact that the unramified cohomology $H^3_{nr}(\mathbb{F}(Y)/Y, \mathbb{Q}_l/\mathbb{Z}_l(2))$ is zero. The vanishing of this unramified cohomology group was raised as a question by Colliot-Thélène ([CT3], [CT4]) in the context of the integral Tate conjecture on 1-cycles. We discuss these results in this article.

1. Quadratic Forms and Galois Cohomology Groups

We now recall some basic facts about the Witt group of quadratic forms over fields. We refer the reader to ([L], [Sc]).

Let k be a field of characteristic not equal to 2. Let q be a quadratic form over k. Then we know that there exists a linear change of variables such that $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2$ for some $a_1, \cdots, a_n \in k$. We say that q is *non-singular* if $a_i \neq 0$ for $1 \leq i \leq n$. Since every anisotropic quadratic form is non-singular and we are interested in anisotropic forms, from now on we assume that by a quadratic form we mean non-singular quadratic form. The quadratic form $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2$ is denoted by $< a_1, \cdots, a_n >$. Let $\mathbb{H} =< 1, -1 >$ be the hyperbolic plane (i.e. a dimension 2 isotropic quadratic form). Let q be a quadratic form over k. Then $q = q_0 \perp \mathbb{H}^r$ for some anisotropic

quadratic form q_0 over k and $r \geq 0$. Witt's cancellation theorem implies that the isometry class of q_0 is uniquely defined by the isometry class of q.

Let $W(k)$ denote the set of isometry classes of anisotropic quadratic forms over k. Let $q_1 \perp q_2 = q_0 \perp \mathbb{H}^r$ for some anisotropic quadratic form q_0 over k. We define the sum of the isometry classes of q_1 and q_2 as the isometry class of q_0. This makes $W(k)$ into an abelian group. The tensor product of two quadratic forms makes $W(k)$ into a commutative ring.

The dimension modulo 2 of a quadratic form defines a ring homomorphism $e_0 : W(k) \to \mathbb{Z}/2$. Let $I(k)$ be the kernel of this homomorphism. Then $I(k)$ is the ideal of $W(k)$ consisting of all the even dimension forms.

Let $q =< a_1, \cdots, a_n >$ be a quadratic form over k. The *discriminant* $\mathrm{disc}(q)$ of q is defined as the class of $(-1)^{n(n+1)/2} a_1 \cdots a_n$ in k^*/k^{*2}. This gives a homomorphism $e_1 : I(k) \to k^*/k^{*2}$. The kernel of this homomorphism is $I^2(k) = I(k)^2$. Since every element in $I(k)$ is represented by an even dimension form, it follows that $I^2(k)$ is generated by the classes of quadratic forms $<1, a><1, b>$ with $a, b \in k^*$.

Let $Br(k)$ be the Brauer group of k and ${}_2Br(k)$ the 2-torsion subgroup of $Br(k)$. For a quadratic form q over k, let $C(q)$ be the Clifford algebra associated to q. Then we have a well defined homomorphism $e_2 : I^2(k) \to {}_2Br(k)$ given by $e_2(q) = C(q)$.

Let k be a field and l a prime not equal to the characteristic of k. Let μ_l be the group of lth roots of unity. For $i \geq 1$, let $\mu_l^{\otimes i}$ be the Galois module given by the tensor product of i copies of μ_l. For $n \geq 0$, let $H^n(k, \mu_l^{\otimes i})$ be the nth Galois cohomology group with coefficients in $\mu_l^{\otimes i}$.

We have the Kummer isomorphism $k^*/k^{*l} \simeq H^1(k, \mu_l)$. For $a \in k^*$, its class in $H^1(k, \mu_l)$ is denoted by (a). If $a_1, \cdots, a_n \in k^*$, the cup product $(a_1) \cdots (a_n) \in H^n(k, \mu_l^{\otimes n})$ is called a *symbol*. We have an isomorphism of $H^2(k, \mu_l)$ with the l-torsion subgroup ${}_lBr(k)$ of the Brauer group of k. We define the *index* of an element $\alpha \in H^2(k, \mu_l)$ to be the index of the corresponding central simple algebra in ${}_lBr(k)$.

Assume that k contains a primitive l^{th} root of unity. We fix a generator ρ for the cyclic group μ_l and identify the group μ_l with $\mathbb{Z}/l\mathbb{Z}$ as Galois modules. This leads to an identification of $H^1(k, \mathbb{Z}/l\mathbb{Z})$ with k^*/k^{*l} and identification of $H^n(k, \mu_l^{\otimes m})$ with $H^n(k, \mathbb{Z}/l\mathbb{Z})$ for all n and m. The element in $H^n(k, \mathbb{Z}/l\mathbb{Z})$ corresponding to the symbol $(a_1) \cdots (a_n) \in H^n(k, \mu_l^{\otimes n})$ through this identification is again denoted by $(a_1) \cdots (a_n)$.

For $a_1, \cdots, a_n \in k^*$, let $<< a_1, \cdots, a_n >>$ denote the n-fold Pfister form given by the tensor product of quadratic forms $< 1, -a_i >$ for $1 \leq i \leq n$. Let P_n be the set of isometry classes of n-fold Pfister forms. There is a well-defined map due to Arason,

$$\tilde{e}_n : P_n(k) \to H^n(k, \mathbb{Z}/2\mathbb{Z})$$

given by

$$\tilde{e}_n(<< a_1, \cdots, a_n >>) = (a_1) \cup (a_2) \cup \cdots \cup (a_n).$$

Milnor's Conjecture: *(Quadratic form version)* The maps $\tilde{e}_n$ extend to homomorphisms

$$e_n : I^n(k) \to H^n(k, \mathbb{Z}/2\mathbb{Z}).$$

which are surjective with kernel $I^{n+1}(k)$.

In other words, the maps (e_n) induce an isomorphism

$$\bigoplus_{n\geq 0} I^n(k)/I^{n+1}(k) \simeq \bigoplus_{n\geq 0} H^n(k, \mathbb{Z}/2\mathbb{Z}).$$

Milnor's conjecture has been proved by Orlov, Voevodsky and Vishik ([OVV]).

We now recall a few basic definitions and facts about Galois cohomology groups and residue homomorphisms. We refer the reader to ([CT1]).

Throughout this article by a discrete valuation we mean a discrete valuation of rank one. Let K be a field with a discrete valuation ν. Let $\kappa(\nu)$ be the residue field at ν. If l is a prime not equal to $\mathrm{char}(\kappa(\nu))$, then there is a residue homomorphism $\partial_\nu : H^n(K, \mu_l^{\otimes m}) \to H^{n-1}(\kappa(\nu), \mu_l^{\otimes m-1})$. Suppose that K contains a primitive l^{th} root of unity. Then $\kappa(\nu)$ also contains a primitive l^{th} root of unity. By fixing a suitable primitive l^{th} root of unity in K and its image in $\kappa(\nu)$, as mentioned above, we identify $H^n(K, \mu_l^{\otimes m})$ with $H^n(K, \mathbb{Z}/l\mathbb{Z})$ and $H^n(\kappa(\nu), \mu_l^{\otimes m})$ with $H^n(\kappa(\nu), \mathbb{Z}/l\mathbb{Z})$.

Let $\mathcal{X}$ be a regular integral scheme of dimension d, with field of fractions K. Let $\mathcal{X}^1$ be the set of codimension one points of $\mathcal{X}$. A point $x \in \mathcal{X}^1$ gives rise to a discrete valuation ν_x on K. The residue field of this discrete valuation ring is denoted by $\kappa(x)$. The corresponding residue homomorphism is denoted by ∂_x. We say that an element $\zeta \in H^n(K, \mu_l^{\otimes m})$ is *unramified* at x if $\partial_x(\zeta) = 0$; otherwise it is said to be *ramified* at x. We define the ramification divisor $\mathrm{ram}_{\mathcal{X}}(\zeta) = \sum x$ as x runs over $\mathcal{X}^1$ where ζ is ramified. The nth unramified cohomology on $\mathcal{X}$, denoted by $H^n_{nr}(K/\mathcal{X}, \mu_l^{\otimes m})$, is defined as the intersection of kernels of the residue homomorphisms $\partial_x : H^n(K, \mu_l^{\otimes m}) \to H^{n-1}(\kappa(x), \mu_l^{\otimes (m-1)})$, x running over $\mathcal{X}^1$. We say that $\zeta \in H^n(K, \mu_l^{\otimes m})$ is *unramified on* $\mathcal{X}$ if $\zeta \in H^n_{nr}(K/\mathcal{X}, \mu_l^{\otimes m})$. Suppose C is an irreducible closed subscheme of $\mathcal{X}$ of codimension 1. Then the generic point x of C belongs to $\mathcal{X}^1$ and we set $\partial_x = \partial_C$. If $\alpha \in H^n(K, \mu_l^{\otimes m})$ is unramified at x, then we say that α is *unramified* at C. The group of elements of $H^n(K, \mu_l^{\otimes m})$ which are unramified at every discrete valuation of K is denoted by $H^n_{nr}(K, \mu_l^{\otimes m})$. If C is an integral curve (not necessarily regular) with function field $\kappa(C)$, $H^n_{nr}(\kappa(C)/C, \mu_l)$ denotes the subgroup of $H^n(\kappa(C), \mu_l)$ consisting of those elements which are unramified at all those discrete valuation which are centered on a closed point of C. If K contains a primitive l^{th} root of unity, then we also denote $H^n_{nr}(K/\mathcal{X}, \mu_l^{\otimes m})$ by $H^n_{nr}(K/\mathcal{X}, \mathbb{Z}/l\mathbb{Z})$ and $H^n_{nr}(K, \mu_l^{\otimes m})$ by $H^n_{nr}(K, \mathbb{Z}/l\mathbb{Z})$.

2. Galois Cohomology Groups of Function Fields of Surfaces

Let $\mathcal{X}$ be a regular, integral surface. Let K be the function field of $\mathcal{X}$. Let l be a prime not equal to the characteristic of K. Suppose that l is a unit in $\mathcal{O}_{\mathcal{X}}$ and K contains a primitive l^{th} root of unity. Let $\mathcal{X}^1$ be the set of all codimension one points of $\mathcal{X}$. For $x \in \mathcal{X}^1$, let $\kappa(x)$ denote the residue field at x and K_x the completion of K with respect to the discrete valuation ν_x given by x.

Suppose that for every irreducible closed curve C on $\mathcal{X}$, $\kappa(C)$ is either a global field or a local field. Then we have the following (cf. [PS4]):

Theorem 2.1. *Let $\zeta \in H^3(K, \mathbb{Z}/l\mathbb{Z})$ and $\alpha \in H^2(K, \mathbb{Z}/l\mathbb{Z})$. Suppose that the central division algebra represented by α has degree l. If for every $x \in \mathcal{X}^1$, there exists $f_x \in K_x^*$ such that $\zeta - (\alpha \cdot (f_x)) \in H^3_{nr}(K_x, \mathbb{Z}/l\mathbb{Z})$, then there exists $f \in K^*$ such that $\zeta - (\alpha \cdot (f)) \in H^3_{nr}(K/\mathcal{X}, \mathbb{Z}/l\mathbb{Z})$.*

Corollary 2.2. *Let k be a p-adic field or a global field of positive characteristic p and K a function field in one variable over k. Let l be a prime not equal to p. Suppose that k contains a primitive l^{th} root of unity. Let $\zeta \in H^3(K, \mathbb{Z}/l\mathbb{Z})$ and $\alpha \in H^2(K, \mathbb{Z}/l\mathbb{Z})$. Suppose that the central division algebra represented by α has degree l. If for every $x \in \mathcal{X}^1$, there exists $f_x \in K_x^*$ such that $\zeta = \alpha \cdot (f_x) \in H^3(K_x, \mathbb{Z}/l\mathbb{Z})$, then there exists $f \in K^*$ such that $\zeta = \alpha \cdot (f) \in H^3(K, \mathbb{Z}/l\mathbb{Z})$.*

Proof. If k is a p-adic field, let $\mathcal{O}$ be the ring of integers in k. If k is a global field of characteristic p, let $\mathcal{O}$ be the field of constants in k. Then there exists an integral, regular surface $\mathcal{X}$ which is projective over $\mathrm{Spec}(\mathcal{O})$ with function field K. Let C be any closed curve on $\mathcal{X}$. Then $\kappa(C)$ is either a p-adic field or a global field of positive characteristic. In both cases, by class field theory, we have $H^2_{nr}(\kappa(C)/C, \mathbb{Z}/l\mathbb{Z}) = 0$. Let $x \in \mathcal{X}^1$. By (2.1), there exists $f \in K^*$ such that $\zeta - (\alpha \cdot (f)) \in H^3_{nr}(K/\mathcal{X}, \mathbb{Z}/l\mathbb{Z})$. By ([CSS], [Ka]), $H^3_{nr}(K/\mathcal{X}, \mathbb{Z}/l\mathbb{Z}) = 0$. Hence $\zeta = \alpha \cdot (f) \in H^3(K, \mathbb{Z}/l\mathbb{Z})$. $\square$

For the function field of a curve over a p-adic field, the above corollary was proved in ([PS3]). This leads to the following (cf. [PS3], [PS4]):

Theorem 2.3. *Let k be a p-adic field or a global field of positive characteristic p. Let l be a prime not equal to p. Assume that k contains a primitive l^{th} root of unity. Let K be a function field in one variable over k. Then every element in $H^3(K, \mathbb{Z}/l\mathbb{Z})$ is a symbol.*

Let k be a p-adic field and K the function field of a p-adic curve over k. Let l be a prime not equal to p. Local class field theory asserts that every element in $H^2(k, \mathbb{Z}/l\mathbb{Z})$ is a symbol and $H^n(k, \mathbb{Z}/l\mathbb{Z}) = 0$ for $n \geq 3$. The following is a higher dimensional analogue of these results for function fields of curves over p-adic fields ([Su]).

Theorem 2.4. *Let k be a p-adic field and K a function field in one variable over k. Let l be a prime not equal to p. Suppose that k contains a primitive l^{th} root of unity. Then every element in $H^2(K, \mathbb{Z}/l\mathbb{Z})$ is a sum of at most 2 symbols.*

The proof of the above theorem uses in a fundamental way a theorem of Saltman on splitting ramification of algebras over such fields ([S3]).

3. The u-invariant

In ([PS3]), we have given sufficient conditions for a field to a have u-invariant at most 8. The following is a slight modification of this result.

Theorem 3.1. *Let K be a field of characteristic not equal to 2. Assume the following:*

1. *Every element in $H^2(K, \mu_2)$ is a sum of at most 2 symbols.*
2. *Every element in $I^3(K)$ is equal to a 3-fold Pfister form.*
3. *If ϕ is a 3-fold Pfister form and q_2 is a quadratic form over K of dimension 2, then $\phi =< 1, f >< 1, g >< 1, h >$ for some $f, g, h \in K^*$ with f a value of q_2.*
4. *If $\phi =< 1, f >< 1, a >< 1, b >$ is a 3-fold Pfister form and q_3 a quadratic form over K of dimension 3, then $\phi =< 1, f >< 1, g >< 1, h >$ for some $g, h \in K^*$ with g a value of q_3.*

Then $u(K) \leq 8$.

Proof. In ([PS3]), we proved this with the additional assumption that $I^4(K) = 0$. We show that condition 3) implies that $I^4(K) = 0$. Let $q =< 1, a >< 1, b >< 1, c >< 1, d >$. By assumption (3), we have $< 1, b >< 1, c >< 1, d >=< 1, f >< 1, g >< 1, h >$ for some $f, g, h \in K^*$ with f a value of the quadratic form $< -1, -a >$. Since f is a value of $< -1, -a >$, we have $< 1, a >< 1, f >= 0 \in W(K)$. In particular, $q =< 1, a >< 1, f >< 1, g >< 1, h >= 0$. Since $I^4(K)$ is generated by elements of the form $< 1, a >< 1, b >< 1, c >< 1, d >$, we have $I^4(K) = 0$. □

In ([PS3], we have also shown that the above conditions, except the first one, are also necessary for a field to have the u-invariant equal to 8.

Let K be a function field of a curve over a p-adic field. Assume that $p \neq 2$. Let ν be a discrete valuation of K and K_ν be the completion of K at ν. Let $\kappa(\nu)$ be the residue field at ν. Then $\kappa(\nu)$ is either a p-adic field or a function field of a curve over a finite field. In both cases we have $u(\kappa(\nu)) = 4$. Using a theorem of Springer, we get that $u(K_\nu) = 8$. Thus K_ν satisfies the conditions in the above theorem except possibly the first one. Using the local-global principle stated in

the previous section, we prove that K satisfies all the conditions of the above theorem except the first condition. A theorem of Saltman ([S1], [S2]) asserts that K also satisfies the first condition. Hence we conclude the following:

Corollary 3.2. *Let k be a p-adic field and K the function field of a curve over k. If $p \neq 2$, then $u(K) = 8$.*

Using the patching methods developed in ([HH]), Harbater-Hartman-Krashen proved the following in ([HHK]).

Theorem 3.3. *Let K be a complete discrete valuated field with residue field κ. Suppose that there exists an integer n such that for every finite extension L of K, $u(L) \leq n$ and for every function field $\kappa(C)$ of a curve C over κ, $u(\kappa(C)) \leq n$. If $char(\kappa) \neq 2$, then for any function field F of a curve over K, $u(F) \leq 2n$.*

Using the above mentioned patching methods, we proved the following local-global principle for isotropy of quadratic forms ([CTPS]):

Theorem 3.4. *Let K be a complete discrete valuated field with residue field κ. Suppose that $char(\kappa) \neq 2$. Let q be a quadratic form over K of dimension at least 3. If q is isotropic over K_ν for every discrete valuation ν of K, then q is isotropic over K.*

Using a theorem of Heath-Brown ([HB1], [HB2]) on the common zeroes of a system of quadratic form over p-adic field, Leep ([L]) proved the following theorem which holds also when $p = 2$.

Theorem 3.5. *Let k be any p-adic field and K a function field in n-variables over k. Then $u(K) = 2^{n+2}$.*

For function fields over p-adic fields this completely solves Kaplansky's conjecture. The next extremely interesting case to explore is the case of function fields of curves over totally imaginary number fields.

4. The Chow Group of 0-cycles

Let X be a smooth, projective, geometrically integral variety over a field k. Let $CH_0(X)$ be the Chow group of 0-cycles modulo rational equivalence. If X is a curve over a number field or a local field, the structure of $CH_0(X)$ is well understood. Very little is known about the structure of this group for general varieties over number fields or local field.

Let k be a field and C a smooth projective geometrically integral curve over k. Let $\pi : X \to C$ be a dominant morphism. Let $\pi_* : CH_0(X) \to CH_0(C)$ be the induced morphism. Let $CH_0(X/C)$ be the kernel of π_*. Since the structure of $CH_0(C)$ is well understood, to study the structure of $CH_0(X)$, one is led to the study of the group $CH_0(X/C)$.

We now describe a characterisation of $CH_0(X/C)$ as a subquotient of the group of units $k(C)^*$ due to Colliot-Thélène and Skorobogatov ([CTS]. Let k be a field of characteristic not equal to 2 and C a smooth projective geometrically integral curve over k. Let $\pi : X \to C$ be an admissible quadric fibration over k (c.f. [CTS]). Let q be a quadratic form over $k(C)$ defining the generic fibre of π. Let $N_q(k(C))$ be the subgroup of $k(C)^*$ generated by ab where a, b are values of the quadratic form q over $k(C)^*$. Let $k(C)_{dn}$ be the subgroup of $k(C)^*$ consisting of elements $f \in k(C)^*$ such that for every closed point P of C, f can be written as a product of a unit at P and an element in $N_q(k(C))$. Then $CH_0(X/C) \simeq k(C)^*_{dn}/k^*N_q(k(C))$ ([CTS]). We have the following ([PS1]):

Theorem 4.1. *Let k be a p-adic field and C a smooth projective curve over k. Let $X \to C$ be an admissible quadric fibration. Then $CH_0(X/C)$ is a finite group.*

The above theorem was proved in ([CTS]) for $\dim(X) = 2, 3$. We have the following:

Theorem 4.2. *Let k be a finitely generated extension of $\mathbb{Q}_p$ of transcendence degree $d \geq 0$ and C a smooth projective curve over k. If $X \to C$ is an admissible quadric fibration and $dim(X) \geq 2^{d+2}$, then $CH_0(X/C) = 0$.*

Proof. Let q be a quadratic form over $k(C)$ defining the generic fibre of the quadric fibration $X \to C$. Since $\dim(X) \geq 2^{d+2}$, we have $\dim(q) \geq 2^{d+2} + 1$. Let $f \in k(C)^*$. Since, by (3.5) $u(k(C)) = 2^{d+3}$, the quadratic form $<1, f> \otimes q$ is isotropic. In particular $f \in N_q(k(C)$. Hence $N_q(k(C) = k(C)^*_{dn} = k(C)^*$ and $CH_0(X/C) = 0$. □

The above result for $p \neq 2$ and $d = 0$ was proved in ([PS2]).

We now discuss connections with the integral Tate conjecture. We refer the reader to Colloit-Thélène's expositions on this topic ([CT3], [CT4])

Let $\mathbb{F}$ be a finite field and X a smooth, projective, geometrically integral variety over $\mathbb{F}$ of dimension d. Let l be a prime not equal to the characteristic of $\mathbb{F}$. We have the *cycle* map $CH^i(X) \otimes_{\mathbb{Z}} \mathbb{Z}_l \to H^{2i}(X, \mathbb{Z}_l(i))$. The integral version of Tate Conjecture states that this map is surjective.

Let C be a smooth, projective, geometrically integral curve over $\mathbb{F}$. Let X be a smooth, projective, geometrically integral variety over $\mathbb{F}$ of dimension $d + 1$ with a flat morphism $X \to C$ with generic fibre $X_\eta/\mathbb{F}(C)$ smooth and geometrically integral. A theorem of Saito/Colliot-Thélène ([Sa], [C2]) asserts that if the cycle map $CH^d(X) \otimes \mathbb{Z}_l \to H^{2d}(X, \mathbb{Z}_l(d))$ is onto, then the Brauer-Manin obstruction is the only obstruction to the local-global principle for the existence of zero-cycles of degree 1 on X_η.

For a certain class $B_{Tate}(\mathbb{F})$ of smooth projective varieties Y, Bruno Kahn ([K]) showed that the cycle map $CH^2(Y) \otimes \mathbb{Z}_l \to H^4(Y, \mathbb{Z}_l(2)$ is onto if and only if $H^3_{nr}(\mathbb{F}(Y)/Y, \mathbb{Q}_l/\mathbb{Z}_l(2)) = 0$. Suppose Y is a smooth projective variety with a

dominant morphism $Y \to X$, where X is a smooth geometrically ruled surface, with generic fibre a smooth conic over $\mathbb{F}(X)$. Then Y belongs to $B_{Tate}(\mathbb{F})$ ([So]).

Let X be a smooth, projective, geometrically integral surface over a finite field $\mathbb{F}$ of characteristic not equal to 2. Let Y be a smooth, projective, geometrically integral variety over $\mathbb{F}$ and a dominant morphism $Y \to X$ with generic fibre a smooth projective conic over $\mathbb{F}(X)$. In [PS4], we have shown that the vanishing of $H^3_{nr}(\mathbb{F}(Y)/Y, \mu_2)$ is equivalent to the local-global principle discussed in (2.1). This leads to the vanishing of $H^3_{nr}(\mathbb{F}(Y)/Y, \mu_2)$, answering a question of Colliot-Thélène ([CT3], [CT4]), leading to the following:

Theorem 4.3. *Let $\mathbb{F}$ be a finite field of characteristic not equal to 2. Let l be a prime not equal to the characteristic of $\mathbb{F}$. Let X be a smooth, projective, geometrically ruled surface over $\mathbb{F}$ and $Y \to X$ a surjective morphism with generic fibre a smooth conic over $\mathbb{F}(X)$. Then the cycle map $CH^2(Y) \otimes \mathbb{Z}_l \to H^4(Y, \mathbb{Z}_l(2))$ is onto.*

Corollary 4.4. *Let $\mathbb{F}$ be a finite field of characteristic not equal to 2. Let C be a smooth, projective, geometrically integral curve over $\mathbb{F}$ and Y a smooth, geometrically integral 3-fold with a dominant morphism $Y \to C \times \mathbf{P}^1$. Let Y_η be the generic fibre of the composite morphism $Y \to C \times \mathbf{P}^1 \to C$. Then the Brauer-Manin obstruction is the only obstruction to the existence of zero-cycles of degree one on Y_η.*

References

[A] Arason, J.K., *Cohomologische Invarianten quadratischer Formen, J. Algbera* **36** (1975), 448–491.

[AEJ] Arason, J.K., Elman, R. and Jacob, B., *Fields of cohomological 2-dimension three, Math. Ann.* **274** (1986), 649–657 .

[CT1] Colliot-Thélène, J.-L., *Birational invariants, purity, and the Gresten conjecture, Proceedings of Symposia in Pure Math.* **55**, Part 1, 1–64.

[CT2] Colliot-Thélène, J.-L., *Conjectures de type local-global sur l'image des groupes de Chow dans la cohomologie étale*, Algebraic K-theory (Seattle, WA, 1997), Proc. Sympos. Pure Math., **67**, 1–12

[CT3] Colliot-Thélène, J.-L., *Local-global principles for zero-cycles of degree one and integral Tate conjecture for 1-cycles*, talk at the workshop Anabelian Geometry Newton Institute, Cambridge, 24-28 August 2008, http://www.math.u-psud.fr/∼colliot/expocambridge240809.pdf

[CT4] Colliot-Thélène, J.-L., *Local-global principle for zero-cycles of degree one and integral Tate conjecture for 1-cycles*, Workshop on motives, Tokyo, 14–18 December 2009, http://www.math.u-psud.fr/∼colliot/beamexpotokyo.pdf

[CTPS] Colliot-Thélène, J.-L., Parimala, R and Suresh, V, Patching and local-global principles for homogeneous spaces over function fields of p-adic fields, to appear in Commentarii Mathematici Helvetici.

[CTSS] Colliot-Thélène, J.-L., Sansuc, J.-J. and Soulé, C.. *Torsion dans le groupe de Chow de codimension deux*, Duke Math.J. **50** (1983) 763–801.

[CTS] Colliot-Thélène, J.-L ét Skorobogatov, A.N., *Groupes de Chow des zéro-cycles des fibrés en quadriques*, Journal of K-theory **7** (1993) 477–500.

[HH] Harbater, D. and Hartmann, J., *Patching over fields*, Israel J. Math. **176** (2010), 61–108.

[HHK] Harbater, D., Hartmann, J. and Krashen, D., *Applications of patching to quadratic forms and central simple algebras*, Invent. Math. **178** (2009), 231–263.

[HB1] Heath-Brown, D.R., *Zeros of systems of p-adic quadratic forms*, to appear in Composito Math.

[HB2] Heath-Brown, D.R., *Artin's Conjecture on Zeros of p-Adic forms*, arxiv:1002.3754v1.

[I] Izhboldin, O.T., *Fields of u-invariant* 9, Ann. of Math. **154** (2001), no. 3, 529–587

[K] Kahn, B., *Équivalences rationnelle et numérique sur certaines variétés de type abélien sur un corps fini*, Ann. Sci. École Norm. Sup. (4) **36** (2003), no. 6, 977–1002 (2004).

[K] Kato, K., *A Hasse principle for two-dimensional global fields, J. reine Angew. Math.* **366** (1986), 142–181.

[L] Lam, T.Y., *Introduction to quadratic forms over fields*, GSM 67, American Mathematical Society, 2004.

[Le] Leep, D.B., *The u-invariant of p-adic function fields*, preprint.

[M] Merkurjev, A.S., *Simple algebras and quadratic forms.* (Russian) Izv. Akad. Nauk SSSR Ser. Mat. **55** (1991), no. 1, 218–224; translation in Math. USSR-Izv. **38** (1992), no. 1, 215–221.

[Mi] Milne, J.S., *Étale Cohomology*, Princeton University Press, Princeton, New Jersey 1980.

[OVV] Orlov, D., Vishik, A. and Voevodsky, V., *An exact sequence for $K_*^M/2$ with applications to quadratic forms*, Ann. of Math. **165** (2007), no. 1, 1–13.

[PS1] Parimala, R. and Suresh, V., *Zero-cycles on quadric fibrations: finiteness theorems and the cycle map*, Invent. Math. **122** (1995), 83–117.

[PS2] Parimala, R. and Suresh, V., *Isotropy of quadratic forms over function fields in one variable over p-adic fields, Publ. de I.H.É.S.* **88** (1998), 129–150.

[PS3] Parimala, R. and Suresh, V., *The u-invariant of the function fields of p-adic curves*, to appear in Annals of Mathematics.

[PS4] Parimala, R. and Suresh, V., *Degree three cohomology of function fields of arithmetic surfaces*, preprint 2010.

[Sa] Saito, S., *Some observations on motivic cohomology of arithmetic schemes*, Invent. Math. **98** (1989), 371–404.

[S1] Saltman, D.J., *Division Algebras over p-adic curves, J. Ramanujan Math. Soc.* **12** (1997), 25–47.

[S2] Saltman, D.J., *Correction to Division algebras over p-adic curves, J. Ramanujan Math. Soc.* **13** (1998), 125–130.

[S3] Saltman, D.J., *Cyclic Algebras over p-adic curves, J. Algebra* **314** (2007) 817–843.

[S4] Saltman, D.J., *Division algebras over surfaces, J. Algebra* **320** (2008), 1543–1585.

[Sc] Scharlau, W., *Quadratic and Hermitian Forms*, Grundlehren der Math. Wiss., Vol. 270, Berlin, Heidelberg, New York 1985.

[Sh] Shafarevich, I. R., *Lectures on Minimal Models and Birational Transformations of two Dimensional Schemes*, Tata Institute of Fundamental Research, (1966).

[So] Soulé, C., *Groupes de Chow et K-théorie de variétés sur un corps fini*, Math. Ann. **268** (1984), no. 3, 317–345.

[Su] Suresh, V., *Bounding the symbol length in the Galois cohomology of function field of p-adic curves*, to appear in Comm. Math. Helv.

[V] Vishik, A., *Fields of u-invariant* 2^r+1, "Algebra, Arithmetic and Geometry - Manin Festschrift", Birkhauser, 2007.

Section 3

Number Theory

Christophe Breuil
The Emerging p-adic Langlands Programme 203

Ralph Greenberg
Selmer Groups and Congruences 231

D.R. Heath-Brown
Artin's Conjecture on Zeros of p-adic Forms 249

Kiran Sridhara Kedlaya
Relative p-adic Hodge Theory and Rapoport-Zink Period Domains 258

Chandrashekhar Khare* **and Jean-Pierre Wintenberger***
Serre's Modularity Conjecture 280

Mark Kisin
The Structure of Potentially Semi-stable Deformation Rings 294

Sophie Morel
The Intersection Complex as a Weight Truncation and an
Application to Shimura Varieties 312

Takeshi Saito
Wild Ramification of Schemes and Sheaves 335

K. Soundararajan
Quantum Unique Ergodicity and Number Theory 357

Akshay Venkatesh* **and Jordan S. Ellenberg**
Statistics of Number Fields and Function Fields 383

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

The Emerging p-adic Langlands Programme

Christophe Breuil*

Abstract

We give a brief overview of some aspects of the p-adic and modulo p Langlands programmes.

Mathematics Subject Classification (2010). Primary 11S80; Secondary 22D12.

Keywords. p-adic Langlands programme, p-adic Hodge theory, $\mathrm{GL}_2(\mathbb{Q}_p)$, (φ, Γ)-modules, completed cohomology.

1. Introduction

Fix p a prime number and $\overline{\mathbb{Q}_p}$ an algebraic closure of the field of p-adic numbers $\mathbb{Q}_p$. Let $\ell \neq p$ be another prime number and $\overline{\mathbb{Q}_\ell}$ an algebraic closure of $\mathbb{Q}_\ell$. If F is a field which is a finite extension of $\mathbb{Q}_p$ and n a positive integer, the celebrated local Langlands programme for GL_n ([48], [37], [38]) establishes a "natural" $1-1$ correspondence between certain $\overline{\mathbb{Q}_\ell}$-linear continuous representations ρ of the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ on n-dimensional $\overline{\mathbb{Q}_\ell}$-vector spaces and certain $\overline{\mathbb{Q}_\ell}$-linear locally constant (or smooth) irreducible representations π of $\mathrm{GL}_n(F)$ on (usually infinite dimensional) $\overline{\mathbb{Q}_\ell}$-vector spaces. This local correspondence is moreover compatible with reduction modulo ℓ ([68]) and with cohomology ([49], [23], [37]). By "compatible with cohomology", we mean here that there exist towers of algebraic (Shimura) varieties $(S(K))_K$ over F of dimension d indexed by compact open subgroups K of $\mathrm{GL}_n(F)$ on which $\mathrm{GL}_n(F)$ acts on the right and such that the natural action of $\mathrm{GL}_n(F) \times \mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ on the inductive limit of ℓ-adic étale cohomology groups:

$$\varinjlim_{K} H^d_{\text{ét}}(S(K) \times_F \overline{\mathbb{Q}_p}, \overline{\mathbb{Q}_\ell}) \tag{1}$$

*I thank David Savitt for his corrections and comments on a first version.
C.N.R.S. & I.H.É.S., 35 route de Chartres, 91440 Bures-sur-Yvette, France.
E-mail: breuil@ihes.fr.

makes it a direct sum of representations $\pi\otimes\rho$ where ρ matches π by the previous local correspondence. (One can also take étale cohomology with values in certain locally constant sheaves of finite dimensional $\overline{\mathbb{Q}_\ell}$-vector spaces.)

Now $\mathrm{GL}_n(F)$ is a topological group (even a p-adic Lie group) and by [69] one can replace the above locally constant irreducible representations π of $\mathrm{GL}_n(F)$ on $\overline{\mathbb{Q}_\ell}$-vector spaces by continuous topologically irreducible representations $\widehat{\pi}$ of $\mathrm{GL}_n(F)$ on ℓ-adic Banach spaces (by a completion process which turns out to be reversible). This is quite natural as it gives now a $1-1$ correspondence between two kinds of continuous ℓ-adic representations. The original aim of the local p-adic Langlands programme is to look for a possible p-adic analogue of this ℓ-adic correspondence, that is:

Can one match certain linear continuous representations ρ of the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ on n-dimensional $\overline{\mathbb{Q}_p}$-vector spaces to certain linear continuous representations $\widehat{\pi}$ of $\mathrm{GL}_n(F)$ on p-adic Banach spaces, in a way that is compatible with reduction modulo p, with cohomology, and also with "p-adic families"?

It turns out that such a nice p-adic correspondence indeed exists between 2-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ and certain continuous representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ on "unitary p-adic Banach spaces" (that is, with an invariant norm) which satisfies all of the above requirements. Based on the work of precursors ([50], [1], [2]) and on the papers [67], [59], [60], [61], the first cases were discovered and studied by the author in [6], [7], [8], [9], [15] and a partial programme was stated for $\mathrm{GL}_2(\mathbb{Q}_p)$ in [8]. The local p-adic correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$, together with its compatibility with "p-adic families" and with reduction modulo p, was then fully developed, essentially by Colmez, in the papers [19], [5], [3], [20], [21] after Colmez discovered that the theory of (φ,Γ)-modules was a fundamental intermediary between the representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ and the representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ (see Berger's Bourbaki talk [4] and [12] for a historical account). These local results already have had important global applications by work of Kisin ([45]) and Emerton ([28]) as, combined with deformations techniques, they provide an almost complete proof of the Fontaine-Mazur conjecture ([31]). Finally, the important compatibility with cohomology is currently being written in [28]. Note that the relevant cohomology in that setting is not (1) but rather its p-adic completion, which is a much more intricate representation. Such p-adically completed cohomology spaces were introduced by Emerton in [24] (although some cases had been considered before, see, e.g., [51]). Their study as continuous representations of $\mathrm{GL}_n(F)\times\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ seems a mammoth task which is sometimes called the "global p-adic Langlands programme" (as these cohomology spaces are of a global nature). We sum up some of these results for $\mathrm{GL}_2(\mathbb{Q}_p)$ in §2.

At about the same time as the p-adic and modulo p theories for $\mathrm{GL}_2(\mathbb{Q}_p)$ were definitely flourishing, the theory modulo p for $\mathrm{GL}_2(F)$ and $F\neq\mathbb{Q}_p$ was discovered in [16], much to the surprise of everybody, to be much more involved. Although nothing really different happens on the Galois side when one goes from

$\mathbb{Q}_p$ to F, the complications on the GL_2 side are roughly twofold: (i) there are infinitely many smooth irreducible (admissible) representations of $\mathrm{GL}_2(F)$ over any finite field containing the residue field of F (whereas when $F = \mathbb{Q}_p$ there is only a finite number of them) and (ii) the vast majority of them are much harder to study than for $F = \mathbb{Q}_p$. In particular (i) has the consequence that there is no possible naive $1-1$ correspondence as for the $F = \mathbb{Q}_p$ case and (ii) has the consequence that no one so far has been able to find an explicit construction of *one single* irreducible representation of $\mathrm{GL}_2(F)$ that isn't a subquotient of a principal series. The p-adic theory shouldn't be expected to be significantly simpler ([54]). And yet, cohomology spaces analogous to (1) are known to exist and to support interesting representations of $\mathrm{GL}_2(F)$ over $\overline{\mathbb{F}_p}$ (where $\overline{\mathbb{F}_p}$ is an algebraic closure of the finite field $\mathbb{F}_p$) as well as related representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$, but the representations of $\mathrm{GL}_2(F)$ occuring there seem to be of a very special type. We report on these phenomena in §3.

We then conclude this non-exhaustive survey in §4 more optimistically by mentioning, among other scattered statements, three theorems or conjectures available for $\mathrm{GL}_n(F)$ that give some kind of (p-adic or modulo p) relations between the $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ side and the $\mathrm{GL}_n(F)$ side. Although they are quite far from any sort of correspondence, these statements are clearly part of the p-adic Langlands programme and will probably play a role in the future.

One word about the title. Strangely, the terminology "p-adic Langlands correspondence/programme" started to spread (at least in the author's memory) only shortly after preprints of [6], [7], [8], [9], [14], [19], [24], [53], [59], [60], [61], [70] were available (that is, around 2004), although of course p-adic considerations on automorphic forms (e.g., congruences modulo p between automorphic forms, p-adic families of automorphic forms) had begun years earlier with the fundamental work of Serre, Katz, Mazur, Hida, Coleman, etc. Maybe one of the reasons was that an important difference between the above more recent references and older ones was the focus on (i) topological group representation theory "à la Langlands" and (ii) purely p-adic aspects in relation with Fontaine's classifications of p-adic Galois representations.

The present status of the p-adic Langlands programme so far is thus the following: almost everything is known for $\mathrm{GL}_2(\mathbb{Q}_p)$ but most of the experts (including the author) are quite puzzled by the apparent complexity of whatever seems to happen for any other group. The only certainty one can have is that much remains to be discovered!

Let us introduce some notations. Recall that $\overline{\mathbb{Q}_p}$ (resp. $\overline{\mathbb{F}_p}$) is an algebraic closure of $\mathbb{Q}_p$ (resp. $\mathbb{F}_p$). If K is a finite extension of $\mathbb{Q}_p$, we denote by $\mathcal{O}_K$ its ring of integers, by ϖ_K a uniformizer in $\mathcal{O}_K$ and by $k_K := \mathcal{O}_K/(\varpi_K\mathcal{O}_K)$ its residue field.

Throughout the text, we denote by F a finite extension of $\mathbb{Q}_p$ inside $\overline{\mathbb{Q}_p}$, by $q = p^f$ the cardinality of k_F and by $e = [F : \mathbb{Q}_p]/f$ the ramification index of F. For $x \in F^\times$, we let $|x| := q^{-\mathrm{val}_F(x)}$ where $\mathrm{val}_F(p) := e$. The Weil group of F is the subgroup of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ of elements w mapping to an integral power

$d(w)$ of the arithmetic Frobenius of $\mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_p)$ (that is, $x \mapsto x^p$) via the map $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F) \to \mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_p)$.

Representations always take values either in E-vector spaces, in $\mathcal{O}_E$-modules or in k_E-vector spaces where E is always a "sufficiently big" finite extension of $\mathbb{Q}_p$. By "sufficiently big", we mean big enough so that we do not have to deal with rationality issues. For instance irreducible always means absolutely irreducible, we always assume $|\mathrm{Hom}(F,E)| = |\mathrm{Hom}(F,\overline{\mathbb{Q}_p})|$, $|\mathrm{Hom}(k_F,k_E)| = |\mathrm{Hom}(k_F,\overline{\mathbb{F}_p})|$, etc.

We normalize the reciprocity map $F^\times \hookrightarrow \mathrm{Gal}(\overline{\mathbb{Q}_p}/F)^{\mathrm{ab}}$ of local class field theory by sending inverses of uniformizers to arithmetic Frobeniuses. Via this map, we consider without comment Galois characters as characters of $F^\times$ by restriction. We denote by $\varepsilon : \mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p) \to \mathbb{Z}_p^\times$ the p-adic cyclotomic character and by ω its reduction modulo p. Seen as a character of $\mathbb{Q}_p^\times$, ε is the identity on $\mathbb{Z}_p^\times$ and sends p to 1.

If A is any $\mathbb{Z}$-algebra, we denote by $B(A)$ (resp. $T(A)$) the upper triangular matrices (resp. the diagonal matrices) in $\mathrm{GL}_n(A)$. We denote by I (resp. I_1) the Iwahori subgroup (resp. the pro-p Iwahori subgroup) of $\mathrm{GL}_n(\mathcal{O}_F)$, that is, the matrices of $\mathrm{GL}_n(\mathcal{O}_F)$ that are upper triangular modulo ϖ_F (resp. upper unipotent modulo ϖ_F).

A smooth representation of a topological group is a representation such that any vector is fixed by a non-empty open subgroup. A smooth representation of $\mathrm{GL}_n(\mathcal{O}_F)$ over a field is admissible if its subspace of invariant elements under any open (compact) subgroup of $\mathrm{GL}_n(\mathcal{O}_F)$ is finite dimensional. We recall that the socle of a smooth representation of a topological group over a field is the (direct) sum of all its irreducible subrepresentations.

We call a Serre weight for $\mathrm{GL}_n(\mathcal{O}_F)F^\times$ any smooth irreducible representation of $\mathrm{GL}_n(\mathcal{O}_F)F^\times$ over k_E. In particular, a Serre weight is finite dimensional, $F^\times$ acts on it by a character and its restriction to $\mathrm{GL}_n(\mathcal{O}_F)$ is irreducible. In other references (e.g., [18] or [39]), a Serre weight is just a smooth irreducible representation of $\mathrm{GL}_n(\mathcal{O}_F)$ over k_E; however, in all representations we consider, $F^\times$ acts by a character, and it is very convenient to extend the action to $\mathrm{GL}_n(\mathcal{O}_F)F^\times$.

2. The Group $\mathrm{GL}_2(\mathbb{Q}_p)$

We assume here $F = \mathbb{Q}_p$. The p-adic Langlands programme for $\mathrm{GL}_2(\mathbb{Q}_p)$ and 2-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ is close to being finished. We sum up below some of the local and global aspects of the theory.

2.1. The modulo p local correspondence. We first describe the modulo p Langlands correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$ (at least in the "generic" case), which is much easier than the p-adic one and which was historically found before.

Let σ be a Serre weight for $\mathrm{GL}_2(\mathbb{Z}_p)\mathbb{Q}_p^\times$ and denote by:

$$\mathrm{c}-\mathrm{Ind}_{\mathrm{GL}_2(\mathbb{Z}_p)\mathbb{Q}_p^\times}^{\mathrm{GL}_2(\mathbb{Q}_p)}\sigma$$

the k_E-vector space of functions $f:\mathrm{GL}_2(\mathbb{Q}_p)\to\sigma$ which have compact support modulo $\mathbb{Q}_p^\times$ and such that $f(kg)=\sigma(k)(f(g))$ for $(k,g)\in\mathrm{GL}_2(\mathbb{Z}_p)\mathbb{Q}_p^\times\times\mathrm{GL}_2(\mathbb{Q}_p)$. We endow this space with the left and smooth action of $\mathrm{GL}_2(\mathbb{Q}_p)$ defined by $(gf)(g'):=f(g'g)$. By a standard result, one has ([2]):

$$\mathrm{End}_{\mathrm{GL}_2(\mathbb{Q}_p)}\Big(\mathrm{c}-\mathrm{Ind}_{\mathrm{GL}_2(\mathbb{Z}_p)\mathbb{Q}_p^\times}^{\mathrm{GL}_2(\mathbb{Q}_p)}\sigma\Big)=k_E[T]$$

for a certain Hecke operator T. One then defines:

$$\pi(\sigma,0):=\Big(\mathrm{c}-\mathrm{Ind}_{\mathrm{GL}_2(\mathbb{Z}_p)\mathbb{Q}_p^\times}^{\mathrm{GL}_2(\mathbb{Q}_p)}\sigma\Big)/(T).$$

One can prove that the representations $\pi(\sigma,0)$ are irreducible and admissible ([6]). The representations $\pi(\sigma,0)$ form the so-called *supersingular* representations of $\mathrm{GL}_2(\mathbb{Q}_p)$.

Let $\chi_i:\mathbb{Q}_p^\times\to k_E^\times$, $i\in\{1,2\}$ be smooth multiplicative characters and define:

$$\begin{array}{rccc}\chi_1\otimes\chi_2 & : & B(\mathbb{Q}_p) & \to & k_E^\times\\ & & \begin{pmatrix}a & b\\ 0 & d\end{pmatrix} & \mapsto & \chi_1(a)\chi_2(d).\end{array}$$

Denote by:

$$\mathrm{Ind}_{B(\mathbb{Q}_p)}^{\mathrm{GL}_2(\mathbb{Q}_p)}\chi_1\otimes\chi_2$$

the k_E-vector space of locally constant functions $f:\mathrm{GL}_2(\mathbb{Q}_p)\to k_E$ such that $f(hg)=(\chi_1\otimes\chi_2)(h)f(g)$ for $(h,g)\in B(\mathbb{Q}_p)\times\mathrm{GL}_2(\mathbb{Q}_p)$. We endow this space with the same left and smooth action of $\mathrm{GL}_2(\mathbb{Q}_p)$ as previously. The representations $\mathrm{Ind}_{B(\mathbb{Q}_p)}^{\mathrm{GL}_2(\mathbb{Q}_p)}\chi_1\otimes\chi_2$ are admissible. They are irreducible if $\chi_1\neq\chi_2$ and have length 2 otherwise ([1], [2]). They form the so-called *principal series*. The supersingular representations together with the Jordan-Hölder factors of the principal series exhaust the smooth irreducible representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ over k_E with a central character ([2], [6]).

Theorem 2.1. *For $\chi_1\neq\chi_2$ and $\chi_1\neq\chi_2\omega^{\pm1}$ the k_E-vector space:*

$$\mathrm{Ext}^1_{\mathrm{GL}_2(\mathbb{Q}_p)}\left(\mathrm{Ind}_{B(\mathbb{Q}_p)}^{\mathrm{GL}_2(\mathbb{Q}_p)}\chi_1\otimes\chi_2\omega^{-1},\mathrm{Ind}_{B(\mathbb{Q}_p)}^{\mathrm{GL}_2(\mathbb{Q}_p)}\chi_2\otimes\chi_1\omega^{-1}\right)$$

has dimension 1.

Proof. This follows for instance from [16, Cor.8.6] but other (and earlier) proofs can be found in [27] and [21, §VII]. □

Note that the assumptions on χ_i imply that both principal series in Theorem 2.1 are irreducible distinct (and hence that any extension between them has their central character) and that $\mathrm{Ext}^1_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)}(\chi_2,\chi_1)$ also has dimension 1.

For g in the inertia subgroup of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$, let:

$$\omega_2(g) := \frac{g(\sqrt[p^2-1]{-p})}{\sqrt[p^2-1]{-p}} \in \mu_{p^2-1}(\overline{\mathbb{Q}_p}) \xrightarrow{\sim} \mathbb{F}_{p^2}^\times \hookrightarrow k_E^\times$$

be Serre's level 2 fundamental character (where the first map is reduction modulo p and where we choose an arbitrary field embedding $\mathbb{F}_{p^2} \hookrightarrow k_E$). For $0 \le r \le p-1$, we denote by σ_r the unique Serre weight for $\mathrm{GL}_2(\mathbb{Z}_p)\mathbb{Q}_p^\times$ such that $\sigma_r(p)=1$ and σ_r has dimension $r+1$ (in fact $\sigma_r|_{\mathrm{GL}_2(\mathbb{Z}_p)} \simeq \mathrm{Sym}^r(k_E^2)$). For $0 \le r \le p-1$, we denote by ρ_r the unique continuous representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ over k_E such that its determinant is ω^{r+1} and its restriction to inertia is $\omega_2^{r+1} \oplus \omega_2^{p(r+1)}$.

The modulo p local correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$ can be defined as follows.

Definition 2.2. (i) For $0 \le r \le p-1$ and $\chi : \mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p) \to k_E^\times$, the representation $\pi(\sigma_r,0)\otimes(\chi\circ\det)$ corresponds to $\rho_r\otimes\chi$.
(ii) For $\chi_1 \notin \{\chi_2, \chi_2\omega^{\pm 1}\}$ the representation associated to the unique non-split (resp. split) extension in:

$$\mathrm{Ext}^1_{\mathrm{GL}_2(\mathbb{Q}_p)}\left(\mathrm{Ind}_{B(\mathbb{Q}_p)}^{\mathrm{GL}_2(\mathbb{Q}_p)}\chi_1\otimes\chi_2\omega^{-1}, \mathrm{Ind}_{B(\mathbb{Q}_p)}^{\mathrm{GL}_2(\mathbb{Q}_p)}\chi_2\otimes\chi_1\omega^{-1}\right)$$

corresponds to the representation associated to the unique non-split (resp. split) extension in $\mathrm{Ext}^1_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)}(\chi_2,\chi_1)$.

For more general 2-dimensional reducible representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$, the corresponding representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ are a bit more subtle to define and we refer the reader to [27] or [21, §VII]. When representations (on both side) are semi-simple, the above correspondence was first defined in [6]. Note that Definition 2.2 requires one to check that whenever there is an isomorphism between the $\mathrm{GL}_2(\mathbb{Q}_p)$-representations involved, the corresponding Galois representations are also isomorphic. The correspondence of Definition 2.2 (without restrictions on the χ_i) can now be realized using the theory of (φ,Γ)-modules (see §2.3), which makes it much more natural.

2.2. Over *E*: first properties. We now switch to continuous representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E and explain the first properties of the p-adic local correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$.

We fix a p-adic absolute value $|\cdot|$ on E extending the one on $F=\mathbb{Q}_p$ and recall that a (p-adic) norm on an E-vector space V is a function $\|\cdot\| : V \to \mathbb{R}_{\ge 0}$

such that $\|v\| = 0$ if and only if $v = 0$, $\|\lambda v\| = |\lambda|\|v\|$ ($\lambda \in E$, $v \in V$) and $\|v+w\| \leq \mathrm{Max}(\|v\|, \|w\|)$ ($v, w \in V$). Any norm on V defines a metric $\|v-w\|$ which in turns defines a topology on V by the usual recipe. A (p-adic) Banach space over E is an E-vector space endowed with a topology coming from a norm and such that the underlying metric space is complete. All norms on a Banach space over E defining its topology are equivalent.

Definition 2.3. (i) A Banach space representation of a topological group G over E is a Banach space B over E together with a linear action of G by continuous automorphisms such that the natural map $G \times B \to B$ is continuous. (ii) A Banach space representation B of G over E is unitary if there exists a norm $\|\cdot\|$ on B defining its topology such that $\|gv\| = \|v\|$ for all $g \in G$ and $v \in B$.

If G is compact, any Banach space representation of G is unitary but this is not true if G is not compact, e.g., $G = \mathrm{GL}_2(\mathbb{Q}_p)$. Let B be a unitary Banach space representation of G and $B^0 := \{v \in B,\ \|v\| \leq 1\}$ the unit ball with respect to an invariant norm on B (giving its topology); then $B^0 \otimes_{\mathcal{O}_E} k_E$ is a smooth representation of G over k_E. A unitary Banach space representation of $\mathrm{GL}_2(\mathbb{Q}_p)$ is said to be admissible if $B^0 \otimes_{\mathcal{O}_E} k_E$ is admissible. This does not depend on the choice of B^0 ([60, §3], [8, §4.6]). The category of unitary admissible Banach space representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E is abelian ([60]).

To any Banach space representation B of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E, one can associate two subspaces $B^{\mathrm{alg}} \subset B^{\mathrm{an}}$ which are stable under $\mathrm{GL}_2(\mathbb{Q}_p)$. We define $B^{\mathrm{an}} \subset B$ (the locally analytic vectors) to be the subspace of vectors $v \in B$ such that the function $\mathrm{GL}_2(\mathbb{Q}_p) \to B$, $g \mapsto gv$ is locally analytic in the sense of [61]. We define $B^{\mathrm{alg}} \subset B^{\mathrm{an}}$ (the locally algebraic vectors) to be the subspace of vectors $v \in B$ for which there exists a compact open subgroup $H \subset \mathrm{GL}_2(\mathbb{Q}_p)$ such that the H-representation $\langle H \cdot v \rangle \subset B|_H$ is isomorphic to a direct sum of finite dimensional (irreducible) algebraic representations of H. In general one has $B^{\mathrm{alg}} = 0$, but if B is admissible as a representation of the compact group $\mathrm{GL}_2(\mathbb{Z}_p)$ it is a major result due to Schneider and Teitelbaum (which holds in much greater generality) that the subspace B^{an} is never 0 and is even dense in B ([62]).

Inspired by the modulo p correspondence of Definition 2.2 and by lots of computations on locally algebraic representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ ([7], [15]), the author suggested in [8, §1.3] (see also [25, §3.3]) the following partial "programme".

Fix V a linear continuous potentially semi-stable 2-dimensional representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ over E with distinct Hodge-Tate weights $w_1 < w_2$. As in §4.1 below, following Fontaine ([30]) one can associate to V a Weil-Deligne representation to which (after semi-simplifying its underlying Weil representation) one can in turn attach a smooth admissible infinite dimensional representation π of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E by the classical local Langlands correspondence (slightly modified as in §2.4 or §4.1 below). We denote by $\overline{V}^{\mathrm{ss}}$ the semi-simplification of $V^0 \otimes_{\mathcal{O}_E} k_E$ where V^0 is any Galois $\mathcal{O}_E$-lattice in V. To V one should be able

to attach an admissible unitary Banach space representation $B(V)$ of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E satisfying the following properties:

(i) $V \simeq V'$ if and only if $B(V) \simeq B(V')$ if and only if $B(V)^{\mathrm{an}} \simeq B(V')^{\mathrm{an}}$;

(ii) if V is irreducible then $B(V)$ is topologically irreducible; if V is reducible and indecomposable (resp. semi-simple) then $B(V)$ is reducible and indecomposable (resp. semi-simple);

(iii) for any unit ball $B^0 \subset B(V)$ preserved by $\mathrm{GL}_2(\mathbb{Q}_p)$, the semi-simplification of $B^0 \otimes_{\mathcal{O}_E} k_E$ corresponds to $\overline{V}^{\mathrm{ss}}$ under the modulo p correspondence of Definition 2.2;

(iv) the $\mathrm{GL}_2(\mathbb{Q}_p)$-subrepresentation $B(V)^{\mathrm{alg}}$ is isomorphic to:

$$\det{}^{w_1} \otimes_E \mathrm{Sym}^{w_2-w_1-1}(E^2) \otimes_E \pi.$$

When V is irreducible, (ii) and (iv) imply that $B(V)$ is a suitable completion of the locally algebraic representation $B(V)^{\mathrm{alg}} = \det{}^{w_1} \otimes_E \mathrm{Sym}^{w_2-w_1}(E^2) \otimes_E \pi$ with respect to an invariant norm. This property is the basic idea which initially motivated the above programme: what is missing to recover V from w_1, w_2 and its associated Weil-Deligne representation, or equivalently from $B(V)^{\mathrm{alg}}$, is a certain weakly admissible Hodge filtration ([22]). This missing data should precisely correspond to an invariant norm on $B(V)^{\mathrm{alg}}$. For instance, when V is irreducible and becomes crystalline over an abelian extension of $\mathbb{Q}_p$, such a filtration turns out to be unique (see, e.g., [32, §3.2]). Correspondingly one finds that there is a unique class of invariant norms on $B(V)^{\mathrm{alg}}$ in that case ([5, §5.3], [55]).

The first instances of $B(V)$ were constructed "by hand" for V semi-stable and small values of $w_2 - w_1$ in [7], [8] and [9]. Shortly after these examples were worked out, Colmez discovered that there was a way to define $B(V)$ directly out of Fontaine's (φ, Γ)-module of V ([19], [5]), thus explaining the above basic idea and also the compatibility (iii) with Definition 2.2 (the latter was checked in detail by Berger [3]). Using the (φ, Γ)-module machinery, Colmez was ultimately able to fulfil the above programme and even to associate a $B(V)$ to *any* linear continuous 2-dimensional representation V of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ over E. It was then recently proved by Paškūnas that these $B(V)$ and their Jordan-Hölder constituents essentially exhaust all topologically irreducible admissible unitary Banach space representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E.

2.3. (φ, Γ)-modules and the theorems of Colmez and of Paškūnas. We first briefly recall what a (φ, Γ)-module is ([30]) and then state the main results on the Banach space representations $B(V)$.

Let $\Gamma := \mathrm{Gal}(\mathbb{Q}_p(\sqrt[p^\infty]{1})/\mathbb{Q}_p)$ and note that the p-adic cyclotomic character ε canonically identifies Γ with $\mathbb{Z}_p^\times$. If $a \in \mathbb{Z}_p^\times$, let $\gamma_a \in \Gamma$ be the unique element such that $\varepsilon(\gamma_a) = a$. Let $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ be the p-adic completion of $\mathcal{O}_E[[X]][\frac{1}{X}]$

equipped with the unique ring topology such that a basis of neighbourhoods of 0 is:

$$\left(p^n\mathcal{O}_E[[X]]\left[\frac{1}{X}\right]^\wedge + X^m\mathcal{O}_E[[X]]\right)_{n\geq 0, m\geq 0}.$$

We endow $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ with the unique $\mathcal{O}_E$-linear continuous Frobenius φ such that $\varphi(X^j) := ((1+X)^p-1)^j$ ($j \in \mathbb{Z}$) and with the unique $\mathcal{O}_E$-linear continuous action of Γ such that ($a \in \mathbb{Z}_p^\times$):

$$\gamma_a(X^j) := ((1+X)^a-1)^j = \left(\sum_{i=1}^{+\infty} \frac{a(a-1)\cdots(a-i+1)}{i!} X^i\right)^j.$$

We extend φ and Γ by E-linearity to the field $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$. Note that the actions of φ and Γ commute and preserve the subring $\mathcal{O}_E[[X]]$.

A (φ,Γ)-module over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ (resp. $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$) is an $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$-module of finite type (resp. an $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$-vector space of finite dimension) D equipped with the topology coming from that on $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ together with a homomorphism $\varphi : D \to D$ such that $\varphi(sd) = \varphi(s)\varphi(d)$ and with a continuous action of Γ such that $\gamma(sd) = \gamma(s)\gamma(d)$ and $\gamma\circ\varphi = \varphi\circ\gamma$ ($s \in \mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ or $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$, $d \in D$, $\gamma \in \Gamma$). A (φ,Γ)-module over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ or $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$ is said to be *étale* if moreover the image of φ generates D, in which case φ is automatically injective. There is a third important $\mathcal{O}_E$-linear map $\psi : D \to D$ on any étale (φ,Γ)-module D defined by $\psi(d) := d_0$ if $d = \sum_{i=0}^{p-1}(1+X)^i\varphi(d_i) \in D$ (any d determines uniquely such $d_i \in D$ as D is étale). The map ψ is surjective, commutes with Γ and satisfies by definition $(\psi\circ\varphi)(d) = d$. The main theorem is the following equivalence of categories due to Fontaine (we won't need more details here, see [29]).

Theorem 2.4. *There is an equivalence of categories between the category of $\mathcal{O}_E$-linear continuous representations of* $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ *on finite type $\mathcal{O}_E$-modules (resp. on finite dimensional E-vector spaces) and étale (φ,Γ)-modules over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ (resp. over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$).*

If T (resp. V) is an $\mathcal{O}_E$-linear continuous representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ on a finite type $\mathcal{O}_E$-module (resp. on a finite dimensional E-vector space), we denote by $D(T)$ (resp. $D(V)$) the corresponding (φ,Γ)-module over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ (resp. over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$).

Let V be any linear continuous 2-dimensional representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ over E and $\chi : \mathbb{Q}_p^\times \to \mathcal{O}_E^\times$ any continuous character. For $d \in D(V)^{\psi=0} := \{d \in D(V),\ \psi(d) = 0\}$, one can prove that the formula:

$$w_\chi(d) := \lim_{n\to+\infty} \sum_{i\in\mathbb{Z}_p^\times \bmod p^n} \chi(i)^{-1}(1+X)^i\gamma_{-i^2}\big(\varphi^n\psi^n\big((1+X)^{-i^{-1}}d\big)\big)$$

converges in $D(V)^{\psi=0}$ and that $w_\chi^2(d)=d$ ([21, §II]). One defines the following E-vector space (recalling that $(1-\varphi\psi)(D(V))\subseteq D(V)^{\psi=0}$):

$$D(V)\boxtimes_\chi \mathbb{P}^1 := \{(d_1,d_2)\in D(V)\times D(V), (1-\varphi\psi)(d_1)=w_\chi\big((1-\varphi\psi)(d_2)\big)\}.$$

Note that $(d_1,d_2)\in D(V)\boxtimes_\chi\mathbb{P}^1$ is determined by $\varphi\psi(d_1)$ and d_2, or by d_1 and $\varphi\psi(d_2)$. One can show that the following formulas define an action of the group $\mathrm{GL}_2(\mathbb{Q}_p)$ on $D(V)\boxtimes_\chi\mathbb{P}^1$ (even if V has dimension ≥ 2):

(i) if $a\in\mathbb{Q}_p^\times$, $\left(\begin{smallmatrix} a & 0\\ 0 & a\end{smallmatrix}\right)(d_1,d_2):=(\chi(a)d_1,\chi(a)d_2)$;

(ii) if $a\in\mathbb{Z}_p^\times$, $\left(\begin{smallmatrix} a & 0\\ 0 & 1\end{smallmatrix}\right)(d_1,d_2):=(\gamma_a(d_1),\chi(a)\gamma_{a^{-1}}(d_2))$;

(iii) $\left(\begin{smallmatrix} 0 & 1\\ 1 & 0\end{smallmatrix}\right)(d_1,d_2):=(d_2,d_1)$;

(iv) $\left(\begin{smallmatrix} p & 0\\ 0 & 1\end{smallmatrix}\right)(d_1,d_2)$ is the unique element (d_1',d_2') of $D(V)\boxtimes_\chi\mathbb{P}^1$ such that $\varphi\psi(d_1'):=\varphi(d_1)$ and $d_2':=\chi(p)\psi(d_2)$;

(v) if $b\in p\mathbb{Z}_p$, $\left(\begin{smallmatrix} 1 & b\\ 0 & 1\end{smallmatrix}\right)(d_1,d_2)$ is the unique element (d_1',d_2') of $D(V)\boxtimes_\chi\mathbb{P}^1$ such that $d_1':=(1+X)^b d_1$ and:

$$\varphi\psi(d_2'):=\chi(1+b)^{-1}(1+X)^{-1}w_\chi\left(\gamma_{1+b}\left((1+X)^b w_\chi\left(1+X)^{(1+b)^{-1}}\varphi\psi(d_2)\right)\right)\right).$$

All of the above mysterious formulas were first discovered in the case V crystalline, where everything can be made very explicit ([5], [19]), and then extended more or less *verbatim* to any V.

For any étale (φ,Γ)-module D over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$, let $D^\natural\subset D$ be the smallest compact $\mathcal{O}_E[[X]]$-submodule which generates D over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ and which is preserved by ψ (one can prove that such a module exists). If D is an étale (φ,Γ)-module over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge[\frac{1}{p}]$, choose any lattice $D_0\subset D$, that is any étale (φ,Γ)-module D_0 which is free over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ and generates D, and let $D^\natural:=D_0^\natural[\frac{1}{p}]$. Going back to our 2-dimensional V, for $(d_1,d_2)\in D(V)\boxtimes_\chi\mathbb{P}^1$ and $n\in\mathbb{Z}_{\geq 0}$, let $(d_1^{(n)},d_2^{(n)}):=\left(\begin{smallmatrix} p^n & 0\\ 0 & 1\end{smallmatrix}\right)(d_1,d_2)\in D(V)\boxtimes_\chi\mathbb{P}^1$. Note that from the iteration of (iv) above and from $\psi\circ\varphi=\mathrm{Id}$, one gets $\psi(d_1^{(n+1)})=d_1^{(n)}$. One then defines the following subspace of $D(V)\boxtimes_\chi\mathbb{P}^1$:

$$D(V)^\natural\boxtimes_\chi\mathbb{P}^1:=\{(d_1,d_2)\in D(V)\boxtimes_\chi\mathbb{P}^1, d_1^{(n)}\in D(V)^\natural \text{ for all } n\in\mathbb{Z}_{\geq 0}\}.$$

Now let $\chi(x)=\chi_V(x):=(x|x|)^{-1}\det(V)(x)$ $(x\in\mathbb{Q}_p^\times)$. It turns out that, for such a χ, $D(V)^\natural\boxtimes_\chi\mathbb{P}^1$ is preserved by $\mathrm{GL}_2(\mathbb{Q}_p)$ inside $D(V)\boxtimes_\chi\mathbb{P}^1$. The stability of the subspace $D(V)^\natural\boxtimes_{\chi_V}\mathbb{P}^1$ by $\mathrm{GL}_2(\mathbb{Q}_p)$ is the most subtle part of the theory and, so far, the only existing proof (following a suggestion of Kisin) is by analytic continuation from the crystalline case (see [21, §II.3]).

We can now state the main theorem giving the local p-adic Langlands correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$ in the case V is irreducible ([21]).

Theorem 2.5. *Assume V is irreducible. Then the quotient:*

$$B(V) := D(V) \boxtimes_{\chi_V} \mathbb{P}^1 / D(V)^\natural \boxtimes_{\chi_V} \mathbb{P}^1$$

together with the induced action of $\mathrm{GL}_2(\mathbb{Q}_p)$ above is naturally an admissible unitary topologically irreducible Banach space representation of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E satisfying properties (i) to (iii) of §2.2. Moreover, $B(V)^{\mathrm{alg}} \neq 0$ if and only if V is potentially semi-stable with distinct Hodge-Tate weights, and $B(V)$ then satisfies property (iv)[1] *of §2.2.*

A unit ball of $B(V)$ is $B(T) := D(T) \boxtimes_{\chi_V} \mathbb{P}^1 / D(T)^\natural \boxtimes_{\chi_V} \mathbb{P}^1$ where $T \subset V$ is any Galois $\mathcal{O}_E$-lattice (one can extend all the previous constructions with $D(T)$ instead of $D(V)$). For the second part of property (i) of §2.2, one has to use that the subspace $B(V)^{\mathrm{an}} \subset B(V)$ of locally analytic vectors admits an analogous construction in terms of the (φ, Γ)-module of V over the Robba ring ([21, §V.2]). When V is reducible, a reducible $B(V)$ can also be constructed as an extension between two continuous principal series in a way analogous to (ii) of Definition 2.2 (see [19] or [27] or [47], see also [46]).

There is a nice functorial way to recover in all cases $D(T)$ from $B(T)$ (and hence $D(V)$ from $B(V)$) as follows. Let $n \in \mathbb{Z}_{>0}$, $T^\vee := \mathrm{Hom}_{\mathcal{O}_E}(T, \mathcal{O}_E)$ and let $\sigma \subset B(T^\vee)/p^n B(T^\vee)$ be any $\mathcal{O}_E$-submodule of finite type that generates $B(T^\vee)/p^n B(T^\vee)$ as a $\mathrm{GL}(\mathbb{Q}_p)$-representation (such a σ exists as a consequence of property (iii) of §2.2). Consider the $\mathcal{O}_E/p^n\mathcal{O}_E$-module:

$$\mathrm{Hom}_{\mathcal{O}_E/p^n\mathcal{O}_E}\left(\sum_{m\geq 0} \begin{pmatrix} p^m & \mathbb{Z}_p \\ 0 & 1 \end{pmatrix} \sigma, \mathcal{O}_E/p^n\mathcal{O}_E\right) \tag{2}$$

where the left entry is the $\mathcal{O}_E/p^n\mathcal{O}_E$-submodule of $B(T^\vee)/p^n B(T^\vee)$ generated by σ under the matrices $\left(\begin{smallmatrix} p^m & a \\ 0 & 1 \end{smallmatrix}\right)$, $a \in \mathbb{Z}_p$, $m \in \mathbb{Z}_{\geq 0}$. The natural action of $\left(\begin{smallmatrix} 1 & \mathbb{Z}_p \\ 0 & 1 \end{smallmatrix}\right)$ (resp. of $\left(\begin{smallmatrix} \mathbb{Z}_p^\times & 0 \\ 0 & 1 \end{smallmatrix}\right)$) on $\sum_{m\geq 0} \left(\begin{smallmatrix} p^m & \mathbb{Z}_p \\ 0 & 1 \end{smallmatrix}\right)\sigma$ makes (2) a module over the Iwasawa algebra $\mathcal{O}_E[[\left(\begin{smallmatrix} 1 & \mathbb{Z}_p \\ 0 & 1 \end{smallmatrix}\right)]] = \mathcal{O}_E[[\mathbb{Z}_p]] = \mathcal{O}_E[[X]]$ where $X := [\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)] - 1$ (resp. endows (2) with an action of $\Gamma \simeq \mathbb{Z}_p^\times$). After tensoring (2) by $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ over $\mathcal{O}_E[[X]]$, one can moreover define a natural Frobenius φ coming from the action of $\left(\begin{smallmatrix} p^{-1} & 0 \\ 0 & 1 \end{smallmatrix}\right)$. The final result turns out to be an étale (φ, Γ)-module over $\mathcal{O}_E[[X]][\frac{1}{X}]^\wedge$ (killed by p^n) which is independent of the choice of σ and isomorphic to $D(T)/p^n D(T)$ ([21, §IV]). One then recovers $D(T)$ by taking the projective limit over n.

This last functor $B(T) \mapsto D(T)$ has revealed itself to be of great importance. For instance it allowed Kisin to give in many cases another construction of $B(V)$ more amenable to deformation theory ([46]) and it was a key ingredient in

[1] Some of the arguments of [21] here rely on the global results of [28], in particular on Theorem 2.7 below. Hence property (iv) might not yet be completely proven in a few cases like $p = 2$ or $\overline{V}^{\mathrm{ss}} \cong \left(\begin{smallmatrix} 1 & 0 \\ 0 & \omega \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ up to twist, etc.

Kisin's or Emerton's proof of almost all cases of the Fontaine-Mazur conjecture ([45], [28]). Together with Kisin's construction, it was also used by Paškūnas to recently prove the following nice theorem ([56]):

Theorem 2.6. *Assume $p \geq 5$, then the above functor $B(T) \mapsto D(T)$ induces (after tensoring by E) a bijection between isomorphism classes of:*

(i) admissible unitary topologically irreducible Banach space representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E which are not subquotients of continuous parabolic inductions of unitary characters;

(ii) irreducible 2-dimensional continuous representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ over E.

Finally, let us mention that this functor has been extended to a more general setting in [64].

2.4. Local-global compatibility. The local correspondence of §2.3 turns out to be realized on suitable cohomology spaces of (towers of) modular curves. This aspect, usually called "local-global compatibility" (as the cohomology spaces have a global origin), is the deepest and most important part of the theory.

Denote by $\mathbb{A}$ the adèles of $\mathbb{Q}$, $\mathbb{A}_f \subset \mathbb{A}$ the finite adèles and $\mathbb{A}_f^p \subset \mathbb{A}_f$ the finite adèles outside p. For any compact open subgroup K_f of $\mathrm{GL}_2(\mathbb{A}_f)$, consider the following complex curve:

$$Y(K_f)(\mathbb{C}) := \mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\mathbb{A})/K_f\mathbb{R}^\times\mathrm{SO}_2(\mathbb{R}).$$

For varying K_f, $(Y(K_f)(\mathbb{C}))_{K_f}$ forms a projective system on which $\mathrm{GL}_2(\mathbb{A}_f)$ naturally acts on the right ($g \in \mathrm{GL}(\mathbb{A}_f)$ maps $Y(K_f)(\mathbb{C})$ to $Y(g^{-1}K_fg)(\mathbb{C})$). Likewise, for each fixed compact open subgroup $K_f^p \subset \mathrm{GL}_2(\mathbb{A}_f^p)$ and varying compact open subgroups $K_{f,p}$ of $\mathrm{GL}_2(\mathbb{Q}_p)$, $(Y(K_f^pK_{f,p})(\mathbb{C}))_{K_{f,p}}$ forms a projective system on which $\mathrm{GL}_2(\mathbb{Q}_p)$ acts on the right. One considers the following "completed cohomology spaces":

$$\begin{aligned}
\widehat{H}^1(K_f^p) &:= \left(\varprojlim_n \varinjlim_{K_{f,p}} H^1\left(Y(K_f^pK_{f,p})(\mathbb{C}), \mathcal{O}_E/p^n\mathcal{O}_E\right)\right) \otimes_{\mathcal{O}_E} E \\
\widehat{H}^1 &:= \varinjlim_{K_f^p} \widehat{H}^1(K_f^p)
\end{aligned}$$

where H^1 is usual Betti cohomology and where $K_{f,p}$ (resp. K_f^p) runs over the compact open subgroups of $\mathrm{GL}(\mathbb{Q}_p)$ (resp. of $\mathrm{GL}(\mathbb{A}_f^p)$). The group $\mathrm{GL}_2(\mathbb{Q}_p)$ (resp. $\mathrm{GL}_2(\mathbb{A}_f)$) acts on $\widehat{H}^1(K_f^p)$ (resp. on $\widehat{H}^1$) and one can prove that each $\widehat{H}^1(K_f^p)$ is an admissible unitary Banach space representation of $\mathrm{GL}_2(\mathbb{Q}_p)$ over E, an open unit ball being given by $\varprojlim \varinjlim H^1\big(Y(K_f^pK_{f,p})(\mathbb{C}), \mathcal{O}_E/p^n\mathcal{O}_E\big)$ (this

result, due to Emerton, actually holds in much greater generality, see [24, §2]). Moreover, all the Betti cohomology spaces $H^1\big(Y(K_f^pK_{f,p})(\mathbb{C}), \mathcal{O}_E/p^n\mathcal{O}_E\big)$ can be identified with étale cohomology spaces, in particular they carry a natural action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. We thus also have a (commuting) action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on $\widehat{H}^1(K_f^p)$ and $\widehat{H}^1$.

Let $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(E)$ be a linear continuous representation (where $\overline{\mathbb{Q}}$ is an algebraic closure of $\mathbb{Q}$) and for each prime number ℓ let ρ_ℓ be the restriction of ρ to a decomposition group at ℓ. By the classical local Langlands correspondence as in [38], if $\ell \neq p$ one can associate to ρ_ℓ (after maybe semi-simplifying the action of Frobenius) a smooth irreducible representation π'_ℓ of $\mathrm{GL}_2(\mathbb{Q}_\ell)$ over E. We slightly modify π'_ℓ as follows: if π'_ℓ is infinite dimensional, we let $\pi_\ell(\rho_\ell) := \pi'_\ell \otimes |\det|^{-\frac{1}{2}}$. If π'_ℓ is finite dimensional (that is, 1-dimensional), we let $\pi_\ell(\rho_\ell)$ be the unique principal series which has $\pi'_\ell \otimes |\det|^{-\frac{1}{2}}$ as unique irreducible quotient ($\pi_\ell(\rho_\ell)$ is a non-split extension of $\pi'_\ell \otimes |\det|^{-\frac{1}{2}}$ by a suitable twist of the Steinberg representation). For $\ell = p$, recall we have the unitary admissible Banach space representation $B(\rho_p)$ of §2.3. The following theorem is currently being proven by Emerton ([28]).

Theorem 2.7. *Let $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(E)$ be a linear continuous representation which is unramified outside a finite set of primes and such that the determinant of one (or equivalently any) complex conjugation is -1. Let $\overline{\rho} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(k_E)$ be the semi-simplification modulo p of ρ. Assume $p > 2$, $\overline{\rho}|_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}(\sqrt[p]{1}))}$ irreducible and $\overline{\rho}|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)} \not\cong \left(\begin{smallmatrix} 1 & * \\ 0 & \omega \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right)$ up to twist. Then the $\mathrm{GL}_2(\mathbb{A}_f)$-representation $\mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(\rho, \widehat{H}^1)$ decomposes as a restricted tensor product:*[2]

$$\mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(\rho, \widehat{H}^1) \simeq B(\rho_p) \otimes_E \big(\otimes'_{\ell \neq p} \pi_\ell(\rho_\ell)\big).$$

Note that this theorem in particular states that $\mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(\rho, \widehat{H}^1)$ is always non-zero. It is thus at the same time a local-global compatibility result *and* a modularity result! When ρ comes from a modular form (so that one already knows $\mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(\rho, \widehat{H}^1) \neq 0$) and when moreover ρ_p is semi-stable, it was proven in [8], [5] and [13] that, for a suitable K_f^p, one has $\mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(\rho, \widehat{H}^1(K_f^p)) \simeq B(\rho_p)$. These results were the first cohomological incarnations of the representations $B(V)$ of §2.3. Note that the case where ρ_p is crystalline and irreducible is easy here. Indeed, as ρ is modular, one knows that the locally algebraic representation $\det^{w_1} \otimes_E \mathrm{Sym}^{w_2-w_1-1}(E^2) \otimes_E \pi_p$ (see property (iv) of §2.2) embeds into $\widehat{H}^1(K_f^p)$ and its closure has to be $B(\rho_p)$ since this Banach space is its only unitary completion (see the end of §2.2).

[2]Depending on normalizations, one may have to replace ρ_p and the ρ_ℓ here by their duals or their Cartier duals.

The proof of Theorem 2.7 uses many ingredients, such as the aforementioned local-global compatibility in the crystalline case, the density in the space of all ρ of those ρ such that ρ_p is crystalline, Serre's modularity conjecture ([44]), Colmez's last functor at the end of §2.3, Mazur's deformation theory, Kisin's construction of $D(V)$ ([46]), etc. In fact, Theorem 2.7 is a consequence of an even stronger result giving a *full* description of the $\mathrm{GL}_2(\mathbb{A}_f)$-representation $\widehat{H}^1_{\overline{\rho}}$ (where $\widehat{H}^1_{\overline{\rho}}$ is the localization of $\widehat{H}^1$ at the maximal Hecke ideal defined by $\overline{\rho}$) and not just of $\mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(\rho, \widehat{H}^1) = \mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(\rho, \widehat{H}^1_{\overline{\rho}})$ ([28]).

3. The Group $\mathrm{GL}_2(F)$

After the group $\mathrm{GL}_2(\mathbb{Q}_p)$, it is natural to look at the group $\mathrm{GL}_2(F)$, where many new phenomena appear and where the theory is thus still in its infancy. We describe below some of these new aspects, starting with the modulo p theory.

3.1. Why the $\mathrm{GL}_2(\mathbb{Q}_p)$ theory cannot extend directly. Let us start with reducible 2-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over k_E. One of the first naive hopes in order to extend the modulo p Langlands correspondence from $\mathrm{GL}_2(\mathbb{Q}_p)$ to $\mathrm{GL}_2(F)$ in that case (see (ii) of Definition 2.2) was the following: since, if $F = \mathbb{Q}_p$, the unique non-split (resp. split) $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$-extension:

$$\begin{pmatrix} \chi_1 & * \\ 0 & \chi_2 \end{pmatrix}$$

corresponds to the unique non-split (resp. split) $\mathrm{GL}_2(\mathbb{Q}_p)$-extension:

$$0 \longrightarrow \mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_2 \otimes \chi_1\omega^{-1} \longrightarrow * \longrightarrow \mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_1 \otimes \chi_2\omega^{-1} \longrightarrow 0$$

(at least in "generic" cases) then for general F the space of extensions:

$$\mathrm{Ext}^1_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)}(\chi_2, \chi_1)$$

(which has generic dimension $[F : \mathbb{Q}_p]$) would hopefully be (canonically) isomorphic to the space of extensions:

$$\mathrm{Ext}^1_{\mathrm{GL}_2(F)}\left(\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_1 \otimes \chi_2\omega^{-1}, \mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_2 \otimes \chi_1\omega^{-1}\right)$$

thus yielding a nice correspondence.

Unfortunately, this turned out to be completely wrong.

Theorem 3.1. *Assume $F \neq \mathbb{Q}_p$. For $\chi_1 \neq \chi_2$ one has:*

$$\mathrm{Ext}^1_{\mathrm{GL}_2(F)}\left(\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_1 \otimes \chi_2\omega^{-1}, \mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_2 \otimes \chi_1\omega^{-1}\right) = 0.$$

Proof. This follows from [16, Thm.8.1] together with [16, Thm.7.16(i)] and [16, Cor.6.6]. □

Remark 3.2. In fact, at least for $\chi_1 \neq \chi_2$ and $\chi_1 \neq \chi_2\omega^{\pm 1}$, one can prove that $\mathrm{Ext}^i_{\mathrm{GL}_2(F)}\big(\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)}\chi_1 \otimes \chi_2\omega^{-1}, \mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)}\chi_2 \otimes \chi_1\omega^{-1}\big) = 0$ for $0 \leq i \leq [F:\mathbb{Q}_p]-1$ and that $\mathrm{Ext}^{[F:\mathbb{Q}_p]}_{\mathrm{GL}_2(F)}\big(\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)}\chi_1 \otimes \chi_2\omega^{-1}, \mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)}\chi_2 \otimes \chi_1\omega^{-1}\big)$ has dimension 1.

Let us now consider irreducible 2-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over k_E. Just as for $F = \mathbb{Q}_p$, we define the following smooth representations of $\mathrm{GL}_2(F)$:

$$\pi(\sigma, 0) := \Big(\mathrm{c}-\mathrm{Ind}_{\mathrm{GL}_2(\mathcal{O}_F)F^\times}^{\mathrm{GL}_2(F)}\sigma\Big)/(T)$$

where σ is a Serre weight for $\mathrm{GL}_2(\mathcal{O}_F)F^\times$ (the definition of T holds for any F). Recall that for $F = \mathbb{Q}_p$ the representations $\pi(\sigma, 0)$ are all irreducible admissible.

Again, this turns out to be wrong for $F \neq \mathbb{Q}_p$.

Theorem 3.3. *Assume $F \neq \mathbb{Q}_p$. For any Serre weight σ the representation $\pi(\sigma, 0)$ is of infinite length and is not admissible.*

Proof. When k_F is strictly bigger than $\mathbb{F}_p$, this can be derived from the results of [16], in particular Theorem 3.4 below. When $k_F = \mathbb{F}_p$, one can prove (by an explicit calculation) that $\pi(\sigma, 0)$ contains $\mathrm{c}-\mathrm{Ind}_{\mathrm{GL}_2(\mathcal{O}_F)F^\times}^{\mathrm{GL}_2(F)}\sigma'$ for some Serre weight σ', which implies both statements as this representation is neither of finite length nor admissible. □

3.2. So many representations of $\mathrm{GL}_2(F)$.

We survey most of the results so far on smooth admissible representations of $\mathrm{GL}_2(F)$ over k_E.

It is not known how to define an irreducible quotient of $\pi(\sigma, 0)$ by *explicit* equations, although we know such quotients exist by an abstract argument using Zorn's lemma ([2]). The classification of all irreducible representations of $\mathrm{GL}_2(F)$ over k_E with a central character remains thus unsettled. But one can prove that there exist many irreducible admissible quotients of $\pi(\sigma, 0)$ with, for instance, a given $\mathrm{GL}_2(\mathcal{O}_F)F^\times$-socle (containing σ). This is enough to show that irreducible representations of $\mathrm{GL}_2(F)$ over k_E are far more "numerous" than irreducible representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ over k_E. This also turns out to be useful as the representations of $\mathrm{GL}_2(F)$ appearing in étale cohomology groups over k_E analogous to (1) are expected to have specific $\mathrm{GL}_2(\mathcal{O}_F)F^\times$-socles (see §3.3 below).

Denote by $\mathcal{N}(F)$ the normalizer of the Iwahori subgroup I inside $\mathrm{GL}_2(F)$, that is, $\mathcal{N}(F)$ is the subgroup of $\mathrm{GL}_2(F)$ generated by I, the scalars $F^\times$ and the matrix $\left(\begin{smallmatrix} 0 & 1 \\ \varpi_F & 0 \end{smallmatrix}\right)$. The following theorem was proved in [16, §9] using and

generalizing constructions of Paškūnas based on the existence and properties of injective envelopes of Serre weights for $\mathrm{GL}_2(\mathcal{O}_F)F^\times$ ([53]).

Theorem 3.4. *Assume $p > 2$. Let D_0 be a finite dimensional smooth representation of $\mathrm{GL}_2(\mathcal{O}_F)F^\times$ over k_E with a central character and $D_1 \subseteq D_0|_{IF^\times}$ a non-zero subrepresentation of $IF^\times$. For each k_E-linear action of $\mathcal{N}(F)$ on D_1 that induces the $IF^\times$-action, there exists a smooth admissible representation π of $\mathrm{GL}_2(F)$ over k_E with a central character such that the following diagram commutes:*

$$\begin{tikzcd} D_0 \arrow[r, hook] & \pi \arrow[d, equal] \\ D_1 \arrow[u, hook] \arrow[r, hook] & \pi \end{tikzcd}$$

(where the two horizontal injections are respectively $\mathrm{GL}_2(\mathcal{O}_F)F^\times$ and $\mathcal{N}(F)$-equivariant), such that π is generated by D_0 under $\mathrm{GL}_2(F)$ and such that :

$$\mathrm{socle}(\pi|_{\mathrm{GL}_2(\mathcal{O}_F)F^\times}) = \mathrm{socle}(D_0).$$

In general, it is not straightforward to construct explicitly such pairs (D_0, D_1) with a compatible action of $\mathcal{N}(F)$ on D_1, but there is one case where it is: the case where the pro-p subgroup I_1 of I acts trivially on D_1, for instance if $D_1 = D_0^{I_1}$ (which is never 0 as I_1 is pro-p). Indeed, in that case, D_1 is just a direct sum of characters of $IF^\times$ (as I/I_1 has order prime to p) and an action of $\left(\begin{smallmatrix} 0 & 1 \\ \varpi_F & 0 \end{smallmatrix}\right)$ is then essentially a certain permutation of order 2 on these characters. Moreover for such pairs (D_0, D_1) the assumption $p > 2$ in Theorem 3.4 is unnecessary. These examples are enough to show that there are infinitely many irreducible admissible non-isomorphic quotients of the representations $\pi(\sigma, 0)$, for instance because there are infinitely many D_0 containing σ for which there exist many non-isomorphic compatible actions of $\mathcal{N}(F)$ on $D_0^{I_1}$ such that any π as in Theorem 3.4 is irreducible and is not a subquotient of a principal series (see [16] when k_F is not $\mathbb{F}_p$).

We now give two series of examples of such pairs (D_0, D_1).

The first examples are very explicit and arise from the generalization of Serre's modularity conjecture in [18] (see also [58]). For these examples we assume F unramified over $\mathbb{Q}_p$. To any linear continuous 2-dimensional representation ρ of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over k_E is associated in [18] a finite set $\mathcal{W}(\rho)$ of Serre weights which generically has $2^{[F:\mathbb{Q}_p]}$ elements. Let $D_0(\rho)$ be a linear representation of $\mathrm{GL}_2(\mathcal{O}_F)F^\times$ over k_E such that:

(i) $\mathrm{soc}_{\mathrm{GL}_2(\mathcal{O}_F)F^\times} D_0(\rho) = \oplus_{\sigma \in \mathcal{W}(\rho)} \sigma$

(ii) the action of $\mathrm{GL}_2(\mathcal{O}_F)$ on $D_0(\rho)$ factors through $\mathrm{GL}_2(\mathcal{O}_F) \twoheadrightarrow \mathrm{GL}_2(k_F)$

(iii) $D_0(\rho)$ is maximal for inclusion with respect to (i) and (ii).

If ρ is sufficiently generic (in a sense that can be made precise, see [16, §11]), one can prove that such a $D_0(\rho)$ exists, is unique, and that $D_1(\rho) := D_0(\rho)^{I_1}$ can be endowed with (many) compatible actions of $\mathcal{N}(F)$. For each such action of $\mathcal{N}(F)$, Theorem 3.4 applied to $(D_0, D_1) := (D_0(\rho), D_1(\rho))$ gives a smooth admissible representation π of $\mathrm{GL}_2(F)$. In fact, based on explicit computations in special cases ([43]), it is expected that the number of isomorphism classes of π as in Theorem 3.4 will be *strictly bigger* than one for *each* action of $\mathcal{N}(F)$ on $D_1(\rho)$ as soon as $F \neq \mathbb{Q}_p$. Denote by $\Pi(\rho)$ the set of isomorphism classes of all π given by Theorem 3.4 for all compatible actions of $\mathcal{N}(F)$ on $D_1(\rho)$. The following result is proved in [16].

Theorem 3.5. *If ρ is (sufficiently generic and) irreducible, then any π in $\Pi(\rho)$ is irreducible. If ρ is (sufficiently generic and) reducible, then any π in $\Pi(\rho)$ is reducible.*

Remark 3.6. When ρ is irreducible, one could replace $D_0(\rho)$ by its subrepresentation $\langle \mathrm{GL}_2(\mathcal{O}_F) \cdot D_1(\rho) \rangle$ as one can prove that any π as in Theorem 3.4 for $(\langle \mathrm{GL}_2(\mathcal{O}_F) \cdot D_1(\rho) \rangle, D_1(\rho))$ contains $D_0(\rho)$ in that case, i.e., is in $\Pi(\rho)$.

In the case ρ is reducible, $\rho \simeq \begin{pmatrix} \chi_1 & * \\ 0 & \chi_2 \end{pmatrix}$, any π in $\Pi(\rho)$ is reducible because it strictly contains the representation $\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_2 \otimes \chi_1 \omega^{-1}$. By Theorem 3.1 (which can be applied as the genericity of ρ entails in particular $\chi_1 \neq \chi_2$), it cannot be an extension of $\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_1 \otimes \chi_2 \omega^{-1}$ by $\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)} \chi_2 \otimes \chi_1 \omega^{-1}$. So what could π look like in this case? Consider the following two propositions, the first one being in [16, §19] and the second one being elementary.

Proposition 3.7. *If ρ is (sufficiently generic and) reducible split, then some of the π in $\Pi(\rho)$ are semi-simple with $[F : \mathbb{Q}_p] + 1$ non-isomorphic Jordan-Hölder factors, two of them being the above two principal series (which are irreducible for ρ sufficiently generic) and the others being irreducible admissible quotients of representations $\pi(\sigma, 0)$.*

Proposition 3.8. *If ρ is (sufficiently generic and) reducible, then the tensor induction of ρ from $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ to $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ is a successive extension of $[F : \mathbb{Q}_p] + 1$ non-isomorphic semi-simple representations, two of them being 1-dimensional.*

If ρ is reducible non-split, then any extension between two consecutive semi-simple representations as in Proposition 3.8 is non-split and the two 1-dimensional representations are the unique irreducible subobject and the unique irreducible quotient of the tensor induction of ρ. If ρ is split, the link between the $[F : \mathbb{Q}_p]+1$ Jordan-Hölder factors of π in Proposition 3.7 and the $[F : \mathbb{Q}_p]+1$ semi-simple representations of the tensor induction of ρ in Proposition 3.8 can be made much more convincing by using (φ, Γ)-modules (see [10]). In particular, the above two propositions suggest that, among all the π in $\Pi(\rho)$ for ρ

reducible and sufficiently generic, some of them (the "good" ones) should have exactly $[F:\mathbb{Q}_p]+1$ distinct Jordan-Hölder factors as in Proposition 3.7 and should be:

(i) semi-simple if ρ is reducible split, or

(ii) uniserial with $\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)}\chi_2\otimes\chi_1\omega^{-1}$ as unique irreducible subobject and $\mathrm{Ind}_{B(F)}^{\mathrm{GL}_2(F)}\chi_1\otimes\chi_2\omega^{-1}$ as unique irreducible quotient if ρ is reducible non-split.

This gives a possible explanation for Theorem 3.1: there is no extension for $F\neq\mathbb{Q}_p$ because we are missing the "middle" Jordan-Hölder factors!

The second examples of pairs (D_0,D_1) are constructed in [42] (no assumption on F here). Let π be an irreducible *not necessarily admissible* representation of $\mathrm{GL}_2(F)$ over k_E with a central character and $\sigma\subset\pi|_{\mathrm{GL}_2(\mathcal{O}_F)F^\times}$ a Serre weight for $\mathrm{GL}_2(\mathcal{O}_F)F^\times$ (which always exists). One first defines an $\mathcal{N}(F)$-subrepresentation $D_1(\pi)$ of π as follows (with notations analogous to (2)):

$$D_1(\pi):=\left(\sum_{m\geq 0}\begin{pmatrix}\varpi_F^m & \mathcal{O}_F\\ 0 & 1\end{pmatrix}\sigma\right)\bigcap\begin{pmatrix}0 & 1\\ \varpi_F & 0\end{pmatrix}\left(\sum_{m\geq 0}\begin{pmatrix}\varpi_F^m & \mathcal{O}_F\\ 0 & 1\end{pmatrix}\sigma\right)$$

(which is checked to be preserved by $\mathcal{N}(F)$ inside π). One can prove that $D_1(\pi)$ does not depend on the choice of the Serre weight σ in π and that it always contains π^{I_1}. One then considers the pair $(D_0(\pi),D_1(\pi))$ with $D_0(\pi):=\langle\mathrm{GL}_2(\mathcal{O}_F)\cdot D_1(\pi)\rangle\subset\pi$.

Theorem 3.9. *If $D_1(\pi)$ is finite dimensional, then there is a unique representation of $\mathrm{GL}_2(F)$ as in Theorem 3.4 with $(D_0,D_1):=(D_0(\pi),D_1(\pi))$ (even if $p=2$) and it is the representation π. In particular π is then admissible.*

This theorem is proved in [42]. In fact, [42] proves more: (i) without any assumption on $D_1(\pi)$ the pair $(D_0(\pi),D_1(\pi))$ *always* uniquely determines π and (ii) $D_1(\pi)$ is finite dimensional if and only if π is of finite presentation (i.e., is a quotient of some $\mathrm{c}-\mathrm{Ind}_{\mathrm{GL}_2(\mathcal{O}_F)F^\times}^{\mathrm{GL}_2(F)}\sigma$ by an invariant subspace which is finitely generated under $\mathrm{GL}_2(F)$). However, if $F\neq\mathbb{Q}_p$ it is not known in general whether $D_1(\pi)$ is or isn't finite dimensional, and it seems quite hard to determine $D_1(\pi)$ explicitly if π is not a subquotient of a principal series. For those π in $\Pi(\rho)$, note that one has the inclusions $D_1(\rho)\subseteq\pi^{I_1}\subseteq D_1(\pi)$ hence also $\langle\mathrm{GL}_2(\mathcal{O}_F)\cdot D_1(\rho)\rangle\subseteq D_0(\pi)$ with equalities if $F=\mathbb{Q}_p$.

3.3. Questions on local-global compatibility. We conclude our discussion of the modulo p theory for $\mathrm{GL}_2(F)$ with questions on local-global compatibility.

Let L be a totally real finite extension of $\mathbb{Q}$ with ring of integers $\mathcal{O}_L$. Assume for simplicity that p is inert in L (i.e., $p\mathcal{O}_L$ is a prime ideal) and let L_p denote the

completion of L at p and $\mathbb{A}^p_{L,f}$ the finite adèles of L outside p. To any quaternion algebra D over L which splits at only one of the infinite places and which splits at p and to any compact open subgroup $K^p_f \subset (D \otimes_L \mathbb{A}^p_{L,f})^\times$, one can associate a tower of Shimura algebraic curves $(S(K^p_f K_{f,p}))_{K_{f,p}}$ over L where $K_{f,p}$ runs over the compact open subgroups of $(D \otimes_L L_p)^\times \simeq \mathrm{GL}_2(L_p)$. Analogously to the case $L = \mathbb{Q}$ and $D = \mathrm{GL}_2$ of §2.4, one would like to understand:

$$\varinjlim_{K_{f,p}} H^1_{\text{ét}}\big(S(K^p_f K_{f,p}) \times_L \overline{\mathbb{Q}}, k_E\big)$$

as a representation of $\mathrm{GL}_2(L_p) \times \mathrm{Gal}(\overline{\mathbb{Q}}/L)$ over k_E. Fix a linear continuous totally odd (i.e., any complex conjugation has determinant -1) irreducible representation:

$$\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/L) \to \mathrm{GL}_2(k_E).$$

One can at least state the following conjecture which generalizes one of the main conjectures of [18].

Conjecture 3.10. *If $\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/L_p)}$ is sufficiently generic (in the sense of [16, §11]) then for each compact open subgroup $K^p_f \subset (D \otimes_L \mathbb{A}^p_{L,f})^\times$ one has:*

$$\mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/L)}\left(\rho, \varinjlim_{K_{f,p}} H^1_{\text{ét}}\big(S(K^p_f K_{f,p}) \times_L \overline{\mathbb{Q}}, k_E\big)\right) \simeq \pi_p^n$$

for some integer $n \geq 0$ and some π_p in the set[3] $\Pi(\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/L_p)})$ *(see §3.2).*

Note that Conjecture 3.10 does not state that the above space of homomorphisms is non-zero, but that, *if* it is non-zero, then it is a number of copies of some π_p in $\Pi(\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/L_p)})$. Conjecture 3.10 is known for $L = \mathbb{Q}$ ([28]). For $L \neq \mathbb{Q}$, some non-trivial evidence for this conjecture and for a variant with 0-dimensional Shimura varieties and H^0 (instead of Shimura curves and H^1) can be found in [18], [57], [33] and [11] (see also [58] and [35]). If Conjecture 3.10 holds, the main crucial questions are then (recalling from §3.2 that $\Pi(\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/L_p)})$ is a huge set if $L_p \neq \mathbb{Q}_p$):

Question 3.11. *Does π_p in Conjecture 3.10 only depend on $\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/L_p)}$? How can one "distinguish" the π_p of Conjecture 3.10 in the purely local set $\Pi(\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/L_p)})$?*

If the answer to the first question is yes, then this will enable one to define a genuine modulo p local Langlands correspondence for $\mathrm{GL}_2(F)$ that is compatible with cohomology. Again, the answer is of course yes if $L = \mathbb{Q}$.

[3] As in Theorem 2.7, depending on normalizations, one may have to replace $\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/L_p)}$ here by its dual or its Cartier dual.

3.4. Over E. The modulo p theory being so involved, it is not surprising that very little is known in characteristic 0. We just state here the main theorem of [54], which shows that one also has too many admissible unitary topologically irreducible Banach space representations of $\mathrm{GL}_2(F)$ over E when $F \neq \mathbb{Q}_p$.

Theorem 3.12. *Let π be a smooth irreducible admissible representation of $\mathrm{GL}_2(F)$ over k_E. Then there exists an admissible unitary topologically irreducible Banach space representation B of $\mathrm{GL}_2(F)$ over E and a unit ball $B^0 \subset B$ preserved by $\mathrm{GL}_2(F)$ such that:*

$$\mathrm{Hom}_{\mathrm{GL}_2(F)}(\pi, B^0 \otimes_{\mathcal{O}_E} k_E) \neq 0.$$

In particular, because of the results of §3.2, one should not expect a naive extension of Theorem 2.6 to hold for $F \neq \mathbb{Q}_p$. The question whether one can always choose B above such that $\pi \simeq B^0 \otimes_{\mathcal{O}_E} k_E$ is open (except for $F = \mathbb{Q}_p$ where the answer is yes and is already essentially in [7]). If such a B does not always exist, maybe one should only consider those π for which it does, i.e., those π which lift to characteristic 0.

All the other results concerning $\mathrm{GL}_2(F)$ over E are very partial so far. In some cases, one can for instance associate to a 2-dimensional semi-stable p-adic representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over E a locally $\mathbb{Q}_p$-analytic strongly admissible (in the sense of [61]) representation of $\mathrm{GL}_2(F)$ over E which generalizes the representation from the $F = \mathbb{Q}_p$ case and that one would wish to find inside completed cohomology spaces analogous to the $\widehat{H}^1(K_f^p)$ of §2.4 (see, e.g., [65] for the non-crystalline case). However, if this holds, it is likely that for $F \neq \mathbb{Q}_p$ this locally $\mathbb{Q}_p$-analytic representation is only a strict subrepresentation of the "correct" (unknown) locally $\mathbb{Q}_p$-analytic representation(s) of $\mathrm{GL}_2(F)$.

4. Other Groups

If not much is known for $\mathrm{GL}_2(F)$, almost nothing is known for groups other than $\mathrm{GL}_2(F)$, even conjecturally, although some non-trivial results start to appear in various cases like $\mathrm{GL}_3(\mathbb{Q}_p)$ ([66]) or quaternion algebras ([36]). We content ourselves here to mention briefly a few results and conjectures that have been stated for $\mathrm{GL}_n(F)$ and that give some kind of "relations" between the $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ side and the $\mathrm{GL}_n(F)$ side. Although these relations are very far from any kind of correspondence, it is plausible that they will play some role in the future.

4.1. Invariant lattices and admissible filtrations. Locally algebraic representations of $\mathrm{GL}_n(F)$ over E (such as the representations $B(V)^{\mathrm{alg}}$ of §2.2) which are "related" to continuous n-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over E (e.g., that appear as subrepresentations in completed cohomology spaces) should have invariant $\mathcal{O}_E$-lattices, as is clear from the $\mathrm{GL}_2(\mathbb{Q}_p)$

case (§2.2). It turns out that a necessary condition for a locally algebraic representation of $\mathrm{GL}_n(F)$ to have invariant lattices is essentially a well-known condition in Fontaine's theory called "weakly admissible".

Let us fix (r, N, D) a Weil-Deligne representation on an n-dimensional E-vector space D where r is the underlying representation of the Weil group of F (which has open kernel) and N the nilpotent endomorphism on D satisfying the usual relation $r(w) \circ N \circ r(w)^{-1} = p^{d(w)} N$ (for w in the Weil group of F, see introduction for $d(w)$).

To (r, N, D), one can associate a smooth irreducible representation π' of $\mathrm{GL}_n(F)$ over E by the classical local Langlands correspondence as in [38] (after semi-simplifying r). We then slightly modify it as in §2.4: if π' is generic, we let $\pi := \pi' \otimes |\det|^{\frac{(1-n)}{2}}$. If π' is not generic, we replace π' by a certain parabolic induction π'' which has π' as unique irreducible quotient (see [17, §4]) and let $\pi := \pi'' \otimes |\det|^{\frac{(1-n)}{2}}$.

For each embedding $\tau : F \hookrightarrow E$, let us fix n integers $i_{1,\tau} < i_{2,\tau} < \cdots < i_{n,\tau}$. We denote by σ_τ the algebraic representation of GL_n over E of highest weight $-i_{1,\tau} - (n-1) \geq -i_{2,\tau} - (n-2) \geq \cdots \geq -i_{n,\tau}$ that we see as a representation of $\mathrm{GL}_n(F)$ via the embedding $\tau : F \hookrightarrow E$. We then set $\sigma := \otimes_\tau \sigma_\tau$. This is a finite dimensional representation of $\mathrm{GL}_n(F)$ over E.

Any p-adic potentially semi-stable representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ on an n-dimensional E-vector space V gives rise to some (r, N, D) and some $(i_{j,\tau})_{j,\tau}$ as follows ([30]). Let F' be a finite Galois extension of F such that $V|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/F')}$ becomes semi-stable and set:

$$D := (\mathrm{B}_{\mathrm{st}} \otimes_{\mathbb{Q}_p} V)^{\mathrm{Gal}(\overline{\mathbb{Q}_p}/F')} \otimes_{F'_0 \otimes E} E$$

where B_{st} is Fontaine's semi-stable period ring, F'_0 is the maximal unramified subfield in F' and $F'_0 \hookrightarrow E$ is any embedding. It is an n-dimensional E-vector space endowed with a nilpotent endomorphism N coming from the one on B_{st}. We define $r(w)$ on D by $r(w) := \varphi^{-d(w)} \circ \overline{w}$ where w is any element in the Weil group of F, $\overline{w}$ its image in $\mathrm{Gal}(F'/F)$ and φ the semi-linear endomorphism coming from the action of the Frobenius on B_{st} (as $\varphi^{-d(w)} \circ \overline{w}$ is $F'_0 \otimes E$-linear, $r(w)$ goes down to D). Finally, the $i_{j,\tau}$ are just the opposite of the various Hodge-Tate weights of V (n weights for each embedding $\tau : F \hookrightarrow E$).

The following conjecture was stated in [17, §4].

Conjecture 4.1. *Fix (r, N, W) and $(i_{j,\tau})_{j,\tau}$ as above. There exists an invariant $\mathcal{O}_E$-lattice on the locally algebraic $\mathrm{GL}_n(F)$-representation $\sigma \otimes_E \pi$ if and only if the data $\big((r, N, W), (i_{j,\tau})_{j,\tau}\big)$ comes from a p-adic n-dimensional potentially semi-stable representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$.*

The following theorem gives one complete direction in the above conjecture. After many cases were proved in [63] and [17], its full proof was given in [41].

Theorem 4.2. *If there exists an invariant $\mathcal{O}_E$-lattice on $\sigma \otimes_E \pi$ then the data $\big((r, N, W), (i_{j,\tau})_{j,\tau}\big)$ comes from a p-adic n-dimensional potentially semi-stable representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$.*

The proof is divided into four steps. (i) It is essentially trivial if π is supercuspidal. Hence one can restrict to the non-supercuspidal cases. (ii) Using a result of Emerton ([26, Lem.4.4.2]), one deduces from the existence of an invariant lattice on $\sigma \otimes_E \pi$ a finite number of inequalities relating the numbers $i_{j,\tau}$ to the "powers of p" in the action of r. (iii) These inequalities are just what is needed so that there exists a weakly admissible filtration on a certain (φ, N)-module naturally associated to (r, N, W). (iv) Such a filtration gives an n-dimensional potentially semi-stable representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ by the main result of [22].

The other direction in Conjecture 4.1 is much harder. Apart from trivial or scattered partial results, the only case which is completely known is again that of $\mathrm{GL}_2(\mathbb{Q}_p)$ ([5]).

4.2. Supersingular modules and irreducible Galois representations. We now state a theorem on Hecke-Iwahori modules for $\mathrm{GL}_n(F)$ over k_E in relation with irreducible n-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over k_E.

Let $\mathcal{H}_1$ be the Hecke algebra of I_1 over k_E, that is, $\mathcal{H}_1 := k_E[I_1 \backslash \mathrm{GL}_n(F)/I_1]$. The usual product of double cosets makes $\mathcal{H}_1$ a non-commutative k_E-algebra of finite type. An $\mathcal{H}_1$-module M over k_E is a k_E-vector space endowed with a linear right action of $\mathcal{H}_1$. By Schur's lemma, the center $\mathcal{Z}_1$ of $\mathcal{H}_1$ acts on a simple (and thus finite dimensional) $\mathcal{H}_1$-module M by a character with values in k_E called the central character of M. The commutative k_E-subalgebra $\mathcal{Z}_1$ is generated by (cosets of) scalars, by certain elements of $k_E[I_1 \backslash I/I_1] = k_E[I/I_1]$ and by $n-1$ cosets $Z_1, \cdots, Z_{n-1}$. A finite dimensional simple $\mathcal{H}_1$-module is said to be supersingular if its central character sends all these Z_i to 0 ([71]).

The following nice numerical coincidence was conjectured in [71] and completely proved in [52].

Theorem 4.3. *The number of simple n-dimensional supersingular $\mathcal{H}_1$-modules over k_E is equal to the number of linear continuous n-dimensional irreducible representations of* $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ *over k_E.*

Let us briefly give the case $n = 3$ as an example. The number of (isomorphism classes of) continuous 3-dimensional irreducible representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over k_E with determinant mapping a fixed choice of Frobenius to $1 \in k_E$ is easily checked to be $\frac{q^3-q}{3}$. The number of 3-dimensional simple supersingular $\mathcal{H}_1$-modules over k_E with central character mapping a fixed choice of uniformizer to $1 \in k_E$ turns out to be:

$$2\left(q - 1 + (q-1)(q-2) + \frac{(q-1)(q-2)(q-3)}{6}\right).$$

The reader can check that these two numbers are just the same (whence the theorem for $n = 3$ by varying the central character/determinant).

For $(n, F) = (2, \mathbb{Q}_p)$, the functor $\pi \mapsto \pi^{I_1}$ induces a bijection between smooth irreducible supersingular representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ over k_E and 2-dimensional simple supersingular $\mathcal{H}_1$-modules over k_E ([70]), but this already completely breaks down when $n = 2$ and $F \neq \mathbb{Q}_p$ (see §3). The meaning of Theorem 4.3 in terms of smooth representations of $\mathrm{GL}_n(F)$ over k_E (if any) thus remains mysterious for $(n, F) \neq (2, \mathbb{Q}_p)$.

4.3. Serre weights and Galois representations. We have seen in §3 that the set of Serre weights $\mathcal{W}(\rho)$ associated in [18] and [58] to a linear continuous 2-dimensional representation of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/F)$ over k_E is expected to be the set of simple summands (forgetting possible multiplicities) of the $\mathrm{GL}_2(\mathcal{O}_F)F^\times$-socle of some smooth admissible representation of $\mathrm{GL}_2(F)$ over k_E. (Without restrictions on ρ, one may indeed have multiplicities in this socle.) This yields a non-trivial link between the weights in Serre-type conjectures and the modulo p Langlands programme for $\mathrm{GL}_2(F)$.

For $\mathrm{GL}_n(\mathbb{Q}_p)$ when $n > 2$ the modulo p Langlands programme is essentially open (although there is recent progress in the classification of "non-supersingular" smooth irreducible admissible representations of $\mathrm{GL}_n(F)$ over k_E, see [40]). But the set of Serre weights $\mathcal{W}(\rho)$ has been generalized by Herzig and Gee in [39] and [34] to linear continuous n-dimensional representations of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ over k_E.

For integers $a_1 \geq a_2 \geq \cdots \geq a_n$ such that $a_i - a_{i+1} \leq p - 1$ for all i we let $F(a_1, \cdots, a_n)$ denote the restriction to $\mathrm{GL}_n(\mathbb{F}_p)$ of the GL_n-socle of the algebraic dual Weyl module for GL_n of highest weight $(t_1, \cdots, t_n) \mapsto t_1^{a_1} t_2^{a_2} \cdots t_n^{a_n}$ (see [39, §3.1]). The $F(a_1, \cdots, a_n)$ exhaust the irreducible representations of $\mathrm{GL}_n(\mathbb{F}_p)$ (equivalently of $\mathrm{GL}_n(\mathbb{Z}_p)$) over k_E.

Let $\rho : \mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p) \to \mathrm{GL}_n(k_E)$ be any linear continuous representation. Its determinant has the form $\omega^m \mathrm{unr}$ where unr is an unramified character of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ and m an integer. We can see unr as a character of $\mathrm{GL}_n(\mathbb{Z}_p)\mathbb{Q}_p^\times$ which is trivial on $\mathrm{GL}_n(\mathbb{Z}_p)$.

Definition 4.4. The set $\mathcal{W}(\rho)$ of Serre weights for $\mathrm{GL}_n(\mathbb{Z}_p)\mathbb{Q}_p^\times$ associated to ρ is the set of $F(a_1, \cdots, a_n) \otimes \mathrm{unr}$ such that ρ has a crystalline lift with Hodge-Tate weights $a_1 + n - 1, a_2 + n - 2, \cdots, a_n$.

Definition 4.4 is quite general but not at all explicit. When ρ is sufficiently generic and semi-simple, a conjectural but much more explicit description of the weights of $\mathcal{W}(\rho)$ has been given in [39] (which was actually written before [34]). The method of [39] is first to associate to ρ (in fact to its restriction to the inertia subgroup of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$) a finite dimensional "Deligne-Lusztig" representation $\sigma(\rho)$ of $\mathrm{GL}_n(\mathbb{F}_p)$ over E. For instance, if $\rho = \oplus_{i=1}^n \chi_i$ is a direct sum of characters of $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$, then:

$$\sigma(\rho) := \mathrm{Ind}_{B(\mathbb{F}_p)}^{\mathrm{GL}_n(\mathbb{F}_p)} \chi$$

where $\chi : B(\mathbb{F}_p) \twoheadrightarrow T(\mathbb{F}_p) \to E^\times$, $(t_1, \cdots, t_n) \mapsto [\chi_1(t_1)\chi_2(t_2)\cdots\chi_n(t_n)]$ (note that $\chi_i|_{1+p\mathbb{Z}_p} = 1$ and $[\cdot]$ is here the multiplicative representative). If ρ is sufficiently generic, all Jordan-Hölder factors $F(a_1, \cdots, a_n)$ of the modulo p semi-simplification $\overline{\sigma}(\rho)^{\mathrm{ss}}$ of $\sigma(\rho)$ (that is, of the semi-simplification of the reduction modulo ϖ_E of any invariant $\mathcal{O}_E$-lattice in $\sigma(\rho)$) are such that $a_i - a_{i+1} \leq p-2$ for all i. If ρ is moreover semi-simple, the set $\mathcal{W}(\rho)$ of Definition 4.4 is then expected to be the set of Serre weights:

$$F\big(a_n + (n-1)(p-2), a_{n-1} + (n-2)(p-2), \cdots, a_2 + p - 2, a_1\big) \otimes \mathrm{unr}$$

for $F(a_1, \cdots, a_n)$ a Jordan-Hölder factor of $\overline{\sigma}(\rho)^{\mathrm{ss}}$.

Changing notations, let:

$$\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_n(k_E)$$

be a linear continuous irreducible odd representation, that is, either $p = 2$ or the eigenvalues of the image of a complex conjugation are:

$$\underbrace{1, \cdots, 1}_{n_+ \text{ times}}, \underbrace{-1, \cdots, -1}_{n_- \text{ times}}$$

with $-1 \leq n_+ - n_- \leq 1$. Let N be the Artin conductor of ρ measuring its ramification at primes other than p and let unr_p be as above the unramified part of $\det(\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)})$. Then the "Serre conjecture" of [39] and [33] states that the Serre weights of $\mathcal{W}(\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)})$ should be exactly those Serre weights $F(a_1, \cdots, a_n) \otimes \mathrm{unr}_p$ for $\mathrm{GL}_n(\mathbb{Z}_p)\mathbb{Q}_p^\times$ such that ρ "arises" from a non-zero Hecke eigenclass in some group cohomology $H^*(\Gamma_1(N), F(a_1, \cdots, a_n))$. Here $\Gamma_1(N) \subset \mathrm{SL}_n(\mathbb{Z})$ is the subgroup of matrices with last row congruent to $(0, \cdots, 0, 1)$ modulo N (see [39, §6] for details).

The results of §3 suggest that the Serre weights of $\mathcal{W}(\rho|_{\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)})$ may form (up to multiplicities) the $\mathrm{GL}_n(\mathbb{Z}_p)\mathbb{Q}_p^\times$-socle of interesting smooth admissible representations of $\mathrm{GL}_n(\mathbb{Q}_p)$ over k_E (that remain to be discovered if $n > 2$). But one should keep in mind the following numbers. Assuming ρ is semi-simple, for $n = 2$ one has generically $|\mathcal{W}(\rho)| = 2$, and for $n = 3$ one should have $|\mathcal{W}(\rho)| = 9$, but then $\mathcal{W}(\rho)$ rapidly grows: $n = 4$ should give $|\mathcal{W}(\rho)| = 88$ and $n = 5$ should give $|\mathcal{W}(\rho)| = 1640$! Also, consider for instance the case $n = 3$ and $\rho = \oplus_{i=1}^3 \chi_i$ with ρ sufficiently generic. Then 6 of the 9 weights of $\mathcal{W}(\rho)$ are easily checked to be the $\mathrm{GL}_3(\mathbb{Z}_p)\mathbb{Q}_p^\times$-socle of 6 natural principal series representations of $\mathrm{GL}_3(\mathbb{Q}_p)$ analogous to the 2 principal series in (ii) of Definition 2.2. But there are 3 remaining Serre weights and their combinatorics suggests that they might form the $\mathrm{GL}_3(\mathbb{Z}_p)\mathbb{Q}_p^\times$-socle of an irreducible admissible representation of $\mathrm{GL}_3(\mathbb{Q}_p)$ that does not occur in any (strict) parabolic induction, i.e., of a supersingular representation of $\mathrm{GL}_3(\mathbb{Q}_p)$. We thus may have a phenomenon analogous to what happens with $\mathrm{GL}_2(F)$ (see Proposition 3.7 and the discussion that follows) except that the possible appearance here of this "extra" supersingular constituent seems now quite mysterious.

References

[1] L. Barthel, R. Livné, *Modular representations of* GL_2 *of a local field: the ordinary, unramified case*, J. of Number Theory **55** (1995), 1–27.

[2] L. Barthel, R. Livné, *Irreducible modular representations of* GL_2 *of a local field*, Duke Math. J. **75** (1994), 261–292.

[3] L. Berger, *Représentations modulaires de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *et représentations galoisiennes de dimension* 2, Astérisque **330** (2010), 263–279.

[4] L. Berger, *La correspondance de Langlands locale p-adique pour* $\mathrm{GL}_2(\mathbb{Q}_p)$, Bourbaki Seminar (2010).

[5] L. Berger, C. Breuil, *Sur quelques représentations potentiellement cristallines de* $\mathrm{GL}_2(\mathbb{Q}_p)$, Astérisque **330** (2010), 155–211.

[6] C. Breuil, *Sur quelques représentations modulaires et p-adiques de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *I*, Compositio Math. **138** (2003), 165–188.

[7] C. Breuil, *Sur quelques représentations modulaires et p-adiques de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *II*, J. Inst. Math. Jussieu **2** (2003), 1–36.

[8] C. Breuil, *Invariant* $\mathcal{L}$ *et série spéciale p-adique*, Annales Scientifiques E.N.S. **37** (2004), 559–610.

[9] C. Breuil, *Invariant* $\mathcal{L}$ *et cohomologie étale complétée*, Astérisque **331** (2010), 65–115.

[10] C. Breuil, *Diagrammes de Diamond et* (φ, Γ)*-modules*, to appear, Israel J. Math.

[11] C. Breuil, *Sur un problème de compatibilité local-global modulo p pour* GL_2 (with an appendix by L. Dembélé), preprint 2009, available at http://www.ihes.fr/∼breuil/PUBLICATIONS/compamodp.pdf

[12] C. Breuil, *Introduction générale*, Astérisque **319** (2008), 1–12.

[13] C. Breuil, M. Emerton, *Représentations p-adiques ordinaires de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *et compatibilité local-global*, Astérisque **331** (2010), 255–315.

[14] C. Breuil, A. Mézard, *Multiplicités modulaires et représentations de* $\mathrm{GL}_2(\mathbb{Z}_p)$ *et de* $\mathrm{Gal}(\overline{\mathbb{Q}_p}/\mathbb{Q}_p)$ *en* $\ell = p$ (with an appendix by G. Henniart), Duke Math. J. **115** (2002), 205–310.

[15] C. Breuil, A. Mézard, *Représentations semi-stables de* $\mathrm{GL}_2(\mathbb{Q}_p)$*, demi-plan p-adique et réduction modulo p*, Astérisque **331** (2010), 117–178.

[16] C. Breuil, V. Paškūnas, *Towards a modulo p Langlands correspondence for* GL_2, to appear, Memoirs of A.M.S.

[17] C. Breuil, P. Schneider, *First steps towards p-adic Langlands funtoriality*, J. Reine Angew. Math. **610** (2007), 149–180.

[18] K. Buzzard, F. Diamond, F. Jarvis, *On Serre's conjecture for mod* ℓ *Galois representations over totally real fields*, to appear, Duke Math. J.

[19] P. Colmez, *La série principale unitaire de* $\mathrm{GL}_2(\mathbb{Q}_p)$, Astérisque **330** (2010), 213–262.

[20] P. Colmez, (φ, Γ)*-modules et représentations du mirabolique de* $\mathrm{GL}_2(\mathbb{Q}_p)$, Astérisque **330** (2010), 61–153.

[21] P. Colmez, *Représentations de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *et* (φ,Γ)*-modules*, Astérisque **330** (2010), 281–509.

[22] P. Colmez, J.-M. Fontaine, *Construction des représentations semi-stables*, Inventiones Math. **140** (2000), 1–43.

[23] P. Deligne, *Letter to Piatetski-Shapiro*, 1973.

[24] M. Emerton, *On the interpolation of systems of eigenvalues attached to automorphic Hecke eigenforms*, Inventiones Math. **164** (2006), 1–84.

[25] M. Emerton, *A local-global compatibility conjecture in the p-adic Langlands programme for* $\mathrm{GL}_{2/\mathbb{Q}}$, Pure and Applied Math. Quarterly **2** (2006), 279–393.

[26] M. Emerton, *Jacquet modules of locally analytic representations of p-adic reductive groups I. Constructions and first properties*, Ann. Scient. É.N.S. **39** (2006), 775–839.

[27] M. Emerton, *Ordinary parts of admissible representations of p-adic reductive groups II*, Astérisque **330** (2010), 403–459.

[28] M. Emerton, *Local-global compatibility in the p-adic Langlands programme for* $\mathrm{GL}_{2/\mathbb{Q}}$, in preparation.

[29] J.-M. Fontaine, *Représentations p-adiques des corps locaux I*, Progr. Math. **87** (1990), 249–309.

[30] J.-M. Fontaine, *Représentations* ℓ*-adiques potentiellement semi-stables*, Astérisque **223** (1994), 321–347.

[31] J.-M. Fontaine, B. Mazur, *Geometric* Gal*ois representations*, Elliptic curves, Modular Forms and Fermat's Last Theorem, International Press (1995), 41–78.

[32] E. Gathe, A. Mézard, *Filtered modules with coefficients*, Trans. of A.M.S. **361** (2009), 2243–2261.

[33] T. Gee, *On the weight of mod p Hilbert modular forms*, preprint 2005.

[34] T. Gee, *Automorphic lifts of precribed types*, preprint 2006.

[35] T. Gee, D. Savitt, *Serre weights for mod p Hilbert modular forms: the totally ramified case*, preprint 2009.

[36] T. Gee, D. Savitt, *Serre weights for quaternion algebras*, preprint 2009.

[37] M. Harris, R. Taylor, *The geometry and cohomology of some simple Shimura varieties*, Ann. Math. Study **151** (2001), Princeton University Press.

[38] G. Henniart, *Une preuve simple des conjectures de Langlands pour* GL_n *sur un corps p-adique*, Inventiones Math. **139** (2000), 439–455.

[39] F. Herzig, *The weight in a Serre-type conjecture for tame n-dimensional Galois representations*, Duke Math. J. **149** (2009), 37–116.

[40] F. Herzig, *The classification of irreducible admissible mod p representations of a p-adic* GL_n, preprint 2009.

[41] Y. Hu, *Normes invariantes et existence de filtrations admissibles*, to appear, J. Reine Angew. Math.

[42] Y. Hu, *Diagrammes canoniques et représentations modulo p de* $\mathrm{GL}_2(F)$, preprint 2009.

[43] Y. Hu, *Sur quelques représentations supersingulières de* $\mathrm{GL}_2(\mathbb{Q}_{p^f})$, to appear at J. Algebra.

[44] C. Khare, J.-P. Wintenberger, *On Serre's conjecture for* 2*-dimensional mod* p *representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, to appear at Annals of Maths.

[45] M. Kisin, *The Fontaine-Mazur conjecture for* GL_2, J. Amer. Math. Soc. **22** (2009), 641–690.

[46] M. Kisin, *Deformations of* $G_{\mathbb{Q}_p}$ *and* $\mathrm{GL}_2(\mathbb{Q}_p)$ *representations*, Astérisque **330** (2010), 511–528.

[47] J. Kohlhaase, *The cohomology of locally analytic representations*, preprint 2007.

[48] R. Langlands, *Problems in the theory of automorphic forms*, Lecture Notes in Maths **170** (1970), Springer, 18–86.

[49] R. Langlands, *Modular forms and* ℓ*-adic representations*, Lecture Notes in Maths **349** (1973), Springer, 361–500.

[50] Y. Morita, *Analytic representations of* SL_2 *over a* $\mathfrak{p}$*-adic number field II*, Progress in Maths **46** (1984), Birkhäuser, 282–297.

[51] M. Ohta, *On cohomology groups attached to towers of algebraic curves*, J. Math. Soc. Japan **45** (1993), 131–183.

[52] R. Ollivier, *Parabolic induction and Hecke modules in characteristic* p *for* p*-adic* $\mathrm{GL}(n)$, preprint 2009.

[53] V. Paškūnas, *Coefficient systems and supersingular representations of* $\mathrm{GL}_2(F)$, Mém. Soc. Math. de France **99** (2004).

[54] V. Paškūnas, *Admissible unitary completions of locally* $\mathbb{Q}_p$*-rational representations of* $\mathrm{GL}_2(F)$, preprint 2008.

[55] V. Paškūnas, *On some crystalline representations of* $\mathrm{GL}_2(\mathbb{Q}_p)$, Algebra and Number Theory J. **3** (2009), 411–421.

[56] V. Paškūnas, *The image of Colmez's Montréal functor*, in preparation.

[57] M. Schein, *Weights of Galois representations associated to Hilbert modular forms*, J. Reine Angew. Math. **622** (2008), 57–94.

[58] M. Schein, *Weights in Serre's conjecture for Hilbert modular forms: the ramified case*, Israel J. Math. **166** (2008), 369–391.

[59] P. Schneider, J. Teitelbaum, $U(\mathfrak{g})$*-finite locally analytic representations*, Represent. Theory **5** (2001), 111–128.

[60] P. Schneider, J. Teitelbaum, *Banach space representations and Iwasawa theory*, Israel J. Math. **127** (2002), 359–380.

[61] P. Schneider, J. Teitelbaum, *Locally analytic distributions and* p*-adic representation theory*, J. Amer. Math. Soc. **15** (2002), 443–468.

[62] P. Schneider, J. Teitelbaum, *Algebras of* p*-adic distributions and admissible representations*, Inventiones Math. **153** (2003), 145–196.

[63] P. Schneider, J. Teitelbaum, *Banach-Hecke algebras and* p*-adic Galois representations*, Documenta Math. J. Coates' sixtieth birthday (2006), 631–684.

[64] P. Schneider, M.-F. Vignéras, *A functor from smooth* $\mathcal{O}$*-torsion representations to* (φ, Γ)*-modules*, preprint 2008.

[65] B. Schraen, *Représentations p-adiques de* $\mathrm{GL}_2(F)$ *et catégories dérivées*, to appear at Israel J. Math.

[66] B. Schraen, *Représentations localement analytiques de* $\mathrm{GL}_3(\mathbb{Q}_p)$ *et invariants* $\mathcal{L}$, to appear at Annales Scientifiques E.N.S.

[67] J. Teitelbaum, *Modular representations of* PGL_2 *and automorphic forms for Shimura curves*, Inventiones Math. 113 (1993), 561–580.

[68] M.-F. Vignéras, *Correspondance de Langlands semi-simple pour* $\mathrm{GL}(n, F)$ *modulo* $\ell \neq p$, Inventiones Math. **144** (2001), 177–223.

[69] M.-F. Vignéras, *ℓ-adic Banach representations of reductive p-adic groups*, Astérisque **330** (2010), 1–11.

[70] M.-F. Vignéras, *Representations of the p-adic group* $\mathrm{GL}(2, F)$ *modulo p*, Compositio Math. **140** (2004), 333–358.

[71] M.-F. Vignéras, *On a numerical Langlands correspondence modulo p with the pro-p-Iwahori Hecke ring*, Math. Annalen **331** (2005), 523–556, erratum Math. Annalen **333** (2005), 699–701.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Selmer Groups and Congruences

Ralph Greenberg*

Abstract

We first introduce Selmer groups for elliptic curves, and then Selmer groups for Galois representations. The main topic of the article concerns the behavior of Selmer groups for Galois representations with the same residual representation. We describe a variety of situations where this behavior can be studied fruitfully.

Mathematics Subject Classification (2010). Primary 11G05, 11R23; Secondary 11G40, 11R34.

Keywords. Selmer groups, Iwasawa invariants, Root numbers, Parity conjecture.

1. Selmer Groups

Suppose that E is an elliptic curve defined over a number field F. Let $E(F)$ denote the set of points on E defined over F. Under a certain simply-defined operation, $E(F)$ becomes an abelian group. The classical Mordell-Weil theorem asserts that $E(F)$ is finitely-generated. One crucial step in proving this theorem is to show that $E(F)/nE(F)$ is a finite group for some integer $n \geq 2$. In essence, one proves this finiteness for any n by defining a map from $E(F)/nE(F)$ to the Selmer group for E over F and showing that the kernel and the image of that map are finite.

We will regard F as a subfield of $\overline{\mathbf{Q}}$, a fixed algebraic closure of $\mathbf{Q}$. The torsion subgroup E_{tors} of $E(\overline{\mathbf{Q}})$ is isomorphic to $(\mathbf{Q}/\mathbf{Z})^2$ as a group. One has a natural action of $\mathrm{G}_F = \mathrm{Gal}(\overline{\mathbf{Q}}/F)$ on E_{tors}. The Selmer group is a certain subgroup of the Galois cohomology group $H^1(\mathrm{G}_F, E_{tors})$. Its definition involves Kummer theory for E and is based on the fact that the group of points on E defined over any algebraically closed field is a divisible group.

As is customary, we will write $H^1(F, E_{tors})$ instead of $H^1(\mathrm{G}_F, E_{tors})$. A similar abbreviation will be used for other Galois cohomology groups. Suppose

*The author's research described in this article has been supported by grants from the National Science Foundation.

Department of Mathematics, University of Washington, Seattle, WA 98195-4350, USA. E-mail: greenber@math.washington.edu.

that $P \in E(F)$ and that $n \geq 1$. Then there exists a point $Q \in E(\overline{\mathbf{Q}})$ such that $nQ = P$. In fact, there are n^2 such points Q, all differing by points in E_{tors} of order dividing n. If $g \in \mathrm{G}_F$ and $Q' = g(Q)$, then $nQ' = P$. Therefore, we have $g(Q) - Q \in E_{tors}$. The map $\varphi : G_F \to E_{tors}$ defined by $\varphi(g) = g(Q) - Q$ is a 1-cocycle and defines a class $[\varphi]$ in $H^1(F, E_{tors})$. In this way, we can define the *"Kummer map"*

$$\kappa : \; E(F) \otimes_{\mathbf{Z}} (\mathbf{Q}/\mathbf{Z}) \;\longrightarrow\; H^1(F, E_{tors}).$$

The image of $P \otimes \left(\frac{1}{n} + \mathbf{Z}\right)$ is defined to be the class $[\varphi]$. The map κ is an injective homomorphism.

If v is any prime of F, we can similarly define the v-adic Kummer map

$$\kappa_v : \; E(F_v) \otimes_{\mathbf{Z}} (\mathbf{Q}/\mathbf{Z}) \;\longrightarrow\; H^1(F_v, E_{tors}),$$

where F_v is the completion of F at v. One can identify G_{F_v} with a subgroup of G_F by choosing an embedding of $\overline{\mathbf{Q}}$ into an algebraic closure of F_v which extends the embedding of F into F_v, and thereby define a restriction map from $H^1(F, E_{tors})$ to $H^1(F_v, E_{tors})$. One has such a map for each prime v of F, even for the archimedean primes. One then defines the Selmer group $\mathrm{Sel}_E(F)$ to be the kernel of the map

$$\sigma : \; H^1(F, E_{tors}) \;\longrightarrow\; \bigoplus_v H^1(F_v, E_{tors}) \big/ \mathrm{im}(\kappa_v),$$

where v runs over all the primes of F. One shows that the image of σ is actually contained in the direct sum and that this definition of $\mathrm{Sel}_E(F)$ does not depend on the choice of embeddings. The image of the Kummer map κ is clearly a subgroup of $\mathrm{Sel}_E(F)$. The corresponding quotient group $\mathrm{Sel}_E(F)/\mathrm{im}(\kappa)$ is the Tate-Shafarevich group for E over F.

The elliptic curve E is determined up to isomorphism over F by the action of G_F on E_{tors}. This result was originally conjectured by Tate and proved by Faltings [10]. If p is a prime and $n \geq 1$, then the p^n-torsion on E will be denoted by $E[p^n]$. The p-primary subgroup of E_{tors} is the union of the groups $E[p^n]$ and will be denoted by $E[p^\infty]$. The inverse limit of the $E[p^n]$'s is the p-adic Tate module $T_p(E)$. It is a free $\mathbf{Z}_p$-module of rank 2, where $\mathbf{Z}_p$ denotes the ring of p-adic integers. All of these objects have a continuous action of G_F. We let $V_p(E) = T_p(E) \otimes_{\mathbf{Z}_p} \mathbf{Q}_p$, a 2-dimensional representation space for G_F over $\mathbf{Q}_p$, the field of p-adic numbers. Faltings proves the following version of Tate's conjecture: The elliptic curve E is determined up to isogeny over F by the isomorphism class of the representation space $V_p(E)$ for G_F. The Tate module $T_p(E)$ determines E up to an isogeny of degree prime to p.

The above theorem of Faltings suggests that arithmetic properties of E which depend only on the isomorphism class of E over F should somehow be determined by the Galois module E_{tors}. In particular, the structure of $E(F)$

should be so determined. It is clear how to determine the torsion subgroup of $E(F)$ in terms of E_{tors}. It is just $H^0(F, E_{tors})$. Now it is conjectured that the Tate-Shafarevich group for an elliptic curve over a number field is always finite. If this is so, then the image of the Kummer map should be precisely the maximal divisible subgroup $\mathrm{Sel}_E(F)_{div}$ of $\mathrm{Sel}_E(F)$. If r is the rank of $E(F)$, then that image is isomorphic to $(\mathbf{Q}/\mathbf{Z})^r$. Thus, at least conjecturally, one can determine r from the structure of $\mathrm{Sel}_E(F)$. And, as we will now explain, one can describe $\mathrm{Sel}_E(F)$ entirely in terms of the Galois module E_{tors}. This is not immediately apparent from the definition given earlier.

Let p be any prime. The p-primary subgroup $\mathrm{Sel}_E(F)_p$ of $\mathrm{Sel}_E(F)$ is a subgroup of $H^1(F, E[p^\infty])$. It can be defined as the kernel of the map

$$\sigma_p : \; H^1(F, E[p^\infty]) \;\longrightarrow\; \bigoplus_v H^1(F_v, E[p^\infty])/\mathrm{im}(\kappa_{v,p}) \quad ,$$

where $\kappa_{v,p}$ is the restriction of κ_v to the p-primary subgroup of $E(F_v)\otimes_{\mathbf{Z}}(\mathbf{Q}/\mathbf{Z})$. Thus, if we can describe the image of $\kappa_{v,p}$ for all primes v of F just in terms of the Galois module $E[p^\infty]$, then we will have such a description of $\mathrm{Sel}_E(F)_p$.

First of all, suppose that v is a nonarchimedean prime and that the residue field for v has characteristic ℓ, where $\ell \neq p$. It is known that $E(F_v)$ is an ℓ-adic Lie group. More precisely, $E(F_v)$ contains a subgroup of finite index which is isomorphic to $\mathbf{Z}_\ell^{[F_v:\mathbf{Q}_\ell]}$. Since that group is divisible by p, one sees easily that $E(F_v) \otimes_{\mathbf{Z}} (\mathbf{Q}_p/\mathbf{Z}_p)$, the p-primary subgroup of $E(F_v) \otimes_{\mathbf{Z}} (\mathbf{Q}/\mathbf{Z})$, actually vanishes. Hence $\mathrm{im}(\kappa_{v,p}) = 0$ if $v \nmid p$. A similar argument shows that the same statement is true if v is archimedean.

Now assume that the residue field for v has characteristic p. We also assume that E has good ordinary reduction at v. Good reduction means that one can find an equation for E over the ring of integers of F such that its reduction modulo v defines an elliptic curve $\overline{E}_v$ over the residue field $\mathbf{F}_v$. The reduction is ordinary if the integer $a_v = 1+|\mathbf{F}_v|-|\overline{E}_v(\mathbf{F}_v)|$ is not divisible by p. Equivalently, ordinary reduction means that $\overline{E}_v[p^\infty]$ is isomorphic to $\mathbf{Q}_p/\mathbf{Z}_p$ as a group. Reduction modulo v then defines a surjective homomorphism from $E[p^\infty]$ to $\overline{E}_v[p^\infty]$. Its kernel turns out to be the group of p-power torsion points on a formal group. We denote that kernel by C_v. It is invariant under the action of G_{F_v} and is isomorphic to $\mathbf{Q}_p/\mathbf{Z}_p$ as a group. We have $E[p^\infty]/C_v \cong \overline{E}_v[p^\infty]$. Remarkably, one has the following description of the image of $\kappa_{v,p}$:

$$\mathrm{im}(\kappa_{v,p}) \;=\; \mathrm{im}\big(H^1(F_v, C_v)_{div} \;\longrightarrow\; H^1(F_v, E[p^\infty])\big).$$

One can characterize C_v as follows: It is a G_{F_v}-invariant subgroup of $E[p^\infty]$ and $E[p^\infty]/C_v$ is the maximal quotient of $E[p^\infty]$ which is unramified for the action of G_{F_v}. Thus, the above description of $\mathrm{im}(\kappa_{v,p})$ just involves the Galois module $E[p^\infty]$, as we wanted.

The above description of $\mathrm{im}(\kappa_{v,p})$ was given in [4]. The argument is not very difficult. If E does not have good ordinary reduction at v, there is still a

description of $\mathrm{im}(\kappa_{v,p})$ in terms of $E[p^\infty]$. This was given by Bloch and Kato in [2]. It involves Fontaine's ring B_{crys}. One defines the subspace $H^1_f\big(F_v, V_p(E)\big)$ of $H^1\big(F_v, V_p(E)\big)$ to be the kernel of the the natural map from $H^1\big(F_v, V_p(E)\big)$ to $H^1\big(F_v, V_p(E) \otimes_{\mathbf{Q}_p} B_{crys}\big)$. One has $V_p(E)/T_p(E) \cong E[p^\infty]$. Then $\mathrm{im}(\kappa_{v,p})$ turns out to be the image of $H^1_f\big(F_v, V_p(E)\big)$ under the natural map from $H^1\big(F_v, V_p(E)\big)$ to $H^1(F_v, E[p^\infty])$.

The fact that $\mathrm{Sel}_E(F)_p$ can be defined solely in terms of the Galois module $E[p^\infty]$ was a valuable insight in the 1980's. It suggested a way to give a reasonable definition of Selmer groups in a far more general context. This idea was pursued in [11] for the purpose of generalizing conjectures of Iwasawa and of Mazur concerning the algebraic interpretation of zeros of p-adic L-functions. It was also pursued by Bloch and Kato in [2] for the purpose of generalizing the Birch and Swinnerton-Dyer conjecture.

Since $\mathrm{Sel}_E(F)_p$ is determined by the Galois module $E[p^\infty]$, one can ask whether $\mathrm{Sel}_E(F)[p]$ is determined by the Galois module $E[p]$. This turns out not to be so. Suppose that E_1 and E_2 are elliptic curves defined over F and that $E_1[p] \cong E_2[p]$ as G_F-modules. It is quite possible for $\mathrm{Sel}_{E_1}(F)[p]$ and $\mathrm{Sel}_{E_2}(F)[p]$ to have different $\mathbf{F}_p$-dimensions. In the next section of this article, we will consider this question in the setting of Iwasawa theory. Thus, we will consider the Selmer group for an elliptic curve E over a certain infinite extension F_∞ of F, the so-called *"cyclotomic $\mathbf{Z}_p$-extension"* of F.

Let μ_{p^∞} denote the group of p-power roots of unity in $\overline{\mathbf{Q}}$. Then F_∞ is the unique subfield of $F(\mu_{p^\infty})$ such that $\mathrm{Gal}(F_\infty/F) \cong \mathbf{Z}_p$. We denote that Galois group by Γ. For each $n \geq 0$, Γ has a unique subgroup Γ_n of index p^n. Thus, $F_n = F_\infty^{\Gamma_n}$ is a cyclic extension of F of degree p^n. One can define the Selmer group for E over F_∞ to be the direct limit of the Selmer groups $\mathrm{Sel}_E(F_n)$ as $n \to \infty$. We will concentrate on its p-primary subgroup $\mathrm{Sel}_E(F_\infty)_p$. Now Γ acts naturally on $\mathrm{Sel}_E(F_\infty)_p$. Regarding $\mathrm{Sel}_E(F_\infty)_p$ as a discrete $\mathbf{Z}_p$-module, the action of Γ is continuous and $\mathbf{Z}_p$-linear. We can then regard $\mathrm{Sel}_E(F_\infty)_p$ as a discrete Λ-module, where $\Lambda = \mathbf{Z}_p[[\Gamma]]$ is the completed $\mathbf{Z}_p$-group algebra for the pro-p group Γ. That is, Λ is the inverse limit of the $\mathbf{Z}_p$-group algebras $\mathbf{Z}_p[\Gamma_n]$ defined by the obvious surjective $\mathbf{Z}_p$-algebra homomorphisms $\mathbf{Z}_p[\Gamma_m] \to \mathbf{Z}_p[\Gamma_n]$ for $m \geq n \geq 0$. One often refers to Λ as the *"Iwasawa algebra"* for Γ (over $\mathbf{Z}_p$). A very useful fact in Iwasawa theory is that Λ is isomorphic (non-canonically) to the formal power series ring $\mathbf{Z}_p[[T]]$ in one variable. Thus, Λ is a complete Noetherian local ring of Krull dimension 2.

Assuming that E has good ordinary reduction at the primes of F lying over p, one has a description of $\mathrm{Sel}_E(F_\infty)_p$ just as above. If v is a prime of F not dividing p, and η is a prime of F_∞ lying over v, then the image of the Kummer map over $F_{\infty,\eta}$ is again trivial. If $v|p$, then the direct limits of the local Galois cohomology groups $H^1(F_{n,\eta}, C_\eta)_{div}$ and $H^1(F_{n,\eta}, C_\eta)$ as $n \to \infty$ turn out to be the same, both equal to $H^1(F_{\infty,\eta}, C_\eta)$. Thus, the image of the Kummer map

over $F_{\infty,\eta}$ coincides with the image of the map

$$\varepsilon_{\infty,\eta} : H^1(F_{\infty,\eta}, C_\eta) \longrightarrow H^1(F_{\infty,\eta}, E[p^\infty]) \quad .$$

A very broad generalization of this fact is proven in [4].

The following conjecture of Mazur will play a fundamental role in most of the results we will describe. It was first stated and discussed in [21]. We let $X_E(F_\infty)$ denote the Pontryagin dual of $\mathrm{Sel}_E(F_\infty)_p$. We can regard $X_E(F_\infty)$ as a compact Λ-module. It turns out to always be finitely-generated as a Λ-module. As in [21], we will say that $\mathrm{Sel}_E(F_\infty)_p$ is a cotorsion Λ-module if $X_E(F_\infty)$ is a torsion Λ-module.

Conjecture. *Suppose that E has good ordinary reduction at the primes of F lying over p. Then $\mathrm{Sel}_E(F_\infty)_p$ is a cotorsion Λ-module.*

The above conjecture is proved in [21] under the assumption that $\mathrm{Sel}_E(F)_p$ is finite. We will later cite a much more recent theorem (due to Kato and Rohrlich) which asserts that $\mathrm{Sel}_E(F_\infty)_p$ is indeed Λ-cotorsion if E is an elliptic curve defined over $\mathbf{Q}$ with good ordinary reduction at p and F is any abelian extension of $\mathbf{Q}$. Such a theorem had already been proven by Rubin [31] in the case where E has complex multiplication.

The above conjecture should be valid under somewhat weaker assumptions about the reduction of E at the primes above p. It should suffice to just assume that E does not have potentially supersingular reduction at any of those primes. That assumption is necessary. If E has potentially supersingular reduction at a prime above p, then one can show that $\mathrm{Sel}_E(F_\infty)_p$ is not Λ-cotorsion. We refer the reader to [34] for a discussion of this issue and a precise conjecture about the rank of $X_E(F_\infty)$ as a Λ-module.

2. Behavior Under Congruences

The results that we describe here are mostly from [14]. We will now take $F = \mathbf{Q}$, partly just to simplify the discussion and partly because the deep theorem of Kato and Rohrlich mentioned above is then available. We will also assume that p is an odd prime. We concentrate entirely on the p-primary subgroup of Selmer groups. Let $\mathbf{Q}_\infty$ denote the cyclotomic $\mathbf{Z}_p$-extension of $\mathbf{Q}$. Suppose that E is defined over $\mathbf{Q}$ and has good ordinary reduction at p.

Let π denote the unique prime of $\mathbf{Q}_\infty$ lying over p. We will write $\overline{E}$ for $\overline{E}_p$. It will be useful to note that the image of $\varepsilon_{\infty,\pi}$ coincides with the kernel of the map $H^1(\mathbf{Q}_{\infty,\pi}, E[p^\infty]) \to H^1(\mathbf{Q}_{\infty,\pi}, \overline{E}[p^\infty])$. That map turns out to be surjective and so $H^1(\mathbf{Q}_{\infty,\pi}, E[p^\infty])/\mathrm{im}(\varepsilon_{\infty,\pi})$ is isomorphic to $H^1(\mathbf{Q}_{\infty,\pi}, \overline{E}[p^\infty])$, which we will denote by $\mathcal{H}_p(\mathbf{Q}_\infty, E[p^\infty])$. We will denote $H^1(\mathbf{Q}_{\infty,\pi}, \overline{E}[p])$ by $\mathcal{H}_p(\mathbf{Q}_\infty, E[p])$.

If ℓ is a non-archimedean prime, and $\ell \neq p$, we define

$$\mathcal{H}_\ell(\mathbf{Q}_\infty, E[p^\infty]) \;=\; \bigoplus_{\eta|\ell} H^1(\mathbf{Q}_{\infty,\eta}, E[p^\infty]),$$

a finite direct sum because ℓ is finitely decomposed in $\mathbf{Q}_\infty/\mathbf{Q}$. We similarly define $\mathcal{H}_\ell(\mathbf{Q}_\infty, E[p])$, just replacing the Galois module $E[p^\infty]$ by $E[p]$. We will ignore the local Galois cohomology groups for archimedean primes. They are trivial since we are assuming that p is odd.

Although the Galois module $E[p]$ still does not determine $\mathrm{Sel}_E(F_\infty)[p]$, a somewhat weaker statement turns out to be true. To formulate it, we introduce *"non-primitive"* Selmer groups. Suppose that Σ_0 is a finite set of non-archimedean primes of $\mathbf{Q}$. We assume that $p \notin \Sigma_0$. The corresponding non-primitive Selmer group will be denoted by $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p$ and differs from the actual Selmer group in that we omit the local conditions for the primes of $\mathbf{Q}_\infty$ lying above primes in Σ_0. To be precise, $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p$ is defined to be the kernel of the following map:

$$H^1(\mathbf{Q}_\infty, E[p^\infty]) \;\longrightarrow\; \bigoplus_{\ell \notin \Sigma_0} \mathcal{H}_\ell(\mathbf{Q}_\infty, E[p^\infty]). \tag{1}$$

If we take Σ_0 to be empty, then $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p = \mathrm{Sel}_E(\mathbf{Q}_\infty)_p$.

Suppose that $E[p]$ is irreducible and that Σ_0 contains the primes where E has bad reduction. The map $H^1(\mathbf{Q}_\infty, E[p]) \to H^1(\mathbf{Q}_\infty, E[p^\infty])[p]$ is an isomorphism. The role of the assumption about Σ_0 is that it implies that the preimage of $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)[p]$ under that isomorphism is precisely the kernel of the map

$$H^1(\mathbf{Q}_\infty, E[p]) \longrightarrow \bigoplus_{\ell \notin \Sigma_0} \mathcal{H}_\ell(\mathbf{Q}_\infty, E[p]). \tag{2}$$

Note that the groups and maps here indeed depend only on the Galois module $E[p]$. This is clear for $\ell \neq p$. For $\ell = p$, it follows because one can characterize $\overline{E}[p]$ as the maximal unramified quotient of $E[p]$ for the action of $G_{\mathbf{Q}_p}$. This is so because p is assumed to be odd and therefore the action of the inertia subgroup of $G_{\mathbf{Q}_p}$ on the kernel of the reduction map $E[p] \to \overline{E}[p]$ is nontrivial. Thus, under the above assumption about Σ_0, we have a description of $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)[p]$ in terms of the Galois module $E[p]$.

The local Galois cohomology groups $\mathcal{H}_\ell(\mathbf{Q}_\infty, E[p^\infty])$ can be studied by using standard results, essentially just local class field theory. One finds that

$$\mathcal{H}_\ell(\mathbf{Q}_\infty, E[p^\infty]) \cong (\mathbf{Q}_p/\mathbf{Z}_p)^{\delta(E,\ell)}$$

for any prime $\ell \neq p$, where $\delta(E, \ell)$ is an easily determined non-negative integer.

A theorem of Kato [18], combined with a theorem of Rohrlich [29], implies that $\mathrm{Sel}_E(\mathbf{Q}_\infty)_{p,div} \cong (\mathbf{Q}_p/\mathbf{Z}_p)^{\lambda(E)}$ for some integer $\lambda(E) \geq 0$. This means that the Pontryagin dual of $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$ is a torsion module over the Iwasawa

algebra $\Lambda = \mathbf{Z}_p[[\Gamma]]$. This was conjectured to be so in [21], as we mentioned in section 1. The integer $\lambda(E)$ is the $\mathbf{Z}_p$-corank of $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$. Under the assumption that $E[p]$ is irreducible as a Galois module, it is reasonable to make the conjecture that the Pontryagin dual of $\mathrm{Sel}_E(\mathbf{Q}_\infty)[p]$ is a torsion module over $\Lambda/p\Lambda$. Equivalently, this would mean that $\mathrm{Sel}_E(\mathbf{Q}_\infty)[p]$ is finite, and hence that the so-called μ-invariant for $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$ vanishes. If so, then one can prove that $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$ is a divisible group and so one would have an isomorphism

$$\mathrm{Sel}_E(\mathbf{Q}_\infty)_p \;\cong\; (\mathbf{Q}_p/\mathbf{Z}_p)^{\lambda(E)}.$$

The fact that $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$ is a cotorsion Λ-module allows one to prove that the map

$$\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p \;\longrightarrow\; \bigoplus_{\ell\in\Sigma_0} \mathcal{H}_\ell(\mathbf{Q}_\infty, E[p^\infty])$$

is surjective. The kernel of that map is $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$, and so we have an isomorphism

$$\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p/\mathrm{Sel}_E(\mathbf{Q}_\infty)_p \;\cong\; \bigoplus_{\ell\in\Sigma_0} \mathcal{H}_\ell(\mathbf{Q}_\infty, E[p^\infty]) \;\cong\; (\mathbf{Q}_p/\mathbf{Z}_p)^{\delta(E,\Sigma_0)} \quad ,$$

where $\delta(E,\Sigma_0) = \sum_{\ell\in\Sigma_0}\delta(E,\ell)$. Let $\lambda(E,\Sigma_0)$ denote the $\mathbf{Z}_p$-corank of $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p$. We then obtain the formula $\lambda(E,\Sigma_0) = \lambda(E) + \delta(E,\Sigma_0)$.

If $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$ is divisible, then it is a direct summand in $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p$, and so we will have

$$\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty) \;\cong\; \mathrm{Sel}_E(\mathbf{Q}_\infty) \;\oplus\; \left(\bigoplus_{\ell\in\Sigma_0} \mathcal{H}_\ell(\mathbf{Q}_\infty, E[p^\infty])\right).$$

Thus, $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p$ will also be divisible, and so its $\mathbf{Z}_p$-corank $\lambda(E,\Sigma_0)$ will be equal to the $\mathbf{F}_p$-dimension of $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)[p]$. A similar statement is true for all of the summands in the above direct sum.

Suppose that E_1 and E_2 are elliptic curves defined over $\mathbf{Q}$, that both have good ordinary reduction at p, and that $E_1[p] \cong E_2[p]$ as $\mathrm{G}_{\mathbf{Q}}$-modules. We think of such an isomorphism as a congruence modulo p between the p-adic Tate modules for E_1 and E_2. We will also assume that $G_{\mathbf{Q}}$ acts irreducibly on $E_1[p]$, and hence on $E_2[p]$. Suppose that Σ_0 is chosen to include all the primes where E_1 or E_2 have bad reduction. Under these assumptions, the above discussion shows that

$$\mathrm{Sel}_{E_1}^{\Sigma_0}(\mathbf{Q}_\infty)[p] \;\cong\; \mathrm{Sel}_{E_2}^{\Sigma_0}(\mathbf{Q}_\infty)[p].$$

Consequently, if $\mathrm{Sel}_{E_1}(\mathbf{Q}_\infty)[p]$ is finite, then so is $\mathrm{Sel}_{E_2}(\mathbf{Q}_\infty)[p]$. Their $\mathbf{F}_p$-dimensions will be equal and one then obtains the formula

$$\lambda(E_1) + \delta(E_1,\Sigma_0) \;=\; \lambda(E_2) + \delta(E_2,\Sigma_0) \quad .$$

Since the quantities $\delta(E_1, \Sigma_0)$ and $\delta(E_2, \Sigma_0)$ can be evaluated, one can then determine $\lambda(E_2)$ if one knows $\lambda(E_1)$.

As an example, consider the two elliptic curves

$$E_1 : y^2 = x^3 + x - 10, \qquad E_2 : y^2 = x^3 - 584x + 5444$$

which have conductors 52 and $364 = 7 \cdot 52$, respectively. We take $p = 5$ and $\Sigma_0 = \{2, 7, 13\}$. One has a congruence modulo 5 between the q-expansions of the modular forms corresponding to E_1 and E_2, ignoring the terms for powers q^n where $7|n$. It follows that $E_1[5] \cong E_2[5]$ as $G_{\mathbf{Q}}$-modules. It turns out that $\mathrm{Sel}_{E_1}(\mathbf{Q}_\infty)_p = 0$. Hence, one has $\lambda(E_1) = 0$. One finds that $\delta(E_1, \Sigma_0) = 5$ and $\delta(E_2, \Sigma_0) = 0$. Consequently, we have $\lambda(E_2) = 5$.

Such isomorphisms $E_1[p] \cong E_2[p]$ are not hard to find for $p = 3$ and $p = 5$. In fact, it is shown in [33] that for $p \leq 5$, and for a fixed elliptic curve E_1 defined over $\mathbf{Q}$, one can explicitly describe equations defining an infinite family of non-isomorphic elliptic curves E_2 over $\mathbf{Q}$ with $E_2[p] \cong E_1[p]$. Such isomorphisms are not common for $p \geq 7$. However, if one considers cusp forms of weight 2, then *"raising the level"* theorems show that such isomorphisms occur for every odd prime p. They can be formulated in terms of the Jacobian variety attached to Hecke eigenforms of weight 2. An isomorphism amounts to a congruence between the q-expansions of two such eigenforms. The results described above extend without any real difficulty to this case.

A somewhat different approach is taken in [9]. That paper considers Selmer groups over $\mathbf{Q}_\infty$ associated to Hecke eigenforms of arbitrary weight which are ordinary in a certain sense. If one fixes the residual representation and bounds the prime-to-p part of the conductor, then such eigenforms occur in Hida families which are parametrized by the set of prime ideals of height 1 in a certain ring R. Such families were constructed by Hida in [17]. If $\mathfrak{a}$ is a minimal prime ideal of R, then the height 1 prime ideals of $R/\mathfrak{a}$ parametrize one *"branch"* in such a family. For each eigenform h, one can associate a Galois representation V_h of dimension 2 over a field $\mathcal{F}$, a finite extension of $\mathbf{Q}_p$. Let $\mathcal{O}$ be the ring of integers in $\mathcal{F}$ and let π be a uniformizing parameter. Let $\mathfrak{f} = \mathcal{O}/(\pi)$. One can choose a Galois-invariant $\mathcal{O}$-lattice T_h in V_h and then define a discrete Galois module $\mathcal{A}_h = V_h/T_h$. The representation is ordinary in the sense that V_h has a 1-dimensional quotient which is unramified for the action of $\mathrm{G}_{\mathbf{Q}_p}$. Hence $\mathcal{A}_h$ has an unramified quotient which has $\mathcal{O}$-corank 1.

One can define a Selmer group $\mathrm{Sel}_{\mathcal{A}_h}(\mathbf{Q}_\infty)_p$ for the Galois module $\mathcal{A}_h$ in essentially the same way as for $E[p^\infty] = V_p(E)/T_p(E)$. It is a subgroup of $H^1(\mathbf{Q}_\infty, \mathcal{A}_h)$ and is defined as the kernel of a map just like (1) (taking Σ_0 to be empty). It suffices to define $\mathcal{H}_\ell(\mathbf{Q}_\infty, \mathcal{A}_h)$ for all primes ℓ. The residual representation is given by the Galois action on $\mathcal{A}_h[\pi]$. If we assume that this is irreducible, as before, then $H^0(\mathbf{Q}_\infty, \mathcal{A}_h) = 0$ and one has an isomorphism

$$H^1(\mathbf{Q}_\infty, \mathcal{A}_h[\pi]) \longrightarrow H^1(\mathbf{Q}_\infty, \mathcal{A}_h)[\pi]. \tag{3}$$

The preimage of $\mathrm{Sel}_{\mathcal{A}_h}(\mathbf{Q}_\infty)[\pi]$ under this isomorphism defines an $\mathfrak{f}$-subspace $\mathcal{S}_h$ of $H^1(\mathbf{Q}_\infty, \mathcal{A}_h[\pi])$. It can be characterized by local conditions. That is, one can

define $\mathcal{S}_h$ as the kernel of a map like (2) (again taking Σ_0 to be empty). However, those conditions will not generally be determined by the Galois module $\mathcal{A}_h[\pi]$.

For a prime ℓ not dividing p in some finite set Σ_0, which would usually be nonempty, and for a prime η of $\mathbf{Q}_\infty$ lying over ℓ, the map

$$H^1(\mathbf{Q}_{\infty,\eta}, \mathcal{A}_h[\pi]) \longrightarrow H^1(\mathbf{Q}_{\infty,\eta}, \mathcal{A}_h)[\pi]$$

may have a nontrivial kernel. Let $\delta_\eta(h)$ denote the $\mathfrak{f}$-dimension of the kernel. An element of $\mathcal{S}_h$ will have a trivial image in $H^1(\mathbf{Q}_{\infty,\eta}, \mathcal{A}_h)[\pi]$ (and hence satisfy the local condition for the prime η which occurs in the definition of $\mathrm{Sel}_{\mathcal{A}_h}(\mathbf{Q}_\infty)_p$), but may have a nontrivial image in $H^1(\mathbf{Q}_{\infty,\eta}, \mathcal{A}_h[\pi])$. Thus, for $\ell \in \Sigma_0$, one should define $\mathcal{H}_\ell(\mathbf{Q}_\infty, \mathcal{A}_h[\pi])$ to be a certain quotient of the direct product of the $H^1(\mathbf{Q}_{\infty,\eta}, \mathcal{A}_h[\pi])$'s for $\eta|\ell$ so that the inclusion $\mathcal{A}[\pi] \to \mathcal{A}_h$ induces an isomorphism from $\mathcal{H}_\ell(\mathbf{Q}_\infty, \mathcal{A}_h[\pi])$ to $\mathcal{H}_\ell(\mathbf{Q}_\infty, \mathcal{A}_h)[\pi]$. If one assumes that $\mathrm{Sel}_{\mathcal{A}_h}(\mathbf{Q}_\infty)[\pi]$ is finite, then it turns out that $\mathrm{Sel}_{\mathcal{A}_h}(\mathbf{Q}_\infty)$ is a divisible $\mathcal{O}$-module. Let $\lambda(h)$ denote its $\mathcal{O}$-corank. Thus, we have $\lambda(h) = \dim_{\mathfrak{f}}(\mathcal{S}_h)$. As shown in [9], the variation in $\dim_{\mathfrak{f}}(\mathcal{S}_h)$ is controlled completely by the $\delta_\eta(h)$'s. They turn out to be constant in each branch of the Hida family, and so $\lambda(h)$ will also be constant in each branch. One also obtains a rather simple formula for the change in the λ-invariant from one branch to another.

What we have described above is the algebraic side of Iwasawa theory. A substantial part of both [14] and [9] is devoted to the analytic side of Iwasawa theory, the existence and properties of p-adic L-functions. One can also associate a λ-invariant to p-adic L-functions. A natural domain of definition for those functions is $\mathrm{Hom}_{cont}(\Gamma, \overline{\mathbf{Q}}_p^\times)$. The λ-invariant is the number of zeros, counting multiplicity. In [14], a non-primitive p-adic L-function plays an important role. In both [14] and [9], the results on the algebraic and on the analytic sides are quite parallel, although the nature of the arguments is quite different. The *"main conjecture"* of Iwasawa theory for elliptic curves (due to Mazur [21]), or for modular forms (as in [11]), relates the algebraic and analytic sides in a precise way. It gives an algebraic interpretation of the zeros of the p-adic L-functions.

If E has complex multiplication, then the main conjecture has been settled by Rubin [32] in a somewhat more general situation than we are considering here. The results in [14] and in [9] together with a theorem of Kato [18] imply that if the main conjecture is valid for one elliptic curve E_1, or for one modular form h_1 in a Hida family, then, under the assumption that a certain μ-invariant vanishes, the main conjecture will also be valid for any other elliptic curve E_2 such that $E_2[p] \cong E_1[p]$, or for any other modular form h_2 in the Hida family. Thus, roughly speaking, and under suitable assumptions, the validity of the main conjecture is preserved by congruences. We also want to mention much more recent work of Skinner and Urban which may go a long way to settling this conjecture completely.

There are also results for elliptic curves with good supersingular reduction at p, and more generally for modular forms of weight 2 which are supersingular

in a certain sense. This topic will be discussed in detail in [13]. The results in that paper are intended to be the analogues of those in [14], despite the fact that the corresponding Selmer groups will not be Λ-cotorsion and the corresponding p-adic L-functions will have infinitely many zeros. The higher weight case is not yet understood. One finds a discussion of that case on the analytic side in [27].

3. Artin Twists

The discussion in section 2 mostly concerns the invariant $\lambda(E)$ associated to $\mathrm{Sel}_E(\mathbf{Q}_\infty)_p$, and the non-primitive analogues $\lambda(E,\Sigma_0)$ and $\mathrm{Sel}_E^{\Sigma_0}(\mathbf{Q}_\infty)_p$ corresponding to a suitable set Σ_0. We will now include another variable, an Artin representation σ. We let the base field F be an arbitrary algebraic number field and denote the cyclotomic $\mathbf{Z}_p$-extension of F by F_∞. Suppose that K is a finite Galois extension of F and that $K \cap F_\infty = F$. The Artin representations to be considered will factor through $\Delta = \mathrm{Gal}(K/F)$. However, if K is allowed to vary over the finite extensions of F contained in some infinite Galois extension $\mathcal{K}$ of F satisfying $\mathcal{K} \cap F_\infty = F$, then σ can vary over all Artin representations over F which factor through $\mathrm{Gal}(\mathcal{K}/F)$. One interesting case is where $\mathrm{Gal}(\mathcal{K}/F)$ is a p-adic Lie group.

If v is a non-archimedean prime of F, let $e_v(K/F)$ denote the ramification index for v in K/F. Let

$$\Phi_{K/F} \;=\; \{v \mid v \nmid p,\ v \nmid \infty, \text{ and } e_v(K/F) \text{ is divisible by } p\ \}.$$

This finite set of primes of F will play an important role in this section. We always will assume that E has good ordinary reduction at the primes of F lying over p.

Assume that X is a free $\mathbf{Z}_p$-module of finite rank $\lambda(X)$ and that there is a $\mathbf{Z}_p$-linear action of Δ on X. Thus, X is a $\mathbf{Z}_p[\Delta]$-module. Suppose that σ is defined over $\mathcal{F}$, a finite extension of $\mathbf{Q}_p$, and that σ is absolutely irreducible. We define $\lambda(X,\sigma)$ to be the multiplicity of σ in $V_{\mathcal{F}} = X \otimes_{\mathbf{Z}_p} \mathcal{F}$, an $\mathcal{F}$-representation space for Δ of dimension $\lambda(X)$. The definition of $\lambda(X,\sigma)$ makes sense if we just assume that X/X_{tors} is a free $\mathbf{Z}_p$-module of finite rank. We let $\mathrm{Irr}_{\mathcal{F}}(\Delta)$ denote the set of irreducible representations of Δ over $\mathcal{F}$, always assuming that $\mathcal{F}$ is large enough so that irreducible $\mathcal{F}$-representations are absolutely irreducible. We extend the definition of $\lambda(X,\cdot)$ to arbitrary finite-dimension representations ρ of Δ by making the map $\lambda(X,\cdot)$ a homomorphism from the Grothendieck group $\mathcal{R}_{\mathcal{F}}(\Delta)$ to $\mathbf{Z}$.

Since $K \cap F_\infty = F$, we can identify Δ with $\mathrm{Gal}(K_\infty/F_\infty)$, where K_∞ is KF_∞, the cyclotomic $\mathbf{Z}_p$-extension of K. Hence there is a natural action of Δ on $\mathrm{Sel}_E(K_\infty)_p$. Assume that the Pontryagin dual of $\mathrm{Sel}_E(K_\infty)_p$ is a torsion Λ-module. If we take X to be that module, then we will denote $\lambda(X,\sigma)$ by $\lambda(E,\sigma)$. Let Σ_0 be a finite set of primes of F not lying above p or ∞. Then

there is also a natural action of Δ on $\mathrm{Sel}_E^{\Sigma_0}(K_\infty)_p$. If we take X to be the Pontryagin dual of $\mathrm{Sel}_E^{\Sigma_0}(K_\infty)_p$ (which will also be a torsion Λ-module), then we will denote $\lambda(X,\sigma)$ by $\lambda(E,\Sigma_0,\sigma)$. Although we will not discuss it here, one can also describe the difference $\lambda(E,\Sigma_0,\sigma)-\lambda(E,\sigma)$ in purely local terms. And so one can reduce the study of the $\lambda(E,\sigma)$'s to studying the $\lambda(E,\Sigma_0,\sigma)$'s for a suitable choice of Σ_0. Proposition 3.1 below concerns these non-primitive λ-invariants and is one of the main results of [12].

If v is a prime of F lying above p, we let $\overline{E}_v$ denote the reduction of E at v, an elliptic curve over the residue field of v. We let k_v denote the residue field for any prime of K lying above v.

Proposition 3.1. *Suppose that E has good ordinary reduction at the primes of F lying above p, that $E(K)[p]=0$, that $\overline{E}_v(k_v)[p]=0$ for all primes v of F lying over p, and that $\mathrm{Sel}_E(K_\infty)[p]$ is finite. Let Σ_0 be a finite set of primes containing $\Phi_{K/F}$, but not containing primes lying over p or ∞. Then the Pontryagin dual of $\mathrm{Sel}_E^{\Sigma_0}(K_\infty)_p$ is a projective $\mathbf{Z}_p[\Delta]$-module.*

The assumption that $\mathrm{Sel}_E(K_\infty)[p]$ is finite implies that $\mathrm{Sel}_E(K_\infty)_p$ is Λ-cotorsion.

A corollary of the above result is that the invariants $\lambda(E,\Sigma_0,\rho)$ behave in the following way. Here we let ρ be an arbitrary representation of Δ over $\mathcal{F}$. Let $\mathcal{O}$ be the integers in $\mathcal{F}$. We can choose a Δ-invariant $\mathcal{O}$-lattice in the underlying representation space for ρ. Reducing modulo the maximal ideal $\mathfrak{m}$ of $\mathcal{O}$, we obtain a representation $\widetilde{\rho}$. Its semisimplification $\widetilde{\rho}^{ss}$ is well-defined. It is a representation over the residue field $\mathfrak{f}=\mathcal{O}/\mathfrak{m}$. Then, under the assumptions in the above proposition, we have the following result:

Corollary 3.2. *Suppose that the assumptions in proposition* 3.1 *are satisfied. Assume that ρ_1 and ρ_2 are representations of Δ such that $\widetilde{\rho}_1^{ss}\cong\widetilde{\rho}_2^{ss}$. Then $\lambda(E,\Sigma_0,\rho_1)=\lambda(E,\Sigma_0,\rho_2)$. That is, we have a linear relation*

$$\sum_\sigma m_1(\sigma)\lambda(E,\Sigma_0,\sigma) \;=\; \sum_\sigma m_2(\sigma)\lambda(E,\Sigma_0,\sigma)$$

where σ varies over the irreducible representations of Δ over $\mathcal{F}$ and $m_i(\sigma)$ denotes the multiplicity of σ in ρ_i for $i=1,2$.

If $\rho_1\ncong\rho_2$, but $\widetilde{\rho}_1^{ss}\cong\widetilde{\rho}_2^{ss}$, then the corresponding linear relation is nontrivial. Such nontrivial relations occur whenever $|\Delta|$ is divisible by p. We also remark that the conclusion in the corollary means that the map $\lambda(E,\Sigma_0,\cdot)$ from $\mathcal{R}_{\mathcal{F}}(\Delta)$ to $\mathbf{Z}$ factors through the reduction map $\mathcal{R}_{\mathcal{F}}(\Delta)\to\mathcal{R}_{\mathfrak{f}}(\Delta)$.

The assumptions in the corollary can be weakened. As shown in [12], one can omit the assumptions about $E(K)[p]$ and $\overline{E}_v(k_v)[p]$ for $v|p$. It suffices to assume that $\mathrm{Sel}_E(K_\infty)[p]$ is finite and that Σ_0 is chosen as in proposition 3.1. The Pontryagin dual of $\mathrm{Sel}_E^{\Sigma_0}(K_\infty)_p$ may fail to be a projective $\mathbf{Z}_p[\Delta]$-module, but

it still turns out to have a weaker property which we call *"quasi-projectivity"*. The linear relation in corollary 3.2 still follows. We call such a linear relation a *"congruence relation"* because it arises from an isomorphism $\widetilde{\rho}_1^{ss} \cong \widetilde{\rho}_2^{ss}$. We think of such an isomorphism as a congruence between the two representations ρ_1 and ρ_2. Note that the semisimplifications of $E[p] \otimes \widetilde{\rho}_1$ and $E[p] \otimes \widetilde{\rho}_2$ will also be isomorphic.

Assume now that Π is a normal subgroup of Δ and that Π is a p-group. Let $\Delta_0 = \Delta/\Pi$. Of course, $\Delta_0 = \mathrm{Gal}(K_0/F)$ for some subfield K_0 of K. Since $\mathfrak{f}$ has characteristic p, one sees easily that every irreducible representation of Δ over $\mathfrak{f}$ factors through Δ_0. A result in modular representation theory implies that if ρ_1 is a representation of Δ over $\mathcal{F}$, then there exists a representation ρ_2 of Δ which factors through Δ_0 such that $\widetilde{\rho}_1^{ss} \cong \widetilde{\rho}_2^{ss}$. Furthermore, one can show that $\mathrm{Sel}_E(K_\infty)[p]$ is finite if and only if $\mathrm{Sel}_E(K_{0,\infty})[p]$ is finite, where $K_{0,\infty}$ is the cyclotomic $\mathbf{Z}_p$-extension of K_0. Thus, it suffices to assume the finiteness of $\mathrm{Sel}_E(K_{0,\infty})[p]$. The corresponding congruence relation from corollary 3.2 then shows that $\lambda(E, \Sigma_0, \rho_1)$ can be expressed just in terms of the $\lambda(E, \Sigma_0, \sigma)$'s for $\sigma \in \mathrm{Irr}_{\mathcal{F}}(\Delta_0)$. Thus, the function $\lambda(E, \Sigma_0, \cdot)$ on $\mathrm{Irr}_{\mathcal{F}}(\Delta)$ is completely determined by its restriction to $\mathrm{Irr}_{\mathcal{F}}(\Delta_0)$.

In the special case where Δ is itself a p-group, one obtains the simple formula $\lambda(E, \Sigma_0, \sigma) = \deg(\sigma)\lambda(E, \Sigma_0, \sigma_0)$, where σ_0 is the trivial representation of Δ. In this case, that formula was proven in [16]. It is stated there in a somewhat different form. One needs to assume that $\mathrm{Sel}_E(F_\infty)[p]$ is finite.

There are results of a similar nature in [3]. They concern irreducible Artin representations σ which factor through $\mathcal{G} = \mathrm{Gal}(\mathcal{K}/F)$, where $\mathcal{K}$ is generated over F by all the p-power roots of some $\alpha \in F^\times$ (subject to some mild restrictions on α). This is called a *"false Tate extension"* of F. Note that $\mathcal{G} = \mathrm{Gal}(\mathcal{K}/F)$ is a non-commutative p-adic Lie group of dimension 2. Since $\mathcal{K}$ contains μ_{p^∞}, the cyclotomic $\mathbf{Z}_p$-extension F_∞ of F is contained in $\mathcal{K}$. Therefore, the earlier assumption that $\mathcal{K} \cap F_\infty = F$ is not satisfied here. So we instead let $\Delta = \mathrm{Gal}(\mathcal{K}/F_\infty)$ and let $\Delta_0 = \mathrm{Gal}(F(\mu_{p^\infty})/F_\infty)$. Note that Δ_0 is cyclic and has order dividing $p-1$, and that the kernel Π of the map $\Delta \to \Delta_0$ is a pro-p group. These facts simplify the representation theory significantly, both in characteristic 0 and in characteristic p.

If σ' is an irreducible representation of Δ, one can define $\lambda(E, \sigma')$ essentially as before. One can then define $\lambda(E, \rho')$ for any representation ρ' of Δ. We define $\lambda(E, \sigma)$ to be $\lambda(E, \sigma|_\Delta)$ for all irreducible Artin representations σ of $\mathcal{G}$. The irreducible representations of Δ_0 are 1-dimensional. They are powers of ω, the p-power cyclotomic character which has order dividing $p-1$. Those characters are restrictions of characters of $\mathrm{Gal}(F(\mu_p)/F)$ to Δ. The results in section 4 of [3] give formulas for $\lambda(E, \sigma)$ in terms of the $\lambda(E, \omega^i)$'s under a certain hypothesis which we state below. Such formulas are also derived in [12], but under a somewhat different hypothesis.

We want to briefly discuss these hypotheses. Let $\mathcal{X}$ denote the Pontryagin dual of $\mathrm{Sel}_E(\mathcal{K})_p$. One can view $\mathcal{X}$ as a module over the completed group ring

$\mathbf{Z}_p[[\mathcal{G}]]$, the Iwasawa algebra for the 2-dimensional p-adic Lie group $\mathcal{G}$. The module $\mathcal{X}$ is finitely-generated over that ring. The key hypothesis in [3] is the following:

1: *$\mathcal{X}/\mathcal{X}_{tors}$ is finitely-generated as a $\mathbf{Z}_p[[\Delta]]$-module.*

Now it is known that $\mathbf{Z}_p[[\mathcal{G}]]$ is Noetherian. It follows that $\mathcal{X}_{tors}$ is killed by a fixed power of p. Note also that $\mathcal{X}/\mathcal{X}_{tors}$ is the Pontryagin dual of $\mathrm{Sel}_E(\mathcal{K})_{p,div}$. Under the above hypothesis, the results in [3] are proved by a K-theoretic approach. It may be possible to prove the results in [12] by such an approach. The proofs there work under the following hypothesis.

2: $\mathrm{Sel}_{E'}\big(F(\mu_{p^\infty})\big)[p]$ *is finite for at least one elliptic curve E' in the F-isogeny class of E.*

One can deduce hypothesis **1** from hypothesis **2**. However, the precise relationship between these hypotheses is not clear at present.

The results described in this section can be reformulated in another way. The analogy with the results mentioned in section 2 then becomes clearer. One can give an alternative definition of $\lambda(E,\sigma)$ as the $\mathcal{O}$-corank of a Selmer group over F_∞ associated to the $\mathcal{F}$-representation space $V_p(E)\otimes\sigma$ for G_{F_∞} One chooses a Galois invariant $\mathcal{O}$-lattice. We denote the corresponding quotient by $E[p^\infty]\otimes\sigma$, which is a discrete, divisible $\mathcal{O}$-module whose $\mathcal{O}$-rank is $2\dim_{\mathcal{F}}(\sigma)$. We then define a Selmer group essentially as for $E[p^\infty]$ itself. It is a subgroup of the Galois cohomology group $H^1(F_\infty, E[p^\infty]\otimes\sigma)$. For primes of F_∞ not lying over p, cocycle classes are required to be locally trivial. For primes π lying over p, cocycle classes are required to have a trivial image in $H^1(F_{\infty,\pi}, \overline{E}_\pi[p^\infty]\otimes\sigma)$.

The proof that $\lambda(E,\sigma)$ coincides with the $\mathcal{O}$-corank of $\mathrm{Sel}_{E[p^\infty]\otimes\sigma}(F_\infty)$ is a straightforward argument using the restriction maps for the global and local H^1's. Note that if ρ_1 and ρ_2 are representations of Δ and $\widetilde{\rho}_1^{ss}\cong\widetilde{\rho}_2^{ss}$, then

$$E[p]\otimes\widetilde{\rho}_1^{ss} \;\cong\; E[p]\otimes\widetilde{\rho}_2^{ss} \quad .$$

Thus, the residual representations for $V_p(E)\otimes\rho_1$ and $V_p(E)\otimes\rho_2$ will at least have isomorphic semisimplifications. Chapter 4 in [12] gives a proof of corollary 3.2 from this point of view, although only under more stringent hypotheses.

4. Parity

Continuing with the situation in section 3, the Birch and Swinnerton-Dyer conjecture for E over the field K asserts that the rank of $E(K)$ and the order of vanishing of the Hasse-Weil L-function $L(E,K,s)$ at $s=1$ should be the same. One can factor $L(E,K,s)$ as a product of L-functions $L(E,\sigma,s)$, where

σ varies over the irreducible representations of $\Delta = \mathrm{Gal}(K/F)$ over $\mathbf{C}$, each with multiplicity $\deg(\sigma)$. A refined form of the Birch and Swinnerton-Dyer conjecture asserts that, for every such σ, the multiplicity $r(E,\sigma)$ of σ in the $\mathbf{C}$-representation space $E(K) \otimes_{\mathbf{Z}} \mathbf{C}$ for Δ and the order of vanishing of $L(E,\sigma,s)$ at $s=1$ should be the same. This refined conjecture is stated in [28] where it is derived from the Birch and Swinnerton-Dyer conjecture and a conjecture of Deligne and Gross.

The functional equation for $L(E,\sigma,s)$ relates that function to $L(E,\check{\sigma},2-s)$, where $\check{\sigma}$ is the contragredient of σ. If σ is self-dual (i.e., $\sigma \cong \check{\sigma}$), then that functional equation will have a root number $W(E,\sigma) \in \{\pm 1\}$ which would determine the parity of the order of vanishing at $s=1$. The analytic continuation and functional equation for the L-functions mentioned above are conjectural in general, but there is a precise definition of $W(E,\sigma)$ due to Deligne [5]. General formulas for $W(E,\sigma)$ are derived in [30]. If one just considers the parity of the multiplicity and the order of vanishing, then one is led to conjecture that $W(E,\sigma) = (-1)^{r(E,\sigma)}$ for any self-dual representation σ of Δ.

It has proved easier to study a Selmer group version of the above conjecture. Fix embeddings of $\overline{\mathbf{Q}}$ into $\mathbf{C}$ and into $\overline{\mathbf{Q}}_p$. We can then realize σ as an irreducible representation of Δ over $\overline{\mathbf{Q}}$, and then over $\overline{\mathbf{Q}}_p$. Let $X_E(K)$ denote the Pontryagin dual of $\mathrm{Sel}_E(K)_p$. Let $s(E,\sigma)$ denote the multiplicity of σ in the $\overline{\mathbf{Q}}_p$-representation space $X_E(K) \otimes_{\mathbf{Z}_p} \overline{\mathbf{Q}}_p$. If the p-primary subgroup of the Tate-Shafarevich group for E over K is finite, then one has $s(E,\sigma) = r(E,\sigma)$. The parity conjecture that we will discuss here is the equality:

$$W(E,\sigma) \;=\; (-1)^{s(E,\sigma)} \tag{4}$$

for any self-dual irreducible representation of Δ. We will assume that p is odd.

There has been significant progress on this conjecture in certain cases. The first results go back to [1], and later [20], [15], and [24]. More recently, Nekovar ([25], [26]) proved the conjecture when E is defined over $\mathbf{Q}$ and has good ordinary reduction at p, and σ is trivial. This is now known for arbitrary elliptic curves over $\mathbf{Q}$. (See [7] and [19].) Subsequently, various results for more general self-dual Artin representations σ have been proved in [7], [8], which are part of a long series of papers, and in [22], [23]. Results in [3] and [12] also have a bearing on this question, as we will explain.

Under the assumption that $\mathrm{Sel}_E(K_\infty)_p$ is Λ-cotorsion, one can define $\lambda(E,\sigma)$ as before. Also, recall that self-dual irreducible representations σ of a finite group are of two types: orthogonal or symplectic. The following result is proved in [12].

Proposition 4.1. *Assume that σ is an irreducible orthogonal representation of Δ. Then we have $\lambda(E,\sigma) \equiv s(E,\sigma) \pmod 2$.*

One can use this result together with the congruence relations in section 3 to show that $W(E,\sigma)$ behaves well under congruences. To be precise, the following result is proven in [12]:

Proposition 4.2. *Assume that E has semistable reduction at the primes of F lying above* 2 *and* 3 *and that* $\mathrm{Sel}_E(K_\infty)[p]$ *is finite. Let σ_1 and σ_2 be irreducible orthogonal representations of Δ. Assume that $\widetilde{\sigma}_1^{ss} \cong \widetilde{\sigma}_2^{ss}$. Then* (4) *holds for $\sigma = \sigma_1$ if and only if* (4) *holds for $\sigma = \sigma_2$.*

There is also a version for arbitrary orthogonal representations ρ of Δ. This kind of result is proven in [12] with a significantly weaker assumption concerning the primes above 2 or 3. It should be possible to eliminate that assumption entirely. As for symplectic irreducible representations σ, one has $W(E,\sigma) = 1$, and so one expects $s(E,\sigma)$ to be even. There seem to be no results known in that direction. However, it is not hard to show that $r(E,\sigma)$ is even if σ is symplectic. Thus, if the p-primary subgroup of the Tate-Shafarevich group for E over K is finite, then $s(E,\sigma)$ is indeed even.

In the rest of this article, we will consider the situation mentioned in section 3 where $\Delta = \mathrm{Gal}(K/F)$ has a normal p-subgroup Π, K_0 is the fixed field for Π, and $\Delta_0 = \mathrm{Gal}(K_0/F)$. We assume that $K \cap F_\infty = F$ and let $K_{0,\infty}$ be K_0F_∞, the cyclotomic $\mathbf{Z}_p$-extension of K_0. One can show that $\mathrm{Sel}_E(K_\infty)[p]$ is finite if and only if $\mathrm{Sel}_E(K_{0,\infty})[p]$ is finite. Let us assumes the finiteness of $\mathrm{Sel}_E(K_{0,\infty})[p]$ and the semistability assumption for primes above 2 and 3 in proposition 4.2. One can then derive the following consequence: *If* (4) *is valid for all irreducible orthogonal representations factoring through Δ_0, then* (4) *is valid for all irreducible orthogonal representations factoring through Δ.*

As an illustration, one can consider subfields of $F(A[p^\infty])$, where A is an elliptic curve defined over F. We will assume that the homomorphism $\mathrm{G}_F \to \mathrm{Aut}_{\mathbf{Z}_p}\big(T_p(A)\big)$ giving the action of G_F on $T_p(A)$ is surjective. Thus, $\mathrm{Gal}\big(F(A[p^\infty])/F\big) \cong \mathrm{GL}_2(\mathbf{Z}_p)$ and so $F(A[p^\infty])$ will contain a tower of subfields K_n such that $\Delta_n = \mathrm{Gal}(K_n/F)$ is isomorphic to $\mathrm{PGL}_2(\mathbf{Z}/p^{n+1}\mathbf{Z})$ for all $n \geq 0$. Let $\mathcal{K} = \cup_n K_n$. We will consider Artin representations over F which factor through $\mathrm{Gal}(\mathcal{K}/F)$, and hence through Δ_n for some $n \geq 0$.

To apply the results in [12] to $K = K_n$ for any $n \geq 0$, one may just assume that $\mathrm{Sel}_E(K_{0,\infty})[p]$ is finite. It turns out that all irreducible representations of $\mathrm{PGL}_2(\mathbf{Z}/p^{n+1}\mathbf{Z})$ are self-dual and orthogonal. Thus, under the assumptions about E in proposition 4.2 (or various alternative hypotheses), it follows that (4) holds for all the irreducible Artin representations factoring through $\mathrm{Gal}(\mathcal{K}/\mathbf{Q})$ if it holds for all irreducible Artin representations factoring through $\Delta_0 = \mathrm{Gal}(K_0/F)$. Two of those Artin representations factoring through Δ_0 are 1-dimensional, two are p-dimensional, and all the other irreducible Artin representations of $\mathrm{Gal}(\mathcal{K}/\mathbf{Q})$ are even dimensional. If one just assumes that (4) is valid for the four odd-dimensional irreducible representations σ just mentioned, then one finds that (4) is valid for a certain infinite family of irreducible Artin representations σ.

Assume that $F = \mathbf{Q}$, that $A = E$, and that the surjectivity hypothesis in the previous paragraph is satisfied. Then (4) is valid for the two 1-dimensional representations of Δ_0. This follows from the results of Nekovar, Kim, and of T. and V. Dokchitser cited above. Under mild hypotheses on the reduction type,

one then obtains (4) for the two p-dimensional Artin representations factoring through Δ_0. This follows from a result proven in [3], and also in [6] under certain stronger hypotheses, establishing (4) when σ is trivial and E is an elliptic curve without complex multiplication which has an isogeny of degree p over the base field. One applies this result to certain subfields of K_0.

The results in [3] about the parity conjecture (4) have a similar form. They concern irreducible orthogonal Artin representations which factor through the Galois group $\mathcal{G} = \mathrm{Gal}(\mathcal{K}/F)$, where $\mathcal{K}$ is a p-adic Lie extension of F. A key assumption in [3] is hypothesis **1** discussed in section 3. One takes $\Delta = \mathrm{Gal}(\mathcal{K}/F_\infty)$. The cases considered in that paper have the following property: Δ has a normal pro-p subgroup Π such that $\Delta_0 = \Delta/\Pi$ is abelian and of order prime to p. One can identify Δ_0 with a quotient of $\mathcal{G}$. In addition to the case where $\mathcal{K}$ is the false Tate extension mentioned in section 3, they consider the case where $\mathcal{K} = F(E[p^\infty])$, E is an elliptic curve without complex multiplication, and E has an isogeny of degree p over F. The result cited in the previous paragraph concerning such an elliptic curve establishes (4) for the self-dual irreducible representations of Δ_0, i.e., the characters of Δ_0 of order 1 or 2. Under some mild additional hypotheses, they can then prove (4) for all the other irreducible orthogonal Artin representations which factor through $\mathcal{G}$.

Mazur and Rubin [22] study the case where $\Delta = \mathrm{Gal}(K/F)$ is a dihedral group of order $2p^n$ for $n \geq 1$. One can then take Δ_0 to be the quotient group of Δ of order 2. Let K_0 be the corresponding quadratic extension of F. All the irreducible representations of Δ are orthogonal. There are two of degree 1, which we call ε_0 and ε_1. They factor through Δ_0. If σ is an irreducible representation of Δ which does not factor through Δ_0, then σ has degree 2. Furthermore, we have $\widetilde{\sigma}^{ss} \cong \widetilde{\varepsilon}_0 \oplus \widetilde{\varepsilon}_1$. Note also that the $\mathbf{Z}_p$-corank of $\mathrm{Sel}_E(K_0)_p$ is equal to $s(E,\varepsilon_0) + s(E,\varepsilon_1)$. The results in [22] are stated under an assumption about the parity of the $\mathbf{Z}_p$-corank of $\mathrm{Sel}_E(K_0)_p$. In essence, and under various rather mild sets of hypotheses, the results in [22] establish (4) for the σ's of degree 2 under the assumption that

$$W(E,\varepsilon_0)W(E,\varepsilon_1) \;=\; (-1)^{s(E,\varepsilon_0)+s(E,\varepsilon_1)} \;.$$

This assumption is somewhat weaker than the assumption that (4) is valid for the irreducible representations factoring through Δ_0, namely ε_0 and ε_1. Mazur and Rubin use such a result to show that the $\mathbf{Z}_p$-corank of $\mathrm{Sel}_E(K)_p$ is large for certain Galois extensions K/F. Such an assertion follows under hypotheses which imply that $W(E,\sigma) = -1$ for many self-dual irreducible representations σ of $\mathrm{Gal}(K/F)$. If $s(E,\sigma)$ is odd, then $s(E,\sigma)$ is positive. This idea is exploited in [23]. It is also pursued in [3] and [12], although much more conditionally.

One can define $W(E,\rho)$ for any self-dual Artin representation ρ over a number field F. One can also extend the definition of $s(E,\cdot)$ to all Artin representations ρ over F. Then (4) can be restated as

$$W(E,\rho) \;=\; (-1)^{s(E,\rho)} \tag{5}$$

for all self-dual Artin representations ρ over F. Following the theme of this article, one would like to prove that the validity of (5) is preserved by congruences. That is, if (5) is valid for ρ_1 and if $\widetilde{\rho}_1^{ss} \cong \widetilde{\rho}_2^{ss}$, then (5) should also be valid for ρ_2. We believe that such a result is approachable. The results in [22] discussed above go a long way in the case where Δ is a dihedral group of order $2p^n$. There are also remarkable results concerning (5) in [7] which go a long way in the case where Δ_0 is abelian and also in the case where ρ is a permutation representation.

References

[1] B. J. Birch, N. Stephens, *The parity of the rank of the Mordell-Weil group*, Topology **5** (1966), 295–299.

[2] S. Bloch, K. Kato, *L-functions and Tamagawa numbers of motives*, in the Grothendieck Festschrift, Progress in Math. I, **86**, Birkhäuser (1990), 333–400.

[3] J. Coates, T. Fukaya, K. Kato, R. Sujatha, *Root numbers, Selmer groups, and non-commutative Iwasawa theory*, J. Algebraic Geom. **19** (2010), 19–97.

[4] J. Coates, R. Greenberg, *Kummer theory for abelian varieties over local fields*, Invent. Math. **124** (1996), 129–174.

[5] P. Deligne, *Les constantes des équations fonctionelles des fonctions L*, in Modular Functions of One Variable II, Lect. Notes in Math. **349** (1973), 501–595.

[6] T. Dokchitser, V. Dokchitser, *Parity of ranks for elliptic curves with a cyclic isogeny*, J. Number Theory **128** (2008), 662–679.

[7] T. Dokchitser, V. Dokchitser, *Regulator constants and the parity conjecture*, Invent. Math. **178** (2009), 23–71.

[8] T. Dokchitser, V. Dokchitser, *On the Birch-Swinnerton-Dyer quotients modulo squares*, to appear in Annals of Math.

[9] M. Emerton, R. Pollack, T. Weston, *Variation of Iwasawa invariants in Hida families*, Invent. Math. **163** (2006), 523–580.

[10] G. Faltings, Endlichkeitssätze für abelsche Varietäten über Zahlkörpern, Invent. Math. **73** (1983), 349–366.

[11] R. Greenberg, *Iwasawa theory for p-adic representations*, Advanced Studies in Pure Math. **17** (1989), 97–137.

[12] R. Greenberg, *Iwasawa theory, projective modules, and modular representations*, to appear in Memoirs of the Amer. Math. Soc.

[13] R. Greenberg, A. Iovita, R. Pollack, *On the Iwasawa invariants of elliptic curves with supersingular reduction*, in preparation.

[14] R. Greenberg, V. Vatsal, *On the Iwasawa invariants of elliptic curves*, Invent. Math. **142** (2000), 17–63.

[15] L. Guo, *General Selmer groups and critical values of Hecke L-functions*, Math. Ann. **297** (1993), 221–233.

[16] Y. Hachimori, K. Matsuno, *An analogue of Kida's formula for the Selmer groups of elliptic curves*, J. Alg. Geom. **8** (1999), 581–601.

[17] H. Hida, *Galois representations into* $\mathbf{Z}_p[[X]]$ *attached to ordinary cusp forms*, Invent. Math. **85** (1986), 545–613.

[18] K. Kato, *p-adic Hodge theory and values of zeta functions of modular curves*, Astérisque **295** (2004), 117–290.

[19] B. D. Kim, *The parity conjecture for elliptic curves at supersingular reduction primes*, Comp. Math. **143** (2007), 47–72.

[20] K. Kramer and J. Tunnell, *Elliptic curves and local* ϵ*-factors*, Comp. Math. **46** (1982), 307–352.

[21] B. Mazur, *Rational points of abelian varieties with values in towers of number fields*, Invent. Math. **18** (1972), 183–266.

[22] B. Mazur, K. Rubin, *Finding large Selmer rank via an arithmetic theory of local constants*, Ann. of Math. **166** (2007), 579–612.

[23] B. Mazur, K. Rubin, *Growth of Selmer rank in nonabelian extensions of number fields*, Duke Math. Jour. **143** (2008) 437–461.

[24] P. Monsky, *Generalizing the Birch-Stephens theorem*, Math. Zeit. **221** (1996), 415–420.

[25] J. Nekovář, *On the parity of ranks of Selmer groups II*, Comptes Rendus de l'Acad. Sci. Paris, Serie I, **332** (2001), No. 2, 99–104

[26] J. Nekovář, Selmer complexes, Astérisque **310** (2006).

[27] R. Pollack, T. Weston, *Mazur-Tate elements of non-ordinary modular forms*, preprint.

[28] D. Rohrlich, *The vanishing of certain Rankin-Selberg convolutions*, in Automorphic Forms and Analytic Number Theory, Les publications CRM, Montreal, 1990, 123–133.

[29] D. Rohrlich, *On L-functions of elliptic curves and cyclotomic towers*, Invent. Math. **75** (1984), 404–423.

[30] D. Rohrlich, *Galois theory, elliptic curves, and root numbers*, Comp. Math. **100** (1996), 311–349.

[31] K. Rubin, *On the main conjecture of Iwasawa theory for imaginary quadratic fields*, Invent. Math. **93** (1988), 701–713.

[32] K. Rubin, *The "main conjectures" of Iwasawa theory for imaginary quadratic fields*, Invent. Math. **103** (1991), 25–68.

[33] K. Rubin and A. Silverberg, *Families of elliptic curves with constant mod p representations*, Elliptic curves, modular forms, and Fermat's last theorem, Hong Kong, International Press, 1993.

[34] P. Schneider, *p-Adic height pairings II*, Invent. Math. **79** (1985), 329–374.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Artin's Conjecture on Zeros of p-adic Forms

D.R. Heath-Brown*

Abstract

This is an exposition of work on Artin's Conjecture on the zeros of p-adic forms. A variety of lines of attack are described, going back to 1945. However there is particular emphasis on recent developments concerning quartic forms on the one hand, and systems of quadratic forms on the other.

Mathematics Subject Classification (2010). Primary 11D88; Secondary 11D72, 11E08, 11E76, 11E95

Keywords. Artin's conjecture, p-adic forms, Quartic forms, Systems of quadratic forms, u-invariant

Artin's Conjecture [1, Preface] is the following statement.

Conjecture . *Let $F(x_1, \ldots, x_n) \in \mathbb{Q}_p[x_1, \ldots, x_n]$ be a form of degree d. Then if $n > d^2$ the equation $F(x_1, \ldots, x_n) = 0$ has a non-trivial solution in $\mathbb{Q}_p^n$.*

Here $\mathbb{Q}_p$ is the p-adic field corresponding to a rational prime p. Artin was led to his conjecture by considerations about C^i-fields, and the above assertion can be re-phased to say that $\mathbb{Q}_p$ is a C^2-field. There are easy examples for every prime p and every degree d to show that one cannot take $n = d^2$ here. The conjecture can be generalized to more general $\mathfrak{p}$-adic fields, and to systems of forms of degrees $d_1, \ldots, d_r$, in which case the condition on n becomes $n > d_1^2 + \ldots + d_r^2$.

One reason for the interest in Artin's Conjecture comes from the study of Local-to-Global Principles. One example is provided by the following theorem of Birch [4].

*Mathematical Institute, 24–29, St Giles', Oxford OX1 3LB, UK.
E-mail: rhb@maths.ox.ac.uk

Theorem 1. *Let $F(x_1, \ldots, x_n) \in \mathbb{Q}[x_1, \ldots, x_n]$ be a non-singular form of degree $d \geq 2$ with $n > (d-1)2^d$. Then*

$$\#\left\{\mathbf{x} \in \mathbb{Z}^n : F(\mathbf{x}) = 0, \max_{i=1,\ldots,n} |x_i| \leq B\right\} = c_F B^{n-d} + o(B^{n-d}) \tag{1}$$

as $B \to \infty$. Moreover the constant c_F is strictly positive providing that the equation $F(x_1, \ldots, x_n) = 0$ has zeros in $\mathbb{R}^n$ and in each p-adic field.

Thus if Artin's Conjecture were true the p-adic condition would hold automatically, since $(d-1)2^d \geq d^2$.

Unfortunately Artin's Conjecture is currently only known in the cases $d = 1$ and 2 (which are classical), and $d = 3$ (due to Lewis [21]). Indeed the conjecture is known to be false in general, the first counterexample having been found by Terjanian [25], for degree $d = 4$. If one sets

$$G(x_1, x_2, x_3) = x_1^4 + x_2^4 + x_3^4 - (x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2) - x_1x_2x_3(x_1 + x_2 + x_3)$$

and

$$\begin{aligned} F(x_1, \ldots, x_{18}) &= G(x_1, x_2, x_3) + G(x_4, x_5, x_6) + G(x_7, x_8, x_9) \\ &\quad + 4G(x_{10}, x_{11}, x_{12}) + 4G(x_{13}, x_{14}, x_{15}) + 4G(x_{16}, x_{17}, x_{18}), \end{aligned}$$

then $F(\mathbf{x})$ is a form in 18 variables with only the trivial zero over $\mathbb{Q}_2$. Subsequent work has produced counter-examples for many values of d, though d is even in every case known.

Question 1. *Can one find any counter-examples to Artin's Conjecture with odd degree?*

The most important general result in the positive direction is that of Ax and Kochen [2].

Theorem 2. *For every $d \in \mathbb{N}$ there is a $p_0(d)$ such that Artin's Conjecture holds whenever $p \geq p_0(d)$.*

The proof uses Mathematical Logic, and is based on the fact that the analogue of Artin's Conjecture is known for the fields $\mathbb{F}_p((t))$. A value for $p_0(d)$ was found by Brown [8]:-

$$2^{2^{2^{2^{2^{d^{11^{4d}}}}}}}! \tag{2}$$

Here the "!" symbol is merely an exclamation mark, and not a factorial sign! Another result by Ax and Kochen [3] shows that the theory of p-adic fields is decidable. Thus for each fixed prime p and each fixed degree d there is, in principle, a procedure for deciding whether the statement

> "Every form $F(x_1, \ldots, x_{d^2+1}) \in \mathbb{Q}_p[x_1, \ldots, x_{d^2+1}]$ has a nontrivial zero over $\mathbb{Q}_p$."

is true or false. It follows that one can, in theory, test every prime up to Brown's bound (2), and hence decide whether or not Artin's Conjecture holds for a given degree d.

A second approach to Artin's Conjecture, developed by Lewis [21] for $d = 3$, Birch and Lewis [5] for $d = 5$, and Laxton and Lewis [16] for $d = 7$ and 11, applies a p-adic "minimization" process to the form F to produce a suitable model over $\mathbb{Z}_p$. One then examines the reduction $\overline{F}[\mathbf{x}] \in \mathbb{F}_p[x_1, \ldots, x_n]$. If this can be shown to have a non-singular zero, Hensel's Lemma will allow us to lift it to a non-trivial zero of F over $\mathbb{Z}_p$. However this "minimization" method has limited applicability. If d can be written as a sum of composite numbers it is possible that $\overline{F}$ factors as $G_1^{e_1} \ldots G_k^{e_k}$ with $\deg G_i \geq 2$ and $e_i \geq 2$ for every i. In this case it is impossible for $\overline{F}$ to have a non-singular zero. The method is therefore doomed to fail for such degrees. In fact $d = 1, 2, 3, 5, 7$ and 11 are the only integers which cannot be written as a sum of composite numbers. However for these values the method works moderately well, and produces results of the type given by Ax and Kochen, but with much smaller values for $p_0(d)$. Thus Leep and Yeomans [20] showed that one may take $p_0(5) = 47$, and Wooley [26], that $p_0(7) = 887$ and $p_0(11) = 8059$ are admissible. These are susceptible to further improvement, and indeed calculations by Heath–Brown have shown that for $d = 5$ Artin's Conjecture holds for $p \geq 17$.

Question 2. *Does Artin's Conjecture hold for $d = 5$, for every prime?*

This is certainly decidable in principle, but whether it is realistic to expect a computational answer with current technology is unclear.

The minimization approach can also be used for systems of forms. It shows (Demyanov [12]) that $n > 8$ suffices for a pair of quadratic forms, for every p, and (Birch and Lewis [6], Schuur [23]) that $n > 12$ suffices for a system of 3 quadratic forms, providing that $p \geq 11$. A very recent application involving forms of differing degrees has been given by Zahid [28], who shows that a quadratic and a cubic form over $\mathbb{Q}_p$ have a common zero if $n > 13 = 2^2 + 3^2$, providing that $p > 293$.

Since Artin's Conjecture is false in general, it is natural to ask about the number $v_d(p)$, defined as the minimal integer such that every form $F(x_1, \ldots, x_n) \in \mathbb{Q}_p[x_1, \ldots, x_n]$ of degree d in $n > v_d(p)$ variables, has a non-trivial p-adic zero. We also write $v_d = \max_p v_d(p)$. Brauer [7] proved a result that implies that v_d is finite for every d.

Theorem 3. *For every degree d there is an integer v_d such that for each prime p, every form $F(x_1, \ldots, x_n) \in \mathbb{Q}_p[x_1, \ldots, x_n]$ of degree d with $n > v_d$ has a non-trivial p-adic zero.*

Brauer's proof involves multiple nested inductions, and did not lead to explicit bounds for v_d. More recent versions of the argument due to Leep and Schmidt [19], and particularly Wooley [27], are vastly more efficient, yielding

$$v_d \leq d^{2^d} \tag{3}$$

in general, but this is still disappointingly large. Brauer's basic idea is to show that for any $m \in \mathbb{N}$, the form F will represent a diagonal form in m variables as soon as n is large enough compared to m. It is not hard to show (Davenport and Lewis [11]) that for every p and every d one can solve diagonal equations

$$c_1 x_1^d + \ldots + c_m x_m^d = 0$$

over $\mathbb{Q}_p$ as soon as $m > d^2$. Thus it suffices that F should represent a diagonal form in $m \geq d^2 + 1$ variables. We therefore seek linearly independent vectors $\mathbf{e}_1, \ldots, \mathbf{e}_m \in \mathbb{Q}_p^n$ such that $F(\lambda_1 \mathbf{e}_1 + \ldots + \lambda_m \mathbf{e}_m)$ is a diagonal form in $\lambda_1, \ldots, \lambda_m$. If we choose the vectors $\mathbf{e}_i$ inductively it is clear that $\mathbf{e}_m$ must be a zero of a collection of forms of degree strictly less than d. Specifically there will be $m-1$ forms of degree $d-1$; $m(m-1)/2$ forms of degree $d-2$; and so on. The induction argument therefore involves the analogue of v_d for systems of forms of differing degrees, and not just for a single form of degree d.

There is an approach to these problems (Heath–Brown [14]) which is intermediate between the method of Lewis, Birch and Lewis, and Laxton and Lewis and that of Brauer, Schmidt and Wooley. In this intermediate approach one does not diagonalize F fully, but removes enough of the coefficients to ensure that there is a multiple of F which has a non-singular zero over $\mathbb{F}_p$, so that Hensel's Lemma can be used. As an example we have the following lemma.

Lemma 1. *Let $p \neq 2, 5$ or 13 be prime and let*

$$H(x, y, z) = Ax^4 + Bxy^3 + Cy^4 + Dxz^3 + Eyz^3 + Fz^4 \in \mathbb{Q}_p[x, y, z].$$

Suppose further that A, C and F are p-adic units. Then H must represent zero non-trivially over $\mathbb{Q}_p$.

In order to produce such forms by the inductive construction above one has to solve a system containing quadratic and linear equations, but not cubics.

The power of this new method is well illustrated by the case $d = 4$, for which a direct application of (3) yields $v_4 \leq 4294967296$. In contrast the new method (Heath–Brown [14, Theorem 2 and Note Added in Proof]) establishes the following bounds.

Theorem 4. *We have*

(i) $v_4(p) \leq 120$ *for* $p \geq 11$,

(ii) $v_4(p) \leq 128$ *for* $p = 3$ *and* $p = 7$,

(iii) $v_4(5) \leq 312$, *and*

(iv) $v_4(2) \leq 4221$.

Thus $v_4 \leq 4221$

One sees that $p = 2$ is the worst case by far. It is fair to say that we have absolutely no idea what the correct value for v_4 is, and it seems natural in particular to ask the following question.

Question 3. *Are there any counter-examples to Artin's Conjecture for quartic forms with $p \neq 2$?*

It is convenient at this point to introduce the following notation. For any field K, let $\beta(r; K)$ be the least integer m such that a system of r quadratic forms over K has a non-trivial common zero in K as soon as the number of variables exceeds m. The case $d = 2$ of Artin's Conjecture, which is known to be true, yields $\beta(1; \mathbb{Q}_p) = 4$, and in general the conjecture would imply that $\beta(r; \mathbb{Q}_p) = 4r$.

The results on $v_4(p)$ from Heath–Brown [14] arise from the estimates

$$v_4(p) \leq \begin{cases} 16 + \beta(8; \mathbb{Q}_p), & p \neq 2, 5, \\ 40 + \beta(12; \mathbb{Q}_p), & p = 5, \\ 537 + \beta(43; \mathbb{Q}_p), & p = 2. \end{cases}$$

together with suitable bounds for $\beta(r; \mathbb{Q}_p)$. It is therefore natural to turn our attention to the question of systems of quadratic forms. For general r it has been shown by Leep [17] that $\beta(r; \mathbb{Q}_p) \leq 2r^2 + 2r$ for all r and p. There have been subsequent small improvements, but in all cases the bound is asymptotic to $2r^2$ as $r \to \infty$. Leep's argument is an elementary induction on r, somewhat in the spirit of the Brauer induction method.

A recent alternative attack (Heath–Brown [15]) starts from the work of Birch and Lewis [6], who used the minimization approach to handle systems of three quadratic forms. In general this leads to a set of forms over $\mathbb{F}_p$ for which one wants to find a non-singular common zero. This is done via a counting argument, so that one requires, amongst other information, an estimate for the overall number of common zeros. The following rather easy lemma suffices.

Lemma 2. *Suppose we have a system of quadratic forms*

$$Q^{(i)}(x_1, \ldots, x_n) \in \mathbb{F}_p[x_1, \ldots, x_n], \quad (1 \leq i \leq r)$$

with N common zeros over $\mathbb{F}_p$. Write N_R for the number of vectors $\mathbf{u} \in \mathbb{F}_p^r$ for which

$$\sum_{i=1}^{r} u_i Q^{(i)}(x_1, \ldots, x_n) \tag{4}$$

has rank R, and assume that such a linear combination vanishes only for $\mathbf{u} = \mathbf{0}$. Then

$$|N - q^{n-I}| \leq \sum_{1 \leq t \leq n/2} q^{n-I-t} N_{2t}.$$

For vectors $\mathbf{u}$ in the algebraic completion $\overline{\mathbb{F}}_p$ the condition that (4) should have rank at most R defines a projective algebraic variety. It is possible to derive a good upper bound for the dimension of suitable components of this set, using the fact that the original p-adic system was minimized. This bound on the dimension leads in turn to a bound for N_R. This enables one to show that the system of quadratic forms over $\mathbb{F}_p$ has a non-singular zero when p is large enough. In particular one can show that $\beta(r;\mathbb{Q}_p) = 4r$ as soon as $p > (2r)^r$.

In contrast to the situation for the original formulation of Artin's Conjecture, we know of no counter-examples for systems of quadratic forms. It is therefore possible that $\beta(r;\mathbb{Q}_p) = 4r$ for every prime p.

Question 4. *Is it true that $\beta(r;\mathbb{Q}_p) = 4r$ for every prime p?*

It is not even known what happens if we restrict the quadratic forms to be diagonal.

The Ax–Kochen result already implies the existence of a bound p_r such that $\beta(r;\mathbb{Q}_p) = 4r$ for $p > p_r$. However the two methods have a very important difference when we come to apply them to finite extensions $\mathbb{Q}_\mathfrak{p}$ of $\mathbb{Q}_p$. Suppose the residue field $F_\mathfrak{p}$ of such an extension has cardinality $q = p^e$. Then the Ax–Kochen theorem yields the existence of a bound $p_{r,e}$ such that $\beta(r;\mathbb{Q}_\mathfrak{p}) = 4r$ for $p > p_{r,e}$. Thus there is a condition on the characteristic of $F_\mathfrak{p}$. For example, the theorem leaves open the possibility that $\beta(r;\mathbb{Q}_\mathfrak{p}) > 4r$ whenever $\mathbb{Q}_\mathfrak{p}$ is a finite extension of $\mathbb{Q}_2$. In contrast, the new method extends to give the following result.

Theorem 5. *We have $\beta(r;\mathbb{Q}_\mathfrak{p}) = 4r$ whenever $\#\mathbb{F}_\mathfrak{p} > (2r)^r$.*

Here there is a condition on the cardinality of $\mathbb{F}_\mathfrak{p}$, rather than its characteristic.

This makes a crucial difference when we consider the u-invariant of function fields of the form $\mathbb{Q}_p(t_1,\dots,t_k)$, as has been shown by Leep [18]. The u-invariant of a field K is the smallest integer n such that any quadratic form over K in more than n variables must have a non-trivial zero over K. Thus $u(\mathbb{R}) = \infty, u(\mathbb{C}) = 1$ and $u(\mathbb{Q}_p) = 4$. It is easy to see that $u(K(x)) \geq 2u(K)$ in general, and hence that $u(\mathbb{Q}_p(t_1,\dots,t_k)) \geq 2^{2+k}$ for all $k \geq 0$. Prior to the appearance of the new results on $\beta(r;\mathbb{Q}_\mathfrak{p})$ just described, the only values of k for which it was known that $u(\mathbb{Q}_p(t_1,\dots,t_k))$ is finite were $k = 0$ and $k = 1$. When $k = 1$, Parimala and Suresh [22] have recently shown that the u-invariant is 8, if $p \neq 2$. The same result has been proved in a different way by Harbater, Hartmann and Krashen [13, Corollary 4.14], who handle function fields of arbitrary curves over finite extensions of $\mathbb{Q}_p$. Indeed Wooley, in unpublished work, has shown how to adapt the circle method to handle quite general problems over $\mathbb{Q}_p(t)$, proving in particular that $u(\mathbb{Q}_p(t)) = 8$ for every prime p.

In order to handle the u-invariant for function fields $\mathbb{Q}_p(\mathbf{t}) = \mathbb{Q}_p(t_1,\dots,t_k)$ in k variables, Leep considers a quadratic form $Q(X_1,\dots,X_n)$ over $\mathbb{Q}_p(\mathbf{t})$, in which the coefficients of Q are polynomials in $t_1,\dots,t_k$ of total degree at most

d, say. One now considers a finite extension $\mathbb{Q}_{\mathfrak{p}}$ of $\mathbb{Q}_p$, whose significance will become apparent later, and considers both Q and the X_i as polynomials in $t_1, \ldots, t_k$ over the new field $\mathbb{Q}_{\mathfrak{p}}$. If we suppose that the X_i are polynomials of total degree at most D then the overall number of coefficients in $X_1, \ldots, X_n$ is

$$N := n(D+k)\ldots(D+1)/k!.$$

One may regard these coefficients as variables $c_1, \ldots, c_N \in \mathbb{Q}_{\mathfrak{p}}$, which one uses to force $Q(X_1, \ldots, X_n)$ to vanish identically. Since $Q(X_1, \ldots, X_n)$ has total degree at most $2D + d$ as a function of $t_1, \ldots, t_k$ there are at most

$$R := (2D+d+k)\ldots(2D+d+1)/k!$$

coefficients which one must arrange to vanish. Each of these is a quadratic form in $c_1, \ldots, c_N$. According to Theorem 5 the corresponding system of quadratic forms has a non-trivial zero $(c_1, \ldots, c_N) \in \mathbb{Q}_{\mathfrak{p}}$ providing that $N > 4R$ and $q > (2R)^R$, where q is the cardinality of the residue field of $\mathbb{Q}_{\mathfrak{p}}$. However it is clear that $N/R \to 2^{-k}n$ as $D \to \infty$. Hence if $n = 1 + 2^{2+k}$ we can choose $D = D(k, d)$ so that $N > 4R$. It follows that $Q(X_1, \ldots, X_n) = 0$ has a non-trivial solution $X_1, \ldots, X_n \in \mathbb{Q}_{\mathfrak{p}}(t_1, \ldots, t_k)$ providing that $q > q_0(k, d)$.

One now calls on a result of Springer [24], which states that if Q is a quadratic form over a field F of characteristic different from 2, which has a non-trivial zero over some extension of F of odd degree, then Q has a non-trivial zero over F itself. Thus to complete the proof it suffices to choose $\mathbb{Q}_{\mathfrak{p}}$ to be an extension of $\mathbb{Q}$ of odd degree, and for which $q > q_0(k, d)$. One may then apply Springer's result with $F = \mathbb{Q}_{\mathfrak{p}}(\mathbf{t})$ to produce a non-trivial zero of Q over the original field $\mathbb{Q}_p(\mathbf{t})$. We therefore have the following result, due to Leep [18].

Theorem 6. *We have $u(\mathbb{Q}_p(t_1, \ldots, t_k)) = 2^{2+k}$ for all $k \in \mathbb{N}$ and all primes p.*

The elegant feature of this argument is the way in which the size constraint on q disappears. It is clear that the actual bound $(2R)^R$ is irrelevant. The reader may note that Leep's argument above is, in effect, the same as that given slightly earlier by Colliot–Thélène, Parimala and Suresh [10, Proposition 2.2].

One can utilise the case $k = 1$ of Theorem 6 to obtain new bounds for $\beta(r; \mathbb{Q}_p)$. For example one has $\beta(3; \mathbb{Q}_p) \leq 16$ and $\beta(4; \mathbb{Q}_p) \leq 24$ for every prime p. These estimates are themselves used in the proof of Theorem 4. It is curious that these results hold even for the case when the residue field is small, even though Theorem 5, from which they derive, requires the residue field to be large.

As a corollary of Theorem 6 one can give an analogous statement for pairs of quadratic forms.

Theorem 7. *Two quadratic forms over $\mathbb{Q}_p(t_1, \ldots, t_k)$, in at least $1 + 2^{3+k}$ variables, have a non-trivial common zero.*

This follows from a result of Brumer [9], which shows that if F is a field of characteristic different from 2, then a pair of quadratic forms over F will have a common zero as soon as the number of variables exceeds $u(F(X))$.

As with Theorem 6, there are examples showing that one cannot reduce the number of variables. Of course both results remain true if we replace $\mathbb{Q}_p$ by a finite extension.

In conclusion we remark that it would be interesting to know what happens for systems of cubic forms over $\mathbb{Q}_{\mathfrak{p}}$. One might hope to show that r cubic forms in $n > 9r$ variables have a common zero when the cardinality q of the residue field is large enough in terms of r. However this is currently known only for $r = 1$, by the result of Lewis [21]. If the general statement were established one could deduce an analogue of Theorem 6 for cubic forms, with the number of variables required to exceed 3^{2+k}. Here Springer's theorem would be replaced by the observation that if F is a field of characteristic zero, then any cubic form with a zero over a quadratic extension of F also has a zero over F itself.

References

[1] E. Artin, *The collected papers of Emil Artin*, (Addison–Wesley, Reading, MA, 1965).

[2] J. Ax and S. Kochen, Diophantine problems over local fields. I, *Amer. J. Math.*, 87 (1965), 605–630.

[3] J. Ax and S. Kochen, Diophantine problems over local fields. II, A complete set of axioms for p-adic number theory, *Amer. J. Math.*, 87 (1965), 631–648.

[4] B.J. Birch, Forms in many variables, *Proc. Roy. Soc. Ser.* A 265 (1961/1962), 245–263.

[5] B.J. Birch and D.J. Lewis, $\mathfrak{p}$-adic forms, *J. Indian Math. Soc. (N.S.)*, 23 (1959), 11–32.

[6] B.J. Birch and D.J. Lewis, Systems of three quadratic forms, *Acta Arith.,* 10 (1964/1965), 423–442.

[7] R. Brauer, A note on systems of homogeneous algebraic equations, *Bull. Amer. Math. Soc.*, 51 (1945), 749–755.

[8] S.S. Brown, Bounds on transfer principles for algebraically closed and complete discretely valued fields, *Mem. Amer. Math. Soc.*, 15 (1978), no. 204, iv+92pp.

[9] A. Brumer, Remarques sur les couples de formes quadratiques, *C. R. Acad. Sci. Paris Sér. A–B,* 286 (1978), no. 16, A679–A681.

[10] J.-L. Colliot–Thélène, R. Parimala and V. Suresh, Patching and local-global principles for homogeneous spaces over function fields of p-adic curves, *Comment. Math. Helv.*, to appear.

[11] H. Davenport and D.J. Lewis, Homogeneous additive equations, *Proc. Roy. Soc. Ser. A,* 274 (1963), 443–460.

[12] V.B. Demyanov, Pairs of quadratic forms over a complete field with discrete norm with a finite field of residue classes, *Izv. Akad. Nauk SSSR. Ser. Mat.,* 20 (1956), 307–324.

[13] D. Harbater, J. Hartmann and D. Krashen, Applications of patching to quadratic forms and central simple algebras, *Invent. Math.*, 178 (2009), 231–263.

[14] D.R. Heath–Brown, Zeros of p-adic forms, *Proc. London Math. Soc. (3)*, 100 (2010), 560–584.

[15] D.R. Heath–Brown, Zeros of systems of $\mathfrak{p}$-adic quadratic forms, *Composito Math.*, to appear.

[16] R.R. Laxton and D.J. Lewis, Forms of degrees 7 and 11 over $\mathfrak{p}$-adic fields, *Proc. Sympos. Pure Math., Vol. VIII*, 16–21, (Amer. Math. Soc., Providence, R.I., 1965).

[17] D.B. Leep, Systems of quadratic forms, *J. Reine Angew. Math.* 350 (1984), 109–116.

[18] D.B. Leep, The u-invariant of p-adic function fields, *preprint*.

[19] D.B. Leep and W.M. Schmidt, Systems of homogeneous equations, *Invent. Math.*, 71 (1983), 539–549.

[20] D.B. Leep and C.C. Yeomans, Quintic forms over p-adic fields, *J. Number Theory*, 57 (1996), 231–241.

[21] D.J. Lewis, Cubic homogeneous polynomials over p-adic number fields, *Ann. of Math.*, (2) 56 (1952), 473–478.

[22] R. Parimala and V. Suresh, The u-invariant of the function fields of p-adic curves, `http://arxiv.org/pdf/0708.3128v1`.

[23] S.E. Schuur, On systems of three quadratic forms, *Acta Arith.*, 36 (1980), 315–322.

[24] T.A. Springer, Sur les formes quadratiques d'indice zéro, *C. R. Acad. Sci. Paris*, 234 (1952), 1517–1519.

[25] G. Terjanian, Un contre-exemple à une conjecture d'Artin, *C. R. Acad. Sci. Paris Sér.* A–B, 262 (1966), A612.

[26] T.D. Wooley, Artin's conjecture for septic and unidecic forms, *Acta Arith.*, 133 (2008), 25–35.

[27] T.D. Wooley, On the local solubility of Diophantine systems, *Compositio Math.*, 111 (1998), 149–165.

[28] J. Zahid, Simultaneous zeros of a cubic and quadratic form, `http://arxiv.org/pdf/1001.1055`.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Relative p-adic Hodge Theory and Rapoport-Zink Period Domains

Kiran Sridhara Kedlaya*

Abstract

As an example of relative p-adic Hodge theory, we sketch the construction of the universal admissible filtration of an isocrystal (ϕ-module) over the completion of the maximal unramified extension of $\mathbb{Q}_p$, together with the associated universal crystalline local system.

Mathematics Subject Classification (2010). Primary 14G22; Secondary 11G25.

Keywords. Relative p-adic Hodge theory, Rapoport-Zink period domains.

Introduction

The subject of *p-adic Hodge theory* seeks to clarify the relationship between various cohomology theories (primarily étale and de Rham) associated to algebraic varieties over p-adic fields, in much the same way as ordinary Hodge theory clarifies the relationship between various cohomology theories (primarily Betti and de Rham) associated to complex algebraic varieties. Only recently, however, has p-adic Hodge theory progressed to the point of dealing comfortably with *families* of p-adic varieties, in the way that one uses variations of Hodge structures to deal with families of complex varieties.

In this lecture, we illustrate one example of relative p-adic Hodge theory: the construction of the universal admissible filtration on an isocrystal of given Hodge-Tate weights, and the corresponding universal crystalline local system. This problem was originally introduced by Rapoport and Zink [41], as part of a generalization of the construction of p-adic symmetric spaces by Drinfel'd [15]; the relevant spaces in this construction are the moduli of filtered isocrystals. For

*Supported by NSF (CAREER grant DMS-0545904), DARPA (grant HR0011-09-1-0048), MIT (NEC Fund, Cecil and Ida Green professorship), IAS (NSF grant DMS-0635607, James D. Wolfensohn Fund).

Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. E-mail: kedlaya@mit.edu.

K_0 an absolutely unramified p-adic field with perfect residue field, an *isocrystal* is a finite-dimensional K_0-vector space equipped with an invertible semilinear Frobenius action. Typical examples are the crystalline cohomology groups of a smooth proper scheme over the residue field of K_0. If the scheme lifts to characteristic 0, one then obtains a *filtered isocrystal* by transferring the Hodge filtration from de Rham cohomology to crystalline cohomology via the canonical isomorphism. One can then pass directly from this filtered isocrystal to the étale cohomology of the scheme, by a recipe of Fontaine.

Given an isocrystal, the possible filtrations on it with jumps at particular indices (i.e., with prescribed *Hodge-Tate weights*) are naturally parametrized by a partial flag variety. Of the points of this variety defined over finite extensions of K_0, one can identify those which give rise to Galois representations: by a theorem of Colmez and Fontaine [13], they are the ones satisfying a simple linear-algebraic condition called *weak admissibility* (analogous to the notion of *semistability* in the theory of vector bundles). Rapoport and Zink conjecture the existence of a rigid analytic subspace of this variety, containing exactly the weakly admissible points, and admitting a local system specializing at each point to the appropriate crystalline Galois representation. What makes this conjecture subtle is that while the definition of weak admissibility suggests a natural analytic structure on the set of weakly admissible points, one cannot construct the local system without modifying the Grothendieck topology. As observed by de Jong [14], this situation is better understood in Berkovich's language of nonarchimedean analytic spaces: the space sought by Rapoport-Zink has the same rigid analytic points as the weakly admissible locus, but is missing some of the nonrigid points.

We construct the Rapoport-Zink space and its associated local system by copying as closely as possible the corresponding construction in equal characteristic given by Hartl [22]. The definition of the space itself, as suggested by Hartl [23], is similar in spirit to the definition of weak admissibility, but it concerns not the original filtered isocrystal but an associated isocrystal over a somewhat larger ring. In the case of a rigid analytic point, this ring is the *Robba ring* appearing in the modern theory of p-adic differential equations; it is the ring of germs of power series (over a certain coefficient field) convergent at the outer boundary of the open unit disc (or more precisely, on some unspecified open annulus with unit outer radius). The classification of isocrystals over the Robba ring, analogous to the classification of rational Dieudonné modules, was introduced by this author in [27]. We generalized this classification [29] in a fashion that allows it to be applied to arbitrary Berkovich-theoretic points of the flag variety. Having identified a candidate for the admissible locus, we imitate Berger's alternate proof of the Colmez-Fontaine theorem [6] to construct an isocrystal (more precisely a (ϕ, Γ)-module) over a relative Robba ring, from which we construct the desired local system by the usual procedure from p-adic Hodge theory (specializing to p-th power roots of unity and then performing Galois descent).

One of the intended applications of this construction is to the study of period morphisms associated to moduli spaces of p-divisible groups (Barsotti-Tate groups). Fix a p-divisible group G over $\mathbb{F}_p^{\mathrm{alg}}$ of height h and dimension d; its rational crystalline Dieudonné module $\mathbb{D}(G)_{K_0}$ is then an isocrystal over $K_0 = \mathrm{Frac}(W(\mathbb{F}_p^{\mathrm{alg}}))$. To any complete discrete valuation ring $\mathfrak{o}_K$ of characteristic 0 with residue field $\mathbb{F}_p^{\mathrm{alg}}$, and any deformation of G to a p-divisible group $\tilde{G}$ over $\mathfrak{o}_K$, Grothendieck and Messing [37] associate an extension

$$0 \to (\mathrm{Lie}\ \tilde{G}^\vee)_K^\vee \to \mathbb{D}(G)_K \to (\mathrm{Lie}\ \tilde{G})_K \to 0.$$

This gives $\mathbb{D}(G)_K$ the structure of a filtered isocrystal with Hodge-Tate weights in $\{0,1\}$, and determines a K-point in the Grassmannian $\mathcal{F}$ of $(h-d)$-dimensional subspaces of $\mathbb{D}(G)_{K_0}$. Grothendieck asked [21] which points of $\mathcal{F}$ can occur in this fashion; Rapoport and Zink proved [41, 5.16] that all such points belong to the image of a certain *period morphism* from the generic fibre of the universal deformation space of G. Using results of Faltings, Hartl [24, Theorem 3.5] has shown that his (and our) admissible locus is exactly the image of the Rapoport-Zink period morphism.

The presentation here is based on a lecture series given in January 2010 during the *Trimestre Galoisien* at Institut Henri Poincaré (Paris). The original lecture notes for that series are available online [33]. We are currently preparing a more detailed manuscript in collaboration with Ruochuan Liu.

1. Nonarchimedean Analytic Spaces

It is convenient to use Berkovich's language of nonarchimedean analytic spaces. Here is the briefest of synopses of [8].

Definition 1.1. Consider the following conditions on a ring A (always assumed to be commutative and unital) and a function $\alpha : A \to [0, +\infty)$.

(a) For all $g, h \in A$, we have $\alpha(g-h) \leq \max\{\alpha(g), \alpha(h)\}$.

(b) We have $\alpha(0) = 0$.

(b′) For all $g \in A$, we have $\alpha(g) = 0$ if and only if $g = 0$.

(c) We have $\alpha(1) = 1$, and for all $g, h \in A$, we have $\alpha(gh) \leq \alpha(g)\alpha(h)$.

(c′) We have $\alpha(1) = 1$, and for all $g, h \in A$, we have $\alpha(gh) = \alpha(g)\alpha(h)$.

We say α is a *(nonarchimedean) seminorm* if it satisfies (a) and (b), and a *(nonarchimedean) norm* if it satisfies (a) and (b′). We say α is *submultiplicative* if it satisfies (c), and *multiplicative* if it satisfies (c′).

Example 1.2. For any ring A, the function sending 0 to 0 and every other element of A to 1 is a nonarchimedean norm, called the *trivial norm*. It is multiplicative if A is an integral domain, and submultiplicative otherwise.

Example 1.3. For R a ring equipped with a (sub)multiplicative (semi)norm $|\cdot|$, the *Gauss (semi)norm* on $R[T]$ takes $\sum_i a_i T^i$ to $\max_i\{|a_i|\}$.

Definition 1.4. Let A be a ring equipped with a submultiplicative norm $|\cdot|$. The *Gel'fand spectrum* $\mathcal{M}(A)$ of A is the set of multiplicative seminorms α on A bounded above by $|\cdot|$, topologized as a closed (hence compact, by Tikhonov's theorem) subspace of the product $\prod_{a\in A}[0,|a|]$. A subbasis of this topology is given by the sets $\{\alpha \in \mathcal{M}(A) : \alpha(f) \in I\}$ for each $f \in A$ and each open interval $I \subseteq \mathbb{R}$. For $\widehat{A}$ the separated completion of A with respect to $|\cdot|$, extension by continuity gives a natural identification of $\mathcal{M}(A)$ with $\mathcal{M}(\widehat{A})$.

For $\alpha \in \mathcal{M}(A)$, the seminorm α induces a multiplicative norm on the integral domain $A/\alpha^{-1}(0)$, and hence also on $\mathrm{Frac}(A/\alpha^{-1}(0))$. The completion of this latter field is the *residue field* of α, denoted $\mathcal{H}(\alpha)$.

Lemma 1.5. *Let A, B be rings equipped with submultiplicative norms $|\cdot|_A$, $|\cdot|_B$.*

(a) Let $\phi : A \to B$ be a ring homomorphism for which $|\phi(a)|_B \leq |a|_A$ for all $a \in A$. Then ϕ induces a continuous map $\phi^* : \mathcal{M}(B) \to \mathcal{M}(A)$ by restriction.

(b) Suppose further that ϕ is injective and A admits an orthogonal complement in B. Then ϕ^* is surjective.

Proof. Part (a) is clear. For (b), note that for any $\alpha \in \mathcal{M}(A)$, $\mathcal{H}(\alpha)$ admits an orthogonal complement in the completed tensor product $\mathcal{H}(\alpha)\widehat{\otimes}_A B$. We may thus apply [8, Theorem 1.2.1] to produce an element $\beta \in \mathcal{M}(\mathcal{H}(\alpha)\widehat{\otimes}_A B)$, whose restriction to B will lie in the fibre of ϕ^* above α. □

Example 1.6. Consider the ring $\mathbb{Z}$ equipped with the trivial norm. In this case, one may describe $\mathcal{M}(\mathbb{Z})$ as the comb

$$\bigcup_{p \text{ prime}} \{(cp^{-1}, cp^{-2}) : c \in [0,1]\} \subseteq \mathbb{R}^2,$$

where

$$(cp^{-1}, cp^{-2})(p^a m) = (1-c)^a \quad (a, m \in \mathbb{Z}; a \geq 0; m \not\equiv 0 \pmod{p}).$$

In particular, each neighborhood of $(0,0)$ (the trivial norm) contains the complement of some finite union of the given segments.

Example 1.7. For K a field complete for a multiplicative norm, equip $K[T]$ with the Gauss norm. The points of $\mathcal{M}(K[T])$ (the *closed unit disc* over K) have been classified by Berkovich [8, §1] when K is algebraically closed (see also [31, §2] for the general case). For instance, for each $z \in K$ with $|z| \leq 1$ and each $r \in [0,1]$, the formula

$$\alpha_{z,r}(f) = \max_i \left\{ r^i \left| \frac{1}{i!}\frac{d^i f}{dT^i}(z) \right| \right\} \tag{1.7.1}$$

defines a point $\alpha_{z,r} \in \mathcal{M}(K[T])$; these comprise all points of $\mathcal{M}(K[T])$ if and only if K is spherically complete and algebraically closed. For applications of this space to dynamical systems on the projective line, see [3].

2. Witt Vectors

We will make extensive use of the Witt vectors over a perfect $\mathbb{F}_p$-algebra. Even if these are familiar, some facts about their Gel'fand spectra may not be.

Definition 2.1. For R a perfect $\mathbb{F}_p$-algebra (i.e., a ring in which $p = 0$ and the p-th power map is a bijection), let $W(R)$ denote the ring of p-typical Witt vectors over R. The definition of $W(R)$ may be reconstructed from the following key properties.

(a) The ring $W(R)$ is p-adically complete and separated, and $W(R)/(p) \cong R$.

(b) For each $r \in R$, there is a unique lift $[r]$ of r to $W(R)$ (the *Teichmüller lift*) having p^n-th roots for all positive integers n.

(c) Each $x \in W(R)$ admits a unique representation $\sum_{i=0}^{\infty} p^i[x_i]$ with $x_i \in R$.

Since the construction of $W(R)$ is functorial in R, $W(R)$ also carries an automorphism ϕ (the *Witt vector Frobenius*) lifting the p-power Frobenius map on R. We equip $W(R)$ with the *normalized p-adic norm*, that is, the norm of a nonzero element $\sum_{i=0}^{\infty} p^i[x_i]$ equals p^{-j} for j the smallest index with $x_j \neq 0$.

If R carries a submultiplicative norm and $R[T]$ carries the corresponding Gauss norm, we obtain a map $\lambda : \mathcal{M}(R) \to \mathcal{M}(R[T])$ taking each seminorm on R to its Gauss extension, and a map $\mu : \mathcal{M}(R[T]) \to \mathcal{M}(R)$ induced by the inclusion $R \to R[T]$. For R a perfect $\mathbb{F}_p$-algebra, we have similar maps between $\mathcal{M}(R)$ and $\mathcal{M}(W(R))$, with the role of the inclusion $R \to R[T]$ played by the multiplicative (but not additive) Teichmüller map.

Lemma 2.2. *Equip R with the trivial norm. For $\alpha \in \mathcal{M}(R)$, the function $\lambda(\alpha) : W(R) \to [0,1]$ given by*

$$\lambda(\alpha)\left(\sum_{i=0}^{\infty} p^i[x_i]\right) = \max_i\{p^{-i}\alpha(x_i)\}.$$

is a multiplicative seminorm bounded by the p-adic norm, and so belongs to $\mathcal{M}(W(R))$.

Proof. Let $x = \sum_{i=0}^{\infty} p^i[x_i], y = \sum_{i=0}^{\infty} p^i[y_i]$ be two general elements of $W(R)$. If we write $x + y = \sum_{i=0}^{\infty} p^i[z_i]$, then each z_i is a polynomial in $x_j^{p^{j-i}}, y_j^{p^{j-i}}$ for $j = 0, \dots, i$, which is homogeneous of degree 1 for the weighting in which x_j, y_j

have degree 1. It follows that $\lambda(\alpha)(x+y) \leq \max\{\lambda(\alpha)(x), \lambda(\alpha)(y)\}$, so $\lambda(\alpha)$ is a seminorm. This in turn implies that

$$\begin{aligned}\lambda(\alpha)(xy) &\leq \max_{i,j}\{\lambda(\alpha)(p^i[x_i]p^j[y_j])\} \\ &\leq \lambda(\alpha)(x)\lambda(\alpha)(y),\end{aligned}$$

so $\lambda(\alpha)$ is submultiplicative. To check multiplicativity, we may safely assume $\lambda(\alpha)(x), \lambda(\alpha)(y) > 0$. Choose the minimal indices j, k for which $\lambda(\alpha)(p^j[x_j])$, $\lambda(\alpha)(p^k[y_k])$ attain their maximal values. For

$$x' = \sum_{i=j}^{\infty} p^i[x_i], \qquad y' = \sum_{i=k}^{\infty} p^i[y_i],$$

on one hand we have $\lambda(\alpha)(x-x') < \lambda(\alpha)(x)$, $\lambda(\alpha)(y-y') < \lambda(\alpha)(y)$. Since $\lambda(\alpha)$ is a submultiplicative seminorm, we get that $\lambda(\alpha)(xy) = \lambda(\alpha)(x'y')$. On the other hand, we may write $x'y' = \sum_{i=j+k}^{\infty} p^i[z_i]$ with $z_{j+k} = x_j y_k$. Therefore $\lambda(\alpha)(x'y') \geq \lambda(\alpha)(x)\lambda(\alpha)(y)$. Putting everything together, we deduce that $\lambda(\alpha)$ is multiplicative. □

Lemma 2.3. *Equip $W(R)$ with the p-adic norm. For $\beta \in \mathcal{M}(W(R))$, the function $\mu(\beta) : R \to [0,1]$ given by*

$$\mu(\beta)(x) = \beta([x])$$

is a multiplicative seminorm bounded by the trivial norm, and so belongs to $\mathcal{M}(R)$.

Proof. Given $x_0, y_0 \in R$, choose any $x, y \in W(R)$ lifting them. For $(z_0, z) = (x_0, x), (y_0, y), (x_0+y_0, x+y)$, for any $\epsilon > 0$, for n sufficiently large (depending on z, z_0, ϵ), we have $\max\{\epsilon, \mu(\beta)(z_0)\} = \max\{\epsilon, \beta(\phi^{-n}(z))^{p^n}\}$ because $\phi^{-n}(z^{p^n})$ converges p-adically to $[z_0]$. Since β is a multiplicative seminorm, we deduce the same for $\mu(\beta)$. □

Theorem 2.4. *Equip R with the trivial norm and $W(R)$ with the p-adic norm. Then the functions $\lambda : \mathcal{M}(R) \to \mathcal{M}(W(R)), \mu : \mathcal{M}(W(R)) \to \mathcal{M}(R)$ are continuous. Moreover, for any $\alpha \in \mathcal{M}(R), \beta \in \mathcal{M}(W(R))$, we have $(\mu \circ \lambda)(\alpha) = \alpha$ and $(\lambda \circ \mu)(\beta) \geq \beta$. (The latter means that for any $x \in W(R)$, $(\lambda \circ \mu)(\beta)(x) \geq \beta(x)$.)*

Proof. For $x = \sum_{i=0}^{\infty} p^i[x_i] \in W(R)$ and $\epsilon > 0$, choose $j > 0$ for which $p^{-j} < \epsilon$; then $\lambda(\alpha)(p^i[x_i]) < \epsilon$ for all $\alpha \in \mathcal{M}(R)$ and all $i \geq j$. We thus have

$$\begin{aligned}\{\alpha \in \mathcal{M}(R) : \lambda(\alpha)(x) > \epsilon\} &= \bigcup_{i=0}^{j-1} \{\alpha \in \mathcal{M}(R) : \alpha(x_i) > p^i\epsilon\} \\ \{\alpha \in \mathcal{M}(R) : \lambda(\alpha)(x) < \epsilon\} &= \bigcap_{i=0}^{j-1} \{\alpha \in \mathcal{M}(R) : \alpha(x_i) < p^i\epsilon\},\end{aligned}$$

and the sets on the right are open. It follows that λ is continuous.

For $x_0 \in R$ and $\epsilon > 0$, we have

$$\{\beta \in \mathcal{M}(W(R)) : \mu(\beta)(x_0) > \epsilon\} = \{\beta \in \mathcal{M}(W(R)) : \beta([x_0]) > \epsilon\}$$
$$\{\beta \in \mathcal{M}(W(R)) : \mu(\beta)(x_0) < \epsilon\} = \{\beta \in \mathcal{M}(W(R)) : \beta([x_0]) < \epsilon\},$$

and the sets on the right are open. It follows that μ is continuous.

The equality $(\mu \circ \lambda)(\alpha) = \alpha$ is evident from the definitions. The inequality $(\lambda \circ \mu)(\beta) \geq \beta$ follows from the definition of λ and the observation that $(\lambda \circ \mu)(\beta)([x_0]) = \beta([x_0])$ for any $x_0 \in R$. □

Example 2.5. Here is a simple example to illustrate that $\lambda \circ \mu$ need not be the identity map. Put $R = \cup_{n=1}^{\infty} \mathbb{F}_p[X^{p^{-n}}]$, so that $W(R)$ is isomorphic to the p-adic completion of $\cup_{n=1}^{\infty} \mathbb{Z}_p[[X]^{p^{-n}}]$. The ring $W(R)/([X] - p)$ is isomorphic to the completion of $\cup_{n=1}^{\infty} \mathbb{Z}_p[p^{p^{-n}}]$ for the unique multiplicative extension of the p-adic norm; let $\beta \in \mathcal{M}(W(R))$ be the induced seminorm.

Note that $\mu(\beta)(X) = \beta([X]) = p^{-1}$ and that $\mu(\beta)(y) = 1$ for $y \in \mathbb{F}_p^\times$. These imply that $\mu(\beta)(y) \leq p^{-p^{-n}}$ whenever $y \in \mathbb{F}_p[X^{p^{-n}}]$ is divisible by $X^{p^{-n}}$, so $\mu(\beta)(y) = 1$ whenever $y \in \mathbb{F}_p^\times + X^{p^{-n}}\mathbb{F}_p[X^{p^{-n}}]$. We conclude that for $y \in R$, $\mu(\beta)(y)$ equals the X-adic norm of y with the normalization $\mu(\beta)(X) = p^{-1}$. In particular, we have a strict inequality $(\lambda \circ \mu)(\beta) > \beta$.

Remark 2.6. There is a strong analogy between the geometry of the fibres of μ and the geometry of closed discs (see Example 1.7). This suggests the possibility of constructing a homotopy between the map $\lambda \circ \mu$ on $\mathcal{M}(W(R))$ and the identity map, which acts within fibres of μ and fixes the image of $\lambda \circ \mu$; such a construction would imply that any subset of $\mathcal{M}(R)$ has the same homotopy type as its inverse image under μ. Such a homotopy does in fact exist; see [34].

3. Filtered Isocrystals and Weak Admissibility

To simplify the exposition, we introduce filtered isocrystals only for the group GL_n. One can generalize to an arbitrary reductive Lie group (see [41, Chapter 1] for the setup), but the general results can be deduced from the GL_n case.

Definition 3.1. Put $K_0 = \operatorname{Frac} W(\mathbb{F}_p^{\mathrm{alg}})$. An *isocrystal* over K_0 is a finite-dimensional K_0-vector space equipped with an invertible semilinear action of the Witt vector Frobenius ϕ. For D a nonzero isocrystal over K_0, the *degree* $\deg(D)$ of D is the p-adic valuation of the determinant of the matrix via which ϕ acts on some (and hence any) basis of D. The *slope* of D is the ratio $\mu(D) = \deg(D)/\operatorname{rank}(D)$.

Definition 3.2. Let K be a complete extension of K_0 (not necessarily discretely valued). A *filtered isocrystal* over K consists of an isocrystal D over

K_0 equipped with an exhaustive decreasing filtration $\{\mathrm{Fil}^i D_K\}_{i\in\mathbb{Z}}$ on $D_K = D \otimes_{K_0} K$. The *Hodge-Tate weights* are then defined as the multiset containing $i \in \mathbb{Z}$ with multiplicity $\dim_K(\mathrm{Fil}^i D_K)/(\mathrm{Fil}^{i+1} D_K)$.

For D an isocrystal over K_0 and H a finite multiset of integers, let $\mathcal{F}_{D,H}$ be the partial flag variety parametrizing exhaustive decreasing filtrations on D with Hodge-Tate weights H. Let $\mathcal{F}^{\mathrm{an}}_{D,H}$ be the analytification of $\mathcal{F}_{D,H}$ in the sense of [8, Theorem 3.4.1].

Beware that our notion of filtered isocrystals is slightly nonstandard; for instance, if K is a finite extension of K_0, one would normally replace K_0 by its maximal unramified extension within K. Since our ultimate goal is to fix an isocrystal structure and vary the filtration, this discrepancy is not so harmful.

Lemma 3.3. *Equip $K_0[T_1^{\pm}, \dots, T_d^{\pm}]$ with the Gauss norm. Then $\mathcal{F}^{\mathrm{an}}_{D,H}$ is covered by finitely many copies of $\mathcal{M}(K_0[T_1^{\pm}, \dots, T_d^{\pm}])$.*

Proof. We first observe that the closed unit disc $\mathcal{M}(K_0[T])$ is covered by $\mathcal{M}(K_0[T^{\pm}])$ and $\mathcal{M}(K_0[(T-1)^{\pm}])$. It follows that $\mathcal{M}(K_0[T_1, \dots, T_d])$ is covered by finitely many copies of $\mathcal{M}(K_0[T_1^{\pm}, \dots, T_d^{\pm}])$.

Let $\mathfrak{F}_{D,H}$ be the partial flag variety over $W(\mathbb{F}_p^{\mathrm{alg}})$ with generic fibre $\mathcal{F}_{D,H}$. Then $(\mathfrak{F}_{D,H})_{\mathbb{F}_p^{\mathrm{alg}}}$ is a partial flag variety over $\mathbb{F}_p^{\mathrm{alg}}$ of dimension d, and so can be covered by finitely many d-dimensional affine spaces (e.g., using Plücker coordinates). Lifting such a covering to the p-adic formal completion of $\mathfrak{F}_{D,H}$, then taking (Berkovich) analytic generic fibres, yields a covering of $\mathcal{F}^{\mathrm{an}}_{D,H}$ by finitely many copies of $\mathcal{M}(K_0[T_1, \dots, T_d])$. By the previous paragraph, this implies the desired result. □

Definition 3.4. Let D be a filtered isocrystal over some complete extension of K_0. Define $t_N(D) = \deg(D)$, and let $t_H(D)$ be the sum of the Hodge-Tate weights of D. We say D is *weakly admissible* if the following conditions hold.

(a) We have $t_N(D) = t_H(D)$.

(b) For any subisocrystal (ϕ-stable subspace) D' of D equipped with the induced filtration, $t_N(D') \geq t_H(D')$.

Weak admissibility is an open condition, in the following sense.

Theorem 3.5. *Let $\mathcal{F}^{\mathrm{wa}}_{D,H}$ be the set of $\alpha \in \mathcal{F}^{\mathrm{an}}_{D,H}$ for which D becomes weakly admissible when equipped with the filtration on $D_{\mathcal{H}(\alpha)}$ induced by the universal filtration over $\mathcal{F}_{D,H}$. Then $\mathcal{F}^{\mathrm{wa}}_{D,H}$ is open in $\mathcal{F}^{\mathrm{an}}_{D,H}$.*

Proof. See [41, Proposition 1.36]. □

Definition 3.6. The tensor product of two filtrations $\mathrm{Fil}_1^{\cdot}, \mathrm{Fil}_2^{\cdot}$ is given by

$$(\mathrm{Fil}_1 \otimes \mathrm{Fil}_2)^k = \sum_{i+j=k} \mathrm{Fil}_1^i \otimes \mathrm{Fil}_2^j .$$

It is true but not immediate that the tensor product of two weakly admissible filtered isocrystals is weakly admissible; this was proved by Faltings [16] and Totaro [42], using ideas from geometric invariant theory. It also follows *a posteriori* from describing weak admissibility in terms of Galois representations, or in terms of isocrystals over the Robba ring (Theorem 4.10).

4. Admissibility at Rigid Analytic Points

For the rest of the paper, fix an isocrystal D over K_0 and a finite multiset H of integers. In this section, we follow Berger's proof of the Colmez-Fontaine theorem [6], forging a link between filtered isocrystals and crystalline Galois representations via Frobenius modules over the Robba ring. We will use this link as the basis for our definition of the admissible locus of $\mathcal{F}^{\mathrm{an}}_{D,H}$. (One could use instead Kisin's variant of Berger's method [36]; see Remark 7.1.)

Definition 4.1. Fix once and for all a completed algebraic closure $\mathbb{C}_{K_0}$ of K_0 and a sequence $\epsilon = (\epsilon_0, \epsilon_1, \dots)$ in $\mathbb{C}_{K_0}$ in which ϵ_i is a primitive p^i-th root of 1 and $\epsilon_{i+1}^p = \epsilon_i$. (This is analogous to fixing a choice of $\sqrt{-1}$ in the complex numbers, in order to specify orientations.) Write $K_0(\epsilon)$ as shorthand for $\cup_{n=1}^{\infty} K_0(\epsilon_n)$.

Definition 4.2. For $r > 0$, define the valuation v_r on $K_0[\pi^{\pm}]$ by the formula

$$v_r\left(\sum_i a_i \pi^i\right) = \min_i \{v_p(a_i) + ir\}.$$

Let $\mathcal{R}^r$ be the Fréchet completion of $K_0[\pi^{\pm}]$ for the valuations v_s for $s \in (0, r]$; we may interpret $\mathcal{R}^r$ as the ring of formal Laurent series over K_0 convergent in the range $v_p(\pi) \in (0, r]$. Put $\mathcal{R} = \cup_{r>0} \mathcal{R}^r$ (the *Robba ring* over K_0) and

$$t = \log(1 + \pi) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} \pi^i \in \mathcal{R}.$$

Let $\mathcal{R}^{\mathrm{bd}}$ be the subring of $\mathcal{R}$ consisting of series with bounded coefficients; then $\mathcal{R}^{\mathrm{bd}}$ is a henselian (but not complete) discretely valued field for the p-adic valuation, with residue field $\mathbb{F}_p^{\mathrm{alg}}((\overline{\pi}))$. Let $\mathcal{R}^{\mathrm{int}}$ be the valuation subring of $\mathcal{R}^{\mathrm{bd}}$.

Let $\phi : \mathcal{R} \to \mathcal{R}$ be the map

$$\phi\left(\sum_i a_i \pi^i\right) = \sum_i \phi(a_i)((1+\pi)^p - 1)^i;$$

note that $\phi(t) = pt$. Define also an action of the group $\Gamma = \mathbb{Z}_p^{\times}$ on $\mathcal{R}$ by the formula

$$\gamma\left(\sum_i a_i \pi^i\right) = \sum_i a_i((1+\pi)^{\gamma} - 1)^i;$$

note that $\gamma(t) = \gamma t$, and that the action of Γ commutes with ϕ.

Definition 4.3. For $r > 0$ and n a positive integer such that $r \geq 1/(p^{n-1}(p-1))$, the series belonging to $\mathcal{R}^r$ converge at $\epsilon_n - 1$. Moreover, t vanishes to order 1 at $\epsilon_n - 1$. We thus have a well-defined homomorphism $\theta_n : \mathcal{R}^r \to K_0(\epsilon_n)[\![t]\!]$ with dense image. We also have a commutative diagram

$$\begin{tikzcd} \mathcal{R}^r \arrow[r, "\phi"] \arrow[d, "\theta_n"] & \mathcal{R}^{r/p} \arrow[d, "\theta_{n+1}"] \\ K_0(\epsilon_n)[\![t]\!] \arrow[r] & K_0(\epsilon_{n+1})[\![t]\!] \end{tikzcd} \tag{4.3.1}$$

whenever $r \leq p/(p-1)$, in which the bottom horizontal arrow acts on K_0 via ϕ, fixes ϵ_n, and carries t to pt.

Let S be a finite étale algebra over $\mathcal{R}^{\mathrm{int}}$. Choose some r for which S can be represented as the base extension of a finite étale algebra S_r over $\mathcal{R}^{\mathrm{int}} \cap \mathcal{R}^r$. For n sufficiently large, we can then form $K_{S,n} = S_r \otimes_{\theta_n} K_0(\epsilon)$. View the right side as a $K_0(\epsilon)$-algebra via the unique extension of ϕ^n fixing all of the ϵ_i; then by (4.3.1), ϕ induces an isomorphism $K_{S,n} \cong K_{S,n+1}$ of $K_0(\epsilon)$-algebras. We thus obtain functorially from S a finite étale algebra K_S over $K_0(\epsilon)$.

Theorem 4.4. *The functor $S \mapsto K_S$ is an equivalence of categories.*

Proof. This is typically deduced from Fontaine's theory of (ϕ, Γ)-modules [19], but it can also be obtained as follows. For full faithfulness, it suffices to check that if S is a field, then so is K_S. For this, choose a uniformizer of the residue field of S, and write down the minimal polynomial over $\mathbb{F}_p^{\mathrm{alg}}[\![\overline{\pi}]\!]$. This polynomial is Eisenstein, so when we lift to $\mathcal{R}^{\mathrm{int}}$ and tensor with θ_n, for n large we get an Eisenstein (and hence irreducible) polynomial over $K_0(\epsilon_n)$. Hence K_S is a field.

For essential surjectivity, it suffices to check that every field L finite over $K_0(\epsilon)$ occurs as a K_S. For some positive integer n, we can write $L = L_n(\epsilon)$ for some finite extension L_n of $K_0(\epsilon_n)$ with $[L_n : K_0(\epsilon_n)] = [L : K_0(\epsilon)]$. Fix some $r \geq 1/(p^{n-1}(p-1))$. Note that any nonzero $a \in K_0(\epsilon_n)$ can be lifted to $\tilde{a} \in \mathcal{R}^{\mathrm{int}} \cap \mathcal{R}^r$ having zero p-adic valuation.

If L_n is tamely ramified over $K_0(\epsilon_n)$, then $L_n = K_0(\epsilon_n)(a^{1/m})$ for some positive integer m not divisible by p and some $a \in K_0(\epsilon_n)$ of positive valuation. In this case, put $\tilde{P}(T) = T^m - a$ for some lift $\tilde{a} \in \mathcal{R}^{\mathrm{int}} \cap \mathcal{R}^r$ of $\phi^{-n}(a)$ having zero p-adic valuation.

If L_n is wildly ramified over $K_0(\epsilon_n)$, choose $\alpha \in L_n$ of positive valuation with $L_n = K_0(\epsilon_n)(\alpha)$ and $\mathrm{Trace}_{L_n/K_0(\epsilon_n)}(\alpha) \neq 0$. (We can enforce this last condition by replacing α by $\alpha + p$ if needed.) Let $P(T) = T^m + \sum_{i=0}^{m-1} a_i T^i$ be the minimal polynomial of α over $K_0(\epsilon_n)$, so that m is divisible by p and $a_0, a_{m-1} \neq 0$. Put $\tilde{P}(T) = T^m + \sum_{i=0}^{m-1} \tilde{a}_i T^i$ where each $\tilde{a}_i \in \mathcal{R}^{\mathrm{int}} \cap \mathcal{R}^r$ is a lift of $\phi^{-n}(a_i)$, and $\tilde{a}_i$ has zero p-adic valuation if $i \in \{0, m-1\}$.

In both cases, it can be shown that one obtains a residually separable polynomial $\tilde{P}(T)$ over $\mathcal{R}^{\mathrm{int}}$ for which $S = \mathcal{R}^{\mathrm{int}}[T]/(\tilde{P}(T))$ satisfies $K_S = L$. $\square$

Definition 4.5. For L a finite extension of K_0, let $\mathbf{B}^\dagger_{\mathrm{rig},L}$ be the finite extension of $\mathcal{R}$ obtained by starting with $L(\epsilon)$, producing a finite étale extension S of $\mathcal{R}^{\mathrm{int}}$ with $K_S = L(\epsilon)$ using Theorem 4.4, and base-extending to $\mathcal{R}$. This ring carries unique extensions of the actions of ϕ and Γ; it admits a ring isomorphism to $\mathcal{R}$, but not in a canonical way (and not respecting the actions of ϕ or Γ).

Although $\mathbf{B}^\dagger_{\mathrm{rig},L}$ does not carry a p-adic valuation, the units in this ring are series with bounded coefficients (by analysis of Newton polygons), and so do have well-defined p-adic valuations. If we then define an *isocrystal* over $\mathbf{B}^\dagger_{\mathrm{rig},L}$ to be a finite free module equipped with a semilinear action of ϕ acting on some (hence any) basis via an invertible matrix, it again makes sense to define *degree* and *slope*. An isocrystal is *étale* if it admits a basis via which ϕ acts via an invertible matrix over S.

Beware that L does not embed into $\mathbf{B}^\dagger_{\mathrm{rig},L}$, as indicated by the following lemma.

Lemma 4.6. *For any finite extension L of K_0, K_0 is integrally closed in* $\mathrm{Frac}(\mathbf{B}^\dagger_{\mathrm{rig},L})$.

Proof. Suppose $r, s \in \mathbf{B}^\dagger_{\mathrm{rig},L}$ and $f = r/s$ is integral over K_0. Then for any maximal ideal $\mathfrak{m}$ of $\mathcal{R}$ away from the support of s, the image of f in $\mathbf{B}^\dagger_{\mathrm{rig},L} \otimes_{\mathcal{R}} \mathcal{R}/\mathfrak{m}$ must be a root of a fixed polynomial over K_0. This implies that all of these images are bounded in norm, so in fact $f \in S[p^{-1}]$. In particular, f generates a finite *unramified* extension of K_0. Since K_0 has algebraically closed residue field, this forces $f \in K_0$. □

Lemma 4.7. *For any finite extension L of K_0 and any open subgroup U of Γ, we have* $(\mathrm{Frac}(\mathbf{B}^\dagger_{\mathrm{rig},L}))^U = K_0$.

Proof. Suppose first that $f \in \mathrm{Frac}(\mathcal{R})$ is Γ-invariant. Choose $r > 0$ for which $f \in \mathrm{Frac}(\mathcal{R}^r)$. For some large n, we can embed $\mathrm{Frac}(\mathcal{R}^r)$ into $K_0(\epsilon_n)((t))$ as in Definition 4.3. This action is Γ-equivariant for the action on $K_0(\epsilon_n)$ via the cyclotomic character (i.e., with $\gamma(\epsilon_n) = \epsilon_n^\gamma$) and the substitution $t \mapsto \gamma t$. It is evident that the fixed subring of $K_0(\epsilon_n)((t))$ under this action is precisely K_0, whence the claim.

In the general case, if $f \in \mathrm{Frac}(\mathbf{B}^\dagger_{\mathrm{rig},L})$ is U-invariant, then it is integral over $\mathrm{Frac}(\mathcal{R})^U$ and hence over $\mathrm{Frac}(\mathcal{R})^\Gamma = K_0$. By Lemma 4.6, this forces $f \in K_0$. □

Theorem 4.8. *Let L be a finite extension of K_0. An isocrystal M over $\mathbf{B}^\dagger_{\mathrm{rig},L}$ is étale if and only if the following conditions hold.*

(a) We have $\deg(M) = 0$.

(b) For any nonzero subisocrystal M' of M, $\deg(M') \geq 0$.

Proof. This is a consequence of slope theory for isocrystals over the Robba ring, as introduced in [27]. See [30, Theorem 1.7.1] for a simplified presentation. (For less detailed expositions, see also [11] and [32, Chapter 16].) □

Definition 4.9. Let M be an isocrystal over $\mathbf{B}^\dagger_{\mathrm{rig},L}$. For n sufficiently large, we may base-extend along θ_n to produce a module $M^{(n)}$ over $(K_0(\epsilon_n) \otimes_{\phi^n, K_0} L)((t))$. The construction is not canonical (it depends on the choice of a model of M over some $\mathcal{R}^r$), but any two such constructions give the same answers for n large. Hence any assertion only concerning the $M^{(n)}$ for n large is well-posed.

The following result of Berger [6, §III] makes the link between weak admissibility and slopes of isocrystals over the Robba ring.

Theorem 4.10 (Berger). *Let L be a finite extension of K_0. Let $(D, \mathrm{Fil}^\cdot D)$ be a filtered isocrystal over L, and view $M = D \otimes_{K_0} \mathbf{B}^\dagger_{\mathrm{rig},L}$ as an isocrystal over $\mathbf{B}^\dagger_{\mathrm{rig},L}$.*

(a) There exists an isocrystal M' over $\mathbf{B}^\dagger_{\mathrm{rig},L}$ and a ϕ-equivariant isomorphism $M[t^{-1}] \cong M'[t^{-1}]$ of modules over $\mathbf{B}^\dagger_{\mathrm{rig},L}[t^{-1}]$, via which for n sufficiently large, the t-adic filtration on $(M')^{(n)}$ coincides with the filtration on $M^{(n)}$ obtained by tensoring the t-adic filtration with the one provided by D.

(b) The isocrystal M' is étale if and only if $(D, \mathrm{Fil}^\cdot D)$ is weakly admissible.

Proof. For (a), Berger gives an algebraic construction of M' [6, §III.1]. An alternative geometric approach is to construct M' as an object in the category of coherent locally free sheaves on an open annulus of outer radius 1. One then uses the fact (essentially due to Lazard) that on an annulus over a complete discretely valued field, any coherent locally free sheaf is generated by finitely many global sections [28, Theorem 3.14].

For (b), observe that $\deg(M') = t_N(D) - t_H(D)$. Hence condition (a) of weak admissibility holds if and only if $\deg(M') = 0$. If condition (b) fails for some D', then $D' \otimes_{K_0} \mathbf{B}^\dagger_{\mathrm{rig},L}$ is a subisocrystal of M' of negative degree, so by Theorem 4.8, M' cannot be étale. Conversely, if M' fails to be étale, then by Theorem 4.8 it has a subisocrystal N' of negative slope, which we may assume to be saturated (otherwise its saturation has even smaller degree). There is a unique saturated subisocrystal N of M for which the isomorphism $M[t^{-1}] \cong M'[t^{-1}]$ induces an isomorphism $N[t^{-1}] \cong N'[t^{-1}]$.

To deduce that D is not weakly admissible, one must check that N arises from a subisocrystal of D. Put $D' = D \cap N$, so that the natural map $D/D' \otimes_{K_0} \mathbf{B}^\dagger_{\mathrm{rig},L} \to M/N$ is surjective. We check that this map is also injective. Suppose the contrary, and choose an element $\sum_{i=1}^n d_i \otimes r_i$ mapping to zero in M/N with n minimal. Then for each $j \in \{1, \dots, n\}$ and each $\gamma \in \Gamma$, $\sum_{i \neq j} d_i \otimes (r_i \gamma(r_j) -$

$r_j\gamma(r_i))$ maps to zero in M/N. By the minimality of n, we have $r_i\gamma(r_j) = r_j\gamma(r_i)$ for all $i, j \in \{1, \dots, n\}$ and all $\gamma \in \Gamma$. By Lemma 4.7, $r_i/r_j \in K_0^\times$ for all $i, j \in \{1, \dots, n\}$, which forces the d_i to be linearly dependent over K_0. But then one can rewrite d_1 in terms of $d_2, \dots, d_n$ to reduce the value of n, a contradiction.

We now conclude that $\dim_{K_0} D' = \operatorname{rank} N$, so $N = D' \otimes_{K_0} \mathbf{B}^\dagger_{\mathrm{rig},L}$. Since $t_N(D') - t_H(D') = \deg(N') < 0$, we conclude that D is not weakly admissible. (See [6, Corollaire III.2.5] for a similar argument using the Lie algebra of Γ.) $\square$

To get from étale isocrystals over the Robba ring to Galois representations, we proceed as in Definition 4.3.

Definition 4.11. Let L be a finite extension of K_0. Let M' be an étale isocrystal over $\mathbf{B}^\dagger_{\mathrm{rig},L}$. Choose a basis of M on which ϕ acts via an invertible matrix over the valuation subring $\mathfrak{o}$ of $\mathbf{B}^\dagger_{\mathrm{rig},L}$. Let N be the $\mathfrak{o}$-span of this basis. For each positive integer n, we obtain a connected finite étale Galois algebra S_n over $\mathfrak{o}$ for which $N \otimes_{\mathfrak{o}} S_n/(p^n)$ admits a ϕ-invariant basis. For n large, we can base-extend S_n via θ_n to obtain a (not necessarily connected) finite étale Galois algebra over $K_0(\epsilon) \otimes_{\phi^n, K_0} L$ (which itself may not be connected). From the Galois action on ϕ-invariant elements of $N \otimes_{\mathfrak{o}} S_n/(p^n)$, we obtain a Galois representation $G_{L(\epsilon)} \to \mathrm{GL}_m(\mathbb{Z}_p/p^n\mathbb{Z}_p)$ for $m = \operatorname{rank} M'$. (The coefficient ring is $\mathbb{Z}_p/p^n\mathbb{Z}_p$ because that is the ϕ-invariant subring of $\mathfrak{o}/p^n\mathfrak{o}$.) Taking the inverse limit, we obtain a Galois representation $G_{L(\epsilon)} \to \mathrm{GL}_m(\mathbb{Z}_p)$; the resulting representation $G_{L(\epsilon)} \to \mathrm{GL}_m(\mathbb{Q}_p)$ does not depend on the original choice of a basis of M'.

Let Γ act on $K_0(\epsilon)$ via the cyclotomic character, and let Γ_L be the open subgroup fixing $L \cap K_0(\epsilon)$. Via θ_n, we have

$$\Gamma_L \cong \mathrm{Gal}(K_0(\epsilon)/(L \cap K_0(\epsilon))) = \mathrm{Gal}(L(\epsilon)/L),$$

while for any finite Galois extension L' of L,

$$\mathrm{Aut}(\mathbf{B}^\dagger_{\mathrm{rig},L'}/\mathbf{B}^\dagger_{\mathrm{rig},L}) \cong \mathrm{Gal}(L'(\epsilon)/L(\epsilon)).$$

Taking the semidirect product yields an action of $\mathrm{Gal}(L'(\epsilon)/L)$ on $\mathbf{B}^\dagger_{\mathrm{rig},L'}$ commuting with ϕ. Similarly, if M' comes with an action of Γ_L commuting with the ϕ-action, then combining this Γ_L-action with the action of $\mathrm{Aut}(S/\mathbf{B}^\dagger_{\mathrm{rig},L})$ provides descent data on the Galois representation previously constructed, yielding a Galois representation $G_L \to \mathrm{GL}_m(\mathbb{Q}_p)$.

In the case arising from Theorem 4.10, we obtain a Γ_L-action by defining such an action on M fixing D. Berger shows [6] that the resulting Galois representation is *crystalline* in Fontaine's sense of having a full set of periods within the crystalline period ring $\mathbf{B}_{\mathrm{crys}}$. Moreover, the passage between the filtered isocrystal and the Galois representation is compatible with Fontaine's construction of the *mysterious functor*. That is, if one starts with the de Rham cohomology of a smooth proper scheme X over $\mathfrak{o}_L$, viewed as a filtered isocrystal

(using the Hodge filtration plus the Frobenius action on crystalline cohomology), the resulting Galois representation may be identified with the p-adic étale cohomology of X over L^{alg}.

5. Admissibility at General Analytic Points

Over a finite extension of K_0, we have now interpreted weak admissibility of filtered isocrystals in terms of isocrystals over Robba rings, and given a mechanism for passing from such isocrystals to Galois representations. As noted by Hartl [23], it seems to be difficult to give a direct analogue of Berger's construction over a general extension of K_0. In the case of Hodge-Tate weights in $\{0, 1\}$, Hartl has given an analogue of the notion of admissibility, using a *field of norms* construction in the manner of Fontaine and Wintenberger [20, 43] and a generalization of the slope theory for isocrystals over the Robba ring introduced in [29]. We take a different approach that applies to arbitrary weights, by reformulating in terms of the universal filtration over the partial flag variety. This requires working in local coordinates on the flag variety; the fact that the final construction does not depend on this choice (and so glues) will follow by a similar argument to the one showing that the constructed local system reproduces the previous construction over rigid analytic points (see Definition 6.5).

Definition 5.1. Let L be a field of characteristic p equipped with a valuation v_L. Let L' be the completion of L^{perf} for the unique extension of v_L. For $r > 0$, let $\tilde{\mathcal{R}}_L^{\mathrm{bd},r}$ be the subring of $W(L')[p^{-1}]$ consisting of $x = \sum_{i=m}^{\infty} p^i[x_i]$ for which $i + rv_L(x_i) \to +\infty$ as $i \to +\infty$. On $\tilde{\mathcal{R}}_L^{\mathrm{bd},r}$, the function $v_r(x) = \min_i\{i + rv_L(x_i)\}$ defines a valuation (that is, $e^{-v_r(\cdot)}$ is a multiplicative norm). Put $\tilde{\mathcal{R}}_L^{\mathrm{bd}} = \cup_{r>0}\tilde{\mathcal{R}}_L^{\mathrm{bd},r}$; this is a field which is henselian (but not complete) for the p-adic valuation.

Let $\tilde{\mathcal{R}}_L^r$ be the Fréchet completion of $\tilde{\mathcal{R}}_L^{\mathrm{bd},r}$ for the v_s for all $s \in (0, r]$. Put $\tilde{\mathcal{R}}_L = \cup_{r>0}\tilde{\mathcal{R}}_L^r$; the Witt vector Frobenius ϕ on $W(L')$ extends continuously to $\tilde{\mathcal{R}}_L$. One defines isocrystals over $\tilde{\mathcal{R}}_L$, and the associated notions of degree, slope, and étaleness, by analogy with $\mathcal{R}$. The analogue of Theorem 4.8 carries over; see [29, Corollary 6.4.3]. One has an analogue of the Dieudonné-Manin classification theorem: if L is algebraically closed, then any étale isocrystal admits a ϕ-invariant basis [29, Theorem 4.5.7].

We now make a relative analogue of the construction of Theorem 4.10, working in local coordinates on the partial flag variety $\mathcal{F}_{D,H}$.

Definition 5.2. Let S be the completion of $K_0[T_1^{\pm}, \dots, T_d^{\pm}]$ for the Gauss norm, for $d = \dim \mathcal{F}_{D,H}$. Let $\mathfrak{o}_S$ be the subring of S of elements of norm at most 1. Extend the Witt vector Frobenius ϕ from K_0 to S continuously so that $\phi(T_i) = T_i^p$ for $i = 1, \dots, d$. Let X be the open unit disc over $\mathcal{M}(S)$, i.e., the set of $\alpha \in \mathcal{M}(S[\pi])$ (for the Gauss norm on $S[\pi]$) for which $\alpha(\pi) < 1$. We may also

identify X with the set of $\alpha \in \mathcal{M}(\mathfrak{o}_S[\![\pi]\!][p^{-1}])$ (for the Gauss norm on $\mathfrak{o}_S[\![\pi]\!]$) for which $\alpha(\pi) < 1$.

The analogue of the commutative diagram (4.3.1) is

$$\begin{tikzcd} \mathfrak{o}_S[\![\pi]\!] \arrow[r, "\phi"] \arrow[d, "\theta_n"] & \mathfrak{o}_S[\![\pi]\!] \arrow[d, "\theta_{n+1}"] \\ \mathfrak{o}_S(\epsilon_n)[\![t]\!] \arrow[r] & \mathfrak{o}_S(\epsilon_{n+1})[\![t]\!] \end{tikzcd} \tag{5.2.1}$$

in which the bottom horizontal arrow acts on $\mathfrak{o}_S$ as ϕ, fixes ϵ_n, and carries t to pt. Define the group $\tilde{\Gamma} \cong \Gamma \ltimes \mathbb{Z}_p^d$ of automorphisms of X in which Γ acts as usual on $\mathcal{R}$ and trivially on S, while $(e_1, \dots, e_d) \in \mathbb{Z}_p^d$ acts as the $\mathcal{R}$-linear substitution sending T_i to $(1+\pi)^{e_i} T_i$ for $i = 1, \dots, d$.

Choose an embedding of $\mathcal{M}(S)$ into $\mathcal{F}_{D,H}^{\mathrm{an}}$, as per Lemma 3.3. Let $\mathcal{E}$ be the pullback of D along the structural morphism $X \to \mathcal{M}(K_0)$. Then there exists a coherent locally free sheaf $\mathcal{E}'$ equipped with an isomorphism $\mathcal{E}[t^{-1}] \cong \mathcal{E}'[t^{-1}]$, such that for each positive integer n, the t-adic filtration on $\mathcal{E}' \otimes_{\theta_n} S(\epsilon_n)((t))$ equals the t-adic filtration on $\mathcal{E} \otimes_{\theta_n} S(\epsilon_n)((t))$ tensored with the filtration provided by the universal filtration over $\mathcal{F}_{D,H}$. (Beware that in this construction, $S(\epsilon_n)$ is to be viewed as an S-algebra via ϕ^n.) More explicitly, we may obtain $\mathcal{E}'$ by first modifying $\mathcal{E}$ along $\pi = 0$, then pulling back by ϕ^n to obtain the appropriate modification along $\phi^n(\pi) = 0$. The local freeness of $\mathcal{E}'$ can be checked most easily by covering $\mathcal{F}_{D,H}$ with the variety parametrizing partial flags with marked basis, on which the verification becomes trivial.

One does not get an action of ϕ on $\mathcal{E}'$ over all of X, because of poles introduced at $\pi = 0$. However, one does have an isomorphism $\phi^* \mathcal{E}' \cong \mathcal{E}'$ away from $\pi = 0$.

Definition 5.3. Equip $\overline{S} = \mathbb{F}_p^{\mathrm{alg}}[T_1^{\pm}, \dots, T_d^{\pm}]$ and $\overline{S}' = \overline{S}[\![\overline{z}]\!]$ with the trivial norm. Then consider the diagram

$$\begin{tikzcd} \mathfrak{o}_S[\![\pi]\!] \arrow[r, "\phi"] \arrow[d, "\theta_0"] & \mathfrak{o}_S[\![\pi]\!] \arrow[r, "\phi"] \arrow[d, "\theta_1"] & \cdots \\ \mathfrak{o}_S(\epsilon_0) \arrow[r, "\phi"] & \mathfrak{o}_S(\epsilon_1) \arrow[r, "\phi"] & \cdots \end{tikzcd} \tag{5.3.1}$$

obtained from (5.2.1). Taking the completed direct limit over the top row gives a map $\mathfrak{o}_S[\![\pi]\!] \to W(\overline{S}'^{,\mathrm{perf}})$ sending π to $[\overline{\pi}+1]-1$ and T_i to $[\overline{T}_i]$; this map restricts to a map $\mathfrak{o}_S \to W(\overline{S}^{\mathrm{perf}})$. By Lemma 1.5, the induced maps $\mathcal{M}(W(\overline{S}^{\mathrm{perf}})) \to \mathcal{M}(\mathfrak{o}_S)$ and $\mathcal{M}(W(\overline{S}'^{,\mathrm{perf}})) \to \mathcal{M}(\mathfrak{o}_S[\![\pi]\!])$ are surjective. Taking the completed direct limit over the bottom row gives a map from $\mathfrak{o}_S$ to the completion of $W(\overline{S}^{\mathrm{perf}})(\epsilon)$; again by Lemma 1.5, the induced map $\mathcal{M}(W(\overline{S}^{\mathrm{perf}})(\epsilon)) \to \mathcal{M}(\mathfrak{o}_S)$

is surjective. In fact, its fibres are permuted transitively by the action of $\tilde{\Gamma}$ on $\mathfrak{o}_S[\![\pi]\!]$.

Definition 5.4. Put $\omega = p^{-p/(p-1)}$. Given $\tilde{\alpha} \in \mathcal{M}(W(\overline{S}^{\mathrm{perf}})(\epsilon))$, let β be the image of $\tilde{\alpha}$ under the map $\theta^* : \mathcal{M}(W(\overline{S}^{\mathrm{perf}})(\epsilon)) \to \mathcal{M}(W(\overline{S}'^{,\mathrm{perf}}))$ induced by the vertical arrows in (5.3.1). Note that $\mu(\beta)(\overline{\pi}) = \omega$.

For $L = \mathcal{H}(\mu(\beta))$, the composition $\mathfrak{o}_S[\![\pi]\!] \to W(\overline{S}'^{,\mathrm{perf}}) \to \tilde{\mathcal{R}}_L$ extends to series in π convergent in an open annulus with outer radius 1. It thus makes sense to form the base extension $\mathcal{E}' \otimes \tilde{\mathcal{R}}_L$, which is finite free over $\tilde{\mathcal{R}}_L$ [29, Theorem 2.8.4]. We say that $\alpha \in \mathcal{M}(S)$ is *admissible* if $\mathcal{E}' \otimes \tilde{\mathcal{R}}_L$ is étale for some (hence any, thanks to the $\tilde{\Gamma}$-action) choice of $\tilde{\alpha} \in \mathcal{M}(W(\overline{S}^{\mathrm{perf}})(\epsilon))$ lifting α.

Theorem 5.5. *The set $\mathcal{M}(S)^{\mathrm{adm}}$ of admissible points of $\mathcal{M}(S)$ is an open subset of $\mathcal{M}(S) \cap \mathcal{F}^{\mathrm{wa}}_{D,H}$ having the same rigid analytic points.*

Proof. Openness will follow from the construction of the universal crystalline local system (Theorem 6.2 and Theorem 6.4). The proof of Theorem 4.10 shows that an arbitrary admissible point must also be weakly admissible.

It remains to check that any weakly admissible rigid analytic point $\alpha \in \mathcal{M}(S)$ is admissible. Put $K = \mathcal{H}(\alpha)$; the lifts $\tilde{\alpha}$ of α can be put in bijection with the components of $K \otimes_{K_0} K_0(\epsilon)$. In particular, for a fixed choice of $\tilde{\alpha}$, the stabilizer $\tilde{\Gamma}_K$ of $\tilde{\alpha}$ in $\tilde{\Gamma}$ is an open subgroup. Let $\mathcal{S}_K$ denote the closure of the image of $S[\pi^{\pm}]$ in $\tilde{\mathcal{R}}_L$. Recall that $\tilde{\Gamma}$ contains $\mathbb{Z}_p^d$ as a normal subgroup; one calculates (as in Lemma 4.7) that the $(\mathbb{Z}_p^d \cap \tilde{\Gamma}_K)$-invariants of $\mathcal{S}_K$ form a copy of $\mathbf{B}^{\dagger}_{\mathrm{rig},K}$. By matching up copies of D, we obtain a (ϕ, Γ_K)-equivariant isomorphism of the $(\mathbb{Z}_p^d \cap \tilde{\Gamma}_K)$-invariant submodule of $\mathcal{E} \otimes \mathcal{S}_K$ with the module M from Theorem 4.10. Since t is invariant under $\mathbb{Z}_p^d$, we also obtain a corresponding isomorphism of primed objects. The claim then follows. $\square$

Remark 5.6. In the case of Hodge-Tate weights in $\{0, 1\}$, Hartl defined the admissible locus $\mathcal{F}^{\mathrm{adm}}_{D,H}$ (using a different but equivalent method), and showed that it is open in $\mathcal{F}^{\mathrm{wa}}_{D,H}$ [23, Corollary 5.3]. He also exhibited some examples where the two spaces differ [23, Example 5.4]; such examples have also been exhibited by Genestier and Lafforgue. Subsequently, Hartl showed (using results of Faltings) that $\mathcal{F}^{\mathrm{adm}}_{D,H}$ is the image of the Rapoport-Zink period morphism [24, Theorem 3.5].

6. The Universal Crystalline Local System

Having identified a suitable candidate for the admissible locus on $\mathcal{F}^{\mathrm{an}}_{D,H}$, we are ready to construct the universal crystalline local system over it. (An *étale local system* of a nonarchimedean analytic space can be viewed as a representation of the étale fundamental group. See [14] for a full development.)

Definition 6.1. For T an open subset of $\mathcal{M}(\overline{S}'^{,\mathrm{perf}})$ and $r > 0$, let $\mathcal{R}_S^r(T)$ be the Fréchet completion of $\mathfrak{o}_S[\![\pi]\!][\pi^{-1}, p^{-1}]$ with respect to the restrictions of the seminorms $\lambda(\gamma^{\log_\omega \rho}) \in \mathcal{M}(W(\overline{S}'^{,\mathrm{perf}}))$ for all $\gamma \in T$ and all ρ with $-\log_p \rho \in (0, r]$. Let $\mathcal{R}_S(T)$ be the union of the $\mathcal{R}_S^r(T)$ over all $r > 0$.

Theorem 6.2. *Suppose $\gamma \in \mathcal{M}(\overline{S}'^{,\mathrm{perf}})$ is such that for $L = \mathcal{H}(\gamma)$, $\mathcal{E}' \otimes \tilde{\mathcal{R}}_L$ is étale. Then there exist an open neighborhood T of γ in $\mathcal{M}(\overline{S}'^{,\mathrm{perf}})$ such that for each positive integer n, there exists a finite étale extension of $\mathcal{R}_S^r(T)$ for some $r > 0$ over which $\mathcal{E}'$ acquires a basis on which ϕ acts via a matrix whose difference from the identity has p-adic valuation at least n. (Note that T is chosen uniformly in n.)*

Proof. This is a calculation following [22, Proposition 1.7.2], which in turn follows [29, Lemma 6.1.1]. □

Definition 6.3. Suppose $\alpha \in \mathcal{M}(S)$ is admissible. Define $\tilde{\alpha}, \beta$ as in Definition 5.4, and apply Theorem 6.2 with $\gamma = \mu(\beta)$. Then $\mu^{-1}(T)$ is open in $\mathcal{M}(W(\overline{S}'^{,\mathrm{perf}}))$ because μ is continuous (Theorem 2.4), so $U = (\theta^*)^{-1}(\mu^{-1}(T))$ is open in $\mathcal{M}(W(\overline{S}^{\mathrm{perf}})(\epsilon))$. Let U' be the union of the $\tilde{\Gamma}$-translates of U, which is again open in $\mathcal{M}(W(\overline{S}^{\mathrm{perf}})(\epsilon))$. Then U and U' have the same image V_0 in $\mathcal{M}(\mathfrak{o}_S)$, but U' is the full inverse image of V_0 in $\mathcal{M}(W(\overline{S}^{\mathrm{perf}})(\epsilon))$. Hence the complement of V_0 is the image of a closed and thus compact set (namely the complement of U'), and so is compact and thus closed. We conclude that V_0 is open in $\mathcal{M}(\mathfrak{o}_S)$, and $V = V_0 \cap \mathcal{M}(S)$ is open in $\mathcal{M}(S)$. Since $\beta \in \mu^{-1}(T)$, we also have $\alpha \in V$.

Theorem 6.4. *Suppose $\alpha \in \mathcal{M}(S)$ is admissible. Define an open neighborhood V of α in $\mathcal{M}(S)$ as in Definition 6.3. Then there exists a $\mathbb{Q}_p$-local system over V whose specialization to any rigid analytic point of V may be identified with the crystalline Galois representation produced by Definition 4.11.*

Proof. Retain notation as in Theorem 6.2 and Definition 6.3, and choose a nonnegative integer n with $r > 1/(p^{n-1}(p-1))$. Let $\alpha' \in V$ be any point, choose $\tilde{\alpha}' \in U$ lifting α', and put $\beta' = \theta^*(\alpha') \in \mu^{-1}(T)$ and $\gamma' = \mu(\beta') \in T$. Then $\mathcal{R}_S^r(T)$ admits the seminorm

$$\lambda((\gamma')^{p^{-n}}) = \lambda(\mu(\beta')^{p^{-n}}) = (\lambda \circ \mu)((\phi^{-n})^*(\beta')),$$

which dominates $(\phi^{-n})^*(\beta')$ by Theorem 2.4. We conclude that the elements of $\mathcal{R}_S^r(T)$ define analytic functions on a subspace of X containing $(\phi^{n*})^{-1}(V)$, under the identification of $\mathcal{M}(S)$ with the subspace $\pi = \epsilon_n - 1$ of X. As in Definition 4.11, by considering ϕ-invariant sections of $\mathcal{E}'$ over finite étale extensions of $\mathcal{R}_S^r(T)$, we obtain a $\mathbb{Q}_p$-local system over $(\phi^{n*})^{-1}(V)$. By also keeping track of the action of $\tilde{\Gamma}$, we obtain descent data yielding a $\mathbb{Q}_p$-local system over V itself. The compatibility at rigid analytic points follows from the proof of Theorem 5.5. □

Definition 6.5. In Theorem 6.4, any point of V is automatically admissible. It follows that there exists an open subset $\mathcal{F}_{D,H}^{\mathrm{adm}}$ of $\mathcal{F}_{D,H}^{\mathrm{an}}$ such that $\mathcal{M}(S)^{\mathrm{adm}} = \mathcal{M}(S) \cap \mathcal{F}_{D,H}^{\mathrm{adm}}$ for any embedding of $\mathcal{M}(S)$ into $\mathcal{F}_{D,H}^{\mathrm{an}}$ as in Lemma 3.3. We call $\mathcal{F}_{D,H}^{\mathrm{adm}}$ the *admissible locus* of $\mathcal{F}_{D,H}^{\mathrm{an}}$.

By an argument similar to the proof of Theorem 6.4, one shows that the definition of the admissible locus, and the construction of the local system, does not depend on the choice of local coordinates. We may thus glue using a cover as in Lemma 3.3 to produce a $\mathbb{Q}_p$-local system over $\mathcal{F}_{D,H}^{\mathrm{adm}}$ specializing to the crystalline representations associated to rigid analytic points. We call this the *universal crystalline local system* on $\mathcal{F}_{D,H}^{\mathrm{adm}}$.

7. Further Remarks

Remark 7.1. In some cases of Hodge-Tate weights equal to $\{0, 1\}$, the space $\mathcal{F}_{D,H}^{\mathrm{adm}}$ receives a period morphism constructed by Rapoport-Zink [41] from the generic fibre of the universal deformation space associated to a suitable p-divisible group. One expects (as in [23, Remark 7.8]) that the pullback of the universal crystalline local system is obtained by extension of scalars from a $\mathbb{Z}_p$-local system which computes the (integral) crystalline Dieudonné module at each rigid analytic point. This appears to follow from work of Faltings (manuscript in preparation).

It should be possible to give an alternative proof, more in the spirit of this lecture, using Kisin's variant of Berger's proof of the Colmez-Fontaine theorem [36]. In Kisin's approach, the role of the highly ramified Galois tower $K_0(\epsilon)$ is played by the non-Galois Kummer tower $\cup_n K_0(p^{-p^n})$. Instead of (ϕ, Γ)-modules, one ends up with modules over $\mathfrak{o}_{K_0}[\![u]\!]$ carrying an action of the Frobenius lift $u \mapsto u^p$, with kernel killed by a power of a certain polynomial. These are particularly well suited for studying *integral* properties of crystalline representations; indeed, their definition is inspired by a construction of Breuil [9] introduced precisely to study such integral aspects (moduli of finite flat group schemes and p-divisible groups). We expect that one can carry out a close analogue of the construction described in this lecture using Kisin's modules.

Remark 7.2. When one considers the cohomologies of smooth proper schemes over a p-adic field which are no longer required to have good reduction, one must broaden the class of allowed Galois representations slightly. The correct class is Fontaine's class of *de Rham* representations, which coincides with the class of *potentially semistable* representations by a theorem of Berger [5]. On the side of de Rham cohomology, one must replace the category of isocrystals by the category of (ϕ, N)-modules with finite descent data. That is, one specifies not only a Frobenius action but also a linear endomorphism N satisfying $N\phi = p\phi N$. (Note that any such N is necessarily nilpotent.) It should be possible to follow the model of this lecture to construct a universal *semistable* local system;

one possible point of concern is that the parameter space replacing the partial flag variety is no longer proper.

Remark 7.3. The techniques of this paper fit into the general philosophy that one can study representations of arithmetic fundamental groups of schemes of finite type over a p-adic field using p-adic analytic methods. This attitude has been convincingly articulated by Faltings in several guises, such as his p-adic analogue of the Simpson correspondence [17]. It has subsequently been taken up by Andreatta and his collaborators (Brinon, Ioviţa), who have made great strides in developing and applying a relative theory of (ϕ, Γ)-modules [1, 2].

Remark 7.4. The relative p-adic Hodge theory in this paper has been restricted to the case where one starts with a fixed isocrystal and varies its filtration, corresponding geometrically to a deformation to characteristic 0 of a fixed scheme in characteristic p. It is also of great interest to consider cases where the isocrystal itself may vary.

On the Galois side, families of Galois representations parametrized by a rigid analytic space occur quite frequently in the theory of p-adic modular forms, dating back to the work of Hida [25] on ordinary families, and continuing in the work of Coleman and Mazur [10] on the eigencurve. More recently, the study of families of representations has become central in the understanding of the p-adic local Langlands correspondence, particularly for the group $\mathrm{GL}_2(\mathbb{Q}_p)$ [12].

A partial analogue of the (ϕ, Γ)-module functor for representations in analytic families has been introduced by Berger and Colmez [7]. Unfortunately, it is less than clear what the essential image of the functor is; see [35] for some discussion. Moreover, proper understanding of families of (ϕ, Γ)-modules is seriously hampered by the lack of a good theory of slopes of Frobenius modules in families; the situation at rigid analytic points is understood thanks to [30], but at nonclassical points things are much more mysterious. Nonetheless, the construction has proved useful in the study of Selmer groups in families, as in the work of Bellaïche [4] and Pottharst [39, 40].

One can also make analogous considerations on the side of Breuil-Kisin modules, as in the work of Pappas and Rapoport [38]. Again, the correspondence from modules back to Galois representations is somewhat less transparent in families, so the moduli stack of Breuil-Kisin modules itself becomes the central object of study; on this stack, Pappas and Rapoport introduce an analogue of the Rapoport-Zink period morphism. One can also introduce an analogue of the admissible locus, as in work of Hellmann (in preparation); in this line of inquiry, there appears to be some advantage in replacing Berkovich's theory of nonarchimedean analytic spaces with Huber's more flexible theory of *adic spaces* [26].

References

[1] F. Andreatta and O. Brinon, *Surconvergence des représentations p-adiques: le cas relatif*, Astérisque **319** (2008), 39–116.

[2] F. Andreatta and A. Ioviţa, *Global applications of relative (ϕ, Γ)-modules, I*, Astérisque **319** (2008), 39–420.

[3] M. Baker and R. Rumely, *Potential theory and dynamics on the Berkovich projective line*, Surveys and Monographs 159, Amer. Math. Soc., Providence, 2010.

[4] J. Bellaïche, *Ranks of Selmer groups in an analytic family*, arXiv:0906.1275v1 (2009).

[5] L. Berger, *Représentations p-adiques et équations différentielles*, Invent. Math. **148** (2002), 219–284.

[6] L. Berger, *Équations différentielles p-adiques et (ϕ, N)-modules filtrés*, Astérisque **319** (2008), 13–38.

[7] L. Berger and P. Colmez, *Familles de représentations de de Rham et monodromie p-adique*, Astérisque **319** (2008), 303–337.

[8] V. Berkovich, *Spectral theory and analytic geometry over non-archimedean fields*, Surveys and Monographs 33, Amer. Math. Soc., Providence, 1990.

[9] C. Breuil, *Groupes p-divisibles, groupes finis et modules filtrés*, Ann. of Math. **152** (2000), 489–549.

[10] R. Coleman and B. Mazur, *The eigencurve*, in *Galois representations in arithmetic algebraic geometry (Durham, 1996)*, London Math. Soc. Lecture Note Series 254, Cambridge Univ. Press, Cambridge, 1998, 1–113.

[11] P. Colmez, *Les conjectures de monodromie p-adiques*, in *Séminaire Bourbaki 2001/2002*, Astérisque **290** (2003), 53–101.

[12] P. Colmez, *Représentations de $\mathrm{GL}_2(\mathbb{Q}_p)$ et (ϕ, Γ)-modules*, Astérisque **330** (2010), 283–511.

[13] P. Colmez and J.-M. Fontaine, *Construction des représentations p-adiques semi-stables*, Invent. Math. **140** (2000), 1–43.

[14] A.J. de Jong, *Étale fundamental groups of non-Archimedean analytic spaces*, Compos. Math. **97** (1995), 89–118.

[15] V.G. Drinfel'd, *Coverings of p-adic symmetric regions*, Funct. Anal. Appl. **10** (1976), 29–40.

[16] G. Faltings, *Mumford-Stabilität in der algebraischen Geometrie*, in *Proceedings of the Intl. Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, Birkhäuser, Basel, 1995, 648–655.

[17] G. Faltings, *A p-adic Simpson correspondence*, Adv. Math. **198** (2005), 847–862.

[18] G. Faltings, *Coverings of p-adic period domains*, preprint (2007) available at `http://www.mpim-bonn.mpg.de/preprints/`.

[19] J.-M. Fontaine, *Représentations p-adiques des corps locaux, I*, in *The Grothendieck Festschrift, Vol. II*, Progr. Math. 87, Birkhäuser, Boston, 1990, 249–309.

[20] J.-M. Fontaine and J.-P. Wintenberger, *Le "corps des normes" de certaines extensions algébriques de corps locaux*, C.R. Acad. Sci. Paris Sér. A-B **288** (1979), A367–A370.

[21] A. Grothendieck, *Groupes de Barsotti-Tate et cristaux*, in *Actes du Congrès International des Mathématiciens (Nice, 1970), Tome 1*, Gauthier-Villars, 1971, 431–436.

[22] U. Hartl, *Period spaces for Hodge structures in equal characteristic*, arXiv:math.NT/0511686v2 (2006).

[23] U. Hartl, *On a conjecture of Rapoport and Zink*, arXiv:math.NT/0605254v1 (2006).

[24] U. Hartl, *On period spaces for p-divisible groups*, arXiv:0709.3444v3 (2008).

[25] H. Hida, *Galois representations into* $\mathrm{GL}_2(\mathbb{Z}_p[\![X]\!])$ *attached to ordinary cusp forms*, Invent. Math. **85** (1986), 545–613.

[26] R. Huber, *Étale cohomology of rigid analytic varieties and adic spaces*, Aspects of Mathematics, E30, Friedr. Vieweg & Sohn, Braunschweig, 1996.

[27] K.S. Kedlaya, *A p-adic local monodromy theorem*, Ann. of Math. **160** (2004), 93–184.

[28] K.S. Kedlaya, *Local monodromy of p-adic differential equations: an overview*, Int. J. Num. Theory **1** (2005), 109–154.

[29] K.S. Kedlaya, *Slope filtrations revisited*, Doc. Math. **10** (2005), 447–525; errata, ibid. **12** (2007), 361–362.

[30] K.S. Kedlaya, *Slope filtrations for relative Frobenius*, Astérisque **319** (2008), 259–301.

[31] K.S. Kedlaya, *Semistable reduction for overconvergent F-isocrystals, IV: Local semistable reduction at nonmonomial valuations*, arXiv:0712.3400v3 (2009).

[32] K.S. Kedlaya, *p-adic differential equations*, Cambridge Univ. Press, Cambridge, 2010.

[33] K.S. Kedlaya, *Slope filtrations and (ϕ, Γ)-modules in families*, lecture notes (2010) available at `http://math.mit.edu/~kedlaya/papers/`.

[34] K.S. Kedlaya, *Nonarchimedean geometry of Witt vectors*, arXiv:1004.0466v1 (2010).

[35] K.S. Kedlaya and R. Liu, *On families of (ϕ, Γ)-modules*, Alg. and Num. Theory, to appear; arXiv:0812.0112v2 (2009).

[36] M. Kisin, *Crystalline representations and F-crystals*, in *Algebraic geometry and number theory*, Progr. Math. 253, Birkhäuser, Boston, 2006, 459–496.

[37] W. Messing, *The crystals associated to Barsotti-Tate groups*, Lecture Notes in Math. 264, Springer-Verlag, Berlin, 1972.

[38] G. Pappas and M. Rapoport, *Φ-modules and coefficient spaces*, Moscow Math. J. **9** (2009), 625–663.

[39] J. Pottharst, *Triangulordinary Selmer groups*, arXiv:0805.2572v1 (2008).

[40] J. Pottharst, *Analytic families of finite-slope Selmer groups*, preprint (2010) available at `http://www2.bc.edu/~potthars/writings/`.

[41] M. Rapoport and T. Zink, *Period spaces for p-divisible groups*, Ann. Math. Studies 141, Princeton Univ. Press, Princeton, 1996.

[42] B. Totaro, *Tensor products in p-adic Hodge theory*, Duke Math. J. **83** (1996), no. 1, 79–104.

[43] J.-P. Wintenberger, *Le corps des normes de certaines extensions infinies des corps locaux; applications*, Ann. Sci. Éc. Norm. Sup. **16** (1983), 59–89.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Serre's Modularity Conjecture

Chandrashekhar Khare* and Jean-Pierre Wintenberger*

Abstract

We state Serre's modularity conjecture, give some hints on its proof and give some consequences.

Mathematics Subject Classification (2010). Primary 11R39; Secondary 11F80.

Keywords. Galois representations. Modular forms.

1. Introduction

Let p be a prime number. Let $G_{\mathbb{Q}} = \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ be the absolute Galois group of $\mathbb{Q}$. Let $\bar{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F})$ be a continuous, absolutely irreducible, two-dimensional, odd ($\det(\bar{\rho}(c)) = -1$ for $c \in G_{\mathbb{Q}}$ a complex conjugation), mod p representation, with $\mathbb{F}$ a finite field of characteristic p. We say that such a representation is of *Serre-type*, or S-type, for short. The continuity just says that $\bar{\rho}$ factors through the Galois group of a finite extension of $\mathbb{Q}$: its image lies in $\mathrm{GL}_2(\mathbb{F})$ for a finite field $\mathbb{F} \subset \overline{\mathbb{F}}_p$.

Let $\overline{\mathbb{Q}}_p$ be an algebraic closure of the field $\mathbb{Q}_p$ of p-adic numbers. Let $N \geq 1$ be an integer and let $\Gamma_1(N) \subset \Gamma_0(N)$ be the congruence subgroup:

$$\Gamma_1(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}), c \equiv 0 \bmod. N, a \equiv d \equiv 1 \bmod. N \right\}.$$

In particular, we have $\Gamma_1(1) = \mathrm{SL}_2(\mathbb{Z})$. Let k be an integer ≥ 1 and let f be a normalized new parabolic eigenform for $\Gamma_1(N)$ of weight k and level N. So f is an holomorphic function on the Poincaré upper half space $\{z \in \mathbb{C}, \mathrm{im}(z) > 0\}$ which satisfies:

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z)$$

*CK partially supported by NSF grants and a Guggenheim fellowship. JPW is member of the Institut Universitaire de France.

Department of Mathematics, UCLA, Los Angeles, CA 90095-1555, U.S.A., Université de Strasbourg, Département de Mathématique, 67084, Strasbourg Cedex, France.
E-mails: shekhar@math.ucla.edu and wintenb@math.unistra.fr.

for $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_1(N)$. The parabolic form f has a Fourier expansion: $f(z) = a_1 q + \sum_{n\geq 2} a_n q^n$, $q = e^{2\pi i z}$. It is normalized by the condition $a_1 = 1$. The Fourier coefficients a_n are in the ring of integers of a finite extension of $\mathbb{Q}$ contained in $\overline{\mathbb{Q}} \subset \mathbb{C}$. We write E_f for the field generated by the a_n and call it the field of coefficients of f. The algebraic integer a_p, for p a prime not dividing N, is the eigenvalue of the Hecke operator T_p acting on the eigenform f, and for p dividing N, is the eigenvalue of the Hecke operator U_p.

Let $\Gamma_0(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}), c \equiv 0 \bmod . N \right\}$. As f is an eigenvector for the diamond operators, there exists a character $\eta : (\mathbb{Z}/N\mathbb{Z})^* \to \mathbb{C}^*$, the Nebentypus, such that:

$$f\left(\frac{az+b}{cz+d}\right) = \eta(d)(cz+d)^k f(z)$$

for $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$. By considering the above functional identity for the matrix $-\mathrm{id}$, we see that $\eta(-1)(-1)^k = 1$.

Eichler, Shimura, Deligne ([9]), and Deligne-Serre ([8]), have associated to f (and an embedding $\iota_p : E_f \hookrightarrow \overline{\mathbb{Q}}_p$) a continuous p-adic Galois representaton $\rho(f)_{\iota_p} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$ which is characterized by the fact that it is unramified outside pN, it is irreducible, and it satisfies the Eichler-Shimura relation:

$$\mathrm{tr}(\rho(f)_{\iota_p}(\mathrm{Frob}_\ell)) = \iota_p(a_\ell),\ \det(\rho(f)_{\iota_p}) = \chi_p^{k-1}\eta.$$

Here ℓ is a prime not dividing pN. We call S_N the set of primes dividing N. We note by $G_{\mathbb{Q},S_N\cup\{p\}}$ the Galois group of the maximal extension $\overline{\mathbb{Q}}_{S_N\cup\{p\}}$ of $\mathbb{Q}$ contained in $\overline{\mathbb{Q}}$ and which is unramified outside $S_N \cup \{p\}$. We call $\mathrm{Frob}_\ell \in G_{\mathbb{Q},S_N\cup\{p\}}$ the Frobenius of a prime over ℓ in $\overline{\mathbb{Q}}_{S_N\cup\{p\}}$. The character $\chi_p : G_{\mathbb{Q}} \to \mathbb{Z}_p^*$ is the cyclotomic character. The character $\eta : (\mathbb{Z}/N\mathbb{Z})^* \to \mathbb{C}^*$ is viewed as a Galois character with values in $\overline{\mathbb{Q}}_p^*$ via the isomorphism $(\mathbb{Z}/N\mathbb{Z})^* \simeq \mathrm{Gal}(\mathbb{Q}(\mu_N)/\mathbb{Q})$ and ι_p. We replace the notation $\rho(f)_{\iota_p}$ by $\rho_p(f)$ when it is not confusing. The representation $\rho_p(f)$ is odd meaning $\det(\rho_p(f)(c)) = -1$ for $c \in G_{\mathbb{Q}}$ a complex conjugation. This follows from the relation $\eta(-1)(-1)^k = 1$ and the formula giving the determinant of $\rho_p(f)$.

By compactness of $G_{\mathbb{Q}}$, one sees that, after conjugation by an element $g \in \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$, one can suppose that $\rho(f)_p$ has image in $\mathrm{GL}_2(O)$, for O the ring of integers of a finite extension of $\mathbb{Q}_p$. By reducing modulo the maximal ideal of O, we get a representation $G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$. Its semi-simplification does not depend up to isomorphism on the choice of g. We call it $\bar{\rho}_p(f)$. It clearly is odd. When it is irreducible, it is of type S.

The following theorem had been conjectured by Serre ([32]) and is proved in ([20], [18], [21], [22]):

Theorem 1.1. *Let $\bar{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be an odd, irreducible, Galois representation. Then there is a modular form f as above and an embedding of the coefficient field of f in $\overline{\mathbb{Q}}_p$ such that $\bar{\rho} \simeq \bar{\rho}_p(f)$.*

We may consider that the above theorem proves Serre's conjecture in its qualitative form. Serre's conjecture is more precise. Given $\bar{\rho}$ of S-type, Serre defines two integers $k(\bar{\rho}) \geq 2$ and $N(\bar{\rho}) \geq 1$, such that one should be able to choose f in the theorem of weight $k(\bar{\rho})$ and level $N(\bar{\rho})$.

The integer $N(\bar{\rho})$ is the prime to p part of the conductor of $\bar{\rho}$. More precisely, if $\ell \neq p$, the exponent of the power of ℓ that exactly divides $N(\bar{\rho})$ is:

$$\sum_{i=0}^{\infty} \frac{1}{(I_0 : I_i)} \dim(V/V_i),$$

where I is the inertia group for a prime over ℓ, (I_i) is the ramification filtration, V is the $\overline{\mathbb{F}}_p$ vector space underlying $\bar{\rho}$ and V_i is the subspace of elements of V which are fixed by I_i. In particular, $N(\bar{\rho})$ is divisible by ℓ if and only if $\bar{\rho}$ is ramified at ℓ.

The integer $k(\bar{\rho})$ only depends on the action of the inertia subgroup I_p at p. If $p = 2$, $k(\bar{\rho}) = 2$ or 4. If $p \neq 2$, $2 \leq k(\bar{\rho}) \leq p^2 - 1$. Let us call $\overline{\chi}_p$ the cyclotomic character giving the action of Galois on p-roots of unity. There is an integer j such that $2 \leq k(\bar{\rho} \otimes \overline{\chi}_p^j) \leq p + 1$. Before the proof of Theorem 1.1 it was known by the work of Ribet, Mazur, Carayol, Gross, Coleman-Voloch, Edixhoven, Diamond that its statement implied the precise form of Serre's conjecture:

Theorem 1.2. *If $p \neq 2$ the qualitative form of the conjecture implies the precise form. If $\bar{\rho}$ arises from a newform f, then there exists f' of weight $k(\bar{\rho})$ and level $N(\bar{\rho})$ such that $\bar{\rho}$ arises from f'.*

It is a consequence of the proof of Theorem 1.1, which offers new perspectives on this implication, that in Theorem 1.2 the assumption $p \neq 2$ can be removed.

One has another definition of the weight, where the weight of $\bar{\rho}$ is a subset of the finite set of isomorphism classes of irreducible representations of $\mathrm{GL}_2(\mathbb{F}_p)$ with values in linear groups with coefficients in $\overline{\mathbb{F}}_p$. One can determine the weights in this sense too, as done in [19],[1]. The work of F. Diamond and R. Taylor ([12]) and [19] allows one to determine all the possible $k \geq 2$ and $N \geq 1$ such that $\bar{\rho}$ arises from a newform f of weight k and level N.

Theorems 1.1 and 1.2, the fact that $k(\bar{\rho})$ is in a finite range and the finiteness of the dimension of the space of modular forms of given weight and level, imply:

Corollary 1.3. *There are finitely many isomorphism classes of representations of type S with given level $N(\bar{\rho})$.*

Let A be an abelian variety over $\mathbb{Q}$. We say that A is of *(primitive)* GL_2-*type* if A is simple and there is a number field L such that $[L : \mathbb{Q}] = \dim(A)$ and an order O of L with an embedding $O \hookrightarrow \mathrm{End}_{\mathbb{Q}}(A)$ ([28]).

Let $X_1(N)$ the modular curve over $\mathbb{Q}$ defined by $\Gamma_1(N)$ and $J_1(N)$ its jacobian variety. By applying Theorems 1.1 and 1.2 to the Galois representations on points of A of λ-torsion for λ in an infinite set of primes of O and a theorem of Faltings ([16]), one obtains the following theorem, which characterizes the simple quotients of the abelian varieties $J_1(N)$:

Theorem 1.4. *Let A be an abelian variety of GL_2-type. Then A is isogenous to a $\mathbb{Q}$-simple factor of $J_1(N)$ for some N.*

This statement generalizes to compatible systems of Galois representations which are odd 2-dimensional and odd motives of dimension 2.

Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{C})$ be a continuous irreducible complex representation. It has finite image. Its projective image is either dihedral, or isomorphic to one of the permutation groups A_4, S_4 or A_5. Hecke in the dihedral case, Langlands and Tunnell in the A_4 and S_4 case, proved Artin's conjecture for ρ that the L-function of ρ is holomorphic. In fact, Langlands and Tunnell proved that ρ arises from an automorphic representation of $\mathrm{GL}_2(\mathbb{A}_{\mathbb{Q}})$ that is limit of discrete series at infinity ([25] and [44]). The Theorems 1.1 and 1.2 imply:

Theorem 1.5. *Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{C})$ be an odd 2-dimensional complex representation with projective image A_5. Let N be its conductor. Then ρ arises from such an automorphic representation. i.e., there exists an eigenform f of weight 1 for $\Gamma_1(N)$ such that ρ is isomorphic to the Galois representation attached to f by Deligne-Serre.*

Theorem 1.5 had been previously proved in many cases in [39].

J.-M. Fontaine and B. Mazur made the following conjecture that characterizes the p-adic representations $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$ that arise from an f of weight ≥ 2:

Conjecture 1.6. *If ρ is odd, irreducible, unramified outside a finite set of primes and potentially semistable with Hodge-Tate weights (a, b), say $a \leq b$, then the cyclotomic twist $\rho(-a)$ arises from an f of weight $b - a + 1$*

A p-adic Galois representation $\rho : G_F \to \mathrm{GL}_d(E)$, F a number field, E a finite extension of $\mathbb{Q}_p$, which is unramified outside a finite set of primes and potentially semistable at primes above p is called *geometric* by J.-M. Fontaine and B. Mazur. A subquotient of the Galois representation given by the p-adic etale cohomology of a projective and smooth algebraic variety X over F is geometric. It is unramified outside p and the primes of bad reduction of X, and it is potentially semistable by the p-adic comparison theorem ([43]). Conversely J.-M. Fontaine and B. Mazur conjecture that a p-adic geometric irreducible representation of G_F comes from such a subquotient. In particular, the p-adic Galois representation attached to an f of weight ≥ 2 is geometric, as it appears in the cohomology of a fiber product of the universal generalized elliptic curve over a modular curve. The Galois representation attached by P. Deligne and J.-P. Serre to an f of weight 1 has finite image hence is also geometric.

For a potentially semistable representation ρ of $G_{\mathbb{Q}_p}$ in a E-vector space of dimension d , one can define the collection $(i_1, \ldots, i_d)$ of its Hodge-Tate weights which are integers $\in \mathbb{Z}$. The Hodge-Tate weights of $\rho_p(f)$ for f of weight k are $(0, k-1)$. We say that a 2-dimensional potentially semistable representation of $G_{\mathbb{Q}_p}$ is of weight $k \geq 1$ if its Hodge-Tate weights are $(0, k-1)$. The p-adic Hodge theory of Galois representations allows also to attach to ρ a Frobenius-semisimple representation τ_p of the Weil-Deligne group WD_p ([17]): τ_p is defined by the action of WD_p on the filtered Dieudonné module attached ρ.

If ρ is now a geometric p-adic representation of $G_{\mathbb{Q}}$, for each prime ℓ, one can attach to ρ a representation τ_ℓ of the Weil-Deligne group WD_ℓ. For $\ell \neq p$, it is given by a theorem of Grothendieck ([10]), and for $\ell = p$, it is given by Fontaine's theory as recalled above. We call τ_ℓ the Weil-Deligne parameter at ℓ. If ρ arises from an f of weight ≥ 1, for each ℓ, τ_ℓ is isomorphic to the Frobenius-semisimple representation of WD_ℓ attached to f by local Langlands correspondence ([7],[29]). In particular, one can define the conductor $N(\rho)$; ℓ does not divide $N(\rho)$ if and only if $\ell \neq p$ and ρ is unramified, or $\ell = p$ and ρ is crystalline at p.

We call a Modularity Lifting Theorem ("MLT") a statement of the following type: for ρ a p-adic representation of the Galois group of a number field with reduction $\bar{\rho}$, $\bar{\rho}$ modular (or reducible), ρ geometric and odd, implies ρ modular. Wiles, Taylor-Wiles were the first to prove such a theorem ([45],[42], [11]).

When we know the modularity of $\bar{\rho}$, such a theorem implies the modularity of ρ. For example, the following theorem is a consequence of Theorem 1.1 and a "MLT" theorem of M. Kisin ([24]):

Theorem 1.7. *Let $p \neq 2$. Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$ be an odd irreducible representation satisfying the hypotheses of Fontaine-Mazur conjecture 1.6, with distinct Hodge-Tate weights ($a \neq b$). Suppose further that the reduced representation $\bar{\rho}$ satisfies:*

1) the restriction of $\bar{\rho}$ to the Galois group $G_{\mathbb{Q}(\mu_p)}$ is irreducible ;

2) the restriction of $\bar{\rho}$ to $G_{\mathbb{Q}_p}$ is not an extension of $\overline{\mathbb{F}}_p(\eta)$ by $\overline{\mathbb{F}}_p(\eta\omega)$ where ω is the mod. p cyclotomic character.

Then the Fontaine-Mazur conjecture is true for ρ.

M. Emerton, with another approach, has a similar theorem with mildly different hypothesis 2)([15]) . Both approaches use p-adic Langlands correspondence for GL_2.

For statements for $p = 2$, see Dickinson ([13]), M. Kisin ([23]) and [21], [22].

For the case where the restriction of $\bar{\rho}$ to $G_{\mathbb{Q}(\mu_p)}$ is reducible, we have the following theorem of C. M. Skinner and A. Wiles ([35],[34]):

Theorem 1.8. *Let $p > 2$. Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$ be a continuous Galois representation, unramified outside a finite set of primes. Suppose that $\bar{\rho}$ restricted to $\mathbb{Q}(\mu_p)$ is reducible. Suppose futhermore that*

-1) $\bar{\rho}_{D_p} \simeq \begin{pmatrix} \eta_1 & * \\ 0 & \eta_2 \end{pmatrix}$ *with* $(\eta_1)_{|D_p} \neq (\eta_2)_{|D_p}$ *;*

-2) $\rho_{|I_p} \simeq \begin{pmatrix} * & * \\ 0 & 1 \end{pmatrix}$

-3) $\det\rho = \Psi\chi_p^{k-1}$ *for* $k \geq 2$, Ψ *has finite image and* $\det\rho$ *is odd.*

Then, the Fontaine-Mazur conjecture is true for ρ*.*

For partial results for the case $a = b$ in the Fontaine-Mazur conjecture (1.6), see [5] and [6].

Before we sketch the strategy of the proof of Theorem 1.1, we have to introduce two ingredients in the proof: existence of lifts of Galois representations with a control on the ramification of the lift, and existence of compatible systems.

2. Lifts with Conditions of Ramification

We first give conditions on ramification that we impose to the lifts.

2.1. Case $\ell = p$. Serre's weight [32]. We give some more hints about the weight $k(\bar{\rho})$ of $\bar{\rho} : G_{\mathbb{Q}_p} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$.

Let $p \neq 2$. One can prove that $k(\bar{\rho})$ is the minimum of the integers k, $k \geq 2$, such that $\bar{\rho}$ lifts to a crystalline Galois representation $\rho : G_{\mathbb{Q}_p} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$ of weight k. It only depends on the restriction of $\bar{\rho}$ to the inertia subgroup $I_p \subset G_{\mathbb{Q}_p}$. Let $\omega : I_p \to \mathbb{F}_p^*$ be the cyclotomic character giving the action of I_p on the p-roots of unity. Let $\omega_2 : I_p \to (\mathbb{F}_{p^2})^*$ be the Kummer character for the extension $\mathbb{Q}_{p^2}(p^{1/(p^2-1)})$ of the unramified extension $\mathbb{Q}_{p^2}$ of $\mathbb{Q}_p$ of degree 2. By Fontaine-Laffaille and Berger-Li-Zhu, we have the following description of the $\bar{\rho}$ such that $2 \leq k(\bar{\rho}) \leq p+1$:

- If $\bar{\rho}$ is reducible, then the restriction of $\bar{\rho}$ to I_p has semisimplification $\omega^a \oplus \omega^b$ with a and b integers in $[1, p-1]$. One has $2 \leq k(\bar{\rho}) \leq p+1$ if and only one of the characters ω^a and ω^b is trivial, say ω^a, and $\bar{\rho}$ has an unramified quotient. Then $k(\bar{\rho}) = b+1 \in [2, p]$, unless $b = 1$ and we are in the case "très ramifié" where $k(\bar{\rho}) = p+1$.

- If $\bar{\rho}$ is irreducible, then the restriction of $\bar{\rho}$ to I_p has semisimplification $\omega_2^{a+pb} \oplus \omega_2^{b+pa}$ with $0 \leq a < b < p$. One has $2 \leq k(\bar{\rho}) \leq p+1$ if and only if $a = 0$ in which case one has $k(\bar{\rho}) = b+1$.

We consider $\bar{\rho}$ with $2 \leq k(\bar{\rho}) \leq p+1$. The first type of condition is that ρ is crystalline of weight $k(\bar{\rho})$.

It is also important for us to consider lifts ρ of $\bar{\rho}$ which are potentially semistable of weight 2 (Hodge-Tate weights $(0, 1)$). Recall that one can associate to a potentially semistable p-adic representation ρ a representation $\tau(\rho)$ of the Weil-Deligne group of $\mathbb{Q}_p$. One considers ρ with $\tau(\rho)$ having conductor 1 (ρ is

crystalline) or p. More precisely, still for $2 \leq k(\bar{\rho}) \leq p+1$, we consider the following condition:

- If $k(\bar{\rho}) \neq p+1$, ρ is of weight 2, becomes crystalline after restriction to $\mathbb{Q}_p(\mu_p)$ and the restriction of the Weil-Deligne parameter $\tau(\rho)$ to I_p is $\omega^{k(\bar{\rho})-2} \oplus$ id.

- If $k(\bar{\rho}) = p+1$, ρ is semistable non-crystalline of weight 2.

If $p = 2$, and $\bar{\rho} : G_{\mathbb{Q}_p} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ is either reducible and we are not in the case "très ramifié", or it is irreducible, then $k(\bar{\rho}) = 2$ and we impose to ρ to be a crystalline representation of weight 2. When $\bar{\rho}$ is reducible and "très ramifié", $k(\bar{\rho}) = 4$. In this case, we impose to ρ to be semistable, non-crystalline of weight 2.

2.2. Case $\ell \neq p$ ([22]). Let $I_\ell \subset D_\ell := G_{\mathbb{Q}_\ell}$ be the ramification subgroup. Let $\bar{\rho} : G_{\mathbb{Q}_\ell} \to \mathrm{GL}_2(\overline{\mathbb{F}_p})$ be a mod. p representation. We suppose given a lift $\rho_0 : G_{\mathbb{Q}_\ell} \to \mathrm{GL}_2(\overline{\mathbb{Z}_p})$. We consider the lifts $\rho : G_{\mathbb{Q}_\ell} \to \mathrm{GL}_2(\overline{\mathbb{Z}_p})$ such that the restrictions of ρ and ρ_0 to I_ℓ are conjugate by an element of $\mathrm{GL}_2(\overline{\mathbb{Q}}_p)$.

We can impose on ρ_0 the condition called "minimal" that in particular implies that ρ and $\bar{\rho}$ have the same conductor.

2.3. Statement. Let $\bar{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be of Serre's type. If $p \neq 2$, we suppose that $2 \leq k(\bar{\rho}) \leq p+1$. Let S be a finite set of places of $\mathbb{Q}$ that contains ∞, p and the primes of ramification of $\bar{\rho}$. We consider the problem of lifting $\bar{\rho}$ to a representation $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$ which satisfies the following ramification conditions for $v \in S$:

- $v = \infty$: ρ is odd ($\det(\bar{\rho}(c)) = -1$ for c a complex conjugation). It is a condition only if $p = 2$.

- $v = p$: (2.1). If $p \neq 2$ ρ is either crystalline of weight $k(\bar{\rho})$ or it is of weight 2, potentially crystalline or semistable. If $p = 2$, ρ has to be of weight 2 and crystalline if $k(\bar{\rho}) = 2$, semistable if $k(\bar{\rho}) = 4$.

– for other places v, we fix the ramification up to conjugacy by giving a lift of $\bar{\rho}|_{D_v}$ as in (2.2).

As the maximal abelian extension of $\mathbb{Q}$ has Galois group that is the direct product of its inertia subgroups for all prime numbers, the conditions on inertia for all primes fix the determinant. For $p = 2$, we have to impose that the conditions on ramification for all primes give an odd determinant.

Theorem 2.1. *(Th. 5.1. of [21]). We assume that $\bar{\rho}$ has non-solvable image when $p = 2$, and that $\bar{\rho}|_{\mathbb{Q}(\mu_p)}$ is absolutely irreducible when $p > 2$. We fix conditions on ramification for $v \in S$. Then $\bar{\rho}$ has a lift ρ which satisfies the ramification conditions.*

R. Ramakrishna was the first to construct geometric lifts under general hypotheses, but had to allow ramification at auxiliary primes ([27]).

We give some hints on the proof, which relies on the work of G. Böckle ([3]) and the potential modularity theorem of R. Taylor ([37],[38]). Let ϕ be the fixed

determinant. Let $\mathcal{O}$ be the ring of integers of a sufficiently big extension E of $\mathbb{Z}_p$ (in particular, we ask that the residue field k of $\mathcal{O}$ is such that $\bar{\rho}$ is conjugate to a representation with coefficients in k). For $v \in S$, let $\bar{\rho}_v$ be the restriction of $\bar{\rho}$ to the decomposition group D_v. For each $v \in S$, one has a complete noetherian local $\mathcal{O}$-algebra $\bar{R}_v^{\square}$ with residue field k, and a Galois representation $D_v \to \mathrm{GL}_2(\bar{R}_v^{\square})$ that satisfies in particular the following condition. Let F' is a finite extension of F with ring of integers $\mathcal{O}'$. Then, the natural map from the set of local $\mathcal{O}$-algebra morphisms from $\bar{R}_v^{\square}$ to $\mathcal{O}'$ to the set of isomorphism classes of lifts $\rho_v : D_v \to \mathrm{GL}_2(\mathcal{O}')$ of $\bar{\rho}_v$ such that $D_v \to \mathrm{GL}_2(F')$ satisfy the ramification condition, is a bijection. The rings $\bar{R}_v^{\square}$ are flat over $\mathcal{O}$, $\bar{R}_v^{\square}[1/p]$ is regular, and is equidimensional of relative dimension 2 if $v = \infty$, 4 if $v = p$ and 3 if $v = \ell \neq p$. We call $\bar{R}_S^{\square,\mathrm{loc}}$ the completed tensor product over $\mathcal{O}$ of the $\bar{R}_v^{\square}$.

We have a complete noetherian local $\mathcal{O}$-algebra $\bar{R}_S$ (resp. $\bar{R}_S^{\square}$) with residue field k such that the local $\mathcal{O}$-algebra morphisms from $\bar{R}_S$ (resp. $\bar{R}_S^{\square}$) to $\mathcal{O}'$ are the data of a lift ρ of $\bar{\rho}$ that satisfies the ramification conditions up to conjugacy (resp. and for each $v \in S$ a basis of the underlying space of ρ). The $\mathcal{O}$-algebra $\bar{R}_S^{\square}$ has a natural structure of $\bar{R}_S^{\square,\mathrm{loc}}$-algebra. The deformation theory links the number of generators and relations of $\bar{R}_S^{\square}$ above $\bar{R}_S^{\square,\mathrm{loc}}$ to Galois cohomolgy. The formula of Wiles (Th. 8.6.20 of [26]), taking into account the dimension of the local deformation rings, gives that the dimension of $\bar{R}_S$ is ≥ 1.

Then, it suffices to prove that $\bar{R}_S$ is a finitely generated $\mathcal{O}$-module. Indeed, as it is of dimension ≥ 1, there will exist $\mathcal{O}'$ and a $\mathcal{O}$-algebra morphism from $\bar{R}_S$ to $\mathcal{O}'$ giving the searched lift.

R. Taylor proves that there exist a finite extension F of $\mathbb{Q}$, totally real, such that the restriction of $\bar{\rho}$ to G_F is modular *i.e.* is isomorphic to the Galois representation associated by him in [40] and [41] to an Hilbert eigenform form for F (or a suitable automorphic representation π_F of $\mathrm{GL}_2(\mathbb{A}_F)$). One defines the deformation problem for $\bar{\rho}_{|G_F}$ with conditions of ramification as the restriction to G_F of the conditions of ramification for $G_{\mathbb{Q}}$. One gets a ring $\bar{R}_{S,F}$. An "MLT" theorem (almost) identifies $\bar{R}_{S,F}$ to the completion of an Hecke-algebra. It follows that $\bar{R}_{S,F}$ is finite over $\mathcal{O}$. An easy lemma of algebra implies that $\bar{R}_S$ is finite over $\mathcal{O}$.

3. Existence of Compatible Systems

Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_d(\overline{\mathbb{Q}}_p)$ be a geometric Galois representation. We saw that for every prime ℓ, we get a F-semisimple representation τ_ℓ of the Weil-Deligne group $\mathrm{WD}_\ell \to \mathrm{GL}_d(\overline{\mathbb{Q}}_p)$.

The following theorem is essentially due to L. Dieulefait ([14], [46]):

Theorem 3.1. *Suppose that ρ and $\bar{\rho}$ are as in the theorem 2.1. Then, there exists a finite extension $E \subset \overline{\mathbb{Q}}$ of $\mathbb{Q}$ and for every prime ℓ and every embedding*

$\iota_\ell : \overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_\ell$ an ℓ-adic Galois representation $\rho_{\iota_\ell} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{Q}}_\ell)$ such that:

1) *there is one ι_p such that $\rho \simeq \rho_{\iota_p}$;*
2) *for every prime ℓ, the representation τ_ℓ of WD_ℓ is rational over E, meaning that it is isomorphic via ι_p to a representation which is defined over E;*
2) *For each prime q and every ι_ℓ, the representation of the Weil-Deligne WD_q defined by ρ_{ι_ℓ} is isomorphic to the one defined by ρ_{ι_p}.*

The proof of the theorem uses ideas of R. Taylor. By his potential modularity theorem, there exists a finite totally real Galois extension F of $\mathbb{Q}$ such that the restriction of $\bar{\rho}$ to G_F is the Galois representation attached by him to an automorphic representation π_F for $\mathrm{GL}_2(\mathbb{A}_F)$ and an embedding ι_p of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$. The field F can be choosen satisfying appropriate properties, in particular, that the restriction of $\bar{\rho}$ to $G_{F(\mu_p)}$ is irreducible. An "MLT" theorem and solvable base change theorem ([25]) then imply that there exists $E \subset \overline{\mathbb{Q}}$ a finite extension of $\mathbb{Q}$ and, for every $F' \subset F$ with $\mathrm{Gal}(F/F')$ solvable, an automorphic representation $\pi_{F'}$ such that:

- for every prime q of F' the representation of the local Weil-Deligne groups of F' at q is rational over $E \subset \overline{\mathbb{Q}}$ considered as a subfield of $\overline{\mathbb{Q}}_p$ via ι_p ;

- there is an embedding ι_p of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$ such that $\rho_{|G_{F'}}$ is isomorphic to the Galois representation attached to $\pi_{F'}$ and ι_p.

Using Brauer theorem, one proves that there exists $n_{F'} \in \mathbb{Z}$ and characters $\chi_{F'}$ of $\mathrm{Gal}(F/F')$, for F' as above, such that

$$\rho = \sum_{F'} n_{F'} \mathrm{Ind}_{\mathrm{G}_{\mathrm{F}'}}^{\mathrm{G}_{\mathbb{Q}}} (\chi_{\mathrm{F}'} \otimes \rho_{\pi_{\mathrm{F}'}, \iota_{\mathrm{p}}}).$$

For ι_ℓ, one defines ρ_{ι_ℓ}, as a virtual representation, by the displayed formula, replacing ι_p by ι_ℓ:

$$\rho_{\iota_\ell} = \sum_{F'} n_{F'} \mathrm{Ind}_{\mathrm{G}_{\mathrm{F}'}}^{\mathrm{G}_{\mathbb{Q}}} (\chi_{\mathrm{F}'} \otimes \rho_{\pi_{\mathrm{F}'}, \iota_\ell}).$$

One checks that ρ_{ι_ℓ} is a true representation, as it has, in the Grothendieck group of p-adic representations of $G_{\mathbb{Q}}$, the same norm 1 and dimension 2 as ρ. The compatibility at all places follows from [7], [29],[30],[33].

4. The Strategy of the Proof

We prove Theorem 1.1 by an inductive argument on level and the prime p. We give the starting points. First:

Theorem 4.1. *(Tate-Serre). Let $\bar{\rho}$ be a continuous representation of $G_{\mathbb{Q}}$ in $\mathrm{GL}_2(\overline{\mathbb{F}}_2)$ (resp. $\mathrm{GL}_2(\overline{\mathbb{F}}_3)$) which is unramified outside 2 (resp. 3). Then $\bar{\rho}$ is reducible.*

This theorem has been proved by Tate for $p = 2$ ([36]), and by Serre for $p = 3$, using the Odlyzko bounds.

We also need the following generalisation of the theorem of Fontaine and Abrashkin that there are no non-zero abelian varieties over $\mathrm{Spec}(\mathbb{Z})$:

Theorem 4.2. *(Brumer-Kramer [4], R. Schoof [31]) Let $q = 2, 3, 5, 7$ or 13. There is no non-trivial abelian variety over $\mathbb{Q}$ which has good reduction for all $\ell \neq q$ and is semistable at q.*

Theorem 1.8 of Skinner-Wiles is also one of the starting points of the induction.

We explain the strategy. Let $\bar{\rho}$ be of S-type. We want to prove that it is modular. If $p \neq 2$, by a twist, we may impose that $2 \leq k(\bar{\rho}) \leq p+1$. We can lift $\bar{\rho}$ to a geometric ρ by Theorem 2.1, then make ρ part of a compatible system (ρ_ι) by Theorem 3.1. We choose a prime $\ell \neq p$ and consider the reduction $\bar{\rho}_\ell$ of an ℓ-dic member ρ_ℓ of the compatible system (ρ_ι). The conditions on ramification of ρ and the choice of ℓ make that either $\bar{\rho}_\ell$ becomes reducible after restriction to $\mathbb{Q}(\mu_\ell)$ and ρ_ℓ satisfies the conditions of Theorem 1.8, or $\bar{\rho}_\ell$ restricted to $\mathbb{Q}(\mu_\ell)$ is irreducible, and it has a cyclotomic twist of conductor and weight in a range where we already know Serre's conjecture. Then, a "MLT" theorem allows us to know that ρ_ℓ is modular, hence every member of the compatible system (ρ_ι) is modular, hence $\bar{\rho}$ is modular. Let us give some examples.

Let us prove Serre's conjecture for $\bar{\rho}$ of level 1 and weight 2. By Theorem 2.1, we can lift $\bar{\rho}$ to a ρ that is crystalline of weight 2 and unramified outside p. Then, we can extend ρ to a compatible system (ρ_ι) of odd geometric Galois representations such that ρ_{ι_ℓ} is unramified outside ℓ and is crystalline of weight 2. We consider $\bar{\rho}_3$, the reduction modulo 3 of a 3-adic representation member of the compatible system. By Serre's theorem, $\bar{\rho}_3$ is reducible. The theory of finite flat group schemes implies that the ρ_3 and $\bar{\rho}_3$ satisfy the hypotheses of Theorem 1.8. It follows that ρ_3 is modular, and thus $\bar{\rho}$. Hence in fact $\bar{\rho}$ does not exist.

Let us suppose that $\bar{\rho}$ is of weight 2 and of conductor $q = 2, 3, 5, 7, 13$ prime to $p > 2$ and semistable at q ($\bar{\rho}(I_q)$ is of order p). By Theorem 2.1, we get as above a lift of $\bar{\rho}$ and a compatible system (ρ_ι) of geometric representations of weight 2 with Weil-Deligne parameter at q of conductor q and semistable. It follows from the potential modularity theorem of R. Taylor and a theorem of G. Faltings ([16]) that the compatible system arises from an abelian variety over $\mathbb{Q}$ which has good reduction outside q and has semistable reduction at q. Such an abelian variety is trivial by Theorem 4.2, hence $\bar{\rho}$ does not exist proving Serre's conjecture in this case.

This also yields the proof of Serre's conjecture when $N(\bar{\rho}) = 1, k(\bar{\rho}) = q+1$ ($q = 2, 3, 5, 7, 13$): by Theorem 2.1, we get a lift of $\bar{\rho}$ and a compatible system (ρ'_ι) of geometric representations of weight $q+1$ and conductor 1. A residual representation arising from it at q, say $\bar{\rho}_q$, has Serre weight either 2 or $q+1$. Then we are done using "MLT" result, either if $\bar{\rho}_q$ is reducible, or otherwise by

lifting $\bar{\rho}_q$ to a compatible system (ρ_ι) of geometric representations of weight 2 with Weil-Deligne parameter at q of conductor q and semistable as above. This again yields by the earlier argument that $\bar{\rho}_q$ is reducible. This *inter alia* proves the level one case of Serre's conjecture for $p \leq 5$.

In the proof of the general level 1 case *i.e.* $\bar{\rho}$ unramified outside p, the induction on the prime p uses lifts that are of weight 2 and prime conductor (not necessarily semistable), and crystalline of weight $k(\bar{\rho})$ and conductor 1. Let $p > 5$ and let $S(k,p)$ be the statement for an integer, $2 \leq k \leq p+1$, that Serre's conjecture is true for $\bar{\rho}$ of characteristic p and weight k. By an argument as above, one first proves that $S(k,p)$ is independent of $p \geq k-1$. Then one lifts $\bar{\rho}$ to a ρ which is of weight 2 and level p, part of a compatible system (ρ_ι). One considers $\bar{\rho}_\ell$ the reduction of a member of the compatible system (ρ_ι) of characteristic ℓ. By the choice of ℓ, one is able to find a lift ρ'_ℓ of $\bar{\rho}_\ell$ and a compatible system $(\rho'_{\iota'})$ whose WD_p parameter at p is such that $\bar{\rho}'_{\iota_p}$ is up to twist in a range where we already know $S(k,p)$.

For the general case, we use an inductive argument on the number of primes that divide $N(\bar{\rho})$. We get a compatible system (ρ_ι) that is of level $N(\bar{\rho})$. We consider a divisor q of $N(\bar{\rho})$ and $\bar{\rho}_q$. By the definition of conductor, the conductor of $\bar{\rho}_q$ is prime to q, hence divides the prime to q part of $N(\rho)$. This observation allows us to do the induction on the number of primes that divide the conductor. We are able, when we make this reduction on the level, to avoid $\bar{\rho}_\ell$ which are reducible over $\mathbb{Q}(\mu_\ell)$ by adding some ramification at an auxiliary prime.

It is plausible that future progress on modularity lifting theorems will allow some simplifications in the strategy of the proof.

References

[1] Kevin Buzzard, Fred Diamond, Frazer Jarvis. On Serre's conjecture for mod. ℓ Galois representations over totally real fields. Preprint.

[2] Laurent Berger, Hanfeng Li, and Hui June Zhu. Construction of some families of 2-dimensional crystalline representations. *Math. Ann.*, 329(2):365–377, 2004.

[3] Gebhard Böckle. Presentations of universal deformation rings, L-functions and Galois representations, London Math. Soc. Lecture Note Ser.,320, 24–58, Cambridge Univ. Press, 2007.

[4] Armand Brumer and Kenneth Kramer. Non-existence of certain semistable abelian varieties, Manuscripta Math., 106, 2001, 3, p. 291–304.

[5] Kevin Buzzard and Richard Taylor Companion forms and weight one forms. Ann. of Math. (2), 149, 1999, 3, p. 905–919

[6] Kevin Buzzard. Analytic continuation of overconvergent eigenforms. J. Amer. Math. Soc., 2003, p. 29–55.

[7] Henri Carayol. Sur les représentations l-adiques associées aux formes modulaires de Hilbert. *Ann. Sci. École Norm. Sup. (4)*, 19(3), 409–468, 1986.

[8] Pierre Deligne et Jean-Pierre Serre. Formes modulaires de poids 1. Ann. Sci. École Norm. Sup. (4), 7, 1974.

[9] Pierre Deligne. Formes modulaires et représentations l-adiques. Séminaire Bourbaki, (1968-1969), Expos No. 355, 34 p.

[10] Pierre Deligne. Les constantes des équations fonctionnelles des fonctions L, Modular functions of one variable, II (Proc. Internat. Summer School, Univ. Antwerp, Antwerp, 1972), p. 501–597. Lecture Notes in Math., Vol. 349, Springer, Berlin, 1973.

[11] Henri Darmon, Fred Diamond, and Richard Taylor. Fermat's last theorem. In *Current developments in mathematics, 1995 (Cambridge, MA)*, pages 1–154. Internat. Press, Cambridge, MA, 1994.

[12] Fred Diamond and Richard Taylor. Lifting modular mod l representations. Duke Math. J., 74, 1994, 2, p. 253–269.

[13] Mark Dickinson. On the modularity of certain 2-adic Galois representations. Duke Math. J. 109 (2001), no. 2, 319–382.

[14] Luis Dieulefait. Existence of families of Galois representations and new cases of the Fontaine-Mazur conjecture. J. Reine Angew. Math. 577 (2004), 147–151.

[15] Matthew Emerton. Local-global compatibility in the p-adic Langlands programme for $\mathrm{GL}_{2,\mathbb{Q}}$. In preparation.

[16] Gerd Faltings. Endlichkeitssätze für abelsche Varietäten über Zahlkörpern. Invent. Math., 73 , 1983, 3, p. 349–366.

[17] Jean-Marc Fontaine. Représentations l-adiques potentiellement semi-stables. Périodes p-adiques (Bures-sur-Yvette, 1988), Astérisque,223,1994, p. 321–347.

[18] Chandrashekhar Khare. Serre's modularity conjecture: the level one case. Duke Math. J. 134 (3) (2006), 534–567.

[19] Chandrashekhar Khare A local analysis of congruences in the (p, p) case (II). Invent. Math.,143,2001,1, p. 129–155.

[20] Chandrashekhar Khare and Jean-Pierre Wintenberger. On Serre's conjecture for 2-dimensional mod p representations of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$. Annals of Math. 169 (1) (2009), 229–253.

[21] Chandrashekhar Khare and Jean-Pierre Wintenberger. Serre's modularity conjecture (I). Invent. Math.,178,2009,3, p. 485–504.

[22] Chandrashekhar Khare and Jean-Pierre Wintenberger. Serre's modularity conjecture (II). Invent. Math.,178,2009,3, p. 505–586.

[23] Mark Kisin. Modularity of 2-adic Barsotti-Tate representations. Invent. Math.,178,2009,3, p. 587–634.

[24] Mark Kisin The Fontaine-Mazur conjecture for GL_2, J. Amer. Math. Soc.,22,2009,3,p. 641–690.

[25] Robert Langlands. Base change for GL_2. Annals of Math. Series, Princeton University Press, 1980.

[26] Jürgen Neukirch, Alexander Schmidt and Kay Wingberg. Cohomology of number fields. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 323, Springer-Verlag, Berlin, 2000.

[27] Ravi Ramakrishna. Deforming Galois representations and the conjectures of Serre and Fontaine-Mazur. Ann. of Math. (2) 156 (2002), no. 1, 115–154.

[28] Kenneth A. Ribet, Abelian varieties over $\mathbb{Q}$ and modular forms, Modular curves and abelian varieties, Progr. Math. 224, p. 241–261, Birkhäuser, Basel, 2004.

[29] Takeshi Saito. Modular forms and p-adic Hodge theory. Invent. Math.,129,1997,3, p. 607–620.

[30] Takeshi Saito. Hilbert modular forms and p-adic Hodge theory. Compos. Math., 145, 2009, 5, p. 1081–1113.

[31] René Schoof. Abelian varieties over $\mathbb{Q}$ with bad reduction in one prime only. Compos. Math., 141, 2005, 4, p. 847–868.

[32] Jean-Pierre Serre. Sur les représentations modulaires de degré 2 de $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. Duke Math. J., 54(1):179–230, 1987.

[33] Christopher Skinner. A note on the p-adic Galois representations attached to Hilbert modular forms. Doc. Math.,14,2009, p. 241–258.

[34] Christopher Skinner and Andrew Wiles. Nearly ordinary deformations of irreducible residual representations. Ann. Fac. Sci. Toulouse Math. (6), 10(1):185–215, 2001.

[35] Christopher Skinner and Andrew Wiles. Residually reducible representations and modular forms. Inst. des Hautes Études Scientifiques, Publications Mathématiques, 89, 1999, p. 5–126.

[36] John Tate. The non-existence of certain Galois extensions of $\mathbf{Q}$ unramified outside 2. Arithmetic geometry (Tempe, AZ, 1993), Contemp. Math, 174, p. 153–156, Amer. Math. Soc. Providence, RI, 1994.

[37] Richard Taylor. Remarks on a conjecture of Fontaine and Mazur. Inst. Math. Jussieu, 1(1):125–143, 2002.

[38] Richard Taylor. On the meromorphic continuation of degree two L-functions. Documenta Math. Extra Volume: John H. Coates' Sixtieth Birthday (2006) 729–779.

[39] Richard Taylor. On icosahedral Artin representations. II. Amer. J. Math., 125(3):549–566, 2003.

[40] Richard Taylor. On Galois representations associated to Hilbert modular forms. Invent. Math., 98(2):265–280, 1989.

[41] Richard Taylor. On Galois representations associated to Hilbert modular forms II. Current developments in mathematics, 1995 (Cambridge, MA), pages 333–340. Internat. Press, Cambridge, MA, 1994.

[42] Richard Taylor and Andrew Wiles. Ring-theoretic properties of certain Hecke algebras. Ann. of Math. (2), 141(3):553–572, 1995.

[43] Takeshi Tsuji. p-adic Hodge theory in the semi-stable reduction case. Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998), Doc. Math.,1998, Extra Vol. II, p. 207–216 (electronic),

[44] Jerrold Tunnell. Artin's conjecture for representations of octahedral type. Bull. Amer. Math. Soc. (N.S.) 5 (1981), no. 2, 173–175.

[45] Andrew Wiles. Modular elliptic curves and Fermat's last theorem. Ann. of Math. (2), 141(3):443–551, 1995.

[46] Jean-Pierre Wintenberger. On p-adic geometric representations of $G_{\mathbb{Q}}$. Documenta Math. Extra Volume: John H. Coates' Sixtieth Birthday (2006) 819–827.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

The Structure of Potentially Semi-stable Deformation Rings

Mark Kisin*

Abstract

Inside the universal deformation space of a local Galois representation one has the set of deformations which are potentially semi-stable of given p-adic Hodge and Galois type. It turns out these points cut out a closed subspace of the deformation space. A deep conjecture due to Breuil-Mézard predicts that part of the structure of this space can be described in terms of the local Langlands correspondence. For 2-dimensional representations the conjecture can be made precise. We explain some of the progress in this case, which reveals that the conjecture is intimately connected to the p-adic local Langlands correspondence, as well as to the Fontaine-Mazur conjecture.

Mathematics Subject Classification (2010). 11F80

Keywords. Galois representations

Introduction

The study of deformations of Galois representations was initiated by Mazur [Ma]. Already in that article Mazur considered deformations satisfying certain local conditions formulated in terms of p-adic Hodge theory. The importance of deformations satisfying such conditions became clear with the formulation of the Fontaine-Mazur conjecture [FM], and the spectacular proof of the Shimura-Taniyama conjecture on modularity of elliptic curves over $\mathbb{Q}$ by Wiles, Taylor-Wiles, and their collaborators [Wi], [TW], [BCDT].

The first question which arises concerns the nature of the subspaces cut out by these conditions: Suppose that $K/\mathbb{Q}_p$ is a finite extension with absolute Galois group G_K, let $\mathbb{F}/\mathbb{F}_p$ be a finite extension, and $V_{\mathbb{F}}$ a finite dimensional $\mathbb{F}$-vector space equipped with a continuous, absolutely irreducible

*It is a pleasure to thank Christophe Breuil, Kevin Buzzard, Fred Diamond, Toby Gee, James Newton, Vytautas Paskunas, David Savitt and Wansu Kim for useful comments on this paper. The author was partially supported by NSF grant DMS-0701123.

Department of Mathematics, Harvard, 1 Oxford st, Cambridge MA 02139, USA.
E-mail: kisin@math.harvard.edu.

G_K-action. Then $V_{\mathbb{F}}$ admits a universal deformation ring $R_{V_{\mathbb{F}}}$. A closed point $x \in \operatorname{Spec} R_{V_{\mathbb{F}}}[1/p]$ gives rise to a deformation L_x of $V_{\mathbb{F}}$, so that $L_x \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ is a representation of G_K on a finite dimensional vector space over a finite extension of $W(\mathbb{F})[1/p]$. One can ask whether the points such that $L_x \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ satisfies the condition are cut out by a closed subspace of $\operatorname{Spec} R_{V_{\mathbb{F}}}[1/p]$.

Of course the answer depends on the condition one imposes. In [Fo 2] Fontaine suggests (at least implicitly) that the answer should be affirmative if one requires the representations to become semi-stable over a fixed extension K'/K and with Hodge-Tate weights in a fixed interval. Attached to any such representation V is a finite dimensional representation of the inertia subgroup $I_K \subset G_K$, which, in some sense, measures the failure of V to be semi-stable. One can sharpen Fontaine's conjecture by fixing a representation τ of I_K, with open kernel, and requiring $L_x \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ to have fixed Hodge-Tate weights and associated I_K-representation τ. That this refined condition cuts out a closed subspace was conjectured in special cases in the papers of Fontaine-Mazur [FM, p191], Breuil-Conrad-Diamond-Taylor [BCDT, Conj. 1.1.1], and suggested more generally by Breuil-Mézard [BM, Conj. 1.1, p214].

After partial results by several people (see section 1.2.5 below for a more detailed discussion) such a result was proved in general in [Ki 4]. Thus, for some finite normal extension $\mathcal{O}$ of $W(\mathbb{F})$ one obtains a quotient $R_{V_{\mathbb{F}}}^{\mathbf{v},\tau}$ of $R_{V_{\mathbb{F}}} \otimes_{W(\mathbb{F})} \mathcal{O}$ whose points in characteristic 0 correspond precisely to deformations of $V_{\mathbb{F}}$ which become semi-stable over some finite extension of K, have the chosen fixed Hodge-Tate weights and associated I_K-representation τ. [1]

The conjectures of Breuil-Mézard predict a deep connection between the structure of $R_{V_{\mathbb{F}}}^{\mathbf{v},\tau}$ and the representation theory of $\mathrm{GL}_d(\mathcal{O}_K)$, where $d = \dim_{\mathbb{F}} V_{\mathbb{F}}$. [2] This can be made precise when $V_{\mathbb{F}}$ is two dimensional, which we assume for the rest of this introduction. In this case, a result of Henniart attaches to τ a smooth, irreducible, finite dimensional representation $\sigma(\tau)$ of $\mathrm{GL}_2(\mathcal{O}_K)$ which is characterized in terms of the local Langlands correspondence. On the other hand, the cocharacter $\mathbf{v}$ gives rise to an algebraic representation $\sigma(\mathbf{v})$ of $\mathrm{GL}_2(\mathcal{O}_K)$. Let $L_{\mathbf{v},\tau} \subset \sigma(\mathbf{v}) \otimes \sigma(\tau)$ be a $\mathrm{GL}_2(\mathcal{O}_K)$ invariant lattice. Then the conjecture predicts the Hilbert-Samuel multiplicity $e(R_{V_{\mathbb{F}}}^{\mathbf{v},\tau}/\pi)$ of $R_{V_{\mathbb{F}}}^{\mathbf{v},\tau}/\pi$ in terms of the multiplicities of the Jordan-Hölder factors of $L_{\mathbf{v},\tau}/\pi L_{\mathbf{v},\tau}$. Here $\pi \in \mathcal{O}$ denotes a uniformizer. Indeed, one can formulate such a conjecture in any dimension assuming an analogue of Henniart's result. When τ is irreducible a higher dimensional analogue of Henniart's result has been proved by Paskunas [Pa], building on the work of Bushnell-Kutzko [BK].

[1] Here the symbol $\mathbf{v}$ indicates a conjugacy class of cocharacters corresponding to the choice of Hodge-Tate weights; we refer to section 1.1.3 below for the precise definition. The choice of $\mathcal{O}$ is related to the field of definition of $\mathbf{v}$ and τ.

[2] Strictly speaking [BM] makes this conjecture in detail for two dimensional representations, $K = \mathbb{Q}_p$ and small Hodge-Tate weights. However, the possibility of this connection holding more generally is suggested on p214 of *loc. cit.*

It is slightly more convenient to work with the quotient $R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}$ of $R^{\mathbf{v},\tau}_{V_{\mathbb{F}}}$ which corresponds to deformations having determinant ψ times the cyclotomic character, for some appropriately chosen [3] ψ. The general shape of such a conjecture is then that

$$e(R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}/\pi) = \sum_{\bar{\sigma}} a(\bar{\sigma})\mu_{\bar{\sigma}}(V_{\mathbb{F}}),$$

where $\bar{\sigma}$ runs over irreducible mod p representations of $\mathrm{GL}_2(k)$, k the residue field of K, $a(\bar{\sigma})$ denotes the multiplicity of $\bar{\sigma}$ as a Jordan-Hölder factor of $L_{\mathbf{v},\tau}/\pi L_{\mathbf{v},\tau}$, and $\mu_{\bar{\sigma}}(V_{\mathbb{F}})$ is a non-negative integer. This equality can be viewed as a system of infinitely many equations (corresponding to the choices of $\mathbf{v}$ and τ) in the finitely many unknowns $\mu_{\bar{\sigma}}(V_{\mathbb{F}})$. One can of course also ask for a version of such a conjecture where the $\mu_{\bar{\sigma}}(V_{\mathbb{F}})$ are given explicitly, as is done in [BM] when $K = \mathbb{Q}_p$.

For two dimensional representations and $K = \mathbb{Q}_p$ most of the Breuil-Mézard conjecture is proved in [Ki 5]. The proof consists of two parts: One uses the p-adic local Langlands correspondence of Breuil and Colmez [Br 1], [Co] to show that $e(R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}/\pi)$ is bounded above by the expected value. A modified form of the Taylor-Wiles patching argument, introduced in [Ki 1], is then used to prove the other inequality. To do this one uses $L_{\mathbf{v},\tau}$-valued automorphic forms on a totally definite quaternion algebra to construct a module M_∞ which is finite of rank ≤ 1 over a formally smooth $R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}$-algebra R_∞. Then

$$e(R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}/\pi) = e(R_\infty/\pi) \geq e(M_\infty/\pi M_\infty)$$

where the final quantity denotes the Hilbert-Samuel multiplicity of the R_∞/π-module $M_\infty/\pi M_\infty$. This multiplicity can in turn be analyzed in terms of the Jordan-Hölder factors of $L_{\mathbf{v},\tau}/\pi L_{\mathbf{v},\tau}$.

The restriction $K = \mathbb{Q}_p$ is used primarily so as to be able to apply the p-adic local Langlands correspondence, which is available for $\mathrm{GL}_2(\mathbb{Q}_p)$ but remains somewhat elusive for $\mathrm{GL}_2(K)$ with $K \neq \mathbb{Q}_p$. Indeed the Breuil-Mézard conjecture may be viewed as an avatar of that correspondence. On the other hand, the modified Taylor-Wiles method can be applied without restrictions on K. It always gives an inequality involving $e(R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}/\pi)$ with equality being essentially equivalent to a modularity lifting theorem for representations which are of type $(\mathbf{v},\tau)$ at primes dividing p. Such lifting theorems are predicted by the Fontaine-Mazur conjecture and generalize the results used to prove the Shimura-Taniyama conjecture. They were the main motivation of [Ki 5].

In particular, one can try to *use* modularity lifting theorems to prove cases of the Breuil-Mézard conjecture for $K \neq \mathbb{Q}_p$. We give an example of such a

[3] In order that the quotient is non-zero, one needs a condition of compatibility between ψ and $(\mathbf{v},\tau)$ (see section 2.2 below) which we assume from now on.

result in §3, using the modularity lifting theorems for potentially Barsotti-Tate representations proved in [Ki 1] and [Ge 1]. The coefficients $\mu_{\bar{\sigma}}(V_{\mathbb{F}})$ are not made explicit in this case. One can hope to do that when $K/\mathbb{Q}_p$ is unramified, assuming the Buzzard-Diamond-Jarvis conjecture [BDJ] on the weights of automorphic forms giving rise to a given 2-dimensional mod p representation. Most of this has been proved by Gee [Ge 2], but one really needs the whole conjecture to determine all the coefficients. Nevertheless, we explain how to use Gee's result to prove the expected lower bound for $e(R_{V_{\mathbb{F}}}^{\mathbf{v},\tau,\psi}/\pi)$ when $V_{\mathbb{F}}$ is absolutely irreducible and satisfies a mild additional restriction.

The paper is organized as follows: In §1 we recall the definition of the rings $R_{V_{\mathbb{F}}}^{\mathbf{v},\tau,\psi}$ and some of their variants. In §2, we formulate the general form of the Breuil-Mézard conjecture and recall the explicit definition of $\mu_{\bar{\sigma}}(V_{\mathbb{F}})$ when $K/\mathbb{Q}_p$ is unramified and $V_{\mathbb{F}}$ is absolute irreducible. In this case these integers are all either 0 or 1, and the explicit description is essentially a reformulation of the conjecture of [BDJ]. Finally, in §3 we prove the two theorems on $e(R_{V_{\mathbb{F}}}^{\mathbf{v},\tau,\psi}/\pi)$ mentioned above.

1. Potentially Semi-stable Deformation Rings

1.1. Potentially semi-stable representations. Let $K/\mathbb{Q}_p$ be a finite extension with residue field k, and fix an algebraic closure $\bar{K}/K$. For a subfield $K' \subset \bar{K}$, containing K, we write $G_{K'} = \mathrm{Gal}(\bar{K}/K')$ and $I_{K'} \subset G_{K'}$ for the inertia subgroup of $G_{K'}$. We denote by K'_0 the maximal absolutely unramified subfield of K', and by $\mathcal{O}_{K'}$ the ring of integers of K'.

Recall Fontaine's [Fo 1] period rings

$$B_{\mathrm{cris}} \subset B_{\mathrm{st}} \subset B_{\mathrm{dR}}.$$

The ring B_{st} is a $\bar{K}_0$-algebra, equipped with a Frobenius endomorphism φ and an operator N satisfying $N\varphi = p\varphi N$, and we have $B_{\mathrm{cris}} = B_{\mathrm{st}}^{N=0}$. The ring B_{dR} is a discrete valuation field with residue field $\widehat{\bar{K}}$. In particular, it carries a filtration given by the valuation. The above inclusions induce inclusions

$$B_{\mathrm{cris}} \otimes_{K_0} K \subset B_{\mathrm{st}} \otimes_{K_0} K \subset B_{\mathrm{dR}}.$$

In particular, the rings $B_{\mathrm{cris}} \otimes_{K_0} K$ and $B_{\mathrm{st}} \otimes_{K_0} K$ are equipped with the filtration induced from B_{dR}.

Suppose that V is a finite dimensional $\mathbb{Q}_p$-vector space equipped with a continuous action of G_K. We set

$$D_{\mathrm{cris}}(V) = (B_{\mathrm{cris}} \otimes_{\mathbb{Q}_p} V)^{G_K}, \quad D_{\mathrm{st}}(V) = (B_{\mathrm{st}} \otimes_{\mathbb{Q}_p} V)^{G_K}.$$

Then $D_{\mathrm{st}}(V)$ is a K_0-vector space of dimension $\leq \dim_{\mathbb{Q}_p} V$ equipped with operators φ and N, with φ a bijection and satisfying $N\varphi = p\varphi N$. We have

$D_{\mathrm{cris}}(V) = D_{\mathrm{st}}(V)^{N=0}$. Moreover,

$$D_{\mathrm{cris}}(V) \otimes_{K_0} K \subset D_{\mathrm{st}}(V) \otimes_{K_0} K \subset D_{\mathrm{dR}}(V) := (B_{\mathrm{dR}} \otimes_{\mathbb{Q}_p} V)^{G_K}. \qquad (1.1.1)$$

So $D_{\mathrm{cris}}(V) \otimes_{K_0} K$ and $D_{\mathrm{st}}(V) \otimes_{K_0} K$ are equipped with a filtration.

A representation V is called *crystalline* (respectively *semi-stable*) if $D_{\mathrm{cris}}(V)$ (resp. $D_{\mathrm{st}}(V)$) has K_0-dimension $\dim_{\mathbb{Q}_p} V$, in which case both (resp. the second) inclusions in (1.1.1) are equalities. We say that V is *potentially crystalline* (resp. *potentially semi-stable*) if $V|_{G_{K'}}$ is crystalline (resp. semi-stable) for some finite extension K'/K.

1.1.2. Fix an algebraic closure $\bar{\mathbb{Q}}_p$ of $\mathbb{Q}_p$ and let $E \subset \bar{\mathbb{Q}}_p$ be a finite extension of $\mathbb{Q}_p$ with ring of integers $\mathcal{O}$. Let V_E be an E-vector space of finite dimension d, equipped with a continuous action of G_K. We assume that V_E is potentially semi-stable (viewed as a $\mathbb{Q}_p$-representation). Then

$$D_{\mathrm{pst}}(V_E) = \varinjlim_{K'} (B_{\mathrm{st}} \otimes_{\mathbb{Q}_p} V_E)^{G_{K'}}$$

is a vector space over $\bar{K}_0$ of dimension $\dim_{\mathbb{Q}_p} V_E$. Note that $D_{\mathrm{pst}}(V_E)$ is a $\bar{K}_0 \otimes_{\mathbb{Q}_p} E$-module equipped with a semi-linear action of G_K, and so with a linear action of I_K. Since φ is a bijection on $D_{\mathrm{pst}}(V_E)$, this is necessarily a free $\bar{K}_0 \otimes_{\mathbb{Q}_p} E$-module, and since the action of φ commutes with that of I_K, we have $\mathrm{tr}(\sigma | D_{\mathrm{pst}}(V_E)) \in E$ for any $\sigma \in I_K$.

Let $\tau : I_K \to \mathrm{GL}_d(\bar{\mathbb{Q}}_p)$ be a representation with open kernel. We say that V_E is of Galois type τ if the I_K-representation $D_{\mathrm{pst}}(V_E)$ is equivalent to τ. That is, $\bar{\mathbb{Q}}_p \otimes_E D_{\mathrm{pst}}(V_E)$, equipped with its I_K action is isomorphic to $\tau \otimes_{\mathbb{Q}_p} \bar{K}_0$. Concretely this means that for any $\sigma \in I_K$, $\mathrm{tr}(\sigma | D_{\mathrm{pst}}(V_E)) = \mathrm{tr}(\tau(\sigma))$.

We can extend this definition to finite local E-algebras B : If V_B is a finite free B-module, equipped with a continuous, potentially semi-stable action of G_K, then $D_{\mathrm{pst}}(V_B)$ gives rise to a representation of I_K on a finite free $\bar{K}_0 \otimes_{\mathbb{Q}_p} B$-module with traces in B. We say that V_B is of Galois type τ if the traces of elements of I_K acting on $D_{\mathrm{pst}}(V_B)$ and τ are equal. If B has residue field E then a potentially semi-stable V_B is of type τ if and only if $V_B \otimes_B E$ is.

1.1.3. Let $\mathbf{v}$ be a conjugacy class of cocharacters of $\mathrm{Res}_{K/\mathbb{Q}_p}\mathrm{GL}_d$ (defined over $\bar{\mathbb{Q}}_p$). Concretely, $\mathbf{v}$ consists of the data of a d-tuple of integers for each embedding $K \hookrightarrow \bar{\mathbb{Q}}_p$. Let $E_{\mathbf{v}} \subset \bar{E}$ denote the *reflex field* of $\mathbf{v}$. That is, $E_{\mathbf{v}}$ is the fixed field of the group of $\sigma \in \mathrm{Gal}(\bar{\mathbb{Q}}_p/\mathbb{Q}_p)$ such that $\sigma^*(\mathbf{v}) = \mathbf{v}$. Then $\mathbf{v}$ has a representative defined over $E_{\mathbf{v}}$.

Now let V_E be as above, and suppose that $E \supset E_{\mathbf{v}}$. We say that V_E has p-adic Hodge type $\mathbf{v}$, if the filtration on the $K \otimes_{\mathbb{Q}_p} E$-module $D_{\mathrm{dR}}(V_E)$ is induced by the *inverse* of a cocharacter in the conjugacy class $\mathbf{v}$. As in section 1.1.2, we can extend this definition to representations of G_K on finite local E-algebras B.

1.1.4. Suppose that V_E is of p-adic Hodge type $\mathbf{v}$, and Galois type τ. An extension of V_E by V_E in the category of G_K-representations can be regarded

as a representation of G_K on a finite free module $V_{E[\epsilon]}$ over the dual numbers $E[\epsilon]$. If $V_{E[\epsilon]}$ is potentially semi-stable it is necessarily of p-adic Hodge type $\mathbf{v}$ and Galois type τ. We can compute the space of such extensions as follows: First observe that

$$\mathrm{ad}D_{\mathrm{pst}}(V_E) \xrightarrow{\sim} D_{\mathrm{pst}}(\mathrm{ad}V_E) \subset D_{\mathrm{dR}}(\mathrm{ad}V_E) \otimes_K \bar{K} \xrightarrow{\sim} \mathrm{ad}D_{\mathrm{dR}}(V_E) \otimes_K \bar{K}$$

where ad denotes the adjoint so that, for example, $\mathrm{ad}V_E = \mathrm{Hom}_E(V_E, V_E)$. Hence

$$(\mathrm{ad}D_{\mathrm{pst}}(V_E))^{G_K} \subset \mathrm{ad}D_{\mathrm{dR}}(V_E). \tag{1.1.5}$$

Suppose for a moment that V_E is potentially crystalline. Then it turns out that the space $\mathrm{Ext}^1_{\mathrm{pcris}}(V_E, V_E)$ of self extensions of V_E which are potentially crystalline is canonically isomorphic to the H^1 of the following complex concentrated in degrees 0 and 1

$$(\mathrm{ad}D_{\mathrm{pst}}(V_E))^{G_K} \xrightarrow{(1-\varphi,\mathrm{can})} (\mathrm{ad}D_{\mathrm{pst}}(V_E))^{G_K} \oplus \mathrm{ad}D_{\mathrm{dR}}(V_E)/\mathrm{Fil}^0\mathrm{ad}D_{\mathrm{dR}}(V_E),$$

where the second component of the map is induced by the inclusion (1.1.5). The kernel of this map is canonically isomorphic to $(\mathrm{ad}V_E)^{G_K}$. In particular, we have

$$\dim_E \mathrm{Ext}^1_{\mathrm{pcris}}(V_E, V_E) = \dim_E \mathrm{ad}D_{\mathrm{dR}}(V_E)/\mathrm{Fil}^0\mathrm{ad}D_{\mathrm{dR}}(V_E) + \dim_E(\mathrm{ad}V_E)^{G_K}. \tag{1.1.6}$$

In particular, if V_E is absolutely irreducible, then the right hand side of (1.1.6) depends only on the p-adic Hodge type, and is equal to $1+w_{\mathbf{v}}^{>0}$, where $w_{\mathbf{v}}^{>0}$ is the dimension of the Lie subalgebra of $\mathrm{Res}_{K/\mathbb{Q}_p}\mathfrak{gl}_d$ on which a fixed representative of $\mathbf{v}$ acts with positive weights.

Now suppose that V_E is potentially semi-stable. Then the space $\mathrm{Ext}^1_{\mathrm{pst}}(V_E, V_E)$ of potentially semi-stable self extensions is canonically isomorphic to H^1 of the total complex (concentrated in degrees $0, 1, 2$) of

$$\begin{array}{ccc}
(\mathrm{ad}D_{\mathrm{pst}}(V_E))^{G_K} & \xrightarrow{\quad 1-\varphi \quad} & (\mathrm{ad}D_{\mathrm{pst}}(V_E))^{G_K} \\
\downarrow{\scriptstyle N,\mathrm{can}} & & \downarrow{\scriptstyle N} \\
(\mathrm{ad}D_{\mathrm{pst}}(V_E))^{G_K} \oplus \mathrm{ad}D_{\mathrm{dR}}(V_E)/\mathrm{Fil}^0\mathrm{ad}D_{\mathrm{dR}}(V_E) & \xrightarrow{p\varphi-1,0} & (\mathrm{ad}D_{\mathrm{pst}}(V_E))^{G_K}
\end{array}$$

If V_E is absolutely irreducible, we deduce that the dimension of $\mathrm{Ext}^1_{\mathrm{pst}}(V_E, V_E)$ is again $1+w_{\mathbf{v}}^{>0}$ provided the H^2 of the above total complex vanishes. In general, this H^2 contains obstructions for the deformation theory of V_E as a potentially semi-stable representation.

1.2. Deformation rings. Now let $\bar{\mathbb{F}}_p$ be the residue field of $\bar{\mathbb{Q}}_p$, and $\mathbb{F} \subset \bar{\mathbb{F}}_p$ a finite extension of $\mathbb{F}_p$. Let $V_{\mathbb{F}}$ be an $\mathbb{F}$-vector space of dimension d

equipped with a continuous action of G_K. Let $\mathfrak{AR}_{W(\mathbb{F})}$ denote the category of Artinian $W(\mathbb{F})$-algebras with residue field $\mathbb{F}$. If A is in $\mathfrak{AR}_{W(\mathbb{F})}$, a *deformation* of $V_{\mathbb{F}}$ to A is a finite free A-module equipped with a continuous action of G_K and a G_K-equivariant isomorphism $V_A \otimes_A \mathbb{F} \xrightarrow{\sim} V_{\mathbb{F}}$. We denote by $D_{V_{\mathbb{F}}}(A)$ the set of isomorphism classes of deformations of $V_{\mathbb{F}}$ to A.

If we fix a basis for $V_{\mathbb{F}}$, then a *framed deformation* is a deformation V_A of $V_{\mathbb{F}}$ to A, together with a lifting to V_A of the chosen basis of $V_{\mathbb{F}}$. We denote by $D^{\square}_{V_{\mathbb{F}}}(A)$ the set of isomorphism classes of framed deformations of $V_{\mathbb{F}}$ to A.

The functor $D^{\square}_{V_{\mathbb{F}}}$ is always pro-representable by a complete local $W(\mathbb{F})$-algebra $R^{\square}_{V_{\mathbb{F}}}$. If $\operatorname{End}_{\mathbb{F}[G_K]} V_{\mathbb{F}} = \mathbb{F}$ then the functor $D_{V_{\mathbb{F}}}$ is pro-representable by a complete local $W(\mathbb{F})$-algebra $R_{V_{\mathbb{F}}}$ [Ma]. In this case the canonical morphism $R_{V_{\mathbb{F}}} \to R^{\square}_{V_{\mathbb{F}}}$ is formally smooth.

Now let $E \subset \bar{\mathbb{Q}}_p$ be a finite extension of $\mathbb{Q}_p$ as before, and assume that the residue field of E contains $\mathbb{F}$. Fix a representation $\tau : I_K \to \mathrm{GL}_d(E)$ with open kernel, and a p-adic Hodge type $\mathbf{v}$ such that $E_{\mathbf{v}} \subset E$. The main result of [Ki 4] is that $R^{\square}_{V_{\mathbb{F}}}$ and $R_{V_{\mathbb{F}}}$ (when it is defined) admit quotients which parameterize potentially semi-stable deformations of $V_{\mathbb{F}}$ of Galois type τ and p-adic Hodge type $\mathbf{v}$.

Theorem 1.2.1. *There exists a p-torsion free quotient $R^{\square,\tau,\mathbf{v}}_{V_{\mathbb{F}}}$ of $R^{\square}_{V_{\mathbb{F}}} \otimes_{W(\mathbb{F})} \mathcal{O}$ such that for any finite local E-algebra B, and any homomorphism $\xi : R^{\square}_{V_{\mathbb{F}}} \to B$, the B-representation of G_K induced by ξ is potentially semi-stable of Galois type τ and p-adic Hodge type $\mathbf{v}$ if and only if ξ factors through $R^{\square,\tau,\mathbf{v}}_{V_{\mathbb{F}}}$.*

The irreducible components of $\operatorname{Spec} R^{\square,\tau,\mathbf{v}}_{V_{\mathbb{F}}}[1/p]$ *are generically reduced and of dimension $d^2 + w_{\mathbf{v}}^{>0}$.*

If $\operatorname{End}_{\mathbb{F}[G_K]} V_{\mathbb{F}} = \mathbb{F}$, *then there exists an analogous quotient $R^{\tau,\mathbf{v}}_{V_{\mathbb{F}}}$ of $R_{V_{\mathbb{F}}}$, except that the components of* $\operatorname{Spec} R^{\tau,\mathbf{v}}_{V_{\mathbb{F}}}[1/p]$ *have dimension $1 + w_{\mathbf{v}}^{>0}$.*

We have a completely analogous statement for potentially crystalline representations, except that one can then make a more precise statement about the local structure of the generic fibres of the corresponding rings:

Theorem 1.2.2. *There exists a p-torsion free quotient $R^{\square,\tau,\mathbf{v}}_{V_{\mathbb{F}},\mathrm{cr}}$ of $R^{\square}_{V_{\mathbb{F}}} \otimes_{W(\mathbb{F})} \mathcal{O}$ such that for any finite local E-algebra B, and any homomorphism $\xi : R^{\square}_{V_{\mathbb{F}}} \to B$, the B-representation of G_K induced by ξ is potentially crystalline of Galois type τ and p-adic Hodge type $\mathbf{v}$ if and only if ξ factors through $R^{\square,\tau,\mathbf{v}}_{V_{\mathbb{F}},\mathrm{cr}}$.*

The irreducible components of $\operatorname{Spec} R^{\square,\tau,\mathbf{v}}_{V_{\mathbb{F}},\mathrm{cr}}[1/p]$ *are formally smooth of dimension $d^2 + w_{\mathbf{v}}^{>0}$.*

If $\operatorname{End}_{\mathbb{F}[G_K]} V_{\mathbb{F}} = \mathbb{F}$, *then there exists an analogous quotient $R^{\tau,\mathbf{v}}_{V_{\mathbb{F}},\mathrm{cr}}$ of $R_{V_{\mathbb{F}}}$, except that the components of* $\operatorname{Spec} R^{\tau,\mathbf{v}}_{V_{\mathbb{F}},\mathrm{cr}}[1/p]$ *have dimension $1 + w_{\mathbf{v}}^{>0}$.*

Note that it is clear that, if the above quotients exist, then they are unique. The reason for taking B a finite local E-algebra, rather than just a finite field extension of E, was to ensure this uniqueness.

1.2.3. For τ trivial, the above results were previously known in special cases: In each of those cases what was actually shown were special cases of the following conjecture of Fontaine [Fo 2]:

Conjecture 1.2.4. *(Fontaine) Let $a \leq b$ be integers and V a continuous representation of G_K on a finite free $\mathbb{Z}_p$-module. Suppose that for $n \geq 1$ V/p^nV is a subquotient of a G_K-stable lattice in a semi-stable (resp. crystalline) representation V_n whose Hodge-Tate weights are in $[a, b]$. Then $V \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ is semi-stable (resp. crystalline) with Hodge-Tate weights in $[a, b]$.*

1.2.5. For crystalline deformations this was shown by Ramakrishna [Ra] when $[a, b] = [0, 1]$, using results of Raynaud, [4] by Fontaine-Lafaille [FL] when $K = K_0$ and $[a, b] = [0, p-2]$, and by Berger [Be] whenever $K = K_0$. For semi-stable representations with $[K : K_0]|b - a| < p - 1$ this is a result of Breuil [Br 2].

The results of [Ki 4], are not proved via Fontaine's conjecture. Rather the quotients $R^{\square,\mathbf{v},\tau}_{V_{\mathbb{F}}}$ are constructed more directly using the results of [Ki 2] on Galois stable lattices in semi-stable representations. On the other hand, T. Liu has also used the theory of [Ki 2] to prove Fontaine's conjecture in general [Li].

2. The Breuil-Mézard Conjecture

2.1. Local Langlands and I_K-representations. From now on we fix a normalization of local class field theory so that the restriction of the cyclotomic character $\chi_{\mathrm{cyc}} : G_K \to \mathbb{Z}_p$ to $\mathcal{O}_K^\times \subset G_K$ is given by the norm $N_{K/\mathbb{Q}_p}$. This corresponds to the normalization of global class field theory which takes uniformizers to geometric Frobenii.

Consider a representation $\tau : I_K \to \mathrm{GL}_2(\bar{\mathbb{Q}}_p)$ with open kernel as in section 1.1.2 We will assume that τ is the restriction to I_K of a 2-dimensional representation of the Weil-Deligne group WD_K of K.

If $\tilde{\tau}$ is any continuous, Frobenius semi-simple 2-dimensional representation of WD_K, we denote by $\pi(\tilde{\tau})$ the representation of $\mathrm{GL}_2(K)$ attached to $\tilde{\tau}$ by the local Langlands correspondence [5] , normalized so that $\pi(\tilde{\tau})$ has central character $\det \tilde{\tau}|_{K^\times} |\cdot|^{-1}$. We have the following result [BM, Appendix].

Theorem 2.1.1. *(Henniart) There exists a finite dimensional, irreducible $\bar{\mathbb{Q}}_p$-representation $\sigma(\tau)$ (resp. $\sigma_{\mathrm{cr}}(\tau)$) of $\mathrm{GL}_2(\mathcal{O}_K)$ such that for any 2-dimensional, Frobenius semi-simple representation $\tilde{\tau}$ of WD_K, $\pi(\tilde{\tau})|_{\mathrm{GL}_2(\mathcal{O}_K)}$ contains $\sigma(\tau)$ (resp. $\sigma_{\mathrm{cr}}(\tau)$) if and only if $\tilde{\tau}|_{I_K} \sim \tau$ (resp. $\tilde{\tau}|_{I_K} \sim \tau$ and $N = 0$ on $\tilde{\tau}$).*

[4] Actually, what Ramakrishna shows is that if V_n arises from a p-divisible group then so does V. It was a later result of Breuil that V arises from a p-divisible group if and only if it is crystalline with Hodge-Tate weights in $[0, 1]$.

[5] If $\tilde{\tau} \sim \chi \oplus \chi|\cdot|$ for some character χ of WD_K, then we take $\pi(\tilde{\tau})$ to be the reducible principal series representation $\chi \circ \det \otimes \mathrm{Ind}_B^{\mathrm{GL}_2(K)} \mathbf{1}$ where $B \subset \mathrm{GL}_2(K)$ is a Borel, rather than the more classical choice of the one dimensional representation $\chi \circ \det$.

The representation $\sigma(\tau)$ (resp. $\sigma_{\mathrm{cr}}(\tau)$) is uniquely determined by this property except possibly [6] *when* $|k| = 2$.

2.1.2. Let $\mathbf{v}$ be a cocharacter of $\mathrm{Res}_{K/\mathbb{Q}_p}\mathrm{GL}_2$ and suppose that E contains the image of all embeddings $K \hookrightarrow \bar{\mathbb{Q}}_p$. In particular, $E_{\mathbf{v}} \subset E$. Concretely, $\mathbf{v}$ consists of the data of a pair of integers $(w_\iota, k_\iota + w_\iota)$ with $k_\iota \geq 0$, for each embedding $\iota : K \hookrightarrow \bar{\mathbb{Q}}_p$. We say that $\mathbf{v}$ is *regular* if $k_\iota \geq 1$ for all ι. For a regular $\mathbf{v}$ we set

$$\sigma(\mathbf{v}) = \otimes_{\iota:K\hookrightarrow E}\iota^*(\mathrm{Sym}^{k_\iota-1}K^2 \otimes \det{}^{w_\iota})$$

Now suppose that τ, $\sigma(\tau)$ and $\sigma_{\mathrm{cr}}(\tau)$ are defined over E. We again denote by $\sigma(\tau)$ and $\sigma_{\mathrm{cr}}(\tau)$ the corresponding E-vector spaces. Then we set $\sigma(\mathbf{v}, \tau) = \sigma(\tau) \otimes_E \sigma(\mathbf{v})$, and $\sigma_{\mathrm{cr}}(\mathbf{v}, \tau) = \sigma_{\mathrm{cr}}(\tau) \otimes_E \sigma(\mathbf{v})$.

2.2. Formulation of the conjecture.

Let ϖ be a uniformizer of K, and χ_ϖ the Lubin-Tate character attached to ϖ. For $\mathbf{v}$ as above we set

$$\chi_{\mathbf{v}} = \prod_{\iota:K\hookrightarrow E} (\iota \circ \chi_\varpi)^{k_\iota+2w_\iota-1}.$$

Now fix τ as in section 2.1 and $\mathbf{v}$ as above. Let $\psi : G_K \to \mathcal{O}^\times$ be a continuous character such that $\psi|_{I_K} = \chi_{\mathbf{v}}|_{I_K} \cdot \det \tau$.

Let $\mathbb{F} \subset \bar{\mathbb{F}}_p$ be the residue field of E, and let $V_{\mathbb{F}}$ be a two dimensional $\mathbb{F}$-vector space equipped with a continuous action of G_K such that the determinant of $V_{\mathbb{F}}$ is equal to the reduction of $\psi\chi_{\mathrm{cyc}}$.

We denote by $R^{\square,\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}$ the quotient of the ring $R^{\square,\mathbf{v},\tau}_{V_{\mathbb{F}}}$ introduced in Theorem 1.2.1 corresponding to deformations with determinant (the image of) $\psi\chi_{\mathrm{cyc}}$. Similarly we have the ring $R^{\square,\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}$ and, when $\mathrm{End}_{\mathbb{F}[G_K]}V_{\mathbb{F}} = \mathbb{F}$, the rings $R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}$ and $R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}$

Let $\pi \subset \mathcal{O}$ be a uniformizer. We want to relate the Hilbert-Samuel multiplicity of the ring $R^{\square,\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}/\pi$ and its variants to the reduction mod π of a $\mathrm{GL}_2(\mathcal{O}_K)$-stable $\mathcal{O}$-lattice $L_{\mathbf{v},\tau} \subset \sigma(\mathbf{v}, \tau)$. To do this we need to recall the irreducible mod p representations of $\mathrm{GL}_2(k)$ [BL].

2.2.1. Let $\underline{n} = \{n_{\bar{\iota}}\}$ and $\underline{m} = \{m_{\bar{\iota}}\}$ be tuples of integers indexed by the embeddings $\bar{\iota} : k \hookrightarrow \mathbb{F}$, with $0 \leq n_{\bar{\iota}}, m_{\bar{\iota}} \leq p-1$ and not all $m_{\bar{\iota}} = p-1$. Then the representations

$$\sigma_{\underline{n},\underline{m}} = \otimes_{\bar{\iota}}\bar{\iota}^*(\mathrm{Sym}^{n_{\bar{\iota}}}k^2 \otimes \det{}^{m_{\bar{\iota}}})$$

are irreducible and pairwise distinct, and any irreducible mod p representation of $\mathrm{GL}_2(k)$ is isomorphic to one of the $\sigma_{\underline{n},\underline{m}}$. These are also the irreducible mod p representations of $\mathrm{GL}_2(\mathcal{O}_K)$.

[6] More precisely, if $|k| = 2$ and $\tau \sim \chi \oplus \chi\varepsilon_0$ with ε_0 a ramified character then there are two such representations. In this case, we take $\sigma(\tau) = \sigma_{\mathrm{cr}}(\tau)$ to be $\chi \circ \det$ times the representation denoted by $u_{N_0}(\varepsilon_0)$ in [He, A.2.2]. We are grateful to Fred Diamond for pointing out that, in fact, the two representations have the same semi-simplified reductions, so that the two possible choices for $\sigma(\tau)$ give rise to the same conjectures below.

2.2.2. Recall that the Hilbert-Samuel multiplicity is an invariant which measures the complexity of a Noetherian, local ring A. If A has dimension d and maximal ideal $\mathfrak{m} \subset A$ then, for sufficiently large n, the function $n \mapsto \ell(A/\mathfrak{m}^{n+1})$ is a polynomial of degree d, where ℓ denotes length. Then the Hilbert-Samuel multiplicity $e(A)$ is defined as $d!$ times the coefficient of X^d in this polynomial. It is necessarily an integer.

More generally, if M is a finite A-module, then for n sufficiently large, $n \mapsto \ell(M/\mathfrak{m}^{n+1})$ is a polynomial of degree at most d. The coefficient of X^d has the form $e_A(M)/d!$ for a non-negative integer $e_A(M)$ which is called the Hilbert-Samuel multiplicity of M.

The following is a natural generalization of the Breuil-Mézard conjecture which is, to some extent, already hinted at in [BM, p214].

Conjecture 2.2.3. *There exist integers $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$ such that for any τ and $\mathbf{v}$, and ψ as above, with $\mathbf{v}$ regular, we have*

$$e\left(R_{V_{\mathbb{F}}}^{\square,\mathbf{v},\tau,\psi}/\pi\right) = \sum_{\underline{n},\underline{m}} a(\underline{n},\underline{m})\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}),$$

where

$$(L_{\mathbf{v},\tau} \otimes_{\mathcal{O}} \mathbb{F})^{\mathrm{ss}} \xrightarrow{\sim} \oplus_{\underline{n},\underline{m}} \sigma_{\underline{n},\underline{m}}^{a(\underline{n},\underline{m})}.$$

Similarly, if $L_{\mathbf{v},\tau}^{\mathrm{cr}}$ is a $\mathrm{GL}_2(\mathcal{O}_K)$-stable lattice in $\sigma_{\mathrm{cr}}(\mathbf{v},\tau)$ then

$$e\left(R_{V_{\mathbb{F}},\mathrm{cr}}^{\square,\mathbf{v},\tau,\psi}/\pi\right) = \sum_{\underline{n},\underline{m}} a(\underline{n},\underline{m})\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}),$$

where

$$\left(L_{\mathbf{v},\tau}^{\mathrm{cr}} \otimes_{\mathcal{O}} \mathbb{F}\right)^{\mathrm{ss}} \xrightarrow{\sim} \oplus_{\underline{n},\underline{m}} \sigma_{\underline{n},\underline{m}}^{a(\underline{n},\underline{m})}.$$

2.2.4. Note that when $V_{\mathbb{F}}$ has trivial endomorphisms, the morphism $R_{V_{\mathbb{F}}}^{\mathbf{v},\tau,\psi} \to R_{V_{\mathbb{F}}}^{\square,\mathbf{v},\tau,\psi}$ (resp. $R_{V_{\mathbb{F}},\mathrm{cr}}^{\mathbf{v},\tau,\psi} \to R_{V_{\mathbb{F}},\mathrm{cr}}^{\square,\mathbf{v},\tau,\psi}$) is formally smooth, so the Hilbert-Samuel multiplicities of these two rings are equal.

The equalities in Conjecture 2.2.3 can be viewed as an infinite number of equations (corresponding to the choices of $\mathbf{v}$ and τ) in the finitely many unknowns $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$. If these equalities hold, then the $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$ may be determined by taking τ trivial, and selecting $\mathbf{v}$ as follows: Choose a subset L of the set of embeddings $K \hookrightarrow E$ such that L maps bijectively onto the set of embeddings $k \hookrightarrow \mathbb{F}$. Define $\mathbf{v}$ by $k_\iota = n_{\bar{\iota}} + 1$ and $w_\iota = m_\iota$ if $\iota \in L$ and $k_\iota = 1, w_\iota = 0$ otherwise. Here $\bar{\iota}$ denotes the reduction of ι. Then $\sigma_{\mathrm{cr}}(\tau)$ is the trivial representation of $\mathrm{GL}_2(\mathcal{O}_K)$ and any $\mathrm{GL}_2(\mathcal{O}_K)$-stable lattice $L_{\mathbf{v},\tau}^{\mathrm{cr}}$ in $\sigma_{\mathrm{cr}}(\mathbf{v},\tau)$, has reduction isomorphic to $\sigma_{\underline{n},\underline{m}}$. So Conjecture 2.2.3 predicts

$$\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) = e\left(R_{\mathrm{cr}}^{\square,\mathbf{v},\tau,\psi}/\pi\right). \tag{2.2.5}$$

2.3. The case of an unramified extension. When $K/\mathbb{Q}_p$ is unramified, the integers on the right hand side of (2.2.5) can be determined in almost all cases, and are usually in $\{0,1,2\}$. In this case, the condition that $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) \neq 0$ is closely related to the Buzzard-Diamond-Jarvis conjecture on when a given two dimensional, mod p *global* Galois representation is modular of weight $\sigma_{\underline{n},\underline{m}}$.

2.3.1. Suppose now that $K/\mathbb{Q}_p$ is unramified. We will give the explicit values of $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$ when $V_{\mathbb{F}}$ is absolutely irreducible.

Let K'/K be the unramified extension of degree 2, so that $I_K = I_{K'} = I_{\mathbb{Q}_p}$. Let k' denote the residue field of K'. Let $n = [K : \mathbb{Q}_p]$ and $\omega_{2n} : I_{\mathbb{Q}_p} \to k'^{\times}$ the fundamental character of level $2n$ and $\omega_n = \omega_{2n}^{p^n+1}$ the fundamental character of level n. We will assume that E contains all embeddings of K' into $\bar{\mathbb{Q}}_p$.

Let J be a subset of the embeddings $k' \hookrightarrow \mathbb{F}$ which bijects onto the set of all embeddings $k \hookrightarrow \mathbb{F}$. We set

$$\omega_J = \prod_{\bar{\iota} \in J} \iota \circ \left(\omega_{2n}^{\frac{n}{\iota}+1} \cdot \omega_n^{\frac{m}{\iota}} \right),$$

where for $\iota \in J$ we again denote by ι the restriction of ι to k. Thus ω_J is a character $I_K \to \mathbb{F}^{\times}$. Similarly, if J' denotes the compliment of J in the set of embeddings $\bar{\iota} : k' \hookrightarrow \mathbb{F}$, we have the character $\omega_{J'}$.

Conjecture 2.3.2. *Suppose $V_{\mathbb{F}}$ is absolutely irreducible. Then Conjecture 2.2.3 holds with $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) = 0$ unless there exists J as above such that*

$$V_{\mathbb{F}}|_{I_K} \sim \begin{pmatrix} \omega_J & 0 \\ 0 & \omega_{J'} \end{pmatrix},$$

in which case $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) = 1$.

3. Theorems

3.1. Statements. We will review some cases when Conjecture 2.2.3 is known as well as sketching some of the arguments. We assume from now on that $p > 2$.

Most of the conjecture is known when $K = \mathbb{Q}_p$. In this case each of $\underline{n}$, $\underline{m}$ consist of a single integer which we denote by n and m respectively, and we write $\mu_{n,m}(V_{\mathbb{F}})$ for $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$. The explicit value of $\mu_{n,m}(V_{\mathbb{F}})$ is known in all cases, except when $n = p-2$ and $V_{\mathbb{F}}$ is scalar. One has the following result [Ki 5], which, in particular includes (most of) the original conjecture stated by Breuil-Mézard (here ω denotes the mod p cyclotomic character).

Theorem 3.1.1. *Suppose that $K = \mathbb{Q}_p$, that $V_{\mathbb{F}} \nsim \left(\begin{smallmatrix} \omega\chi & * \\ 0 & \chi \end{smallmatrix}\right)$ for any character χ, and that if $V_{\mathbb{F}}$ has scalar semi-simplification then it is scalar.*

Then Conjecture 2.2.3 holds for any regular $\mathbf{v}$ and any τ.

3.1.2. The proof uses the p-adic local Langlands correspondence for $\mathrm{GL}_2(\mathbb{Q}_p)$ to prove that the left hand side in the equalities in Conjecture 2.2.3 is bounded above by the right hand side. To each two dimensional E-representation V_E of $G_{\mathbb{Q}_p}$, this correspondence attaches a certain representation of $\mathrm{GL}_2(\mathbb{Q}_p)$ on a p-adic Banach space $\Pi(V)$. A key ingredient in the proof is the fact that the p-adic local Langlands correspondence is compatible with the usual local Langlands correspondence, in the sense that, if V_E is potentially semi-stable with p-adic Hodge type $\mathbf{v}$ and Galois type τ, then the locally algebraic vectors in $\Pi(V)$ contain a copy of the $\mathrm{GL}_2(\mathbb{Z}_p)$-representation $\sigma(\mathbf{v},\tau)$. This was proved by Colmez and Berger-Breuil [Co 2], [BB] when τ arises from an *abelian* representation of the Weil group, and by Colmez [Co] in general, using Emerton's work on the local-global compatibility of the p-adic Langlands correspondence [Em].

The opposite inequality is proved by a Taylor-Wiles style patching argument. Indeed, this patching argument shows that Conjecture 2.2.3 is very closely related to the conjecture of Fontaine-Mazur on the modularity of geometric Galois representations. One can attempt to run this argument in reverse and deduce Conjecture 2.2.3 from a modularity lifting theorem. For potentially Barsotti-Tate representations such a theorem was proved in [Ki 1] and generalized by Gee [Ge 1]. Using it one can show that for any $K/\mathbb{Q}_p$ we have

Theorem 3.1.3. *Denote by $\mathbf{v}_0$ the cocharacter corresponding to $k_\iota - 1 = w_\iota = 0$ for all ι. If $V_{\mathbb{F}}$ is absolutely irreducible, then there exist non-negative integers $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$ such that for any τ,*

$$e\left(R^{\square,\mathbf{v}_0,\tau,\psi}_{\mathrm{cr}}/\pi\right) = \sum_{\underline{n},\underline{m}} a(\underline{n},\underline{m})\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}),$$

where

$$\left(L^{\mathrm{cr}}_{\mathbf{v}_0,\tau} \otimes_{\mathcal{O}} \mathbb{F}\right)^{\mathrm{ss}} \xrightarrow{\sim} \oplus_{\underline{n},\underline{m}} \sigma^{a(\underline{n},\underline{m})}_{\underline{n},\underline{m}}.$$

3.1.4. Now return to the case where $K/\mathbb{Q}_p$ is unramified. We assume that $V_{\mathbb{F}}$ is absolutely irreducible, and we now take $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$ to be defined as in Conjecture 2.3.2, so that $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$ is non-zero if and only if there exists J such that $V_{\mathbb{F}}|_{I_K} \sim \left(\begin{smallmatrix} \omega_J & 0 \\ 0 & \omega_{J'} \end{smallmatrix}\right)$ in which case $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) = 1$.

We will say that $\mathbf{v}$ is *paritious* if the integers $k_\iota + 2w_\iota$ are independent of ι. We will say that $V_{\mathbb{F}}$ is regular, if there exists $(\underline{n},\underline{m})$ with $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) \neq 0$ and $2 \leq n_\iota \leq p-4$ for all ι.

Theorem 3.1.5. *Suppose that $K/\mathbb{Q}_p$ is unramified, that $\mathbf{v}$ is paritious and that $V_{\mathbb{F}}$ is absolutely irreducible and regular. Then*

$$e(R^{\mathbf{v},\tau,\psi}/\pi) \geq \sum_{\underline{n},\underline{m}} a(\underline{n},\underline{m})\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}),$$

where

$$(L_{\mathbf{v},\tau} \otimes_{\mathcal{O}} \mathbb{F})^{\mathrm{ss}} \xrightarrow{\sim} \oplus_{\underline{n},\underline{m}} \sigma_{\underline{n},\underline{m}}^{a(\underline{n},\underline{m})},$$

and similarly for $e(R_{\mathrm{cr}}^{\mathbf{v},\tau,\psi}/\pi)$.

3.2. A sketch of the proofs. We now give a sketch of some of the methods which are used to prove Theorems 3.1.3 and 3.1.5. These involve relating the Hilbert-Samuel multiplicities in the conjectures to those of certain spaces of automorphic forms.

It ought to be possible to extend these methods to prove Conjecture 2.2.3 for $e(R_{\mathrm{cr}}^{\square,\mathbf{v},\tau,\psi}/\pi)$ with an explicit collection of integers $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}})$, when $\mathbf{v} = \mathbf{v}_0$ and $K/\mathbb{Q}_p$ is unramified. This is work in progress with Toby Gee.

3.2.1. Let F be a totally real number field and D a totally definite quaternion algebra over F, which is unramified at all primes $v|p$ of F. Denote by $\mathbb{A}_F^f \subset \mathbb{A}_F$ the finite adeles. For each finite place v of F we will denote by $\pi_v \in F_v$ a uniformizer. Fix a maximal order $\mathcal{O}_D \subset D$, and an isomorphism $(\mathcal{O}_D)_v \xrightarrow{\sim} M_2(\mathcal{O}_{F_v})$ for each finite place where D is unramified. Let $U = \prod_v U_v \subset (D \otimes_F \mathbb{A}_F^f)^\times$ be a compact open subgroup contained in $\prod_v (\mathcal{O}_D)_v^\times$. We assume that $U_v = \mathrm{GL}_2(\mathcal{O}_{F_v})$ for $v|p$.

For each $v|p$, we fix a continuous representation $\sigma_v : U_v \to \mathrm{Aut}(W_{\sigma_v})$ on a finite $\mathcal{O}$-module. Write $W_\sigma = \otimes_{v|p,\mathcal{O}} W_{\sigma_v}$ and denote by $\sigma : \prod_{v|p} U_v \to \mathrm{Aut}(W_\sigma)$ the corresponding representation. We regard σ as being a representation of U by letting U_v act trivially if $v \nmid p$. Finally, assume there exists a continuous character $\psi : (\mathbb{A}_F^f)^\times / F^\times \to \mathcal{O}^\times$ such that σ on $U \cap (\mathbb{A}_F^f)^\times$ is given by multiplication by ψ. Fix such a ψ, and extend the action of U on W_σ to $U(\mathbb{A}_F^f)^\times$, by letting $(\mathbb{A}_F^f)^\times$ act via ψ.

Let $S_{\sigma,\psi}(U)$ denote the set of continuous functions

$$f : D^\times \backslash (D \otimes_F \mathbb{A}_F^f)^\times \to W_\sigma$$

such that for $g \in (D \otimes_F \mathbb{A}_F^f)^\times$ we have $f(gu) = \sigma(u)^{-1} f(g)$ for $u \in U$, and $f(gz) = \psi^{-1}(z) f(g)$ for $z \in (\mathbb{A}_F^f)^\times$.

We consider the left action of $(D \otimes_F \mathbb{A}_F^f)^\times$ on W_σ-valued functions on $(D \otimes_F \mathbb{A}_F^f)^\times$ given by the formula $(gf)(z) = f(zg)$. Then for any finite prime v, the double cosets of U_v in $(D \otimes_F \mathbb{A}_F^f)^\times$ act naturally on $S_{\sigma,\psi}(U)$. Denote by $\mathbb{T}_{\sigma,\psi}(U)$ the $\mathcal{O}$-algebra generated by the endomorphisms S_v and T_v of $S_{\sigma,\psi}(U)$ corresponding to $U_v \left(\begin{smallmatrix} \pi_v & 0 \\ 0 & \pi_v \end{smallmatrix}\right) U_v$ and $U_v \left(\begin{smallmatrix} \pi_v & 0 \\ 0 & 1 \end{smallmatrix}\right) U_v$ respectively, where $v \nmid p$ runs over primes at which D is unramified. If U_v is maximal compact in $(D \otimes_F F_v)^\times$, then these operators do not depend on the choice of π_v.

3.2.2. Now fix an algebraic closure $\bar{F}$ of F and let S be a finite set of primes of F, containing the infinite primes, the primes dividing p, the primes where D is ramified, and the primes where U_v is not maximal compact in $(D \otimes_F F_v)^\times$. Let $F_S \subset \bar{F}$ be the maximal extension of F unramified outside S, and set $G_{F,S} = \mathrm{Gal}(F_S/F)$.

Let $\mathfrak{m} \subset \mathbb{T}_{\sigma,\psi}(U)$ be a maximal ideal. Such an ideal is called *Eisenstein* if $T_v - 2 \in \mathfrak{m}$ for all but finitely many primes $v \notin S$ which split completely in some fixed abelian extension of F. After possibly replacing $\mathcal{O}$ by an extension we may assume that $\mathfrak{m}$ has residue field $\mathbb{F}$. If $\mathfrak{m}$ is a non-Eisenstein ideal, then the work of Carayol [Ca] and Taylor [Ta], together with the Jacquet-Langlands correspondence, implies that there exists a unique representation

$$\rho_{\mathfrak{m}} : G_{F,S} \to \mathrm{GL}_2(\mathbb{T}_{\sigma,\psi}(U)_{\mathfrak{m}})$$

such that if $v \notin S$ is a prime of F, and Frob_v denotes an arithmetic Frobenius at v then $\rho_{\mathfrak{m}}(\mathrm{Frob}_v)$ has trace T_v. We denote by $\bar{\rho}_{\mathfrak{m}}$ the reduction of $\rho_{\mathfrak{m}}$ modulo $\mathfrak{m}$. As $\mathfrak{m}$ is non-Eisenstein $\bar{\rho}_{\mathfrak{m}}$ is absolutely irreducible.

3.2.3. Now suppose we are given $\mathbf{v}$ and τ as in section 2.1.2 with $\mathbf{v}$ paritious and an absolutely irreducible representation $V_{\mathbb{F}}$ of G_K. Then we choose F such that there is a unique prime $\mathfrak{p}|p$ of F and $F_{\mathfrak{p}} \xrightarrow{\sim} K$. Fix an embedding $\bar{F} \hookrightarrow \bar{K}$, extending this isomorphism. We choose the character $\psi : (\mathbb{A}_F^f)^\times / F^\times \to \mathcal{O}^\times$ so that $\psi|_{I_K} = \chi_{\mathbf{v}}|_{I_K} \det \tau$, and we apply the above constructions with σ a $\mathrm{GL}_2(\mathcal{O}_K)$-stable $\mathcal{O}$-lattice $L^{\mathrm{cr}}_{\mathbf{v},\tau}$ in $\sigma_{\mathrm{cr}}(\mathbf{v},\tau)$.

Using CM forms, one can find $\mathfrak{m}$ such that $\bar{\rho}_{\mathfrak{m}}|_{G_K} \sim V_{\mathbb{F}}$, and we again denote by $V_{\mathbb{F}}$ the underlying $\mathbb{F}$-vector space of $\bar{\rho}_{\mathfrak{m}}$.

Let $R_{F,S}$ and $R_{\mathfrak{p}}$ denote the the universal deformation rings of $V_{\mathbb{F}}$ and $V_{\mathbb{F}}|_{G_K}$ respectively. We denote by $R^{\psi}_{F,S}$ the quotient of $R_{F,S}$ which parameterizes deformations of determinant $\psi\chi_{\mathrm{cyc}}$, where χ_{cyc} now denotes the p-adic cyclotomic character on $G_{F,S}$. Set

$$R^{\mathbf{v},\tau,\psi}_{F,S} = R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}} \otimes_{R_{\mathfrak{p}}} R^{\psi}_{F,S}.$$

The map

$$R_{F,S} \to \mathbb{T}_{\sigma,\psi}(U)_{\mathfrak{m}},$$

induced by $\rho_{\mathfrak{m}}$, factors through $R^{\mathbf{v},\tau,\psi}_{F,S}$. (See for example [Ki 4, §4].)

Under some technical restrictions on the choice of F, D and U, which can always be arranged for a given representation $V_{\mathbb{F}}$ of G_K, a Taylor-Wiles patching argument, as modified by Diamond [Di] and Fujiwara, and in [Ki 1, §3], [Ki 5, §2], shows that there exist an $\mathcal{O}$-algebra R_∞, maps of $\mathcal{O}$-algebras

$$\mathcal{O}[\![y_1,\ldots,y_h]\!] \to R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}[\![x_1,\ldots,x_{h-d}]\!] \twoheadrightarrow R_\infty, \tag{3.2.4}$$

and an R_∞-module M_∞ satisfying the following properties:

(1) $h \geq d = \dim R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}/\pi = [K : \mathbb{Q}_p]$.

(2) There is an isomorphism of $R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}$ algebras $R_\infty/(y_1,\ldots,y_h) \xrightarrow{\sim} R^{\mathbf{v},\tau,\psi}_{F,S}$.

(3) M_∞ is a finite free $\mathcal{O}[\![y_1,\ldots,y_h]\!]$-module and has rank at most 1 on any irreducible component on $\mathrm{Spec}\, R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}[\![x_1,\ldots,x_{h-d}]\!]$.

(4) There is an isomorphism of $R_{F,S}^{\mathbf{v},\tau,\psi}$-modules

$$M_\infty/(y_1,\dots,y_h)M_\infty \xrightarrow{\sim} S_{\sigma,\psi}(U)_{\mathfrak{m}}.$$

Now let

$$\{0\} = M^0 \subset M^1 \subset \cdots \subset M^s = L^{\mathrm{cr}}_{\mathbf{v},\tau}/\pi$$

be a filtration such that M^{i+1}/M^i is an irreducible representation of $\mathrm{GL}_2(k)$. Then we can enhance the above construction (see [Ki 5, 2.2.9]) in such a way that there exists a filtration

$$\{0\} = M^0_\infty \subset M^1_\infty \subset \cdots \subset M^s_\infty = M_\infty/\pi M_\infty$$

by R_∞-modules such that

(5) $M^i_\infty/M^{i-1}_\infty$ is a finite free $\mathbb{F}[\![y_1,\dots,y_h]\!]$-module.

(6) If $M^i/M^{i-1} \xrightarrow{\sim} \sigma_{\underline{n},\underline{m}}$ then the isomorphism in (4) above induces an isomorphism

$$M^i_\infty/M^{i-1}_\infty \otimes_{R_\infty} R_\infty/(y_1,\dots,y_h) \xrightarrow{\sim} S_{\sigma_{\underline{n},\underline{m}},\psi}(U)_{\mathfrak{m}}.$$

Moreover this construction can be made so that, as an $R_{\mathfrak{p}}[\![x_1,\dots,x_{h-d}]\!]$-module, $M^i_\infty/M^{i-1}_\infty$ depends only on $\sigma_{\underline{n},\underline{m}}$ and $\mathfrak{m}$, and not on the choice of $\mathbf{v}$ and τ. More precisely this module is made by an analogous patching argument but with $\sigma_{\underline{n},\underline{m}}$ in place of $L^{\mathrm{cr}}_{\mathbf{v},\tau}$. We denote this module by $M^{\underline{n},\underline{m}}_\infty$.

Set $R'_\infty = R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}[\![x_1,\dots,x_{h-d}]\!]$, and let $a(\underline{n},\underline{m})$ be the multiplicity with which $\sigma_{\underline{n},\underline{m}}$ appears as a Jordan-Hölder factor in $L^{\mathrm{cr}}_{\mathbf{v},\tau}/\pi$. Using (3) and (5) and standard facts about Hilbert-Samuel multiplicities one obtains

$$e(R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}/\pi) = e(R'_\infty/\pi R'_\infty) \geq e_{R'_\infty/\pi}(M_\infty/\pi M_\infty) = \sum_{\underline{n},\underline{m}} a(\underline{n},\underline{m}) e_{R'_\infty/\pi}(M^{\underline{n},\underline{m}}_\infty). \tag{3.2.5}$$

with equality if and only if $\operatorname{Spec} R'_\infty[1/p]$ is contained in the support of the R'_∞-module M_∞ (cf. [Ki 5, Lem. 2.2.11]). Note that the freeness condition in (3) implies that this support is a union of irreducible components of $\operatorname{Spec} R'_\infty[1/p]$ as the dimensions of $\mathcal{O}[\![y_1,\dots,y_h]\!]$ and R'_∞ coincide by (1). This also implies that $e_{R'_\infty/\pi}(M^{\underline{n},\underline{m}}_\infty)$ depends only on the image of $R_{\mathfrak{p}}[\![x_1,\dots,x_{h-d}]\!]$ in $\operatorname{End} M^{\underline{n},\underline{m}}_\infty$ and not on R'_∞, and is therefore independent of $\mathbf{v}$ and τ.

3.2.6. *Proof of Theorem 3.1.5.* In this case $K/\mathbb{Q}_p$ is unramified and $V_{\mathbb{F}}$ is assumed regular. We have to show that

$$e_{R'_\infty/\pi}(M^{\underline{n},\underline{m}}_\infty) \geq \mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}). \tag{3.2.7}$$

By definition, the term on the right is 0 or 1, and in the former case there is nothing to prove. Suppose $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) = 1$. As above, the condition (5) implies

that the support of $M_\infty^{\underline{n},\underline{m}}$ has dimension equal to $\dim R'_\infty/\pi$. Hence it suffices to show that $M_\infty^{\underline{n},\underline{m}} \neq \{0\}$. By (6) it suffices to show that $S_{\sigma_{\underline{n},\underline{m}},\psi}(U)_{\mathfrak{m}} \neq \{0\}$. This follows from Gee's proof [Ge 2] of the Buzzard-Diamond-Jarvis conjecture for regular weights. Namely our condition on the regularity of $V_{\mathbb{F}}$ implies that any $\sigma_{\underline{n},\underline{m}}$ such that $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) \neq 0$ is regular in the sense of [Ge 2].

This completes the proof of Theorem 3.1.5 for $R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}$ and the proof for $R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}}}$ is identical, replacing $L^{\mathrm{cr}}_{\mathbf{v},\tau}$ by a $\mathrm{GL}_2(\mathcal{O}_K)$-invariant lattice in $\sigma(\mathbf{v},\tau)$. □

3.2.8. *Proof of Theorem 3.1.3:* Let $\mathbf{v} = \mathbf{v}_0$, and set $\mu_{\underline{n},\underline{m}}(V_{\mathbb{F}}) = e_{R'_\infty}(M_\infty^{\underline{n},\underline{m}})$. To prove the theorem we have to show that the inequality in (3.2.5) is an equality. It is enough to show that M_∞ is a faithful R'_∞-module.

The following lemma will be useful.

Lemma 3.2.9. *The following are equivalent*

(1) The support of $S_{\sigma,\psi}(U)_{\mathfrak{m}}$ contains $\operatorname{Spec} R^{\mathbf{v},\tau,\psi}_{F,S}[1/p]$ *and $R^{\mathbf{v},\tau,\psi}_{F,S}$ is a finite $\mathcal{O}$-algebra.*

(2) M_∞ is a faithful R'_∞-module.

Proof. (2) $\Longrightarrow$ (1): If M_∞ is a faithful R'_∞-module then $R'_\infty = R_\infty$ and both are finite over $\mathcal{O}[\![y_1,\ldots,y_h]\!]$. Then (1) follows from conditions (2) and (4) in (3.2.3).

(1) $\Longrightarrow$ (2): One can use an argument of Khare-Wintenberger [KW 2, Cor. 4.7] to show that the second condition in (1) implies that the image of $\operatorname{Spec} R^{\mathbf{v},\tau,\psi}_{F,S}[1/p]$ in $\operatorname{Spec} R^{\mathbf{v},\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}[1/p]$ meets every irreducible component of the latter scheme. Hence the first condition implies that the support of $S_{\sigma,\psi}(U)_{\mathfrak{m}}$ meets every irreducible component of R'_∞. Since the support of M_∞ is a union of irreducible components of $\operatorname{Spec} R'_\infty[1/p]$, it must contain all of $\operatorname{Spec} R'_\infty[1/p]$ by condition (4) in (3.2.3). Finally as R'_∞ is flat over $\mathcal{O}$ with formally smooth (so in particular reduced) generic fibre, this implies that M_∞ is a faithful R'_∞-module. □

3.2.10. We return to the proof of Theorem 3.1.3. Since $\mathbf{v} = \mathbf{v}_0$ the main result of [Ki 1] and [Ge 1] shows that the support of $S_{\sigma,\psi}(U)_{\mathfrak{m}}$ contains $\operatorname{Spec} R^{\mathbf{v},\tau,\psi}_{F,S}[1/p]$.

Moreover the proof in *loc. cit* (cf. also [Ki 3, §1]) together with an argument of Khare-Wintenberger [KW 1, Prop. 3.8] shows that that $R^{\mathbf{v},\tau,\psi}_{F,S}$ is a finite $\mathcal{O}$-algebra. More precisely, the argument in [Ki 1, §3.4] carries out a patching argument analogous to the one sketched here, but over a finite, solvable, totally real extension F'/F. In that situation the analogue of the ring $R^{\mathbf{v}_0,\tau,\psi}_{V_{\mathbb{F}},\mathrm{cr}}$ turns out to be a domain. This implies that the analogue of the condition (2) in Lemma 3.2.9 is automatically satisfied, and hence so is the condition (1). This is enough to imply the finiteness of $R^{\mathbf{v},\tau,\psi}_{F,S}$ itself. □

References

[BB] L. Berger, C. Breuil, *Sur quelques représentations potentiellement cristallines de* $\mathrm{GL}_2(\mathbb{Q}_p)$, Astérisque, 330, 155–211, 2010.

[BCDT] C. Breuil, B. Conrad, F. Diamond, R. Taylor, *On the modularity of elliptic curves over* $\mathbb{Q}$ *: wild* 3*-adic exercises,* J. Amer. Math. Soc. 14, 843–939, 2001.

[BDJ] K. Buzzard, F. Diamond, F. Jarvis, *On Serre's conjecture for mod* l *Galois representations over totally real fields,* Duke Math. J, to appear, 2010.

[Be] L. Berger, *Limites des représentations cristallines,* Compositio Math. 140, 1473–1498, 2004.

[BK] C. Bushnell, P. Kutzko *The admissible dual of* $\mathrm{GL}(N)$ *via open compact subgroups,* Annals of Math. Studies 129, Princeton Univeristy Press, 1993.

[BL] L. Barthel, R. Livne, *Irreducible modular representations of* GL_2 *of a local field,* Duke Math. J, 75, 261–292, 1994.

[BM] C. Breuil, A. Mézard, *Multiplicités modulaires et représentations de* $\mathrm{GL}_2(\mathbb{Z}_p)$ *et de* $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ *en* $l = p$*,* Duke Math. J. 115, 205–310, 2002, with an appendix by G. Henniart.

[Br 1] C. Breuil, *Introduction Générale,* Astérisque 319, 1–12, 2008.

[Br 2] C. Breuil, *Représentations semi-stables et modules fortement divisibles,* Invent. Math. 136, 89–122, 1999.

[Ca] H. Carayol, *Sur les représentations* l*-adiques associes aux formes modulaires de Hilbert,* Ann. Sci. École Norm. Sup, 19, 409–468, 1986.

[Co] P. Colmez, *Représentations de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *et* (φ, Γ)*-modules,* Astérisque, 330, 283–511, 2010.

[Co2] P. Colmez *La série principal unitaire pour* $\mathrm{GL}_2(\mathbb{Q}_p)$. Astérisque, 330, 213–262, 2010.

[Di] F. Diamond, *The Taylor-Wiles construction and multiplicity one,* Invent. Math. 128, 379–391, 1997.

[Em] M. Emerton *Local-global compatibility in the* p*-adic Langlands program for* $\mathrm{GL}_{2/\mathbb{Q}}$, In preparation.

[FL] J-M. Fontaine, G. Laffaille, *Construction de répresentations* p*-adiques,* Ann. Scient. de l'E.N.S. 15, 547–608, 1982.

[FM] J.M. Fontaine, B. Mazur, *Geometric Galois Representations,* Elliptic curves, modular forms, and Fermat's last theorem (Hong Kong 1993), 41–78, Internat. Press, Cambridge MA, 1995.

[Fo 1] J.M. Fontaine, *Représentations* p*-adiques semi-stables,* Périodes p-adiques, Astérisque 223, 11–184, Société Mathématique de France, 1994.

[Fo 2] J.M. Fontaine, *Deforming semi-stable Galois representations,* Proc. Natl. Acad. Sci. USA, 94 11138–11141, 1997.

[Ge 1] T. Gee, *A modularity lifting theorem for weight two modular forms,* Math. Res. Lett. 13, 805–811, 2006.

[Ge 2] T. Gee, *On the weights of mod p Hilbert modular forms,* preprint, 2006.

[Ki 1] M. Kisin, *Moduli of finite flat group schemes and modularity,* Ann. of Math. 170(3), 1085–1180, 2009.

[Ki 2] M. Kisin, *Crystalline representations and F-crystals,* Algebraic geometry and number theory. In honor of Vladimir Drinfeld's 50th birthday, Prog. Math. 253, 459–496, Birkhäuser, 2006.

[Ki 3] M. Kisin, *Modularity of some geometric Galois representations, L*-functions and Galois representations (Durham 2004), LMS 320, 438–470, Cambridge Univ. Press, 2007.

[Ki 4] M. Kisin, *Potentially semi-stable deformation rings,* J. AMS, 21, 513–546, 2008.

[Ki 5] M. Kisin, *The Fontaine-Mazur conjecture for* GL_2, J. AMS, 22, 641–690, 2009.

[KW 1] C. Khare, J.P. Wintenberger, *On Serre's conjecture for* 2*-dimensional mod p representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, Ann. of Math, 169, 229–253, 2009.

[KW 2] C. Khare, J.P. Wintenberger, *Serre's modularity conjecture (II),* Invent. Math, 178(3), 505–586, 2009.

[Li] T. Liu, *Torsion p-adic Galois representations and a conjecture of Fontaine,* Ann. Sci. École Norm. Sup, 40(4), 633–674, 2007.

[Ma] B. Mazur, *Deforming Galois representations,* Galois groups over $\mathbb{Q}$ (Berkeley, CA, 1987), Math. Sci. Res. Inst. Publ. 16, 395–437, Springer, New York-Berlin, 1989.

[Pa] V. Paskunas *Unicity of types for supercuspidal representations of* GL_N, Proc. LMS (3) 91, 623–654, 2005.

[Ra] R. Ramakrishna, *On a variation of Mazur's deformation functor,* Compositio Math. 87, 269–286, 1993.

[Ta] R. Taylor, *On Galois representations associated to Hilbert modular forms,* Invent. Math. 98, 265–280, 1989.

[Wi] A. Wiles, *Modular elliptic curves and Fermat's last theorem,* Ann. of Math. 141(3), 443–551, 1995.

[TW] R. Taylor, A. Wiles, *Ring theoretic properties of certain Hecke algebras,* Ann. of Math. 141(3), 553–572, 1995.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

The Intersection Complex as a Weight Truncation and an Application to Shimura Varieties

Sophie Morel*

Abstract

The purpose of this talk is to present an (apparently) new way to look at the intersection complex of a singular variety over a finite field, or, more generally, at the intermediate extension functor on pure perverse sheaves, and an application of this to the cohomology of noncompact Shimura varieties.

Mathematics Subject Classification (2010). Primary 11F75; Secondary 11G18, 14F20.

Keywords. Shimura varieties, intersection cohomology, Frobenius weights

1. Shimura Varieties

1.1. The complex points. In their simplest form, Shimura varieties are just locally symmetric varieties associated to certain connected reductive groups over $\mathbb{Q}$. So let $\mathbf{G}$ be a connected reductive group over $\mathbb{Q}$ satisfying the conditions in 1.5 of Deligne's article [17]. To be precise, we are actually fixing $\mathbf{G}$ and a morphism $h : \mathbb{C}^\times \longrightarrow \mathbf{G}(\mathbb{R})$ that is algebraic over $\mathbb{R}$. Let us just remark here that these conditions are quite restrictive. For example, they exclude the group $\mathbf{GL}_n$ as soon as $n \geq 3$. The groups $\mathbf{G}$ that we want to think about are, for example, the group $\mathbf{GSp}_{2n}$ (the general symplectic group of a symplectic space of dimension $2n$ over $\mathbb{Q}$) or the general unitary group of a hermitian space over a quadratic imaginary extension of $\mathbb{Q}$. The conditions on $\mathbf{G}$ ensure that the symmetric space $\mathcal{X}$ of $\mathbf{G}(\mathbb{R})$ is a hermitian symmetric domain; so $\mathcal{X}$ has a

*This text was written while I was working as a Professor at the Harvard mathematics department and supported by the Clay Mathematics Institute as a Clay Research Fellow. I would like to thank the referee for their useful comments about the first version of this paper.

Department of Mathematics, Harvard University, One Oxford Street, Cambridge, MA 02138, USA. E-mail: morel@math.harvard.edu.

canonical complex structure. Remember that $\mathcal{X} = \mathbf{G}(\mathbb{R})/\mathrm{K}'_\infty$, where K'_∞ is the centralizer in $\mathbf{G}(\mathbb{R})$ of $h(\mathbb{C}^\times)$. In the examples we consider, K'_∞ is the product of a maximal compact subgroup K_∞ of $\mathbf{G}(\mathbb{R})$ and of $\mathrm{A}_\infty := \mathbf{A}(\mathbb{R})^0$, where $\mathbf{A}$ is the maximal $\mathbb{Q}$-split torus of the center of $\mathbf{G}$. (To avoid technicalities, many authors assume that the maximal $\mathbb{R}$-split torus in the center of $\mathbf{G}$ is also $\mathbb{Q}$-split. We will do so too.)

The locally symmetric spaces associated to $\mathbf{G}$ are the quotients $\Gamma \setminus \mathbf{G}(\mathbb{R})$, where Γ is an *arithmetic subgroup* of $\mathbf{G}(\mathbb{Q})$, that is, a subgroup of $\mathbf{G}(\mathbb{Q})$ such that, for some (or any) $\mathbb{Z}$-structure on $\mathbf{G}$, $\Gamma \cap \mathbf{G}(\mathbb{Z})$ is of finite index in Γ and in $\mathbf{G}(\mathbb{Z})$. If Γ is small enough (for example, if it is torsion-free), then $\Gamma \setminus \mathcal{X}$ is a smooth complex analytic variety. In fact, by the work of Baily and Borel ([4]), it is even a quasi-projective algebraic variety.

In this text, we prefer to use the adelic point of view, as it leads to somewhat simpler statements. So let K be a compact open subgroup of $\mathbf{G}(\mathbb{A}_f)$, where $\mathbb{A}_f = \widehat{\mathbb{Z}} \otimes_{\mathbb{Z}} \mathbb{Q}$ is the ring of finite adeles of $\mathbb{Q}$. This means that K is a subgroup of $\mathbf{G}(\mathbb{A}_f)$ such that, for some (or any) $\mathbb{Z}$-structure on $\mathbf{G}$, $\mathrm{K} \cap \mathbf{G}(\widehat{\mathbb{Z}})$ is of finite index in K and in $\mathbf{G}(\widehat{\mathbb{Z}})$. Set

$$S^{\mathrm{K}}(\mathbb{C}) = \mathbf{G}(\mathbb{Q}) \setminus (\mathcal{X} \times \mathbf{G}(\mathbb{A}_f)/\mathrm{K}),$$

where $\mathbf{G}(\mathbb{Q})$ acts on $\mathcal{X} \times \mathbf{G}(\mathbb{A}_f)/\mathrm{K}$ by the formula $(\gamma, (x, g\mathrm{K})) \longmapsto (\gamma \cdot x, \gamma g \mathrm{K})$.

This space $S^{\mathrm{K}}(\mathbb{C})$ is related to the previous quotients $\Gamma \setminus \mathcal{X}$ in the following way. By the strong approximation theorem, $\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A}_f)/\mathrm{K}$ is finite. Let $(g_i)_{i \in I}$ be a finite family in $\mathbf{G}(\mathbb{A}_f)$ such that $\mathbf{G}(\mathbb{A}_f) = \coprod_{i \in I} \mathbf{G}(\mathbb{Q}) g_i \mathrm{K}$. For every $i \in I$, set $\Gamma_i = \mathbf{G}(\mathbb{Q}) \cap g_i \mathrm{K} g_i^{-1}$. Then the Γ_i are arithmetic subgroups of $\mathbf{G}(\mathbb{Q})$, and

$$S^{\mathrm{K}}(\mathbb{C}) = \coprod_{i \in I} \Gamma_i \setminus \mathcal{X}.$$

In particular, we see that, if K is small enough, then $S^{\mathrm{K}}(\mathbb{C})$ is the set of complex points of a smooth quasi-projective complex algebraic variety, that we will denote by S^{K}. These are the *Shimura varieties* associated to $\mathbf{G}$ and $h : C^\times \longrightarrow \mathbf{G}(\mathbb{R})$ (over $\mathbb{C}$). From now on, we will assume always that the group K is small enough.

Remark 1. If $\mathbf{G} = \mathbf{GL}_2$, then S^{K} is a modular curve, or rather, a finite disjoint union of modular curves; it parametrizes elliptic curves with a certain level structure (depending on K). Higher-dimensional generalizations of this are the Shimura varieties for the symplectic groups $\mathbf{G} = \mathbf{GSp}_{2n}$; they are called the Siegel modular varieties, and parametrize principally polarized abelian varieties with a level structure (depending on K). Some other Shimura varieties have been given a name. For example, if $\mathbf{G}$ is the general unitary group of a 3-dimensional hermitian vector space V over an imaginary quadratic extension of $\mathbb{Q}$ such that V has signature $(2, 1)$ at infinity, then S^{K} is called a Picard modular surface.

1.2. The projective system and Hecke operators. If $\mathrm{K}' \subset \mathrm{K}$ are two open compact subgroups of $\mathbf{G}(\mathbb{A}_f)$, then there is an obvious projection $S^{\mathrm{K}'}(\mathbb{C}) \longrightarrow S^{\mathrm{K}}(\mathbb{C})$, and it defines a finite étale morphism $S^{\mathrm{K}'} \longrightarrow S^{\mathrm{K}}$; if K' is normal in K, then this morphism is Galois, with Galois group K/K'. So we can see the Shimura varieties S^{K} as a projective system $(S^{\mathrm{K}})_{\mathrm{K}\subset\mathbf{G}(\mathbb{A}_f)}$ indexed by (small enough) open compact subgroups of $\mathbf{G}(\mathbb{A}_f)$, and admitting a right continuous action of $\mathbf{G}(\mathbb{A}_f)$.

More generally, if K', K are two open compact subgroups of $\mathbf{G}(\mathbb{A}_f)$ and $g \in \mathbf{G}(\mathbb{A}_f)$, then we get a correspondence $[\mathrm{K}'g\mathrm{K}] : S^{\mathrm{K}\cap g^{-1}\mathrm{K}'g} \longrightarrow S^{\mathrm{K}} \times S^{\mathrm{K}'}$ in the following way. The first map is the obvious projection $S^{\mathrm{K}\cap g^{-1}\mathrm{K}'g} \longrightarrow S^{\mathrm{K}}$, and the second map is the composition of the obvious projection $S^{\mathrm{K}\cap g^{-1}\mathrm{K}'g} \longrightarrow S^{g^{-1}\mathrm{K}'g}$ and of the isomorphism $S^{g^{-1}\mathrm{K}'g} \xrightarrow{\sim} S^{\mathrm{K}'}$. This is the *Hecke correspondence* associated to g (and K, K').

Let H^* be a cohomology theory with coefficients in a ring A that has good fonctoriality properties (for example, Betti cohomology with coefficients in A) and K be an open compact subgroup of $\mathbf{G}(\mathbb{A}_f)$. Then the Hecke correspondences define an action of the Hecke algebra at level K, $\mathcal{H}_{\mathrm{K}}(A) := C(\mathrm{K}\setminus\mathbf{G}(\mathbb{A}_f)/\mathrm{K}, A)$ (of bi-K-invariant functions from $\mathbf{G}(\mathbb{A}_f)$ to A, with the algebra structure given by the convolution product), on the cohomology $\mathrm{H}^*(S^{\mathrm{K}})$. For every $g \in \mathbf{G}(\mathbb{A}_f)$, we make $\mathbb{1}_{\mathrm{K}g\mathrm{K}} \in \mathcal{H}_{\mathrm{K}}(A)$ act by the correspondence $[\mathrm{K}g^{-1}\mathrm{K}]$.

Let $\mathcal{H}(A) = \bigcup_{\mathrm{K}} \mathcal{H}_{\mathrm{K}}(A) = C_c^\infty(\mathbf{G}(\mathbb{A}_f), A)$ (the algebra of locally constant functions $\mathbf{G}(\mathbb{A}_f) \longrightarrow A$ with compact support) be the full Hecke algebra, still with the product given by convolution. Then we get an action of $\mathcal{H}(A)$ on the limit $\varinjlim_{\mathrm{K}} \mathrm{H}^*(S^{\mathrm{K}})$. So the A-module $\varinjlim_{\mathrm{K}} \mathrm{H}^*(S^{\mathrm{K}})$ admits an action of the group $\mathbf{G}(\mathbb{A}_f)$.

1.3. Canonical models. Another feature of Shimura varieties is that they have so-called *canonical models*. That is, they are canonically defined over a number field E, called the *reflex field*, that depends only on $\mathbf{G}$ and the morphism $h : \mathbb{C}^\times \longrightarrow \mathbf{G}(\mathbb{R})$ (in particular, it does not depend on the open compact subgroup K of $\mathbf{G}(\mathbb{A}_f)$). We will use the same notation S^{K} for the model over E. Here "canonically" means in particular that the action of $\mathbf{G}(\mathbb{A}_f)$ on the projective system $(S^{\mathrm{K}})_{\mathrm{K}}$ is defined over E. The theory of canonical models was begun by Shimura, and then continued by Deligne, Borovoi, Milne and Moonen (cf [17], [18], [13], [46], [47], [51]).

So, if the cohomology theory H^* happens to make sense for varieties over E (for example, it could be ℓ-adic étale cohomology, with or without supports), then the limit $\varinjlim_{\mathrm{K}} \mathrm{H}^*(S^{\mathrm{K}})$ admits commuting actions of $\mathbf{G}(\mathbb{A}_f)$ and of $\mathrm{Gal}(\overline{E}/E)$. Another way to look at this is to say that the cohomology group at finite level, $\mathrm{H}^*(S^{\mathrm{K}})$, admits commuting actions of $\mathcal{H}_{\mathrm{K}}(A)$ and of $\mathrm{Gal}(\overline{E}/E)$.

The goal is now to understand the decomposition of those cohomology groups as representations of $\mathbf{G}(\mathbb{A}_f) \times \mathrm{Gal}(\overline{E}/E)$ (or of $\mathcal{H}_{\mathrm{K}}(A) \times \mathrm{Gal}(\overline{E}/E)$).

1.4. Compactifications and the choice of cohomology theory. If the Shimura varieties S^{K} are projective, which happens if and only if the group $\mathbf{G}$ is anisotropic over $\mathbb{Q}$, then the most natural choice of cohomology theory is simply the étale cohomology of S^{K}. There is still the question of the coefficient group A. While the study of cohomology with torsion or integral coefficients is also interesting, very little is known about it at this point, so we will restrict ourselves to the case $A = \overline{\mathbb{Q}}_\ell$, where ℓ is some prime number.

Things get a little more complicated when the S^{K} are not projective, and this is the case we are most interested in here. We can still use ordinary étale cohomology or étale cohomology with compact support, but it becomes much harder to study (among other things, because we do not have Poincaré duality or the fact that the cohomology is pure - in Deligne's sense - any more). Nonetheless, it is still an interesting problem.

Another solution is to use a cohomology theory on a compactification of S^{K}. The author of this article knows of two compactifications of S^{K} as an algebraic variety over E (there are many, many compactifications of $S^{\mathrm{K}}(\mathbb{C})$ as a topological space, see for example the book [11] of Borel and Ji):

(1) The *toroidal compactifications*. They are a family of compactifications of S^{K}, depending on some combinatorial data (that depends on K); they can be chosen to be very nice (i.e. projective smooth and with a boundary that is a divisor with normal crossings).

(2) The *Baily-Borel (or minimal Satake, or Satake-Baily-Borel) compactification* $\overline{S}^{\mathrm{K}}$. It is a canonical compactification of S^{K}, and is a projective normal variety over E, but it is very singular in general.

See the book [3] by Ash, Mumford, Rapoport and Tai for the construction of the toroidal compactifications over $\mathbb{C}$, the article [4] of Baily and Borel for the construction of the Baily-Borel compactification over $\mathbb{C}$, and Pink's dissertation [55] for the models over E of the compactifications.

The problem of using a cohomology theory on a toroidal compactification is that the toroidal compactifications are not canonical, so it is not easy to make the Hecke operators act on their cohomology. On the other hand, while the Baily-Borel compactification is canonical (so the Hecke operators extend to it), it is singular, so its cohomology does not behave well in general. One solution is to use the intersection cohomology (or homology) of the Baily-Borel compactification. In the next section, we say a little more about intersection homology, and explain why it might be a good choice.

2. Intersection Homology and L^2 Cohomology

2.1. Intersection homology. Intersection homology was invented by Goresky and MacPherson to study the topology of singular spaces (cf [24], [25]). Let X be a complex algebraic (or analytic) variety of pure dimension n,

possibly singular. Then the singular homology groups of X (say with coefficients in $\mathbb{Q}$) do not satisfy Poincaré duality if X is not smooth. To fix this, Goresky and MacPherson modify the definition of singular homology in the following way. First, note that X admits a Whitney stratification, that is, a locally finite decomposition into disjoint connected smooth subvarieties $(S_i)_{i\in I}$ satisfying the Whitney condition (cf [24] 5.3). For every $i \in I$, let $c_i = n - \dim(S_i)$ be the (complex) codimension of S_i. Let $(C_k(X))_{k\in\mathbb{Z}}$ be the complex of simplicial chains on X with coefficients in a commutative ring A. The *complex of intersection chains* $(IC_k(X))_{k\in\mathbb{Z}}$ is the subcomplex of $(C_k(X))_{k\in Z}$ consisting of chains $c \in C_k(X)$ satisfying the allowability condition: For every $i \in I$, the real dimension of $c \cap S_i$ is less than $k - c_i$, and the real dimension of $\partial c \cap S_i$ is less than $k - 1 - c_i$. The *intersection homology groups* $\mathrm{IH}_k(X)$ of X are the homology groups of $(IC_k(X))_{k\in\mathbb{Z}}$. (Note that this is the definition of middle-perversity intersection homology. We can get other interesting intersection homology groups of X by playing with the bounds in the definition of intersection chains, but they will not satisfy Poincaré duality.)

Intersection homology groups satisfy many of the properties of ordinary singular homology groups $\mathrm{H}_k(X)$ on smooth varieties. Here are a few of these properties:

- They depend only on X, and not on the stratification $(S_i)_{i\in I}$.
- If X is smooth, then $\mathrm{IH}_k(X) = \mathrm{H}_k(X)$.
- If X is compact, then the $\mathrm{IH}_k(X)$ are finitely generated.
- If the coefficients A are a field, the intersection homology groups satisfy the Künneth theorem.
- If $U \subset X$ is open, then there are relative intersection homology groups $\mathrm{IH}_k(X, U)$ and an excision long exact sequence.
- It is possible to define an intersection product on intersection homology, and, if X is compact and A is a field, this will induce a nondegenerate linear pairing
$$\mathrm{IH}_k(X) \times \mathrm{IH}_{2n-k}(X) \longrightarrow A.$$
(I.e., there is a Poincaré duality theorem for intersection homology.)
- Intersection homology satisfies the Lefschetz hyperplane theorem and the hard Lefschetz theorem (if A is a field for hard Lefschetz).

Note however that the intersection homology groups are not homotopy invariants (though they are functorial for certain maps of varieties, called placid maps).

2.2. L^2 cohomology of Shimura varieties and intersection homology. Consider again a Shimura variety $S^{\mathrm{K}}(\mathbb{C})$ as in section 1 (or rather, the complex manifold of its complex points). For every $k \geq 0$, we write $\Omega^k_{(2)}(S^{\mathrm{K}}(\mathbb{C}))$ for the space of smooth forms ω on $S^{\mathrm{K}}(\mathbb{C})$ such that ω and $d\omega$ are L^2. The L^2 cohomology groups $\mathrm{H}^*_{(2)}(S^{\mathrm{K}}(\mathbb{C}))$ of $S^{\mathrm{K}}(\mathbb{C})$ are the cohomology groups of the complex $\Omega^*_{(2)}$. These groups are known to be finite-dimensional and to satisfy Poincaré duality, and in fact we have the following theorem (remember that $\overline{S}^{\mathrm{K}}$ is the Baily-Borel compactification of S^{K}):

Theorem 2.1. *There are isomorphisms*

$$\mathrm{H}^k_{(2)}(S^{\mathrm{K}}(\mathbb{C})) \simeq \mathrm{IH}_{2d-k}(\overline{S}^{\mathrm{K}}(\mathbb{C}), \mathbb{R}),$$

where $d = \dim(S^{\mathrm{K}})$. Moreover, these isomorphisms are equivariant under the action of $\mathcal{H}_{\mathrm{K}}(\mathbb{R})$. (The Hecke algebra acts on intersection homology because the Hecke correspondences extend to the Baily-Borel compactifications and are still finite, hence placid.)

This was conjectured by Zucker in [67], and then proved (independently) by Looijenga ([44]), Saper-Stern ([61]) and Looijenga-Rapoport ([45]).

So now we have some things in favour of intersection homology of the Baily-Borel compactification: it satisfies Poincaré duality and is isomorphic to a natural invariant of the Shimura variety. We will now see another reason why L^2 cohomology of Shimura varieties (hence, intersection homology of their Baily-Borel compactification) is easier to study than ordinary cohomology: it is closely related to automorphic representations of the group $\mathbf{G}$. (Ordinary cohomology of Shimura varieties, or cohomology with compact support, is also related to automorphic representations, but in a much more complicated way, see the article [22] of Franke.)

2.3. L^2 cohomology of Shimura varieties and discrete automorphic representations. For an introduction to automorphic forms, we refer to the article [10] of Borel and Jacquet and the article [54] of Piatetski-Shapiro. Let $\mathbb{A} = \mathbb{A}_f \times \mathbb{R}$ be the ring of adeles of $\mathbb{Q}$. Very roughly, an *automorphic form* on $\mathbf{G}$ is a smooth function $f : \mathbf{G}(\mathbb{A}) \longrightarrow \mathbb{C}$, left invariant under $\mathbf{G}(\mathbb{Q})$, right invariant under some open compact subgroup of $\mathbf{G}(\mathbb{A}_f)$, K_∞-finite on the right (i.e., such that the right translates of f by elements of K_∞ generate a finite dimensional vector space; remember that K_∞ is a maximal compact subgroup of $\mathbf{G}(\mathbb{R})$) and satisfying certain growth conditions. The group $\mathbf{G}(\mathbb{A})$ acts on the space of automorphic forms by right translations on the argument. Actually, we are cheating a bit here. The group $\mathbf{G}(\mathbb{A}_f)$ does act that way, but $\mathbf{G}(\mathbb{R})$ does not; the space of automorphic forms is really a Harish-Chandra $(\mathfrak{g}, \mathrm{K}_\infty)$-module, where $\mathfrak{g}$ is the Lie algebra of $\mathbf{G}(\mathbb{C})$. An *automorphic representation* of $\mathbf{G}(\mathbb{A})$ (or, really, $\mathbf{G}(\mathbb{A}_f) \times (\mathfrak{g}, \mathrm{K}_\infty)$) is an irreducible representation that appears in the space of automorphic forms as an irreducible subquotient.

Note that there is also a classical point of view on automorphic forms, where they are seen as smooth functions on $\mathbf{G}(\mathbb{R})$, left invariant by some arithmetic subgroup of $\mathbf{G}(\mathbb{Q})$, K_∞-finite on the right and satisfying a growth condition. From that point of view, it may be easier to see that automorphic forms generalize classical modular forms (for modular forms, the group $\mathbf{G}$ is $\mathbf{GL}_2$). The two points of view are closely related, cf. [10] 4.3 (in much the same way that the classical and adelic points of view on Shimura varieties are related). In this article, we adopt the adelic point of view, because it makes it easier to see the action of Hecke operators.

Actually, as we are interested only in discrete automorphic representations (see below for a definition), we can see automorphic forms as L^2 functions on $\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A})$. We follow Arthur's presentation in [1]. First, a word of warning: the quotient $\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A})$ does not have finite volume. This is due to the presence of factors isomorphic to $\mathbb{R}_{>0}$ in the center of $\mathbf{G}(\mathbb{R})$. As in 1.1, let $\mathrm{A}_\infty = \mathbf{A}(\mathbb{R})^0$, where $\mathbf{A}$ is the maximal $\mathbb{R}$-split torus in the center of $\mathbf{G}$. Then $\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A})/\mathrm{A}_\infty$ does have finite volume, and we will consider L^2 functions on this quotient, instead of $\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A})$.

So let $\xi : \mathrm{A}_\infty \longrightarrow \mathbb{C}^\times$ be a character (not necessarily unitary). Then ξ extends to a character $\mathbf{G}(\mathbb{A}) \longrightarrow \mathbb{C}^\times$, that we will still denote by ξ (cf. I.3 of Arthur's introduction to the trace formula, [2]). Let $L^2(\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A}), \xi)$ be the space of measurable functions $f : \mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A}) \longrightarrow \mathbb{C}$ such that:

(1) for every $z \in \mathrm{A}_\infty$ and $g \in \mathbf{G}(\mathbb{A})$, $f(zg) = \xi(z)f(g)$;

(2) the function $\xi^{-1}f$ is square-integrable on $\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A})/\mathrm{A}_\infty$.

Then the group $\mathbf{G}(\mathbb{A})$ acts on $L^2(\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A}), \xi)$ by right translations on the argument. By definition, a *discrete automorphic representation* of $\mathbf{G}$ is an irreducible representation of $\mathbf{G}(\mathbb{A})$ that appears as a direct summand in $L^2(\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A}), \xi)$. It is known that the multiplicity of a discrete automorphic representation π in $L^2(\mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A}), \xi)$ is always finite; we denote it by $m(\pi)$. We also denote by $\Pi_{disc}(\mathbf{G}, \xi)$ the set of discrete automorphic representations on which A_∞ acts by ξ. For the fact that discrete automorphic representations are indeed automorphic representations in the previous sense, see [10] 4.6. (The attentive reader will have noted that automorphic representations are not actual representations of $\mathbf{G}(\mathbb{A})$ - because $\mathbf{G}(\mathbb{R})$ does not act on them - while discrete automorphic representations are. How to make sense of our statement that discrete automorphic representations are automorphic is also explained in [10] 4.6.)

Now, given the definition of discrete automorphic representations and the fact that $S^{\mathrm{K}}(\mathbb{C}) = \mathbf{G}(\mathbb{Q}) \setminus \mathbf{G}(\mathbb{A})/(\mathrm{A}_\infty \mathrm{K}_\infty \times \mathrm{K})$, it is not too surprising that the L^2 cohomology of the Shimura variety $S^{\mathrm{K}}(\mathbb{C})$ should be related to discrete automorphic representations. Here is the precise relation:

Theorem 2.2. *(Borel-Casselman, cf. [9] theorem 4.5) Let* K *be an open compact subgroup of* $\mathbf{G}(\mathbb{A}_f)$*. Then there is a* $\mathcal{H}_{\mathrm{K}}(\mathbb{C})$*-equivariant isomorphism*

$$\mathrm{H}^*_{(2)}(S^{\mathrm{K}}(\mathbb{C})) \otimes_{\mathbb{R}} \mathbb{C} \simeq \bigoplus_{\pi \in \Pi_{disc}(\mathbf{G},1)} \mathrm{H}^*(\mathfrak{g}, \mathrm{A}_\infty \mathrm{K}_\infty; \pi_\infty)^{m(\pi)} \otimes \pi_f^{\mathrm{K}}.$$

(This is often called Matsushima's formula when $S^{\mathrm{K}}(\mathbb{C})$ is compact.)

We need to explain the notation. First, the "1" in $\Pi_{disc}(\mathbf{G}, 1)$ stands for the trivial character of A_∞. (We have chosen to work with the constant sheaf on S^{K}, in order to simplify the notation. In general, for a non-trivial coefficient system on $S^{\mathrm{K}}(\mathbb{C})$, other characters of A_∞ would appear.) Let $\pi \in \Pi_{disc}(\mathbf{G}, 1)$. Then π is an irreducible representation of $\mathbf{G}(\mathbb{A}) = \mathbf{G}(\mathbb{R}) \times \mathbf{G}(\mathbb{A}_f)$ so it decomposes as a tensor product $\pi_\infty \otimes \pi_f$, where π_∞ (resp. π_f) is an irreducible representation of $\mathbf{G}(\mathbb{R})$ (resp. $\mathbf{G}(\mathbb{A}_f)$). We denote by π_f^{K} the space of K-invariant vectors in the space of π_f; it carries an action of the Hecke algebra $\mathcal{H}_{\mathrm{K}}(\mathbb{C})$. Finally, $\mathrm{H}^*(\mathfrak{g}, \mathrm{A}_\infty \mathrm{K}_\infty; \pi_\infty)$, the $(\mathfrak{g}, \mathrm{A}_\infty \mathrm{K}_\infty)$-cohomology of π_∞ (where $\mathfrak{g}$ is as before the Lie algebra of $\mathbf{G}(\mathbb{C})$), is defined in chapter I of the book [12] by Borel and Wallach.

This gives another reason to study the intersection homology of the Baily-Borel compactifications of Shimura varieties: it will give a lot of information about discrete automorphic representations of $\mathbf{G}$. (Even if only about the ones whose infinite part has nontrivial $(\mathfrak{g}, \mathrm{A}_\infty \mathrm{K}_\infty)$-cohomology, and that is a pretty strong condition.)

Note that there is an issue we have been avoiding until now. Namely, in 1.3, we wanted the cohomology theory on the Shimura variety to also have an action of $\mathrm{Gal}(\overline{E}/E)$, where E is the reflex field (i.e., the field over which the varieties S^{K} have canonical models). It is not clear how to endow the L^2 cohomology of $S^{\mathrm{K}}(\mathbb{C})$ with such an action. As we will see in the next section, this will come from the isomorphism of $\mathrm{H}^*_{(2)}(S^{\mathrm{K}}(\mathbb{C}))$ with the intersection homology of $\overline{S}^{\mathrm{K}}(\mathbb{C})$ and from the sheaf-theoretic interpretation of intersection homology (because this interpretation will also make sense in an étale ℓ-adic setting).

3. Intersection (Co)Homology and Perverse Sheaves

We use again the notation of section 2.

3.1. The sheaf-theoretic point of view on intersection homology. Intersection homology of X also has a sheaf-theoretical interpretation. (At this point, we follow Goresky and MacPherson and shift from the homological to the cohomological numbering convention.) For every open U in X, let $\mathbf{IC}^k(U)$ be the group of $(2n-k)$-dimensional intersection chains on U with closed support. If $U' \subset U$, then we have a map $\mathbf{IC}^k(U) \longrightarrow \mathbf{IC}^k(U')$

given by restriction of chains. In this way, we get a sheaf $\mathbf{IC}^k$ on X. Moreover, the boundary maps of the complex of intersection chains give maps of sheaves $\delta : \mathbf{IC}^k \longrightarrow \mathbf{IC}^{k+1}$ such that $\delta \circ \delta = 0$, so the $\mathbf{IC}^k$ form a complex of sheaves $\mathbf{IC}^*$ on X. This is the *intersection complex* of X. Its cohomology with compact support gives back the intersection homology groups of X:

$$\mathrm{H}^k_c(X, \mathbf{IC}^*(X)) = \mathrm{IH}_{2n-k}(X).$$

Its cohomology groups $\mathrm{IH}^k(X) := \mathrm{H}^k(X, \mathbf{IC}^*(X))$ are (by definition) the *intersection cohomology groups* of X.

3.2. Perverse sheaves. This point of view has been extended and generalized by the invention of perverse sheaves. The author's favourite reference for perverse sheaves is the book by Beilinson, Bernstein and Deligne ([6]).

To simplify, assume that the ring of coefficients A is a field. Let $D(X)$ be the derived category of the category of sheaves on X. This category is obtained from the category of complexes of sheaves on X by introducing formal inverses of all the quasi-isomorphisms, i.e. of all the morphisms of complexes that induce isomorphisms on the cohomology sheaves. (This is a categorical analogue of a ring localization.) Note that the objects of $D(X)$ are still the complexes of sheaves, we just added more morphisms. The homological functors on the category of complexes of sheaves (such as the various cohomology functors and the Ext and Tor functors) give functors on $D(X)$, and a morphism in $D(X)$ is an isomorphism if and only if it is an isomorphism on the cohomology sheaves.

This category $D(X)$ is still a little big, and we will work with the full subcategory $D^b_c(X)$ of bounded constructible complexes. If C^* is a complex of sheaves, we will denote its cohomology sheaves by $\mathrm{H}^k C^*$. Then C^* is called *bounded* if $\mathrm{H}^k C^* = 0$ for $k << 0$ and $k >> 0$. It is called *constructible* if its cohomology sheaves $\mathrm{H}^k C^*$ are constructible, that is, if, for every $k \in \mathbb{Z}$, there exists a stratification $(S_i)_{i \in I}$ of X (by smooth subvarieties) such that $\mathrm{H}^k C^*_{|S_i}$ is locally constant and finitely generated for every i.

For every point x of X, we denote by i_x the inclusion of x in X.

Definition 1. A complex of sheaves C^* in $D^b_c(X)$ is called a *perverse sheaf* if it satisfies the following support and cosuport conditions:

(1) Support: for every $k \in \mathbb{Z}$,

$$\dim_{\mathbb{C}}\{x \in X | \mathrm{H}^k(i_x^* C^*) \neq 0\} \leq -k.$$

(2) Cosupport: for every $k \in \mathbb{Z}$,

$$\dim_{\mathbb{C}}\{x \in X | \mathrm{H}^k(i_x^! C^*) \neq 0\} \leq k.$$

We denote by $P(X)$ the category of perverse sheaves on X.

Remark 2. Let $x \in X$. There is another way to look at the groups $i_x^* \mathrm{H}^k C^*$ and $i_x^! \mathrm{H}^k C^*$. Choose an (algebraic or analytic) embedding of a neighbourhood of x into an affine space $\mathbb{C}^p$, and let B_x denote the intersectioon of this neighbourhood and of a small enough open ball in $\mathbb{C}^p$ centered at x. Then

$$\mathrm{H}^k(i_x^* C^*) = \mathrm{H}^k(B_x, C^*)$$

$$\mathrm{H}^k(i_x^! C^*) = \mathrm{H}_c^k(B_x, C^*).$$

Remark 3. As before, we are only considering one perversity, the middle (or self-dual) perversity. For other perversities (and much more), see [6].

Note that perverse sheaves are not sheaves but complexes of sheaves. However, the category of perverse sheaves satisfies many properties that we expect from a category of sheaves, and that are not true for $D_c^b(X)$ (or $D(X)$). For example, $P(X)$ is an abelian category, and it is possible to glue morphisms of perverse sheaves (more precisely, categories of perverse sheaves form a stack, say on the open subsets of X, cf. [6] 2.1.23).

3.3. Intermediate extensions and the intersection complex.

Now we explain the relationship with the intersection complex. First, the intersection complex is a perverse sheaf on X once we put it in the right degree. In fact:

Proposition 3.1. *The intersection complex $\mathbf{IC}^*(X)$ is an object of $D_c^b(X)$ (i.e., it is a bounded complex with constructible cohomology sheaves), and:*

(1) For every $k \neq 0$,

$$\dim_{\mathbb{C}}\{x \in X | \mathrm{H}^k(i_x^* \mathbf{IC}^*(X)) \neq 0\} < n - k.$$

(2) For every $k \neq 2n$,

$$\dim_{\mathbb{C}}\{x \in X | \mathrm{H}^k(i_x^! \mathbf{IC}^*(X)) \neq 0\} < k - n.$$

(3) If U is a smooth open dense subset of X, then $\mathbf{IC}^*(X)_{|U}$ is quasi-isomorphic (i.e., isomorphic in $D_c^b(X)$) to the constant sheaf on U.

Moreover, the intersection complex is uniquely characterized by these properties (up to unique isomorphism in $D_c^b(X)$).

In particular, $\mathbf{IC}^*(X)[n]$ (that is, the intersection complex put in degree $-n$) is a perverse sheaf on X.

Even better, it turns out that every perverse sheaf on X is, in some sense, built from intersection complexes on closed subvarieties of X. Let us be more precise. Let $j : X \longrightarrow Y$ be a locally closed immersion. Then there is a functor $j_{!*} : P(X) \longrightarrow P(Y)$, called the *intermediate extension functor*, such that, for

every perverse sheaf K on X, the perverse sheaf $j_{!*}K$ on Y is uniquely (up to unique quasi-isomorphism) characterized by the following conditions:

(1) For every $k \in \mathbb{Z}$,

$$\dim_{\mathbb{C}}\{x \in Y - X | \mathrm{H}^k(i_x^* j_{!*}K)) \neq 0\} < -k.$$

(2) For every $k \in \mathbb{Z}$,

$$\dim_{\mathbb{C}}\{x \in Y - X | \mathrm{H}^k(i_x^! j_{!*}K) \neq 0\} < k.$$

(3) $j^* j_{!*}K = K$.

Remark 4. Let us explain briefly the name "intermediate extension". Although it is not clear from the way we defined perverse sheaves, there are "perverse cohomology" functors ${}^p\mathrm{H}^k : D_c^b(X) \longrightarrow P(X)$. In fact, it even turns out that $D_c^b(X)$ is equivalent to the derived category of the abelian category of perverse sheaves (this is a result of Beilinson, cf. [5]). We can use these cohomology functors to define perverse extension functors ${}^p j_!$ and ${}^p j_*$ from $P(X)$ to $P(Y)$. (For example, ${}^p j_! = {}^p\mathrm{H}^0 j_!$, where $j_! : D_c^b(X) \longrightarrow D_c^b(Y)$ is the "extension by zero" functor between the derived categories; likewise for ${}^p j_*$). It turns out that, from the perverse point of view, the functor $j_! : D_c^b(Y) \longrightarrow D_c^b(X)$ is right exact and the functor $j_* : D_c^b(Y) \longrightarrow D_c^b(X)$ is left exact (that, if K is perverse on X, ${}^p\mathrm{H}^k j_! K = 0$ for $k > 0$ and ${}^p\mathrm{H}^k j_* K = 0$ for $k < 0$). So the morphism of functors $j_! \longrightarrow j_*$ induces a morphism of functors ${}^p j_! \longrightarrow {}^p j_*$. For every perverse sheaf K on X, we have:

$$j_{!*}K = \mathrm{Im}({}^p j_! K \longrightarrow {}^p j_* K).$$

Now we come back to the description of the category of perverse sheaves on X. Let F be a smooth connected locally closed subvariety of X, and denote by i_F its inclusion in X. If $\mathcal{F}$ is a locally constant sheaf on F, then it is easy to see that $\mathcal{F}[\dim F]$ is a perverse sheaf on F; so $i_{F!*}\mathcal{F}[\dim F]$ is a perverse sheaf on X (it has support in $\overline{F}$, where $\overline{F}$ is the closure of F in X). If the locally constant sheaf $\mathcal{F}$ happens to be irreducible, then this perverse sheaf is a simple object in $P(X)$. In fact:

Theorem 3.2. *The abelian category $P(X)$ is artinian and noetherian (i.e., every object has finite length), and its simple objects are all of the form $i_{F!*}\mathcal{F}[\dim F]$, where F is as above and $\mathcal{F}$ is an irreduible locally constant sheaf on F.*

Finally, here is the relationship with the intersection complex. Let $i_F : F \longrightarrow X$ be as above. Then, if $\mathcal{F}$ is the constant sheaf on F, the restriction to $\overline{F}$ of the perverse sheaf $i_{F!*}\mathcal{F}[\dim F]$ is isomorphic to $\mathbf{IC}^*(\overline{F})[\dim F]$. In fact, we could define the intersection complex on a (possibly singular) variety Y with coefficients in some locally constant sheaf on the smooth locus of Y, and then the simple objects in $P(X)$ would all be intersection complexes on closed subvarieties of X.

3.4. ℓ-adic perverse sheaves. Now we come at last to the point of this section (to make the Galois groups $\mathrm{Gal}(\overline{E}/E)$ act on the intersection (co)homology of $\overline{S}^{\mathrm{K}}(\mathbb{C})$).

Note that the definitions of the category of perverse sheaves and of the intermediate extension in 3.2 and 3.3 would work just as well in a category of étale ℓ-adic sheaves. So now we take for X a quasi-separated scheme of finite type over a field k, we fix a prime number ℓ invertible in k and we consider the category $D_c^b(X, \overline{\mathbb{Q}}_\ell)$ of bounded ℓ-adic complexes on X. (To avoid a headache, we will take k to be algebraically closed or finite, so the simple construction of [6] 2.2.14 applies.) Then we can define an abelian subcategory of perverse sheaves $P(X)$ in $D_c^b(X, \overline{\mathbb{Q}}_\ell)$ and intermediate extension functors $j_{!*} : P(X) \longrightarrow P(Y)$ as before (see [6] 2.2). In particular, we can make the following definition:

Definition 2. Suppose that X is purely of dimension n, and let $j : U \longrightarrow X$ be the inclusion of the smooth locus of X in X. Then the (ℓ-adic) intersection complex of X is

$$\mathbf{IC}^*(X) = (j_{!*}\overline{\mathbb{Q}}_{\ell,U}[n])[-n],$$

where $\overline{\mathbb{Q}}_{\ell,U}$ is the constant sheaf $\overline{\mathbb{Q}}_\ell$ on U. The ℓ-adic intersection cohomology $\mathrm{IH}^*(X, \overline{\mathbb{Q}}_\ell)$ of X is the cohomology of $\mathbf{IC}^*(X)$.

3.5. Application to Shimura varieties. We know that the Shimura variety S^{K} and its Baily-Borel compactification $\overline{S}^{\mathrm{K}}$ are defined over the number field E. So we can form the ℓ-adic intersection cohomology groups $\mathrm{IH}^*(\overline{S}^{\mathrm{K}}_{\overline{E}}, \overline{\mathbb{Q}}_\ell)$. They admit an action of $\mathrm{Gal}(\overline{E}/E)$. Moreover, if we choose a field isomorphism $\overline{\mathbb{Q}}_\ell \simeq \mathbb{C}$, then the comparison theorems between the étale topology and the classical topology will give an isomorphism $\mathrm{IH}^*(\overline{S}^{\mathrm{K}}_{\overline{E}}, \overline{\mathbb{Q}}_\ell) \simeq \mathrm{IH}^*(\overline{S}^{\mathrm{K}}(\mathbb{C}), \mathbb{C})$ (cf. chapter 6 of [6]).

The isomorphism of 2.2 between intersection homology of $\overline{S}^{\mathrm{K}}(\mathbb{C})$ and L^2 cohomology of $S^{\mathrm{K}}(\mathbb{C})$, as well as the duality between intersection homology and intersection cohomology (cf. 3.1), thus give an isomorphism

$$\mathrm{IH}^*(\overline{S}^{\mathrm{K}}_{\overline{E}}, \overline{\mathbb{Q}}_\ell) \simeq \mathrm{H}^*_{(2)}(S^{\mathrm{K}}(\mathbb{C})) \otimes \mathbb{C},$$

and this isomorphism is equivariant under the action of $\mathcal{H}_{\mathrm{K}}(\mathbb{C})$. We know what L^2 cohomology looks like as a representation of $\mathcal{H}_{\mathrm{K}}(\mathbb{C})$, thanks to the theorem of Borel and Casselman (cf. 2.3).

Using this theorem and his own trace invariant formula, Arthur has given a formula for the trace of a Hecke operator on $\mathrm{H}^*_{(2)}(S^{\mathrm{K}}(\mathbb{C})) \otimes \mathbb{C}$ (cf. [1]). This formula involves global volume terms, discrete series characters on $\mathbf{G}(\mathbb{R})$ and orbital integrals on $\mathbf{G}(\mathbb{A}_f)$.

The problem now is to understand the action of the Galois group $\mathrm{Gal}(\overline{E}/E)$. We have a very precise conjectural description of the intersection cohomology of $\overline{S}^{\mathrm{K}}$ as a $\mathcal{H}_{\mathrm{K}}(\mathbb{C}) \times \mathrm{Gal}(\overline{E}/E)$-module, see for example the articles [34] of Kottwitz and [7] of Blasius and Rogawski.

In the next sections, we will explain a strategy to understand how at least part of the Galois group $\mathrm{Gal}(\overline{E}/E)$ acts.

4. Counting Points on Shimura Varieties

We want to understand the action of the Galois group $\mathrm{Gal}(\overline{E}/E)$ on the intersection cohomology groups $\mathrm{IH}^*_{\mathrm{K}} := \mathrm{IH}^*(\overline{S}^{\mathrm{K}}_{\overline{E}}, \overline{\mathbb{Q}}_\ell)$. It is conjectured that this action is unramified almost everywhere. Thus, by the Chebotarev density theorem, it is theoretically enough to understand the action of the Frobenius automorphisms at the places of E where the action is unramified, and one way to do this is to calculate the trace of the powers of the Frobenius automorphisms at these places. However, for some purposes, it is necessary to look at the action of the decomposition groups at other places. This is part of the theory of bad reduction of Shimura varieties, and we will not talk about this here, nor will we attempt to give comprehensive references to it. (Let us just point to the book [31] of Harris and Taylor.)

In general, intersection cohomology can be very hard to calculate. First we will look at simpler objects, the cohomology groups with compact support $\mathrm{H}^*_{c,\mathrm{K}} := \mathrm{H}^*_c(S^{\mathrm{K}}_{\overline{E}}, \overline{\mathbb{Q}}_\ell)$. Assume that the Shimura varieties and their compactifications (the Baily-Borel compactifications and the toroidal compactifications) have "good" models over an open subset U of $\operatorname{Spec} \mathcal{O}_E$, and write $\mathcal{S}^{\mathrm{K}}$ for the model of S^{K}. (It is much easier to imagine what a "good" model should be than to write down a precise definition. An attempt has been made in [49] 1.3, but it is by no means optimal.) Then, by the specialization theorem (SGA 4 III Exposé XVI 2.1), and also by Poincaré duality (cf. SGA 4 III Exposé XVIII), for every finite place $\mathfrak{p}$ of E such that $\mathfrak{p} \in U$ and $\mathfrak{p} \not| \ell$, there is a $\mathrm{Gal}(\overline{E}_{\mathfrak{p}}/E_{\mathfrak{p}})$-equivariant isomorphism

$$\mathrm{H}^*_{c,\mathrm{K}} = \mathrm{H}^*_c(S^{\mathrm{K}}_{\overline{E}}, \overline{\mathbb{Q}}_\ell) \simeq \mathrm{H}^*_c(\mathcal{S}^{\mathrm{K}}_{\overline{\mathbb{F}}_{\mathfrak{p}}}, \overline{\mathbb{Q}}_\ell),$$

where $\mathbb{F}_{\mathfrak{p}}$ is the residue field of $\mathcal{O}_E$ at $\mathfrak{p}$. In particular, the $\mathrm{Gal}(\overline{E}/E)$-representation $\mathrm{H}^*_{c,\mathrm{K}}$ is unramified at $\mathfrak{p}$.

Now, by Grothendieck's fixed point formula (SGA 4 1/2 Rapport), calculating the trace of powers of the Frobenius automorphism on $\mathrm{H}^*_c(\mathcal{S}^{\mathrm{K}}_{\overline{\mathbb{F}}_{\mathfrak{p}}}, \overline{\mathbb{Q}}_\ell)$ is the same as counting the points of $\mathcal{S}^{\mathrm{K}}$ over finite extensions of $\mathbb{F}_{\mathfrak{p}}$.

Langlands has given a conjectural formula for this number of points, cf. [40] and [34]. Ihara had earlier made and proved a similar conjecture for Shimura varieties of dimension 1. Although this conjecture is not known in general, it is easier to study for a special class of Shimura varieties, the so-called PEL Shimura varieties. These are Shimura varieties that can be seen as moduli spaces of abelian with certain supplementary structures (P: polarizations, E: endomorphisms, i.e. complex multiplication by certain CM number fields, and L: level structures). For PEL Shimura varieties of types A and C (i.e., such

that the group $\mathbf{G}$ is of type A or C), Langlands's conjecture had been proved by Kottwitz in [35]. Note that all the examples we gave in 1.1 are of this type. Conveniently enough, the modular interpretation of PEL Shimura varieties also gives a model of the Shimura variety over an explicit open subset of $\operatorname{Spec} \mathcal{O}_E$.

In fact, Kottwitz has done more than counting points; he has also counted the points that are fixed by the composition of a power of the Frobenius automorphism and of a Hecke correspondence (with a condition of triviality at $\mathfrak{p}$). So, using Deligne's conjecture instead of Grothendieck's fixed point formula, we can use Kottwitz's result to understand the commutating actions of $\mathrm{Gal}(\overline{E}/E)$ and of $\mathcal{H}_{\mathrm{K}}(\overline{\mathbb{Q}}_\ell)$ on $\mathrm{H}^*_{c,\mathrm{K}}$. (Deligne's conjecture gives a simple formula for the local terms in the Lefschetz fixed formula if we twist the correspondence by a high power of the Frobenius. It is now a theorem and has been proved independently by Fujiwara in [23] and Varshavsky in [63]. In the case of Shimura varieties, it also follows from an earlier result of Pink in [57].)

Using his counting result, Kottwitz has proved the conjectural description of $\mathrm{IH}^*_{\mathrm{K}}$ for some simple Shimura varieties (cf. [36]). Here "simple" means that the Shimura varieties are compact (so intersection cohomology is cohomology with compact support) and that the phenomenon called "endoscopy" (about which we are trying to say as little as possible) does not appear.

One reason to avoid endoscopic complications was that a very important and necessary result when dealing with endoscopy, the so-called "fundamental lemma", was not available at the time. It now is, thanks to the combined efforts of many people, among which Kottwitz ([33]), Clozel ([15]), Labesse ([38], [16]), Hales ([30]), Laumon, Ngo ([43], [53]), and Waldspurger ([64], [65], [66]).

Assuming the fundamental lemma, the more general case of compact PEL Shimura varieties of type A or C (with endoscopy playing a role) was treated by Kottwitz in [34], admitting Arthur's conjectures on the descripton of discrete automorphic representations of $\mathbf{G}$. Actually, Kottwitz did more: he treated the case of the (expected) contribution of $\mathrm{H}^*_{c,\mathrm{K}}$ to $\mathrm{IH}^*_{\mathrm{K}}$. Let us say a word about Arthur's conjectures. Arthur has announced a proof of a suitable formulation of his conjectures for classical groups (that is, symplectic and orthogonal groups), using the stable twisted trace formula. His proof is expected to adapt to the case of unitary groups (that is, the groups that give PEL Shimura varieties of type A), but this adaptation will likely require a lot of effort.

Let us also note that the case of compact PEL Shimura varieties of type A should be explained in great detail in the book project led by Michael Harris ([8]).

This does not tell us what to do in the case where S^{K} is not projective. First note that the modular interpretation gives us integral models of the Shimura varieties but not of their compactifications. So this is the first problem to solve. Fortunately, it has been solved: See the article [21] of Deligne and Rapoport for the case of modular curves, the book [14] by Chai and Faltings for the case of Siegel modular varieties, Larsen's article [42] for the case of Picard modular varieties, and Lan's dissertation [39] for the general case of PEL Shimura

varieties of type A or C. This allows us to apply the specialization theorem to intersection cohomology. In particular, we get the fact that the $\mathrm{Gal}(\overline{E}/E)$-representation $\mathrm{IH}^*_{c,\mathrm{K}}$ is unramified almost everywhere, and, at the finite places $\mathfrak{p}$ where it is unramified, we can study it by considering the reduction modulo $\mathfrak{p}$ of the Shimura variety and its compactifications.

Next we have to somehow describe the intersection complex. If the group $\mathbf{G}$ has semi-simple $\mathbb{Q}$-rank 1, so it has only one conjugacy class of rational parabolic subgroups, then the Baily-Borel compactification is simpler (it only has one kind of boundary strata) and we can obtain the intersection complex by a simple truncation process from the direct image on $\overline{S}^{\mathrm{K}}$ of the constant sheaf on S^{K}. The conjectural description of $\mathrm{IH}^*_{\mathrm{K}}$ is know for the cases $\mathbf{G} = \mathbf{GL}_2$ (see the book [20]) and the case of Picard modular surfaces, i.e., $\mathbf{G} = \mathbf{GU}(2,1)$ (see the book [41]). In the general case of semi-simple $\mathbb{Q}$-rank 1, Rapoport has given in [58] a formula for the trace of a power of the Frobenius automorphism (at almost every place) on the stalks of the intersection complex.

In the general case, the intersection complex is obtained from the direct image of the constant sheaf on S^{K} by applying several nested truncations (cf. [6] 2.1.11), and it is not clear how to see the action of Frobenius on the stalks of this thing. We will describe a solution in the next section.

5. Weighted Cohomology

In this section, j will be the inclusion of S^{K} in its Baily-Borel compactification $\overline{S}^{\mathrm{K}}$, and j_* will be the derived direct image functor. Here is the main idea: instead of seeing the intersection complex $IC^*(\overline{S}^{\mathrm{K}})$ as a truncation of $j_*\overline{\mathbb{Q}}_{\ell,S^{\mathrm{K}}}$ by the cohomology degree (on various strata of $\overline{S}^{\mathrm{K}} - S^{\mathrm{K}}$), we want to see it as a truncation by Frobenius weights (in the sense of Deligne). This idea goes back to the construction by Goresky, Harder and MacPherson of the weighted cohomology complexes in a topological setting (i.e., on a non-algebraic compactification of the set of complex points $S^{\mathrm{K}}(\mathbb{C})$).

5.1. The topological case. As we have mentioned before, the manifold $S^{\mathrm{K}}(\mathbb{C})$ has a lot of non-algebraic compactifications (these compactifications are defined for a general locally symmetric space, and not just for a Shimura variety). The one used in the construction of weighted cohomology is the reductive Borel-Serre compactification $S^{\mathrm{K}}(\mathbb{C})^{RBS}$ (cf. [11] III.6 and III.10; the reductive Borel-Serre compactification was originally defined by Zucker in [67], though not under that name). The reductive Borel-Serre compactification admits a map $\pi : S^{\mathrm{K}}(\mathbb{C})^{RBS} \longrightarrow \overline{S}^{\mathrm{K}}(\mathbb{C})$ that extends the identity on $S^{\mathrm{K}}(\mathbb{C})$; we also denote by $\widetilde{j}$ the inclusion of $S^{\mathrm{K}}(\mathbb{C})$ in $S^{\mathrm{K}}(\mathbb{C})^{RBS}$.

The boundary $S^{\mathrm{K}}(\mathbb{C})^{RBS} - S^{\mathrm{K}}(\mathbb{C})$ of $S^{\mathrm{K}}(\mathbb{C})^{RBS}$ has a very pleasant description. It is a union of strata, each of which is a locally symmetric space for

the Levi quotient of a rational parabolic subgroup of $\mathbf{G}$; moreover, the closure of a stratum is its reductive Borel-Serre compactification. (A lot more is known about the precise geometry of the strata, see, e.g., [27] 1D).

The weighted cohomology complexes are bounded constructible complexes W^μ of $\mathbb{C}$ or $\mathbb{Q}$-vector spaces on $S^{\mathrm{K}}(\mathbb{C})^{RBS}$ extending the constant sheaf on $S^{\mathrm{K}}(\mathbb{C})$, constructed by Goresky, Harder and MacPherson in [27] (they give two constructions, one for $\mathbb{C}$-coefficients and one for $\mathbb{Q}$-coefficients, and then show that the two constructions agree). They depend on a weight profile μ (which is a function from the set of relative simple roots of $\mathbf{G}$ to $\mathbb{Z}+\frac{1}{2}$). The basic idea of weighted cohomology is to consider the complex $\widetilde{j}_*\mathbb{C}$ (or $\widetilde{j}_*\mathbb{Q}$) on $S^{\mathrm{K}}(\mathbb{C})^{RBS}$ and to truncate it, not by the cohomology degree as for the intersection complex, but by the weights of certain tori. More precisely, on a strata S corresponding to a Levi subgroup $\mathbf{M}$, we truncate by the weights of the $\mathbb{Q}$-split torus $\mathbf{A}_M$ in the center of $\mathbf{M}$ (the group $\mathbf{A}_M(\mathbb{Q})$ acts on $\widetilde{j}_*\mathbb{C}_{|S}$ by what Goresky, Harder and MacPherson call Looijenga Hecke correspondences). The weight profile specifies, for every strata, which weights to keep.

Of course, it is not that simple. The complex $\widetilde{j}_*\mathbb{C}$ is an object in a derived category (which is not abelian but triangulated), and it is not so easy to truncate objects in such a category. To get around this problem, the authors of [27] construct an incarnation of $\widetilde{j}_*\mathbb{C}$, that is, an explicit complex that is quasi-isomorphic to $\widetilde{j} * \mathbb{C}$ and on which the tori $\mathbf{A}_M(\mathbb{Q})$ still act. (In fact, they construct two incarnations, one of $\widetilde{j}_*\mathbb{C}$ and one of $\widetilde{j}_*\mathbb{Q}$).

The upshot (for us) is that the functor $\pi_* : D^b_c(S^{\mathrm{K}}(\mathbb{C})^{RBS}) \longrightarrow D^b_c(\overline{S}^{\mathrm{K}}(\mathbb{C}))$ sends two of these weighted cohomology complexes to the intersection complex on $\overline{S}^{\mathrm{K}}(\mathbb{C})$ (they are the complexes corresponding to the lower and upper middle weight profiles). On the other hand, the weighted cohomology complexes are canonical enough so that the Hecke algebra acts on their cohomology, and explicit enough so that it is possible to calculate the local terms when we apply the Lefschetz fixed point formula to them. This is possible but by no means easy, and is the object of the article [26] of Goresky and MacPherson. Then, in the paper [29], Goresky, Kottwitz and MacPherson show that the result of [26] agrees with the result of Arthur's calculation in [1].

The problem, from our point of view, is that this construction is absolutely not algebraic, so it is unclear how to use it to understand the action of $\mathrm{Gal}(\overline{E}/E)$ on $\mathrm{IH}^*(S^{\mathrm{K}}, \overline{\mathbb{Q}}_\ell)$.

Remark 5. There is another version of weighted cohomology of locally symmetric spaces: Franke's weighted L^2 cohomology, defined in [22]. In his article [52], Nair has shown that Franke's weighted L^2 cohomology groups are weighted cohomology groups in the sense of Goresky-Harder-MacPherson.

5.2. Algebraic construction of weighted cohomology. First, the reductive Borel-Serre compactification is not an algebraic variety, so what we are really looking for is a construction of the complexes $\pi_* W^\mu$, directly on

the Baily-Borel compactification. This looks difficult for several reasons. The Baily-Borel compactification is very singular, which is one of the reasons why Goresky, Harder and MacPherson use the less singular reductive Borel-Serre compactification in the first place. Besides, the boundary strata in $\overline{S}^{\mathrm{K}}$ correspond to maximal rational parabolic subgroups of $\mathbf{G}$, and several strata in $S^{\mathrm{K}}(\mathbb{C})^{RBS}$ can be (rather brutally) contracted to the same stratum in $\overline{S}^{\mathrm{K}}(\mathbb{C})$. It is possible to give a description of the stalks of $\pi_* W^\mu$ (see the article [28] of Goresky, Harder, MacPherson and Nair), but it is a rather complicated description, much more complicated than the simple description of the stalks of W^μ.

The idea is that the action of the Looijenga Hecke correspondences should correspond in some way to the action of the Frobenius automorphism in an algebraic setting. This is actually a very natural ideal. Looijenga himself uses the fact that the eigenspaces of the Looijenga Hecke correspondences are pure in the sense of mixed Hodge theory (cf. [44] 4.2), and we know that the weight filtration of Hodge theory corresponds to the filtration by Frobenius weights in ℓ-adic cohomology (cf. for example [6] 6.2.2). So the correct algebraic analogue of the truncations of [27] should be a truncation by Frobenius weights (in the sense of Deligne's [19], see also chapter 5 of [6]). As a consequence, the most natural place to define the algebraic analogues of the weighted cohomology complexes is the reduction modulo $\mathfrak{p}$ of an integral model of $\overline{S}^{\mathrm{K}}$, where $\mathfrak{p}$ is a finite place of E where good integral models exist. (But see the remark at the end of this subsection.)

In fact, it turns out that we can work in a very general setting. Let $\mathbb{F}_q$ be a finite field, and X be a quasi-separated scheme of finite type over $\mathbb{F}_q$. Then we have the category of mixed ℓ-adic complexes $D^b_m(X, \overline{\mathbb{Q}}_\ell)$ on X, cf. [6] 5.1. (Here "mixed" refers to the weights of the complexes, and the weights are defined by considering the action of the Frobenius automorphisms on the stalks of the complexes; for more details, see [19] or [6] 5). In particular, we get a category $P_m(X)$ of mixed ℓ-adic perverse sheaves on X as a subcategory of $D^b_m(X, \overline{\mathbb{Q}}_\ell)$. One important result of the theory is that mixed perverse sheaves admit a canonical weight filtration. That is, if K is an object in $P_m(X)$, then it has a canonical filtration $(w_{\leq a}K)_{a\in\mathbb{Z}}$ such that each $w_{\leq a}K$ is a subperverse sheaf of K of weight $\leq a$ and such that $K/w_{\leq a}K$ is of weight $> a$.

This functor $w_{\leq a}$ on mixed perverse sheaves does not extend to $D^b_m(X, \overline{\mathbb{Q}}_\ell)$ in the naïve way; that is, the inclusion functor from the category of mixed sheaves of weight $\leq a$ to $D^b_m(X, \overline{\mathbb{Q}}_\ell)$ does not admit a right adjoint. But we can extend $w_{\leq a}$ in another way. Consider the full subcategory ${}^wD^{\leq a}$ of $D^b_m(X, \overline{\mathbb{Q}}_\ell)$ whose objects are the complexes K such that, for every $k \in \mathbb{Z}$, the k-th perverse cohomology sheaf ${}^p\mathrm{H}^k K$ is of weight $\leq a$. (If we wanted to define the complexes of weight $\leq a$, we would require ${}^p\mathrm{H}^k K$ to be of weight $\leq a+k$.) Then ${}^wD^{\leq a}$ is a triangulated subcategory of $D^b_m(X, \overline{\mathbb{Q}}_\ell)$, and the inclusion ${}^wD^{\leq a} \subset D^b_m(X, \overline{\mathbb{Q}}_\ell)$ does admit a right adjoint, which we denote by $w_{\leq a}$ (because it extends the previous $w_{\leq a}$). Likewise, we can define a full triangulated subcategory ${}^wD^{\geq a}$

of $D^b_m(X, \overline{\mathbb{Q}}_\ell)$, whose inclusion into $D^b_m(X, \overline{\mathbb{Q}}_\ell)$ admits a left adjoint $w_{\geq a}$ (extending the functor $K \longmapsto K/w_{\leq a-1}K$ on mixed perverse sheaves). This is explained in section 3 of [48]. Then the analogue of the theorem that $\pi_* W^\mu$ is the intersection complex (for a well-chosen weight profile μ) is the:

Theorem 5.1. *([48] 3.1.4) Let $j : U \longrightarrow X$ a nonempty open subset of X and K be a pure perverse sheaf of weight a on U. Then there are canonical isomorphisms:*

$$j_{!*}K \simeq w_{\leq a} j_* K \simeq w_{\geq a} j_! K.$$

More generally, if we have a stratification on X, we can choose to truncate by different weights on the different strata (cf. [48] 3.3); in this way, we get analogues of the other weighted cohomology complexes, or rather of their images on the Baily-Borel compactification. We also get somewhat more explicit formulas for $w_{\leq a}$, and hence the intersection complex ([48] 3.3.4 and 3.3.5), analogous to the formula of [6] 2.1.11, but where all the truncations by the cohomology degree have been replaced by weight truncations. The reason this makes such a big difference is that the weight truncation functors $w_{\leq a}$ and $w_{\geq a}$ are exact in the perverse sense. (Interestingly enough, it turns out that, in this setting, the weighted cohomology complexes are canonically defined and have nothing to do with Shimura varieties. In fact, there is another application of these ideas, to Schubert varieties, see [50].)

Remark 6. We want to make a remark about the construction of the weighted cohomology complexes on the canonical models $\overline{S}^{\mathrm{K}}$ (and not their reduction modulo a prime ideal). The construction of [48] 3 is very formal and will apply in every category that has a notion of weights and a weight truncation on "perverse" objects. For example, it should apply without any changes to Saito's derived category of mixed Hodge modules. In fact, Arvind Nair has just informed the author that he has indeed been able to construct weighted cohomology complexes in the category of mixed Hodge modules, and to prove that the weighted cohomology complexes he obtained on the Baily-Borel compactification of a Shimura variety are the pushforwards of the Goresky-Harder-MacPherson weighted cohomology complexes on the reductive Borel-Serre compactification. As an application of this, he was able to prove that Franke's spectral sequence ([22] 7.4) is a spectral sequence of mixed Hodge structures (for the locally symmetric spaces that are Shimura varieties).

Now suppose that X is a quasi-separated scheme of finite type over a number field. We can define ℓ-adic perverse sheaves on X, and we can also define a notion of weights for ℓ-adic complexes on X (cf. Deligne's [19] 1.2.2 and Huber's article [32]). The problem is that mixed perverse sheaves on X do not have a weight filtration in general (because number fields have more Galois cohomology than finite fields). To circumvent this problem, we could try to work in the derived category of the abelian category of mixed perverse sheaves on X admitting a weight filtration. Then it is not obvious how to construct the 4/5/6 operations on these categories. It might be possible to copy Saito's approach in [59] (where

he constructs and studies the derived category of mixed Hodge modules); see also Saito's preprint [60]. As far as the author knows, this has not been worked out anywhere.

5.3. Application to the cohomology of Shimura varieties. Once we have the interpretation of the intermediate extension functor given in the previous subsection, it becomes surprisingly easy to calculate the trace of Frobenius automorphisms on the stalks of $\mathbf{IC}^*(\overline{S}^{\mathrm{K}})$. We should mention that one reason it is so easy is that one of the main ingredients, a description of the restriction to the boundary strata of the complex $j_*\overline{\mathbb{Q}}_\ell$ (where j is again the inclusion of S^{K} in $\overline{S}^{\mathrm{K}}$) has been provided by Pink in [56]. And of course, the whole calculation rests on Kottwitz's calculations for the cohomology with compact support (in [35]). Including Hecke correspondences in the picture is just a matter of bookkeeping, and the final result of the Lefschetz trace formula appears in [49] 1.7.

This is not the end of the story. It still remains to compare the result of the Lefschetz fixed point formula with Arthur's invariant trace formula, in order to try to prove the result conjectured in 10.1 of Kottwitz's article [34]. This is basically a generalization of part I of [34] to include the non-elliptic terms. Given the work done by Kottwitz in [34] and [37], it requires no new ideas, but still takes some effort. In the case of general unitary groups over $\mathbb{Q}$, it is the main object of the book [49] (along with some applications).

Even then, we are not quite done. If we want to prove the conjectural description of $\mathrm{IH}^*(\overline{S}^{\mathrm{K}}, \overline{\mathbb{Q}}_\ell)$ given in [34] or [7], we still need to know Arthur's conjectures.

Some applications that do not depend on Arthur's conjectures are worked out in the book [49] (subsection 8.4). They use a weak form of base change from unitary groups to general linear groups, for the automorphic representations that appear in the L^2 cohomology of Shimura varieties. (If we knew full base change, then we would probably also know Arthur's conjectures.) Let us mention the two main applications:

- The logarithm of the L-function of the intersection complex is a linear combination of logarithms of L-functions of automorphic representations of general linear groups ([49] corollary 8.4.5). In fact, we can even get similar formulas for the L-functions of the $\mathcal{H}_{\mathrm{K}}(\overline{\mathbb{Q}}_\ell)$-isotypical components of the intersection cohomology, as in [49] 7.2.2. However, the coefficients in these linear combinations are not explicit, and in particular [49] does not show that they are integers.

- We can derive some cases of the global Langlands correspondence (cf. [49] 8.4.9, 8.4.10). Note however that one of the conclusions of [49] is that, in the end, we do not get more Galois representations in the cohomology of noncompact unitary varieties than we would in the cohomology of compact unitary Shimura varieties. In particular, the cases of the Langlands

correspondence that are worked out in [49] can also be obtained using compact Shimura varieties and gluing of Galois representations (cf. the last chapters of the book project [8] or the article [62] of Shin; note that Shin also considers places of bad reduction).

References

[1] J. Arthur, *The L^2-Lefschetz numbers of Hecke operators*, Inv. Math. 97 (1989), pp. 257–290.

[2] J. Arthur, *An introduction to the trace formula*, in *Harmonic analysis, the trace formula, and Shimura varieties*, pp. 1–263, Clay Math. Proc., 4, Amer. Math. Soc., Providence, RI, 2005.

[3] A. Ash, D. Mumford, M. Rapoport et Y. Tai, *Smooth compactification of locally symmetric spaces*, Lie groups: history, frontiers and applications vol. 4 (1975).

[4] W. Baily et A. Borel, *Compactification of arithmetic quotients of bounded symmetric domains*, Ann. of Math. (2) 84 (1966), pp. 442–528.

[5] A. Beilinson, *On the derived category of perverse sheaves*, in *K-theory, arithmetic and geometry* (Moscow, 1984–1986), pp. 27–41, Lecture Notes in Math., 1289 (1987).

[6] A. Beilinson, J. Bernstein et P. Deligne, *Analyse et topologie sur les espaces singuliers (I)*, Astérique 100 (1982).

[7] D. Blasius and J. Rogawski, *Zeta functions of Shimura varieties*, in *Motives* (Seattle, WA, 1991), pp. 525–571, Proc. Sympos. Pure Math., 55, Part 2, Amer. Math. Soc., Providence, RI, 1994.

[8] *Book Project: Stabilisation de la formule des traces, variétés de Shimura, et applications arithmétiques*, http://www.institut.math.jussieu.fr/projets/fa/bp0.html

[9] A. Borel and W. Casselman, *L^2-cohomology of locally symmetric manifolds of finite volume*, Duke Math. J. 50 (1983), pp. 625–647.

[10] A. Borel and H. Jacquet, *Automorphic forms and automorphic representations*, in *Automorphic forms, representations, and L-functions* (Proc. Symposia in Pure Math., volume 33, 1977), part 1, pp. 189–202.

[11] A. Borel and L. Ji, *Compactifications of symmetric and locally symmatric spaces*, Mathematics: Theory and Applications, Birkäuser (2006).

[12] A. Borel and N. Wallach, *Continuous cohomology, discrete subgroups, and representations of reductive groups. Second edition*, Mathematical surveys and monographs 57, AMS (2000).

[13] M. Borovoi, *Langlands's conjecture concerning conjugation of connected Shimura varieties*, Selecta Math. Soviet. 3 (1983/84), no. 1, pp. 3–39.

[14] C.-L. Chai et G. Faltings, *Degenerations of abelian varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete 22, Springer (1980).

[15] L. Clozel, *The fundamental lemma for stable base change*, Duke Math. J. 61 (1990), n1, pp. 255–302.

[16] L. Clozel and J.-P. Labesse, *Changement de base pour les représentations cohomologiques de certains groupes unitaires*, Astérisque 257 (1999), pp. 119–133.

[17] P. Deligne, *Travaux de Shimura*, Séminaire Bourbaki, exposé 389, février 1971.

[18] P. Deligne, *Variétés de Shimura: Interprétation modulaire, et techniques de construction de modèles canoniques*, Proc. Sympos. Pure Math., Part 2, 33, pp. 247–290, Amer. Math. Soc., Providence, RI, 1979.

[19] P. Deligne, *La conjecture de Weil. II.*, Publications Mathématiques de l'IHES 52 (1981), pp. 137–251.

[20] P. Deligne and W. Kuyk (editors), *Modular functions of one variable II*, Lecture Notes in Mathematics, vol. 349, Springer (1973).

[21] P. Deligne and M. Rapoport, *Les schémas de modules de courbes elliptiques*, in [20], pp. 143–316.

[22] J. Franke, *Harmonic analysis in weighted L_2-spaces*, Ann. Sci. École Norm. Sup. (4) 31 (1998), no. 2, pp. 181–279.

[23] K. Fujiwara, *Rigid geometry, Lefschetz-Verdier trace formula and Deligne's conjecture*, Invent. Math. 127, pp. 489–533 (1997).

[24] M. Goresky and R. MacPherson, *Intersection homology theory*, Topology 19 (1980), no. 2, pp. 135–162.

[25] M. Goresky and R. MacPherson, *Intersection homology II*, Invent. Math. 72 (1983), no. 1, pp. 77–129.

[26] M. Goresky and R. MacPherson, *The topological trace formula*, J. Reine Angew. Math. 560 (2003), pp. 77–150.

[27] M. Goresky, G. Harder and R. MacPherson, *Weighted cohomology*, Invent. math. 166 (1994), pp. 139–213.

[28] M. Goresky, G. Harder, R. MacPherson et A. Nair, *Local intersection cohomology of Baily-Borel compactifications*, Compositio Math. 134 (2002), n3, pp. 243–268.

[29] M. Goresky, R. Kottwitz and R. MacPherson, *Discrete series characters and the Lefschetz formula for Hecke operators*, Duke Math. J. 89 (1997), pp. 477–554 and Duke Math. J. 92 (1998), no. 3, pp. 665–666.

[30] T. Hales, *On the fundamental lemma for standard endoscopy: reduction to unit elements*, Can. J. Math. 47 (1995), n5, pp. 974–994.

[31] M. Harris and R. Taylor, *The geometry and cohomology of some simple Shimura varieties*, Annals of Mathematics Studies 151, Princeton University Press (2001).

[32] A. Huber, *Mixed perverse sheaves for schemes over number fields*, Comp. Math. 108 (1997), pp. 107–121.

[33] R. Kottwitz, *Base change for units of Hecke algebras*, Compositio Math. 60 (1986), pp. 237–250.

[34] R. Kottwitz, *Shimura varieties and λ-adic representations*, in *Automorphic forms, Shimura varieties and L-functions*, Proceedings of the Ann Arbor conference, editors L. Clozel et J. Milne (1990), volume I, pp. 161–209.

[35] R. Kottwitz, *Points on some Shimura varieties over finite fields*, Journal of the AMS, Vol. 5, n2 (1992), pp. 373–444.

[36] R. Kottwitz, *On the λ-adic representations associated to some simple Shimura varieties*, Inv. Math. 108 (1992), pp. 653–665.

[37] R. Kottwitz, unpublished

[38] J.-P. Labesse, *Fonctions élémentaires et lemme fondamental pour le changement de base stable*, Duke Math. J. 61 (1990), 2, pp. 519–530.

[39] K.-W. Lan, *Arithmetic compactifications of PEL-type Shimura varieties*, Ph.D. thesis, Harvard University (2008), http://www.math.princeton.edu/ klan/articles/cpt-PEL-type-thesis.pdf

[40] R. Langlands, *Some contemporary problems with origins in the Jugendtraum*, Proceedings of Symposia in Pure Mathematics 28 (1974), pp. 401–418.

[41] R. Langlands and D. Ramakrishnan (editors), *The zeta function of Picard modular surfaces*, publications du CRM (1992), Montréal.

[42] M. Larsen, *Arithmetic compactification of some Shimura surfaces*, in [LR], pp. 31–46.

[43] G. Laumon and B.-C. Ngo, *Le lemme fondamental pour les groupes unitaires*, Annals of Math. 168 (2008), no. 2, pp. 477–573.

[44] E. Looijenga, *L^2-cohomology of locally symmetric varieties*, Compositio Math. 67 (1988), n1, pp. 3–20.

[45] E. Looijenga et M. Rapoport, *Weights in the local cohomology of a Baily-Borel compactification*, Proceedings of Symposia in Pure Mathematics 53 (1991), pp. 223–260.

[46] J. S. Milne, *The action of an automorphism of $\mathbb{C}$ on a Shimura variety and its special points*, Progr. Math., 35, pp. 239–265, Birkhäuser, Boston, 1983.

[47] J. S. Milne, *Descent for Shimura varieties*, Michigan Math. J. 46 (1999), no. 1, 203–208.

[48] S. Morel, *Complexes pondérés sur les compactifications de Baily-Borel. Le cas des variétés de Siegel*, Journal of the AMS 21 (2008), no. 1, pp. 23–61.

[49] S. Morel, *On the cohomology of certain non-compact Shimura varieties*, Annals of Mathematics Studies 173, Princeton University Press (2010).

[50] S. Morel, *Note sur les polynômes de Kazhdan-Lusztig*, to appear in Mathematische Zeitschrift.

[51] B. Moonen, *Models of Shimura varieties in mixed characteristics*, *Galois representations in arithmetic algebraic geometry* (Durham, 1996), 267–350, London Math. Soc. Lecture Note Ser., 254, Cambridge Univ. Press, Cambridge, 1998.

[52] A. Nair, *Weighted cohomology of arithmetic groups* Ann. of Math. (2) 150 (1999), no. 1, pp. 1–31.

[53] B. C. Ngo, *Le lemme fondamental pour les algèbres de Lie*, submitted, arXiv:0801.0446

[54] I. Piatetski-Shapiro, *Classical and adelic automorphic forms. An introduction*, in *Automorphic forms, representations, and L-functions* (Proc. Symposia in Pure Math., volume 33, 1977), part 1, pp. 185–188.

[55] R. Pink, *Arithmetical compactification of mixed Shimura varieties*, dissertation, Bonner Mathematische Schriften 209 (1989).

[56] R. Pink, *On ℓ-adic sheaves on Shimura varieties and their higher direct images in the Baily-Borel compactification*, Math. Ann. 292 (1992), pp. 197–240.

[57] R. Pink, *On the calculation of local terms in the Lefschetz-Verdier trace formula and its application to a conjecture of Deligne*, Annals of Math., 135 (1992), pp. 483–525.

[58] M. Rapoport, *On the shape of the contribution of a fixed point on the boundary: The case of $\mathbb{Q}$-rank one*, in [LR], p 479-488, with an appendix by L. Saper and M. Stern, pp. 489–491.

[59] M. Saito, *On the derived category of mixed Hodge modules*, Proc. Japan Acad. Ser. A Math. Sci. 62 (1986), no. 9, pp. 364–366.

[60] M. Saito, *On the Formalism of Mixed Sheaves*, preprint RIMS n784 (1991), http://arxiv.org/abs/math/0611597

[61] L. Saper et M. Stern, *L^2-cohomology of arithmetic varieties*, Annals of Math. 132 (1990), n1, pp. 1–69.

[62] S. W. Shin, *Galois representations arising from some compact Shimura varieties*, to appear in Annals of Math.

[63] Y. Varshavsky, *A proof of a generalization of Deligne's conjecture*, Electron. Res. Announc. Amer. Math. Soc. 11 (2005), pp. 78–88.

[64] J.-L.Waldspurger, *Le lemme fondamental implique le transfert*, Comp. Math. 105 (1997), n°2, pp. 153–236.

[65] J.-L.Waldspurger, *Endoscopie et changement de caractéristique*, J. Inst. Math. Jussieu 5 (2006), n°3, pp. 423–525.

[66] J.-L.Waldspurger, *L'endoscopie tordue n'est pas si tordue*, Mem. Amer. Math. Soc. 194 (2008), no. 908.

[67] S. Zucker, *L^2 cohomology of warped products and arithmetic groups*, Invent. Math. 70 (1982/83), no. 2, pp. 169–218.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Wild Ramification of Schemes and Sheaves

Takeshi Saito*

Abstract

We discuss recent developments on geometric theory of ramification of schemes and sheaves. For invariants of ℓ-adic cohomology, we present formulas of Riemann-Roch type expressing them in terms of ramification theoretic invariants of sheaves. The latter invariants allow geometric computations involving some new blow-up constructions.

Mathematics Subject Classification (2010). Primary 14F20; Secondary 11G25, 11S15.

Keywords. Conductor, ℓ-adic sheaf, wild ramification, Grothendieck-Ogg-Shafarevich formula, Swan class, characteristic class.

Introduction

For an extension of a number field, the discriminant is an invariant of fundamental importance, in the classical theory of algebraic integers. The celebrated conductor-discriminant formula [40, Chapitre VI Section 3 Corollaire 1] expresses the discriminant as the product of local invariants of ramification, called the conductor. The conductor is defined for a Galois representation, as a measure of the wild ramification. The relation of the conductor of a Galois representation with the level of corresponding modular form plays a crucial role in the quantitative formulation of the Langlands correspondences.

In arithmetic geometry, the conductor showed up in the 60's in the following scenes among others. For an ℓ-adic sheaf on a curve over an algebraically closed field of positive characteristic different from ℓ, the Grothendieck-Ogg-Shafarevich formula [20] computes the Euler number, in geometric terms. The

*The research is partially supported by JSPS grant-in-aid (B) 18340002
Department of Mathematical Sciences, University of Tokyo, Tokyo, 153-8914, Japan.
E-mail: t-saito@ms.u-tokyo.ac.jp.

conductor appears in the formula as the local contribution of ramification. The formula is a sheaf theoretic variant of the Riemann-Hurwitz formula, which is a geometric counterpart of the conductor-discriminant formula, with the analogy between the discriminant of a number field and the genus of a curve. Grothendieck raised a question to find a formula of Riemann-Roch type computing the Euler number in higher dimension, which generalizes the GOS formula. Deligne and Laumon deduced a generalization for surfaces using fibration [17] [32], a method different from that taken in this article.

For an elliptic curve over a local field, the Tate-Ogg formula [34] expresses the relation between the discriminant and the conductor of the elliptic curve. In the seminal paper [11], Bloch found a correct generalization to arithmetic schemes in general dimension and proved it for curves. His crucial insight is that the ramification should give rise to an invariant as a 0-cycle class, although the ramification does occur in codimension 1. Kato developed this idea in [28].

In this article, we discuss recent developments on geometric theory of ramification, inspired by the insight of Bloch. Some of related results were already discussed 20 years ago in Kyoto by Kato [27, Section 4]. We will not touch on arithmetic applications of ramifications, including canonical subgroups of abelian varieties [3], [42], explicit computation of local Fourier transform [8], finite flat group schemes [22] etc., although they should not be ignored. We will not discuss either the p-adic approach using p-adic $\mathcal{D}$-modules [10], see e.g. [9], [33].

The article consists of two parts. In the first part, we introduce an invariant, called the Swan class, as a measure of the wild ramification of a covering of schemes or a sheaf. We present formulas of Riemann-Roch type computing the Euler number or the conductor of cohomology of an ℓ-adic sheaf in terms of the Swan class, as generalizations of the GOS formula and Bloch's formula. In the geometric case, the characteristic class of an ℓ-adic sheaf is defined as a cohomology class and is shown to equal the cycle class of the Swan class. This gives a refinement of the generalized GOS formula.

In the second part, we discuss a new geometric method to study the wild ramification, blowing-up at the ramification locus in the diagonal. A traditional approach in the study of ramification of a sheaf, taken in the first part of the article, is to kill the ramification by taking a ramified covering. The new approach replaces ramified coverings by blowing-ups. It grew out of the definition of the upper numbering filtration of ramification groups of the absolute Galois group of a local field with not necessarily perfect residue field. By globalizing the construction, we have a geometric method to study the ramification of a sheaf along the boundary.

At the end of the article, we introduce the characteristic cycle of an ℓ-adic sheaf satisfying a certain condition and show that it computes the characteristic class and hence the Euler number. An analogy of the wild ramification of ℓ-adic sheaves with the irregularities of $\mathcal{D}$-modules has been observed by

many mathematicians e.g [16], [28]. The author expects that the new geometric approach shed more light on it.

The author would like to thank his coauthors Kazuya Kato and Ahmed Abbes for long time and fruitful collaborations. Large parts of Sections 1.3 and 2.2 are based on papers in preparation coauthored with them, respectively. It should be evident from the article that a considerable part of the contents is due to them.

1. ℓ-Adic Riemann-Roch Formulas

In Sections 1.1 and 1.2, we consider the geometric case where the base field is a perfect field of positive characteristic. We introduce in Section 1.1 the Swan class of an ℓ-adic sheaf and state a formula for the Euler number, as a generalization of the Grothendieck-Ogg-Shafarevich formula. We define the characteristic class in Section 1.2 as a refinement of the Euler number and gives a relation with the Swan class. In Section 1.3, we consider the arithmetic case where the base field is a p-adic field with perfect residue field and formulate results analogous to those in Section 1.1.

1.1. Euler numbers. Let k be a perfect field and U be a smooth separated scheme of finite type of dimension d over k. For a separated scheme X of finite type over k, the Chow group $CH_0(X)$ denotes the group of 0-cycles modulo rational equivalence. We define

$$CH_0(\partial U) = \varprojlim CH_0(X \setminus U), \quad CH_0(\partial U)_{\mathbb{Q}} = \varprojlim (CH_0(X \setminus U) \otimes_{\mathbb{Z}} \mathbb{Q})$$

to be the projective limits with respect to proper schemes X containing U as a dense open subscheme and proper push-forward. The degree maps $CH_0(X \setminus U) \to CH_0(\mathrm{Spec}\ k) = \mathbb{Z}$ induce $\deg_k \colon CH_0(\partial U) \to \mathbb{Z}$.

For a finite etale Galois covering $V \to U$ of Galois group G, we define the Swan character class

$$s_{V/U}(\sigma) \in CH_0(\partial V)_{\mathbb{Q}}$$

for $\sigma \in G$. We refer to [31, Definition 4.1] for the definition in the general case that requires alteration [15], causing the denominator. Here we only give a definition of the image in $CH_0(Y \setminus V)$, for a smooth compactification Y of V satisfying certain good properties.

Assume Y is a proper smooth scheme containing V as the complement of a divisor D with simple normal crossings. Let $D_1, \ldots, D_n$ be the irreducible components of D and let $(Y \times_k Y)' \to Y \times_k Y$ be the blow-up at $D_i \times_k D_i$ for every $i = 1, \ldots, n$. Namely the blow-up by the product of the ideal sheaves $\mathcal{I}_{D_i \times_k D_i} \subset \mathcal{O}_{Y \times_k Y}$. We call the complement $Y *_k Y \subset (Y \times_k Y)'$ of the proper transform of $(D \times_k Y) \cup (Y \times_k D)$ the log product. The diagonal map $\delta \colon Y \to Y \times_k Y$ is uniquely lifted to a closed immersion $\tilde{\delta} \colon Y \to Y *_k Y$ called

the log diagonal. The log products and the log diagonal seem to have been first introduced by Faltings [21] and by Pink [36] apparantly independently. For an explicit local description, see Example 2.2 in Section 2.2. For more intrinsic definition in the language of log geometry, we refer to [30, Section 4]. We introduce the log product in order to focus on the wild ramification. A heuristic reason for this is that a tamely ramified covering can be regarded as an unramified covering in log geometry.

Let $\sigma \in G$ be an element different from the identity and let Γ be a closed subscheme of $Y *_k Y$ of dimension $d = \dim Y$ such that the intersection $\Gamma \cap (V \times_k V)$ is equal to the graph Γ_σ of σ. By the assumption that V is etale over U, the intersection $\Gamma_\sigma \cap \Delta_V$ with the diagonal $\Delta_V = \delta(V) \subset V \times_k V$ is empty. Hence the intersection product $(\Gamma, \Delta_Y^{\log})_{Y *_k Y}$ with the log diagonal $\Delta_Y^{\log} = \tilde{\delta}(Y) \subset Y *_k Y$ is defined in $CH_0(Y \setminus V)$. The intersection product $(\Gamma, \Delta_Y^{\log})_{Y *_k Y}$ is shown to be independent of the choice of Γ under the assumption that $V \to U$ is extended to a map $Y \to X$ to a proper scheme X over k containing U as the complement of a Cartier divisor B and that the image of Γ in the log product $X *_k X$ defined with respect to B is contained in the log diagonal $\Delta_X^{\log}$.

The Swan chararacter class $s_{V/U}(\sigma) \in CH_0(Y \setminus V)$ for $\sigma \neq 1$ is defined by

$$s_{V/U}(\sigma) = -(\Gamma, \Delta_Y^{\log})_{Y *_k Y}. \tag{1}$$

For $\sigma = 1$, it is defined by requiring $\sum_{\sigma \in G} s_{V/U}(\sigma) = 0$. For $\sigma \neq 1$, we have

$$\sum_{q=0}^{2 \dim V} (-1)^q \mathrm{Tr}(\sigma^* \colon H_c^q(V_{\bar{k}}, \mathbb{Q}_\ell)) = -\deg_k s_{V/U}(\sigma) \tag{2}$$

by a Lefschetz trace formula for open varieties [31, Theorem 2.3.4] for a prime number ℓ different from the characteristic of k.

Example 1.1. Assume that V is a curve. Then, Y is unique and we have $CH_0(\partial V) = CH_0(Y \setminus V) = \bigoplus_{y \in Y \setminus V} \mathbb{Z}$. For $\sigma \neq 1, \in G$, we have

$$s_{V/U}(\sigma) = -\sum_{y \in \{y \in Y \mid \sigma(y) = y\}} \text{length } \mathcal{O}_y \Big/ \left(\frac{\sigma(a)}{a} - 1; a \in \mathcal{O}_y, \neq 0 \right) \cdot [y]. \tag{3}$$

Let ℓ be a prime number different from $p = \mathrm{char}\ k > 0$. We consider a smooth ℓ-adic sheaf $\mathcal{F}$ on U and define the Swan class $\mathrm{Sw}_U \mathcal{F} \in CH_0(\partial U)_{\mathbb{Q}(\zeta_{p^\infty})}$. We refer to [31, Definition 4.2.2] for the definition in the general case that requires reduction modulo ℓ and Brauer traces [23]. Here we only give a definition assuming that there exists a finite etale Galois covering $f \colon V \to U$ trivializing $\mathcal{F}$. Let G denote the Galois group $\mathrm{Gal}(V/U)$ and M be the representation of

G corresponding to $\mathcal{F}$. Then, the Swan class is defined by

$$\mathrm{Sw}_U\mathcal{F} = \frac{1}{|G|}\sum_{\sigma\in G} f_* s_{V/U}(\sigma)\cdot \mathrm{Tr}(\sigma : M). \tag{4}$$

By the equality (3), this is an immediate generalization of the classical definition of the Swan conductor [41, Partie III], see also Example 1.3 in Section 1.3. For the Swan class, we expect that the Hasse-Arf theorem [40, Chapitre VI Section 2 Théorème 1] can be generalized as follows:

Conjecture 1.1. *The Swan class* $\mathrm{Sw}_U\mathcal{F} \in CH_0(\partial U)_{\mathbb{Q}(\zeta_{p^\infty})}$ *is in the image of* $CH_0(\partial U)$.

Conjecture 1.1 implies a conjecture of Serre on the integrality of the Artin character for an isolated fixed point [39] in the geometric case. By the standard argument using Brauer induction, it is reduced to the rank 1 case. By computing the Swan class in the rank 1 case using Theorem 2.12, Conjecture 1.1 is proved in [31, Corollary 5.1.7] for U of dimension 2. The conjecture of Serre for surfaces is proved earlier in [29].

For a smooth ℓ-adic sheaf $\mathcal{F}$ on U, the Euler number $\chi_c(U_{\bar{k}}, \mathcal{F})$ is defined as the alternating sum $\sum_{q=0}^{2\dim U}(-1)^q \dim H^q_c(U_{\bar{k}}, \mathcal{F})$. The Lefschetz trace formula for open varieties (2) implies the following generalization of the Grothendieck-Ogg-Shafarevich formula:

Theorem 1.2 ([31, Theorem 4.2.9]). *Let U be a separated smooth scheme of finite type over k. For a smooth ℓ-adic sheaf $\mathcal{F}$ on U, we have*

$$\chi_c(U_{\bar{k}}, \mathcal{F}) = \mathrm{rank}\ \mathcal{F}\cdot \chi_c(U_{\bar{k}}, \mathbb{Q}_\ell) - \deg_k \mathrm{Sw}_U\mathcal{F}. \tag{5}$$

1.2. Characteristic classes. We recall from [6, Definition 2.1.1] the definition of the characteristic class of an ℓ-adic sheaf on a separated scheme of finite type over a field k of characteristic different from ℓ. Although it is not stated explicitly, essential ingredients in the definition are contained in [19], see also [24, Section 9.1].

Let X be a separated scheme of finite type over a field k. As a coefficient ring Λ, we consider a ring finite over $\mathbb{Z}/\ell^n\mathbb{Z}$, $\mathbb{Z}_\ell$ or $\mathbb{Q}_\ell$ for a prime number $\ell \neq \mathrm{char}\ k$. Let $a\colon X \to \mathrm{Spec}\ k$ denote the structure map and $K_X = Ra^!\Lambda$ denote the dualizing complex. If X is smooth of dimension d over k, we have $K_X = \Lambda(d)[2d]$.

Let $\mathcal{F}$ be a constructible sheaf of flat Λ-modules on X and consider the object

$$\mathcal{H} = R\mathcal{H}om(\mathrm{pr}_2^*\mathcal{F}, R\mathrm{pr}_1^!\mathcal{F})$$

of the derived category $D_{\mathrm{ctf}}(X\times_k X, \Lambda)$ of constructible sheaves of Λ-modules of finite tor-dimension on the product $X\times_k X$. If X is smooth of dimension d over k and if $\mathcal{F}$ is smooth, we have a canonical isomorphism $\mathcal{H} \to \mathcal{H}om(\mathrm{pr}_2^*\mathcal{F}, \mathrm{pr}_1^*\mathcal{F})(d)[2d]$.

A canonical isomorphism

$$End(\mathcal{F}) \to H^0_X(X \times_k X, \mathcal{H}) \tag{6}$$

is defined in [19]. Hence, we may regard the identity $\mathrm{id}_{\mathcal{F}}$ as a cohomology class $\mathrm{id}_{\mathcal{F}} \in H^0_X(X \times_k X, \mathcal{H})$ supported on the diagonal $X \subset X \times_k X$. Let $\delta\colon X \to X \times_k X$ be the diagonal map. Further in [19], a canonical map $\delta^*\mathcal{H} \to K_X$ is defined as the trace map. The characteristic class

$$C(\mathcal{F}) \in H^0(X, K_X)$$

is defined as the image of the pull-back $\delta^*\mathrm{id}_{\mathcal{F}} \in H^0(X, \delta^*\mathcal{H})$ by the induced map $H^0(X, \delta^*\mathcal{H}) \to H^0(X, K_X)$. If X is smooth and if $\mathcal{F}$ is smooth, we have $C(\mathcal{F}) = \mathrm{rank}\ \mathcal{F} \cdot (X, X)_{X \times_k X}$ where $(X, X)_{X \times_k X}$ denotes the self-intersection in the product $X \times_k X$. The Lefschetz trace formula [19] asserts that, if X is proper, the trace map $H^0(X, K_X) \to \Lambda$ sends the characteristic class $C(\mathcal{F})$ to the Euler number $\chi(X_{\bar{k}}, \mathcal{F})$. In other words, the characteristic class is a geometric refinement of the Euler number.

By a standard devissage, the computation of the characteristic classes is reduced to that of the difference $C(j_!\mathcal{F}_U) - \mathrm{rank}\ \mathcal{F}_U \cdot C(j_!\Lambda)$ where $j\colon U \to X$ is the immersion of a dense open subscheme $U \subset X$ smooth over k and $\mathcal{F}_U$ is a smooth sheaf of flat Λ-modules on U. Under a certain mild technical assumption on $\mathcal{F}$, the difference is computed by the Swan class as follows.

Theorem 1.3 ([6, Theorem 3.3.1]). *Let U be a smooth dense open subscheme of a separated scheme X of finite type over k. Let Λ be a finite extension of $\mathbb{Q}_\ell$ and $\mathcal{F}_U$ be a smooth Λ-sheaf on U. Assume that there exists a finite etale covering $V \to U$ such that the pull-back $\mathcal{F}_V$ is of Kummer type with respect to the normalization Y of X in V.*

Then, we have

$$C(j_!\mathcal{F}_U) - \mathrm{rank}\ \mathcal{F}_U \cdot C(j_!\Lambda) = -[\mathrm{Sw}_U \mathcal{F}_U] \tag{7}$$

in $H^0(X, K_X)$, where $[\ ]$ denotes the cycle class.

We refer to [6, Definition 3.1.1] for the definition of being of Kummer type. We only remark here that the purity theorem of Zariski-Nagata and Abhyankar's lemma [37] imply that the assumption on $\mathcal{F}$ is satisfied if we admit a strong form of resolution of singularities for Y. One can also deduce Theorem 1.2 unconditionally from Theorem 1.3.

Problem 1. Find a definition of the characteristic class of an ℓ-adic sheaf on a separated scheme of finite type over a complete discrete valuation ring with perfect residue field and prove a relation similar to Theorem 1.3 with the Swan class defined in Section 1.3.

1.3. Conductor formula. Let K be a complete discrete valuation field with perfect residue field $F = \mathcal{O}_K/\mathfrak{m}_K$. For simplicity, we will assume that the characteristic of K is 0. We consider constructions and formulas for schemes over K analogous to those in Section 1.1. For a scheme X of finite type over $S = \operatorname{Spec} \mathcal{O}_K$, let $G(X)$ denote the Grothendieck group of coherent $\mathcal{O}_X$-modules and $F_\bullet$ be the increasing filtration of $G(X)$ defined by the dimension of support.

Let U be a smooth separated scheme of finite type of dimension $d-1$ over K. We define

$$F_0G(\partial_F U) = \varprojlim F_0G(X_F), \quad F_0G(\partial_F U)_{\mathbb{Q}} = \varprojlim (F_0G(X_F) \otimes_{\mathbb{Z}} \mathbb{Q})$$

to be the projective limits with respect to schemes X proper over S containing U as a dense open subscheme and proper push-forward. For a morphism $f\colon U \to V$ of separated smooth schemes of finite type over K, the push-forward maps define a map $f_!\colon F_0G(\partial_F U) \to F_0G(\partial_F V)$. In particular, for $V = \operatorname{Spec} K$, the degree map $\deg_F\colon F_0G(\partial_F U) \to \mathbb{Z}$ is defined.

For a finite etale Galois covering $V \to U$ of Galois group G, we define the Swan character class

$$s_{V/U}(\sigma) \in F_0G(\partial_F V)_{\mathbb{Q}}$$

for $\sigma \in G$. Here we only sketch the definition of the image in $F_0G(Y_F)$, for a smooth compactification Y of V satisfying certain good properties similarly as in Section 1.1.

Assume Y is a proper regular flat scheme over S containing V as the complement of a divisor D with simple normal crossings. Then, we define the log product $Y *_S Y$ similarly to $Y *_k Y$. The diagonal map $\delta\colon Y \to Y \times_S Y$ is uniquely lifted to a closed immersion $\tilde{\delta}\colon Y \to Y *_S Y$ called the log diagonal.

Let $\sigma \in G$ be an element and let Γ be a closed subscheme of $Y *_S Y$ of dimension $d = \dim Y$ such that the intersection $\Gamma \cap (V \times_S V)$ is equal to the graph Γ_σ of σ. The localized K-theoretic intersection product $((\Gamma, \Delta_Y^{\log}))_{Y *_S Y} \in F_0G(Y_F)$ is then defined by

$$((\Gamma, \Delta_Y^{\log}))_{Y *_S Y} = (-1)^q \left([\mathcal{T}or_q^{\mathcal{O}_{Y *_S Y}}(\mathcal{O}_\Gamma, \mathcal{O}_{\Delta_Y^{\log}})] - [\mathcal{T}or_{q+1}^{\mathcal{O}_{Y *_S Y}}(\mathcal{O}_\Gamma, \mathcal{O}_{\Delta_Y^{\log}})] \right) \tag{8}$$

for $q \geq d = \dim Y$ [30, Section 3]. It is a non-trivial fact that the right hand side is independent of the choice of Γ or $q \geq d$, under an assumption similar to the corresponding one in Section 1.1. We write $((\Gamma_\sigma, \Delta_V))$ for $((\Gamma, \Delta_Y^{\log}))_{Y *_S Y}$.

The Swan chararacter class $s_{V/U}(\sigma) \in F_0G(Y_F)$ for $\sigma \neq 1$ is defined by

$$s_{V/U}(\sigma) = -((\Gamma_\sigma, \Delta_V)). \tag{9}$$

For $\sigma = 1$, it is defined by requiring $\sum_{\sigma \in G} s_{V/U}(\sigma) = 0$.

Example 1.2. Assume that $V = \operatorname{Spec} L$ for a totally ramified extension of K. Then, $Y = \operatorname{Spec} \mathcal{O}_L$ is unique and we have $F_0G(\partial_F V) = F_0G(\operatorname{Spec} F) = \mathbb{Z}$.

The log product $Y *_S Y$ is Spec $\mathcal{O}_L \otimes_{\mathcal{O}_K} \mathcal{O}_L[U^{\pm 1}]/(1 \otimes t - t \otimes 1 \cdot U)$ for a prime element t of L, by definition. The minimal polynomial $f \in \mathcal{O}_K[T]$ of t is an Eisenstein polynomial. We have $Y *_S Y = \mathrm{Spec}\ \mathcal{O}_L[U^{\pm 1}]/(f(tU))$.

Assume L is a Galois extension and let σ be an element of $G = \mathrm{Gal}(L/K)$. We define $g_\sigma \in \mathcal{O}_L[U]$ by $f(tU) = g_\sigma \cdot (U - \sigma(t)/t)$ and put $A = \mathcal{O}_L[U^{\pm 1}]/(f(tU))$, $\mathcal{O}_\sigma = A/(U - \sigma(t)/t)$. Then we have a periodic free resolution

$$\cdots \longrightarrow A \xrightarrow{\times(U-\sigma(t)/t)} A \xrightarrow{\times g_\sigma} A \xrightarrow{\times(U-\sigma(t)/t)} A \longrightarrow \mathcal{O}_\sigma \to 0.$$

Hence, for $\sigma \neq 1$, we have

$$\mathcal{T}or_q^A(\mathcal{O}_\sigma, \mathcal{O}_1) = \begin{cases} \mathcal{O}_L/(\sigma(t)/t - 1) & \text{if } q \text{ is even,} \\ 0 & \text{if } q \text{ is odd.} \end{cases}$$

Consequently, we have

$$s_{V/U}(\sigma) = -\mathrm{ord}_L \left(\frac{\sigma(t)}{t} - 1 \right). \tag{10}$$

For a smooth ℓ-adic sheaf $\mathcal{F}$ on U, the Swan class $\mathrm{Sw}_U \mathcal{F} \in F_0 G(\partial_F U)_{\mathbb{Q}(\zeta_{p^\infty})}$ is defined. Under the same simplifying assumptions, the Swan class is defined by the same formula (4) as in the geometric case. By the equality (10), this is an immediate generalization of the classical definition of the Swan conductor as follows.

Example 1.3. We consider $\mathcal{F}$ on $U = \mathrm{Spec}\ K$ corresponding to an ℓ-adic repreresentation M of the absolute Galois group $G_K = \mathrm{Gal}(\bar{K}/K)$ factoring through a finite quotient $G = \mathrm{Gal}(L/K)$. Then, the Swan class $\mathrm{Sw}_U \mathcal{F}$ is nothing but the Swan conductor $\mathrm{Sw}_K M \in \mathbb{Z} = CH_0(\partial_F U) = CH_0(\mathrm{Spec}\ F)$ defined by

$$\mathrm{Sw}_K M = \frac{1}{|G|} \sum_{\sigma \in G} s_{L/K}(\sigma) \cdot \mathrm{Tr}(\sigma : M). \tag{11}$$

For the Swan class defined above, we also expect that the Hasse-Arf theorem can be generalized as in Conjecture 1.1. As in the geometric case, it is proved for a curve U over K. A proof of the conjecture of Serre [39] in the corresponding case is announced in [27], see also [1], [2].

Let $U_1 \subset U$ be a regular closed subscheme and $i \colon U_1 \to U$ and $j \colon U_0 = U \setminus U_1 \to U$ denote the immersions. Then, we have an excision formula

$$\mathrm{Sw}_U \mathcal{F} = j_! \mathrm{Sw}_{U_0} \mathcal{F}|_{U_0} + i_! \mathrm{Sw}_{U_1} \mathcal{F}|_{U_1}. \tag{12}$$

This enables us to extend the definition of the Swan classes to constructible sheaves. For a smooth sheaf $\mathcal{F}$ on U, we define a variant of the Swan class by

$$\overline{\mathrm{Sw}}_U \mathcal{F} = -\mathrm{rank}\ \mathcal{F} \cdot ((\Delta_U, \Delta_U)) + \mathrm{Sw}_U \mathcal{F}.$$

This is also extended to constructible sheaves by the excision formula.

For the variant, we have the following formula of Riemann-Roch type:

Theorem 1.4. *Assume K is of characteristic 0. For a morphism $f\colon U \to V$ of separated schemes of finite type over K and for a constructible ℓ-adic sheaf $\mathcal{F}$ on U, we have*

$$\overline{\mathrm{Sw}}_V Rf_!\mathcal{F} = f_!\overline{\mathrm{Sw}}_U\mathcal{F}. \tag{13}$$

The outline of the proof is as follows. By standard devissage using the excision formula, it is reduced to the relative curve case. Then, we take an alteration and apply a logarithmic Lefschetz trace formula for an open variety over a local field generalizing [30, Theorem 6.5.1], to conclude the proof.

In the case where $V = \mathrm{Spec}\ K$, Theorem 1.4 gives a conductor formula. For a smooth ℓ-adic sheaf $\mathcal{F}$ on a separated scheme U of finite type over K, let $\mathrm{Sw}_K H^*_c(U_{\bar{K}}, \mathcal{F})$ denote the alternating sum $\sum_{q=0}^{2\dim U}(-1)^q \mathrm{Sw}_K H^q_c(U_{\bar{K}}, \mathcal{F})$ of the Swan conductor.

Corollary. *Let U be a separated smooth scheme of finite type of dimension $d-1$ over K.*

1. *For a smooth ℓ-adic sheaf $\mathcal{F}$ on U, we have*

$$\mathrm{Sw}_K H^*_c(U_{\bar{K}}, \mathcal{F}) = \mathrm{rank}\ \mathcal{F} \cdot \mathrm{Sw}_K H^*_c(U_{\bar{K}}, \mathbb{Q}_\ell) + \deg_F\ \mathrm{Sw}_U \mathcal{F}. \tag{14}$$

2. ([30, Theorem 6.2.3]) *Assume U is proper over K and let X be a proper regular flat scheme such that $U = X_K$. Assume that the reduced closed fiber $X_{F,\mathrm{red}}$ is a divisor with simple normal crossings. Then, we have*

$$\chi_c(X_{\bar{K}}, \mathbb{Q}_\ell) - \chi_c(X_{\bar{F}}, \mathbb{Q}_\ell) + \mathrm{Sw}_K H^*_c(X_{\bar{K}}, \mathbb{Q}_\ell) = (-1)^{d-1} \deg_F c_d(\Omega^1_{X/S}), \tag{15}$$

where $c_d(\Omega^1_{X/S}) \in CH_0(X_F)$ denotes the localized Chern class.

The formula (15) is conjectured by Bloch in [11] and proved there for curves. For surfaces $\dim U = 2$, we can remove the assumption on the reduced closed fiber $X_{F,\mathrm{red}}$, since the strong resolution of singularity is now obtained by blow-up in dimension 2 [14]. In a geometric case, a formula analogous to (14) is obtained by using a localized refinement of the characteristic class in [43], assuming resolution of singularities.

2. Geometric Ramification Theory

We recall the geometric definition of the filtration by ramification groups of Galois groups of local fields in Section 2.1. We globalize it in Section 2.2 and study the ramification of a Galois covering of a smooth scheme over a perfect field of characteristic $p > 0$ along the boundary. We compute in Section 2.3 the characteristic class using the construction in Section 2.2 and introduce the characteristic cycle of an ℓ-adic sheaf that enables one to compute the Euler number.

2.1. Ramification groups of a local field. Let K be a complete discrete valuation field with not necessarily perfect residue field $F = \mathcal{O}_K/\mathfrak{m}_K$. For a finite Galois extension L over K, the Galois group $G = \mathrm{Gal}(L/K)$ has two decreasing filtrations, the lower numbering filtration $(G_i)_{i\in\mathbb{N}}$ and the upper numbering filtration $(G^r)_{r\in\mathbb{Q},>0}$. In the classical case where the residue field is perfect, they are the same up to renumbering by the Herbrand function [40, Chapitre IV Section 3]. However, their properties make good contrasts. The lower one has an elementary definition and is compatible with subgroups while the upper one has more sophisticated definition and is compatible with quotients. The lower one is simply defined by $G_i = \mathrm{Ker}(G \to \mathrm{Aut}(\mathcal{O}_L/\mathfrak{m}_L^i))$. More geometrically, it is rephrased as follows.

Take a presentation $\mathcal{O}_K[X_1,\dots,X_n]/(f_1,\dots,f_n) \to \mathcal{O}_L$ of the integer ring of L. We consider the n-dimensional closed disk D^n defined by $\|x\| \le 1$ over K in the sense of rigid geometry and the morphism of disks $f\colon D^n \to D^n$ defined by $f_1,\dots,f_n$. Then the Galois group G is identified with the inverse image $f^{-1}(0)$ of the origin $0 \in D^n$. In other words, we have a cartesian diagram

$$\begin{tikzcd} G \arrow[r] \arrow[d] & D^n \arrow[d, "f"] \\ \{0\} \arrow[r] & D^n. \end{tikzcd} \tag{16}$$

The subgroups G_i and G^r are defined to consist of the points of G that are *close* to the identity in certain senses. For the lower one, the closeness is simply measured by the distance. Namely, the lower numbering subgroup $G_i \subset G$ consists of the points $\sigma \in G$ satisfying $d(\sigma, \mathrm{id}) \le |\pi_L^i|$ for a prime element π_L of L.

To define the upper numbering filtration, we consider, for a rational number $r > 0$, the inverse image $V_r = \{x \in D^n \mid d(f(x), 0) \le |\pi_K|^r\} \subset D^n$ of the closed subdisk of radius $|\pi_K|^r$, as an affinoid subdomain containing G. The upper numbering subgroup G^r consists of the points in G contained in the same geometric connected component of V_r as the identity.

Theorem 2.1 ([4, Theorems 3.3, 3.8]). *Let L be a finite Galois extension over K of Galois group $G = \mathrm{Gal}(L/K)$.*

1. *For a rational number $r > 0$, the subset $G^r \subset G$ defined above is independent of the choice of presentation $\mathcal{O}_K[X_1,\dots,X_n]/(f_1,\dots,f_n) \to \mathcal{O}_L$ and is a normal subgroup of G. Further the inclusion $G \to V_r$ induces a bijection $G/G^r \to \pi_0(V_r)$ to the set of geometric connected components.*

2. *There exist rational numbers $0 = r_0 < r_1 < \ldots < r_m$ such that $G^r = G^{r_i}$ for $r \in (r_{i-1}, r_i] \cap \mathbb{Q}$ and $i = 1,\dots,m$ and $G^r = 1$ for $r > r_m$.*

3. *For a subfield $M \subset L$ Galois over K and for a rational number $r > 0$, the subgroup $\mathrm{Gal}(M/K)^r \subset \mathrm{Gal}(M/K)$ is the image of $G^r = \mathrm{Gal}(L/K)^r$.*

The definition of the upper numbering filtration in the general residue field case is first found using rigid geometry, as described above. The use of rigid geometry is quite essential. For example, the proof of 2. in Theorem 2.1 relies on the reduced fiber theorem in rigid geometry [12]. However, an alternative scheme theoretic approach described below turned out to be quite powerful as well.

In the following, we give a definition of a logarithmic variant of the upper numbering filtration, that seems more essential than the non-logarithmic one and is defined using the natural log structure of the integer rings. For the generality on log schemes, we refer to [26] and [30, Section 4]. In the classical case where the residue field is perfect, the two filtrations are the same up to the shift by 1. In the general residue field case, there is no simple relation among them. We emphasize here that both filtrations have scheme theoretic descriptions.

We regard $S = \operatorname{Spec} \mathcal{O}_K$ as a log scheme with the canonical log structure defined by the closed point $D_S = \operatorname{Spec} F$. Let Q be a regular flat separated scheme of finite type over S. Assume that the reduced closed fiber $D_Q = (Q \times_S D_S)_{\mathrm{red}}$ is regular and that the log scheme Q endowed with the log structure defined by D_Q is log smooth over S. For example, $Q = \operatorname{Spec} \mathcal{O}_K[T_1, \ldots, T_d, U^{\pm 1}]/(\pi - UT_1^m)$ for integers $d, m \geq 1$ and a prime element π of $\mathcal{O}_K$.

Let L be a finite Galois extension of K. We put $T = \operatorname{Spec} \mathcal{O}_L$ and $D_T = (T \times_S D_S)_{\mathrm{red}}$. We consider a closed immersion $i\colon T \to Q$ that is exact in the sense that $D_T = D_Q \times_Q T$. Let P be a separated smooth scheme of finite type over S, $s\colon S \to P$ be a section and $f\colon Q \to P$ be a finite and flat morphism over S such that the diagram

$$\begin{tikzcd} T \arrow[r, "i"] \arrow[d] & Q \arrow[d, "f"] \\ S \arrow[r, "s"] & P \end{tikzcd} \tag{17}$$

is cartesian. This diagram should be regarded as a scheme theoretic counterpart of (16).

We consider a finite separable extension K' of K containing L as a subextension, in order to make a base change. We put $S' = \operatorname{Spec} \mathcal{O}_{K'}$, $F' = \mathcal{O}_{K'}/\mathfrak{m}_{K'}$ and let $e = e_{K'/K}$ be the ramification index.

Let $r > 0$ be a rational number and assume that $r' = er$ is an integer. We regard the divisor $R' = r'D_{S'} = \operatorname{Spec} \mathcal{O}_{K'}/\mathfrak{m}_{K'}^{r'}$ of S' as a closed subscheme of $P_{S'} = P \times_S S'$ by the section $s'\colon S' \to P_{S'}$ induced by $s\colon S \to P$. We consider the blow-up of $P_{S'}$ at the center R' and let $P_{S'}^{(r)}$ denote the complement of the proper transform of the closed fiber $P_{S'} \times_{S'} D_{S'}$. The scheme $P_{S'}^{(r)}$ is smooth over S' and the closed fiber $P_{S'}^{(r)} \times_{S'} D_{S'}$ is the vector bundle $\Theta_{F'}^{(r)}$ over F' such that the set of F'-valued points is the F'-vector space

$Hom_{F'}(\mathfrak{m}_{K'}^{r'}/\mathfrak{m}_{K'}^{r'+1}, \mathcal{I}_{s(S)}/\mathcal{I}_{s(S)}^2 \otimes_{\mathcal{O}_K} F')$ where $\mathcal{I}_{s(S)} \subset \mathcal{O}_{P_S}$ denotes the ideal sheaf.

We consider the normalizations $\bar{Q}_{S'}^{(r)}$ and $\bar{T}_{S'}$ of $Q \times_P P_{S'}^{(r)}$ and of $T \times_S S'$ respectively. Then, the diagram (17) induces a diagram

$$\begin{tikzcd} \bar{T}_{S'} \arrow[r, "i^{(r)}"] \arrow[d] & \bar{Q}_{S'}^{(r)} \arrow[d, "f^{(r)}"] \\ S' \arrow[r, "s^{(r)}"] & P_{S'}^{(r)}. \end{tikzcd} \tag{18}$$

By the assumption that K' contains L, the scheme $\bar{T}_{S'}$ is isomorphic to the disjoint union of finitely many copies of S' and the geometric fiber $\bar{T}_{\bar{F}} = \bar{T}_{S'} \times_{S'} \bar{F}$ is identified with $\mathrm{Gal}(L/K)$.

By Epp's theorem [18], after replacing K' by some finite separable extension, the geometric closed fiber $\bar{Q}_{\bar{F}}^{(r)} = \bar{Q}_{S'}^{(r)} \times_{S'} \mathrm{Spec}\, \bar{F}$ is reduced and the formation of $\bar{Q}_{S'}^{(r)}$ commutes with further base change. We call such $\bar{Q}_{S'}^{(r)}$ a stable integral model. The finite map $i^{(r)} \colon \bar{T}_{S'} \to \bar{Q}_{S'}^{(r)}$ induces surjections

$$\begin{tikzcd} \bar{T}_{\bar{F}} = \mathrm{Gal}(L/K) \arrow[r, "i_*^{(r+)}"] \arrow[dr, "i_*^{(r)}"'] & f^{(r)-1}(0) \arrow[d] \\ & \pi_0(\bar{Q}_{\bar{F}}^{(r)}) \end{tikzcd} \tag{19}$$

to the set of geometric connected components and to the inverse image of the origin $0 \in P_{\bar{F}}^{(r)} = \Theta_{\bar{F}}^{(r)}$.

Theorem 2.2 ([4, Theorems 3.11, 3.16], [38, Section 1.3]). *Let L be a finite Galois extension over K of Galois group G and we consider a diagram* (17) *as above.*

1. *For a rational number $r > 0$, we take a finite separable extension K' of K containing L such that $e_{K'/K} r$ is an integer and that $Q_{\bar{S}'}^{(r)}$ is a stable integral model.*

   Then, the inverse image $i_*^{(r)-1}(i_*^{(r)}(1)) = G_{\log}^r \subset G$ is independent of the choice of diagram (17) *or an extension K' and is a normal subgroup of G. Further the surjection $i_*^{(r)}$* (19) *induces a bijection $G/G_{\log}^r \to \pi_0(\bar{Q}_{\bar{F}}^{(r)})$.*

2. *Let the notation be as in* 1. *Then, there exist rational numbers $0 = r_0 < r_1 < \ldots < r_m$ such that $G_{\log}^r = G_{\log}^{r_i}$ for $r \in (r_{i-1}, r_i] \cap \mathbb{Q}$ and $i = 1, \ldots, m$ and $G_{\log}^r = 1$ for $r > r_m$.*

3. *For a subfield $M \subset L$ Galois over K and for a rational number $r > 0$, the subgroup $\mathrm{Gal}(M/K)_{\log}^r \subset \mathrm{Gal}(M/K)$ is the image of $G^r = \mathrm{Gal}(L/K)_{\log}^r$.*

For a rational number $r \geq 0$, we put $G^{r+}_{\log} = \bigcup_{s>r} G^s_{\log}$. Then, under the same assumption as in Theorem 2.2.1., the surjection $i_*^{(r+)}$ (19) induces a bijection $G/G^{r+}_{\log} \to f^{(r)-1}(0)$.

Example 2.1 ([25], [7]). If K is of characteristic $p > 0$, a cyclic extension L of degree p^{m+1} is defined by a Witt vector by the isomorphism $W_{m+1}(K)/(F-1) \to H^1(K, \mathbb{Z}/p^{m+1}\mathbb{Z})$ of Artin-Schreier-Witt theory. An increasing filtration on $W_{m+1}(K)$ is defined in [13] by

$$F^n W_{m+1}(K) = \{(a_0, \ldots, a_m) \in W_{m+1}(K) \mid p^{m-i} v_K(a_i) \geq -n \text{ for } i = 0, \ldots, m\}.$$

The filtration induced by the surjection $W_{m+1}(K) \to H^1(K, \mathbb{Z}/p^{m+1}\mathbb{Z})$ is considered in [25]. For $G = \mathrm{Gal}(L/K)$, the filtration $(G^n_{\log})_{n \geq 0}$ indexed by integers is the dual of the restriction to $Hom(\mathrm{Gal}(L/K), \mathbb{Z}/p^{m+1}\mathbb{Z}) \subset H^1(K, \mathbb{Z}/p^{m+1}\mathbb{Z})$. Namely, we have $G^n_{\log} = \{\sigma \in G \mid c(\sigma) = 0 \text{ if } c \in F^n H^1(K, \mathbb{Z}/p^{m+1}\mathbb{Z})\}$. Further, for a rational number $r \in (n-1, n] \cap \mathbb{Q}$, we have $G^r_{\log} = G^n_{\log}$.

Problem 2. Prove that, for an abelian extension in the mixed characteristic case, the filtration $(G^n_{\log})_{n \geq 0}$ is the same as that defined by Kato in [25] and show $G^r_{\log} = G^n_{\log}$ for $r \in (n-1, n] \cap \mathbb{Q}$.

Definition 2.3. *Let L be a finite etale K-algebra and $r > 0$ (resp. $r \geq 0$) be a rational number. Let M be a finite Galois extension of K such that $L \otimes_K M$ is isomorphic to the product of copies of M. Then, we say that the log ramification of L over K is bounded by r (resp. by $r \geq 0$) if the action of the subgroup* $\mathrm{Gal}(M/K)^r_{\log}$ *(resp.* $\mathrm{Gal}(M/K)^{r+}_{\log}$*) on the finite set $Hom_K(L, M)$ is trivial.*

It is interpreted geometrically as follows.

Proposition 2.4 ([38, Lemma 1.13]). *Let $r > 0$ be a rational number and we consider a diagram* (17) *and a finite separable extension K' of K as in Theorem* 2.2 *such that $\bar{Q}^{(r)}_{S'}$ is a stable integral model.*

1. *The following conditions are equivalent:*
   (1) *The log ramification of L over K is bounded by r.*
   (2) *The finite covering $\bar{Q}^{(r)}_{F'} \to P^{(r)}_{F'}$ is a split etale covering.*
2. *The following conditions are equivalent:*
   (1) *The log ramification of L over K is bounded by $r+$.*
   (2) *The finite map $\bar{Q}^{(r)}_{S'} \to P^{(r)}_{S'}$ is etale on a neighborhood of the closed fiber $\bar{Q}^{(r)}_{F'}$.*

2.2. Ramification along a divisor. Let X be a smooth separated scheme of finite type over a perfect field k of characteristic $p > 0$ and $U = X \setminus D$ be the complement of a divisor D with simple normal crossings. We consider a finite etale G-torsor V over U for a finite group G and study the ramification of V along D. The ramification of V along D will be measured by linear combinations $R = \sum_i r_i D_i$ with rational coefficients $r_i \geq 0$ of irreducible components of D as follows.

We consider the log product $P = X *_k X \subset (X \times_k X)'$ and the log diagonal $\tilde{\delta}\colon X \to P = X *_k X$ as in Section 1.1. We define a relatively affine scheme $P^{(R)}$ over P. If the coefficients of $R = \sum_i r_i D_i$ are integers, the scheme $P^{(R)}$ is the complement of the proper transforms of $P \times_X R$ in the blow-up of P at the center $R \subset X$ embedded by the log diagonal map $\tilde{\delta}\colon X \to P$. In other words, $P^{(R)}$ is the relatively affine scheme over P defined by the quasi-coherent $\mathcal{O}_P$-algebra $\mathcal{O}_P[\mathcal{I}_X(R_P)] = \sum_{n \geq 0} \mathcal{I}_X^n(nR_P)$ where $\mathcal{I}_X \subset \mathcal{O}_P$ is the ideal sheaf defining the log diagonal and the divisor $R_P \subset P$ is the pull-back of $R \subset X$. The base change $P^{(R)} \times_X R$ with respect to the projection $P^{(R)} \to X \supset R$ is the twisted tangent bundle $\Theta^{(R)} = \mathbf{V}(\Omega^1_X(\log D)(R)) \times_X R$ where $\mathbf{V}(\Omega^1_X(\log D)(R))$ denotes the vector bundle defined by the symmetric algebra of the locally free $\mathcal{O}_X$-module $\Omega^1_X(\log D)(R)$.

For general R, it is defined by the quasi-coherent $\mathcal{O}_P$-algebra $\sum_{n \geq 0} \mathcal{I}_X^n(\lfloor nR_P \rfloor)$ where $\lfloor nR_P \rfloor$ denotes the integral part. The log diagonal $\tilde{\delta}\colon X \to P = X *_k X$ is uniquely lifted to an immersion $\delta^{(R)}\colon X \to P^{(R)}$. The open immersion $U \times_k U \to X *_k X = P$ is lifted to an open immersion $j^{(R)}\colon U \times_k U \to P^{(R)}$.

Example 2.2. Assume $X = \operatorname{Spec} k[T_1, \dots, T_d]$ and D is defined by $T_1 \cdots T_n$ for $0 \leq n \leq d$. Then, the log product $P = X *_k X$ is the spectrum of

$$A = k[T_1, \dots, T_d, S_1, \dots, S_d, U_1^{\pm 1}, \dots, U_n^{\pm 1}]/(S_1 - U_1 T_1, \dots, S_n - U_n T_n) \quad (20)$$

and the log diagonal $\tilde{\delta}\colon X \to P = X *_k X$ is defined by $U_1 = \cdots = U_n = 1$ and $T_{n+1} = S_{n+1}, \ \dots, \ T_d = S_d$.

Further assume that the coefficients of $R = \sum_{i=1}^n r_i D_i$ are integral. Then, if we put $T^R = T_1^{r_1} \cdots T_n^{r_n}$, the scheme $(X *_k X)^{(R)}$ is the spectrum of

$$\begin{aligned} &A[V_1, \dots, V_d]/(U_1 - 1 - V_1 T^R, \dots, U_n - 1 - V_n T^R, \\ &\qquad\qquad S_{n+1} - T_{n+1} - V_{n+1} T^R, \dots, S_d - T_d - V_d T^R) \qquad (21) \\ &= k[T_1, \dots, T_d, V_1, \dots, V_d, (1 + V_1 T^R)^{-1}, \dots, (1 + V_n T^R)^{-1}]. \end{aligned}$$

The immersion $\delta^{(R)}\colon X \to (X *_k X)^{(R)}$ is defined by $V_1 = \cdots = V_d = 0$.

Let V be a G-torsor over U for a finite group G. We consider the quotient $(V \times_k V)/\Delta G$ by the diagonal $\Delta G \subset G \times G$ as a finite etale covering of $U \times_k U$ and let Z be the normalization of $(X *_k X)^{(R)}$ in the quotient $(V \times_k V)/\Delta G$. The diagonal map $V \to V \times_k V$ induces a closed immersion $U = V/G \to (V \times_k V)/\Delta G$ on the quotients and is extended to a closed immersion $e\colon X \to Z$.

Definition 2.5. *Let V be a G-torsor over U for a finite group G. We say that the ramification of V over U is bounded by $R+$ if the normalization Z of $(X *_k X)^{(R)}$ in the quotient $(V \times_k V)/\Delta G$ is etale over $(X *_k X)^{(R)}$ on a neighborhood of the image of $e\colon X \to Z$.*

The following is an immediate consequence of Proposition 2.4.2.

Lemma 2.6. *Assume D is irreducible and let $K = \operatorname{Frac} \widehat{\mathcal{O}}_{X,\xi}$ be the fraction field of the completion of the local ring $\mathcal{O}_{X,\xi}$ at the generic point ξ of D. Then, for a rational number $r \geq 0$, the following conditions are equivalent:*

(1) *The log ramification of the etale K-algebra $\Gamma(V \times_U \operatorname{Spec} K, \mathcal{O})$ is bounded by $r+$.*

(2) *There exists an open neighborhood X' of ξ such that the ramification of $V \cap X'$ over $U \cap X'$ is bounded by $r(D \cap X')+$.*

By the following lemma, the general case is reduced to the case where the coefficients of R are integral.

Lemma 2.7. *Let $f\colon X' \to X$ be a morphism of separated smooth schemes of finite type over k. Let $U \subset X$ and $U' \subset X'$ be the complements of divisors with simple normal crossings respectively satisfying $U' \subset f^{-1}(U)$.*

Let $V \to U$ be a G-torsor for a finite group G and $V' = V \times_U U' \to U'$ be the pull-back. Let $R = \sum_i r_i D_i \geq 0$ be an effective divisor with rational coefficients and $R' = f^* R$ be the pull-back. We consider the following conditions.

(1) *The ramification of V is bounded by $R+$.*

(2) *The ramification of V' is bounded by $R'+$.*

We always have an implication (1) $\Rightarrow$ (2). Conversely, if $f\colon X' \to X$ is log smooth and is faithfully flat and if $U' = f^{-1}(U)$, we have the other implication (2) $\Rightarrow$ (1).

The main result is the following.

Theorem 2.8. *Let X be a separated smooth scheme of finite type over k and $U = X \backslash D$ be the complement of a divisor with simple normal crossings. Assume that the coefficients of $R = \sum_i r_i D_i \geq 0$ are integral. Let V be a G-torsor over U for a finite group G and $Z_0 \subset Z$ be the maximum open subscheme etale over $(X *_k X)^{(R)}$ of the normalization Z of $(V \times_k V)/\Delta G$. Let $e\colon X \to Z$ be the section induced by the diagonal.*

Then, the base change $Z_{0,R} = Z_0 \times_X R$ with respect to the projection $Z_0 \to (X *_k X)^{(R)} \to X \supset R$ has a natural structure of smooth commutative group scheme over R such that the map $e_R\colon X_R \to Z_{0,R}$ induced by $e\colon X \to Z$ is the unit. Further the etale map $Z_{0,R} \to \Theta^{(R)} = (X *_k X)^{(R)} \times_X R$ induced by the canonical map $Z \to (X *_k X)^{(R)}$ is a group homomorphism.

For every point $x \in R$, the connected component $Z^0_{0,x}$ of the fiber $Z_{0,x}$ is isomorphic to the product of finitely many copies of the additive group $\mathbf{G}_{a,x}$ and the map $Z^0_{0,x} \to \Theta^{(R)}_x$ is an etale isogeny.

Problem 3. Prove an analogous result for schemes over a discrete valuation rings with perfect residue field.

Theorem 2.8 has the following application to the filtration by ramification groups in the equal characteristic case.

Let K be a complete discrete valuation field of characteristic $p > 0$ and assume that the residue field F has a finite p-basis. Let $\Omega^1_{\mathcal{O}_K}(\log)$ denote the $\mathcal{O}_K$-submodule of the K-vector space Ω^1_K generated by $\Omega^1_{\mathcal{O}_K}$ and $d\log\pi$ for a prime element π of K. By abuse of notation, let $\Omega^1_F(\log)$ denote the F-vector space $\Omega^1_{\mathcal{O}_K}(\log)\otimes_{\mathcal{O}_K} F$. Then, we have an exact sequence $0 \to \Omega^1_F \to \Omega^1_F(\log) \overset{\mathrm{res}}{\to} F \to 0$ of F-vector spaces of finite dimension. We extend the normalized discrete valuation v of K to a separable closure $\bar{K}$ and, for a rational number r, we put $\mathfrak{m}^r_{\bar{K}} = \{a \in \bar{K} \mid v(a) \geq r\}$ and $\mathfrak{m}^{r+}_{\bar{K}} = \{a \in \bar{K} \mid v(a) > r\}$. The $\bar{F}$-vector space $\mathfrak{m}^r_{\bar{K}}/\mathfrak{m}^{r+}_{\bar{K}}$ is of dimension 1.

Corollary ([5, Theorem 2.15], [38, Theorem 1.24, Corollary 1.25]). *Let K be a complete discrete valuation field of characteristic $p > 0$ and L be a finite Galois extension of Galois group G. Then, for a rational number $r > 0$, the graded quotient $\mathrm{Gr}^r_{\log}G = G^r_{\log}/G^{r+}_{\log}$ is abelian and killed by p.*

If F has a finite p-basis, there exists a canonical injection

$$Hom(\mathrm{Gr}^r_{\log}G, \mathbb{F}_p) \to Hom_{\bar{F}}(\mathfrak{m}^r_{\bar{K}}/\mathfrak{m}^{r+}_{\bar{K}}, \Omega^1_F(\log) \otimes_F \bar{F}). \tag{22}$$

For a non-trivial character $\chi \in Hom(\mathrm{Gr}^r_{\log}G, \mathbb{F}_p)$, we call the image $\mathrm{rsw}\chi \in Hom_{\bar{F}}(\mathfrak{m}^r_{\bar{K}}/\mathfrak{m}^{r+}_{\bar{K}}, \Omega^1_F(\log) \otimes_F \bar{F})$ the refined Swan character of χ.

For K of mixed characteristic, one has an analogous result. The proof is similar but technically more difficult.

Problem 4. Determine the union of the images of the injections (22) for all finite Galois extensions $L \subset \bar{K}$ over K.

In the classical case where the residue field is perfect, the union is the whole.

Example 2.3 ([25], [7]). We keep the notation in Example 2.1. We define a canonical map $F^m d\colon W_{m+1}(K) \to \Omega^1_K$ by sending $(a_0, \dots, a_m)$ to $a_0^{p^m-1}da_0 + \cdots + da_m$. It maps $F^n W_{m+1}(K)$ to $F^n\Omega^1_K = \mathfrak{m}_K^{-n}\Omega^1_{\mathcal{O}_K}(\log)$ for $n \in \mathbb{Z}$ and induces an injection

$$\mathrm{Gr}^n H^1(K, \mathbb{Z}/p^{m+1}\mathbb{Z}) \to \mathrm{Gr}^n\Omega^1_K = Hom_F(\mathfrak{m}^n_K/\mathfrak{m}^{n+1}_K, \Omega^1_F(\log)) \tag{23}$$

for $n > 0$.

Let L be a cyclic extension of degree p^{m+1} corresponding to a character $\chi \in H^1(K, \mathbb{Z}/p^{m+1}\mathbb{Z})$. The smallest integer $n \geq 0$ such that $\chi \in$

$F^n H^1(K, \mathbb{Z}/p^{m+1}\mathbb{Z})$ is called the conductor of χ and is equal to the smallest rational number r such that the ramification of L is bounded by $r+$. The character is ramified if and only if the conductor is > 0. For a ramified character χ of conductor $n > 0$, the image of the class of χ by the injection (23) in $Hom_F(\mathfrak{m}_K^n/\mathfrak{m}_K^{n+1}, \Omega_F^1(\log)) \subset Hom_{\bar{F}}(\mathfrak{m}_{\bar{K}}^n/\mathfrak{m}_{\bar{K}}^{n+}, \Omega_F^1(\log) \otimes_F \bar{F})$ is the refined Swan character $\mathrm{rsw}\chi$.

2.3. Characteristic cycles. We keep the notation in Section 2.2 that X is a separated smooth scheme of dimension d over a perfect field k of characteristic $p > 0$ and $U = X \setminus D$ is the complement of a divisor with simple normal crossings. For each irreducible component D_i of D, let ξ_i be the generic point of D_i and $K_i = \mathrm{Frac}\, \widehat{\mathcal{O}}_{X,\xi_i}$ be the local field. Recall that, for a divisor $R = \sum_i r_i D_i$ with rational coefficients $r_i \geq 0$, we have a cartesian diagram

$$\begin{tikzcd} U \arrow[r, "j"] \arrow[d] & X \arrow[d, "\delta^{(R)}"] \\ U \times_k U \arrow[r, "j^{(R)}"] & (X *_k X)^{(R)}. \end{tikzcd}$$

For a smooth sheaf on U, we make a definition similar to Definition 2.5. As a coefficient ring Λ, we consider a ring finite over $\mathbb{Z}/\ell^n\mathbb{Z}$, $\mathbb{Z}_\ell$ or $\mathbb{Q}_\ell$ for a prime number $\ell \neq \mathrm{char}\, k$ as in Section 1.2.

Definition 2.9 ([38, Definition 2.19]). *Let $\mathcal{F}$ be a smooth sheaf of Λ-modules on U and we put $\mathcal{H}_0 = \mathcal{H}om(\mathrm{pr}_2^*\mathcal{F}, \mathrm{pr}_1^*\mathcal{F})$ on $U \times_k U$. We say that the ramification of $\mathcal{F}$ is bounded by $R+$ if the identity $\mathrm{id}_{\mathcal{F}}$ is in the image of the base change map*

$$\Gamma(X, \delta^{(R)*} j_*^{(R)} \mathcal{H}_0) \to \Gamma(X, j_* \mathcal{E}nd(\mathcal{F})) = End(\mathcal{F}). \tag{24}$$

The following is an immediate consequence of Definition.

Lemma 2.10. *The following conditions are equivalent:*

(1) *The ramification of $\mathcal{F}$ is bounded by $R+$.*

(2) *The base change map $\delta^{(R)*} j_*^{(R)} \mathcal{H}_0 \to j_* \mathcal{E}nd(\mathcal{F})$ is an isomorphism.*

Example 2.4 ([6, Proposition 4.2.2]). Let $\mathcal{F}$ be a smooth sheaf of rank 1 corresponding to a character $\chi\colon \pi_1(U)^{\mathrm{ab}} \to \Lambda^\times$. For each irreducible component D_i, let K_i be the local field and n_i be the conductor of the p-part of the character $\chi_i\colon G_{K_i}^{\mathrm{ab}} \to \Lambda^\times$. We put $R = \sum_i n_i D_i$. Then, the ramification of $\mathcal{F}$ is bounded by $R+$. Further, $j_*^{(R)}\mathcal{H}_0$ is a smooth sheaf of Λ-modules of rank 1 on $(X *_k X)^{(R)}$. For a component with $r_i > 0$, the restriction of $j_*^{(R)}\mathcal{H}_0$ to the fiber $\Theta_{\xi_i}^{(R)}$ is the Artin-Schreier sheaf defined by the refined Swan character $\mathrm{rsw}\chi_i$ regarded as a linear form on $\Theta_{\xi_i}^{(R)}$.

In the remaining part of the article, we present a computation of the characteristic class $C(j_!\mathcal{F}) \in H^{2d}(X, \Lambda(d))$ for a smooth ℓ-adic sheaf $\mathcal{F}$ on U whose ramification is bounded by $R+$ using the geometric ramification theory under a certain assumption. We assume that the coefficients of $R = \sum_i r_i D_i$ are integral for simplicity. For the general case, we refer to [38, Section 3].

For each irreducible component D_i of D such that $r_i > 0$, we consider the ℓ-adic representation M_i of the local field $K_i = \mathrm{Frac}\, \widehat{\mathcal{O}}_{X,\xi_i}$ defined by $\mathcal{F}$. By the assumption that the ramification of $\mathcal{F}$ is bounded by $R+$, the subgroup $G^{r_i+}_{K_i,\log}$ acts trivially on M_i. We assume the following condition:

- The $G^{r_i}_{K_i,\log}$-fixed part of M_i is 0.

This condition means that, for each irreducible component, the wild ramification of $\mathcal{F}$ at the generic point is homogeneous. The restriction to $G^{r_i}_{K_i,\log}$ is then decomposed as the sum $M_i|_{G^{r_i}_{K_i,\log}} = \bigoplus_j \chi_{ij}^{\oplus m_{ij}}$ of non-trivial characters of $\mathrm{Gr}^{r_i}_{\log} G_{K_i}$. By the assumption that r_i is an integer, the refined Swan character defines a non-trivial $\overline{F}_i$-linear homomorphism

$$\mathrm{rsw}\ \chi_{ij} \colon \mathfrak{m}^{r_i}_{K_i}/\mathfrak{m}^{r_i+1}_{K_i} \otimes_{F_i} \overline{F}_i \to \Omega^1_X(\log D)_{\xi_i} \otimes \overline{F}_i,$$

where $F_i = \kappa(\xi_i)$ is the function field of D_i. Let E_{ij} be a finite extension of F_i such that rsw χ_{ij} is defined and let T_{ij} be the normalization of D_i in E_{ij}.

We assume further the following condition:

(C) The refined Swan character rsw χ_{ij} defines a locally splitting injection

$$\mathrm{rsw}\ \chi_{ij} \colon \mathcal{O}_X(-R) \otimes_{\mathcal{O}_X} \mathcal{O}_{T_{ij}} \to \Omega^1_X(\log D) \otimes_{\mathcal{O}_X} \mathcal{O}_{T_{ij}}.$$

This condition says that for each irreducible component, the wild ramification of $\mathcal{F}$ is controlled at the generic point. In the rank one case, the condition (C) is called the cleanness condition and studied in [28]. The key ingredient in the proof of the following computation is Theorem 2.8.

Proposition 2.11 ([38, Corollary 3.3]). *Assume the condition* (C) *above is satisfied. Let* $\pi^{(R)} \colon (X *_k X)^{(R)} \to X \times_k X$ *denote the canonical map. We put* $\mathcal{H}_0 = \mathcal{H}om(\mathrm{pr}_2^*\mathcal{F}, \mathrm{pr}_1^*\mathcal{F})$ *on* $U \times_k U$ *and* $\mathcal{H} = R\mathcal{H}om(\mathrm{pr}_2^* j_!\mathcal{F}, R\mathrm{pr}_1^! j_!\mathcal{F})$ *on* $X \times_k X$. *We regard the identity* $\mathrm{id}_{\mathcal{F}}$ *as an element of* $H^0_X(X \times_k X, \mathcal{H})$ *and of* $H^0(X, \delta^{(R)*} j_*^{(R)} \mathcal{H}_0)$ *by the isomorphisms* $End(j_!\mathcal{F}) \to H^0_X(X \times_k X, \mathcal{H})$ (6) *and* $H^0(X, \delta^{(R)*} j_*^{(R)} \mathcal{H}_0) \to End(\mathcal{F})$ (24).

Then, the image of the identity $\mathrm{id}_{\mathcal{F}}$ *by the pull-back map*

$$H^0_X(X \times_k X, \mathcal{H}) \longrightarrow H^{2d}_{\pi^{(R)-1}(X)}((X *_k X)^{(R)}, j_*^{(R)} \mathcal{H}_0(d))$$

is equal to the image of the cup-product $[X] \cup \mathrm{id}_{\mathcal{F}} \in H^{2d}_X((X *_k X)^{(R)}, j_*^{(R)} \mathcal{H}_0(d))$ *of the cycle class* $[X] \in H^{2d}_X((X *_k X)^{(R)}, \Lambda(d))$ *with* $\mathrm{id}_{\mathcal{F}} \in H^0(X, \delta^{(R)*} j_*^{(R)} \mathcal{H}_0)$.

Corollary ([38, Theorem 3.4]). *For the characteristic class, we have an equality*

$$C(j_!\mathcal{F}) = \text{rank } \mathcal{F} \cdot (X, X)_{(X *_k X)^{(R)}}.$$

Consequently, if X is proper, we have

$$\chi_c(U_{\bar{k}}, \mathcal{F}) = \text{rank } \mathcal{F} \times \deg(X, X)_{(X *_k X)^{(R)}}.$$

We keep the assumptions and define the characteristic cycle in order to describe the computation in Corollary more geometric terms. We call the vector bundle over X defined by the symmetric $\mathcal{O}_X$-algebra of the dual module $\Omega^1_X(\log D)^*$ the logarithmic cotangent bundle $T^*X(\log D)$. Let L denote the line bundle over X defined by the symmetric $\mathcal{O}_X$-algebra of $\mathcal{O}_X(R)$. By the condition (C), the refined Swan character rsw χ_{ij} defines a linear map

$$r_{ij} \colon L \times_X T_{ij} \to T^*X(\log D) \times_X T_{ij}.$$

We define the characteristic cycle $CC(\mathcal{F})$ by

$$CC(\mathcal{F}) = (-1)^d \left(\text{rank } \mathcal{F} \cdot [X] + \sum_i \sum_j \frac{m_{ij}}{[E_{ij} : F_i]} \text{pr}_{1*} r_{ij*}[L \times_X T_{ij}] \right)$$

as a dimension d-cycle of $T^*X(\log D)$. In the first term of the right hand side, $[X]$ denotes the class of the 0-section. In the second term, $\text{pr}_{1*} r_{ij*}[L \times_X T_{ij}]$ denote the image of the class $[L \times_X T_{ij}]$ by the composition $L \times_X T_{ij} \to T^*X(\log D) \times_X T_{ij} \to T^*X(\log D)$. The reason why the characteristic cycle defined above is determined by points of codimension 1 is the condition (C).

As a consequence of Corollary, we have the following.

Theorem 2.12 ([38, Theorem 3.7]). *Assume the condition* (C).

1. *The characteristic class $C(j_!\mathcal{F})$ is equal to the pull-back by the* 0*-section* $0 \colon X \to T^*X(\log D)$ *of the cycle class of the characteristic cycle $CC(\mathcal{F})$:*

$$C(j_!\mathcal{F}) = 0^*[CC(\mathcal{F})].$$

2. *Assume further that X is proper. Then the Euler number $\chi_c(U_{\bar{k}}, \mathcal{F})$ is equal to the intersection number of the* 0*-section with the characteristic cycle:*

$$\chi_c(U_{\bar{k}}, \mathcal{F}) = (X, CC(\mathcal{F}))_{T^*X(\log D)}.$$

Problem 5. Find an intrinsic definition of the characteristic cycle.

References

[1] A. Abbes, *Cycles on arithmetic surfaces*, Compositio Math. 122 (2000), no. 1, 23–111.

[2] ———, *The Grothendieck-Ogg-Shafarevich formula for arithmetic surfaces*, J. Algebraic Geom. 9 (2000), no. 3, 529–576.

[3] A. Abbes, A. Mokrane, *Sous-groupes canoniques et cycles évanescents p-adiques pour les variétés abéliennes*, Publ. Math. IHES 101, (2004), 117–162

[4] A. Abbes, T. Saito, *Ramification of local fields with imperfect residue fields*, Amer. J. of Math. **124** (2002), 879–920

[5] ———, *Ramification of local fields with imperfect residue fields II*, Documenta Math., Extra Volume K. Kato (2003), 3–70.

[6] ———, *The characteristic class and ramification of an ℓ-adic étale sheaf*, Invent. Math., 168 (2007) 567–612.

[7] ———, *Analyse micro-locale ℓ-adique en caractéristique $p > 0$: Le cas d'un trait*, Publ. RIMS 45–1 (2009) 25–74.

[8] ———, *Local Fourier transform and epsilon factors*, `arXiv:0809.0180` accepted for publication at Compo. Math.

[9] T. Abe, *Comparison between Swan conductors and characteristic cycles*, accepted for publication at Compo. Math.

[10] P. Berthelot, *Introduction à la théorie arithmétique des $\mathcal{D}$-modules*, Cohomologies p-adiques et applications arithmétiques, II. Astérisque 279 (2002), 1–80.

[11] S. Bloch, *Cycles on arithmetic schemes and Euler characteristics of curves*, Algebraic geometry, Bowdoin, 1985, 421–450, Proc. Sympos. Pure Math. 46, Part 2, Amer. Math. Soc., Providence, RI, 1987.

[12] S. Bosch, W. Lütkebohmert, and M. Raynaud, *Formal and rigid geometry. IV. The reduced fibre theorem,* Invent. Math. 119 (1995), 361–398.

[13] J.-L. Brylinski, *Théorie du corps de classes de Kato et revêtements abéliens de surfaces*, Ann. Inst. Fourier 33 (1983), 23–38.

[14] V. Cossart, U. Jannsen and S. Saito, *Canonical embedded and non-embedded resolution of singularities for excellent two-dimensional schemes*, Preprint 2009, `arXiv:math.AG/0905.2191`

[15] A. J. de Jong, *Families of curves and alterations*, Ann. Inst. Fourier (Grenoble) 47 (1997), no. 2, 599–621.

[16] P. Deligne, *Équations différentielles à points singuliers réguliers,* LNM **163,** Springer-Verlag, Berlin-New York, 1970.

[17] ———, Letter to Illusie, Nov. 4, 1976

[18] H. P. Epp, *Eliminating wild ramification*, Invent. Math. 19 (1973), 235–249.

[19] A. Grothendieck, rédigé par L. Illusie, *Formule de Lefschetz,* exposé III, SGA 5, Springer LNM **589** (1977) 73–137.

[20] A. Grothendieck, rédigé par I. Bucur, *Formule d'Euler-Poincaré en cohomologie étale,* exposé X, SGA 5, Springer LNM **589** (1977) 372–406.

[21] G. Faltings, *Crystalline cohomology and p-adic Galois-representations,* Algebraic analysis, geometry, and number theory (Baltimore, MD, 1988), 25–80, Johns Hopkins Univ. Press, Baltimore, MD, 1989.

[22] S. Hattori, *Tame characters and ramification of finite flat group schemes*, J. of Number theory 128 (2008), 1091–1108

[23] L. Illusie, *Théorie de Brauer et caractéristique d'Euler-Poincaré, d'après P. Deligne,* Astérisques 82–83, SMF, (1981), 161–172.

[24] M. Kashiwara, P. Schapira, Sheaves on manifolds, Springer-Verlag (1990).

[25] K. Kato, *Swan conductors for characters of degree one in the imperfect residue field case*, Algebraic K-theory and algebraic number theory (Honolulu, HI, 1987), 101–131, Contemp. Math., 83, Amer. Math. Soc., Providence, RI, 1989.

[26] ———, *Logarithmic structures of Fontaine-Illusie,* Algebraic analysis, geometry, and number theory, (J.-I. Igusa ed.), Johns Hopkins UP, Baltimore, 1989, 191–224.

[27] ———, *Generalized class field theory,* Proceedings of ICM (Kyoto, 1990), 419–428, Math. Soc. Japan, Tokyo, 1991.

[28] ———, *Class field theory, D-modules, and ramification of higher dimensional schemes, Part I,* American J. of Math., 116 (1994), 757–784.

[29] K. Kato, S. Saito, T. Saito, *Artin characters for algebraic surfaces,* American J. of Math. 109 (1987), 49–76.

[30] K. Kato, T. Saito, *On the conductor formula of Bloch,* Publ. Math. IHES 100, (2004), 5–151.

[31] ———, *Ramification theory for varieties over a perfect field*, Ann. of Math., 168 (2008), 33–96.

[32] G. Laumon, *Caractéristique d'Euler-Poincaré des faisceaux constructibles sur une surface,* Astérisques, 101–102 (1982), 193–207.

[33] X. Liang, Preprints on his webpage
`http://math.uchicago.edu/~lxiao/papers.htm`

[34] A. P. Ogg, *Elliptic curves and wild ramification*, Amer. J. Math. 89 (1967) 1–21.

[35] F. Orgogozo, *Conjecture de Bloch et nombres de Milnor,* Annales de l'Institut Fourier 53, no. 6 (2003) 1739–1754

[36] R. Pink, *On the calculation of local terms in the Lefschetz-Verdier trace formula and its application to a conjecture of Deligne,* Ann. of Math. (2) 135 (1992), no. 3, 483–525.

[37] Mme M. Raynaud (d'après notes inédites de A. Grothendieck), *Propreté cohomologique des faisceaux d'ensembles et des faisceaux de groupes non commutatifs*, exposé XIII, SGA 1, Springer LNM **224** (1971), Édition recomposée SMF (2003).

[38] T. Saito, *Wild ramification and the characteristic cycle of an ℓ-adic sheaf*, Journal de l'Institut de Mathematiques de Jussieu, (2009) 8(4), 769–829

[39] J-P. Serre, *Sur la rationalité des représentations d'Artin,* Ann. of Math. 72 (1960), 406–420.

[40] ———, Corps Locaux, Hermann, Paris, France, 1962.

[41] ———, Représentations linéaires des groupes finis, Hermann, Paris, France, 1967.

[42] Y. Tian, *Canonical subgroups of Barsotti-Tate groups*, accepted for publication at Ann. of Math.

[43] T. Tsushima, *On localization of the characteristic class of ℓ-adic sheaves and relative Kato-Saito conductor formula in characteristic $p > 0$*, preprint.

[44] I. Vidal, *Théorie de Brauer et conducteur de Swan*, J. Algebraic Geom. 13 (2004), 349–391.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Quantum Unique Ergodicity and Number Theory

K. Soundararajan*

Abstract

A fundamental problem in the area of quantum chaos is to understand the distribution of high eigenvalue eigenfunctions of the Laplacian on certain Riemannian manifolds. A particular case which is of interest to number theorists concerns hyperbolic surfacess arising as a quotient of the upper half-plane by a discrete "arithmetic" subgroup of $SL_2(\mathbb{R})$ (for example, $SL_2(\mathbb{Z})$, and in this case the corresponding eigenfunctions are called Maass cusp forms). In this case, Rudnick and Sarnak have conjectured that the high energy eigenfunctions become equi-distributed. I will discuss some recent progress which has led to a resolution of this conjecture, and also on a holomorphic analog for classical modular forms

Mathematics Subject Classification (2010). Primary 11F11, 11F67, 11M99, 11N64.

Keywords. Quantum unique ergodicity, modular surface, Hecke operators, subconvexity problem, L-functions, multiplicative functions, sieve methods.

1. Introduction

Given a Riemannian manifold, it is of great interest to study the behavior of eigenfunctions of the Laplacian in the limit as the eigenvalue tends to infinity. In particular if the eigenvalue is large, we may ask if the L^2-mass of the eigenfunction is spread out evenly over the manifold, or if it can accumulate in sub-regions. A general result of Shnirelman [47], Colin de Verdiere [2], and Zelditch [53] states that when the geodesic flow (which corresponds to the classical dynamics on this space) is ergodic, a typical eigenfunction of large eigenvalue does get evenly distributed (more precisely, one has equidistribution for

*The author is partially supported by a grant from the National Science Foundation (DMS 0500711).

Department of Mathematics, Stanford University, Stanford, CA 94305.
E-mail: ksound@math.stanford.edu.

a density one subsequence of eigenfunctions). This result is known as *quantum ergodicity*, and one is led to wonder if every eigenfunction gets equi-distributed in the large eigenvalue limit; more precisely, what are the possible weak-$*$ limits of the eigenfunctions? This problem is known as *quantum unique ergodicity*. A recent result of Hassell [17] shows that in the case of "stadium billiards" the answer is negative.

In contrast, for (strictly) negatively curved compact manifolds (where the geodesic flow is chaotic) Rudnick and Sarnak [41] have conjectured that QUE holds. In this generality, the conjecture remains wide open, but recently Anantharaman [1] made significant progress by showing that any limiting measure must have positive entropy.

The QUE conjecture has been established for a class of manifolds arising in number theory, and our aim in this article is to describe some of these developments; for other accounts see [32], [35], [43] and [44]. Specifically let us consider the surfaces $\Gamma\backslash\mathbb{H}$ of constant negative curvature, where $\mathbb{H}$ denotes the upper half-plane, and Γ a discrete subgroup of $SL_2(\mathbb{R})$ with the quotient having finite area. The proto-typical example of an arithmetic group is $\Gamma = SL_2(\mathbb{Z})$ and we shall focus largely on this case. Note here that the quotient $SL_2(\mathbb{Z})\backslash\mathbb{H}$ is not compact, but of finite area; recall that the area measure on $\mathbb{H}$ is given by $\frac{dx\ dy}{y^2}$, and that a fundamental domain for $SL_2(\mathbb{Z})\backslash\mathbb{H}$ is the region $\{z = x+iy:\ -1/2 < x \le 1/2, |z| \ge 1\}$ (with the additional constraint that $x \ge 0$ if $|z| = 1$) which has area $\pi/3$. Other examples of arithmetic surfaces are the quotients of congruence subgroups of $SL_2(\mathbb{Z})$, or the quotients of groups arising from quaternion algebras (and these quotients are compact). The distinguishing feature of these arithmetic surfaces (and which is responsible for the success in this case) is the presence of a large family of commuting, self-adjoint operators known as the *Hecke operators*.

The spectrum of the hyperbolic Laplacian $\Delta = -y^2(\frac{d^2}{dx^2} + \frac{d^2}{dy^2})$ acting on $\Gamma\backslash\mathbb{H}$ falls into three types: the constant function, unitary Eisenstein series $E(z, \frac{1}{2} + it)$ with $t \in \mathbb{R}$ (these are not in L^2 and constitute the continuous spectrum), and a discrete spectrum of Maass forms ϕ. The Maass forms are square-integrable and decay rapidly at infinity. The existence of Maass forms is not evident, but the Selberg trace formula demonstrates this, and also counts the number of Maass forms with eigenvalue up to T. Given a Maass form ϕ of eigenvalue λ normalized to have $\int_{\Gamma\backslash\mathbb{H}} |\phi(z)|^2 \frac{dx\ dy}{y^2} = 1$ we denote by $\mu_\phi(z)$ the measure $|\phi(z)|^2 \frac{dx\ dy}{y^2}$. Quantum ergodicity (see [54]) tells us that for typical eigenfunctions ϕ, as $\lambda \to \infty$ the measures μ_ϕ tend to the uniform distribution measure $\frac{3}{\pi}\frac{dx\ dy}{y^2}$. The Rudnick-Sarnak QUE conjecture asserts that this holds for every sequence of eigenfunctions; that is, no other weak-$*$ limit is possible. One can also formulate a version of QUE for the Eisenstein series $E(z, \frac{1}{2} + it)$ and this was established by Luo and Sarnak [34] and by Jakobson [29] for a stronger lifted version on $SL_2(\mathbb{Z})\backslash\mathbb{H}$.

The space of functions on $\Gamma\backslash\mathbb{H}$ has a natural inner product, the Petersson inner product, given by $\langle f, g\rangle = \int_{\Gamma\backslash\mathbb{H}} f(z)\overline{g(z)}\frac{dx\ dy}{y^2}$. For each natural number

n, the arithmetic surface $\Gamma\backslash\mathbb{H}$ has a Hecke operator T_n defined by

$$(T_n f)(z) = \frac{1}{\sqrt{n}} \sum_{ad=n} \sum_{b \pmod d} f\left(\frac{az+b}{d}\right).$$

These Hecke operators satisfy $T_m T_n = \sum_{d|(m,n)} T_{mn/d^2}$, so that they commute, and are self-adjoint with respect to the Petersson inner product. The unitary Eisenstein series $E(z, \frac{1}{2}+it)$ are eigenfunctions of all the Hecke operators, and we may also diagonalize the space of Maass forms to get a basis of eigenfunctions for all Hecke operators. We call such eigenfunctions Hecke-Maass forms. Numerical experiments suggest that the spectrum of the Laplacian on $\Gamma\backslash\mathbb{H}$ is simple, so that every Maass form would automatically be an eigenfunction of all Hecke operators, but this is very far from being known. From the point of view of number theory, in studying QUE for $\Gamma\backslash\mathbb{H}$ it is natural to restrict to Hecke-Maass forms, and it is this problem that has now been solved.

Lindenstrauss [31] made great progress towards QUE for Hecke-Maass forms. He considers micro-local lifts of the Hecke-Maass forms to the unit tangent bundle $SL_2(\mathbb{Z})\backslash SL_2(\mathbb{R})$, and these lifts are known to be approximately invariant under the geodesic flow. Using results from measure rigidity, he then shows that the only possible limiting measures are of the form $\frac{3}{\pi}c\frac{dx\ dy}{y^2}$ (restricting ourselves to the modular surface $SL_2(\mathbb{Z})\backslash\mathbb{H}$; Lindenstrauss's result holds for the larger space $SL_2(\mathbb{Z})\backslash SL_2(\mathbb{R})$) where $0 \le c \le 1$. In other words, the measures get equi-distributed except for the possibility that some of the mass escapes into the cusp at infinity. Recently I showed [49] that escape of mass is not possible here, so that $c = 1$, and the proof of QUE for Hecke-Maass forms on $SL_2(\mathbb{Z})\backslash\mathbb{H}$ is complete. It is conceivable that QUE fails for a different basis of Maass forms, but as noted before that seems unlikely.

Theorem 1. *For any sequence of L^2-normalized Hecke-Maass eigenforms ϕ_j, the measures $|\phi_j|^2 \frac{dx\ dy}{y^2}$ tend weakly to the measure $\frac{3}{\pi}\frac{dx\ dy}{y^2}$ as the Laplace eigenvalue of ϕ_j tends to infinity.*

In addition to the Maass forms discussed above, the complex structure of $\mathbb{H}$ allows for a rich theory of holomorphic functions which transform nicely under the action of $SL_2(\mathbb{Z})$. The classical theory of modular forms of weight k (an even positive integer) considers holomorphic functions $f : \mathbb{H} \to \mathbb{C}$ satisfying $f(\gamma z) = (cz+d)^k f(z)$ for all $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$. If we also require f to be holomorphic and decay rapidly "at the cusp at infinity" then we get the theory of cusp forms. The most famous example of a cusp form is Ramanujan's Δ-function given by

$$\Delta(z) = q\prod_{n=1}^{\infty}(1-q^n)^{24}, \qquad q = e^{2\pi i z}.$$

The space of cusp forms of weight k comes equipped with an inner product: $\langle f, g\rangle_k = \int_{\Gamma\backslash\mathbb{H}} y^k f(z)\overline{g(z)}\frac{dx\ dy}{y^2}$. Normalizing a cusp form to have L^2 norm

$\|f\|^2 = \langle f, f\rangle_k = 1$, we may ask whether the holomorphic analog of QUE holds: that is, whether the measure $\mu_f := y^k|f(z)|^2 \frac{dx\ dy}{y^2}$ tends to the equidistribution measure $\frac{3}{\pi}\frac{dx\ dy}{y^2}$. Here we see that some care must be taken: The space $S_k(SL_2(\mathbb{Z}))$ of cusp forms of weight k for $SL_2(\mathbb{Z})$ is a vector space of dimension about $k/12$, and contains elements such as $\Delta(z)^{k/12}$ (if $12|k$, and where Δ is Ramanujan's cusp form given above) for which the measure will not tend to uniform distribution. Therefore one must restrict attention to a particular basis of cusp forms, and it is natural to consider the basis of eigenfunctions of all the Hecke operators. Analogously to the Maass form case we may define the n-th Hecke operator acting on modular forms of weight k by

$$(T_n f)(z) = \frac{1}{n^{\frac{k+1}{2}}} \sum_{ad=n} a^k \sum_{b \pmod d} f\left(\frac{az+b}{d}\right).$$

The Hecke operators commute, with $T_m T_n = \sum_{d|(m,n)} T_{mn/d^2}$, and are self-adjoint with respect to the Petersson inner product. Thus we may choose a basis for the space of cusp forms of weight k consisting of eigenfunctions for all the Hecke operators.

The Rudnick-Sarnak conjecture in this context states that as $k \to \infty$, for *every* Hecke eigencuspform f the measure μ_f tends to the uniform distribution measure. Luo and Sarnak [33] have shown that equidistribution holds for most Hecke eigenforms, and Sarnak [42] has shown that it holds in the special case of dihedral forms (these are not present for $SL_2(\mathbb{Z})$ but arise in the case of congruence subgroups). There seems to be no apparent way to define a microlocal lift of a holomorphic form with invariance under the geodesic flow, and so it does not seem clear how to adapt Lindenstrauss's method to the holomorphic setting.

Theorem 2. *For any sequence of L^2-normalized Hecke eigencusp forms f of weight k, the measures $\mu_f = y^k|f(z)|^2 \frac{dx\ dy}{y^2}$ tend weakly to $\frac{3}{\pi}\frac{dx\ dy}{y^2}$ as $k \to \infty$.*

The proof of this holomorphic analog of QUE combines two different approaches developed independently by Holowinsky [23] and Soundararajan [50]. At their heart, both approaches rely on an understanding of mean-values of multiplicative functions. Either of these approaches is capable of showing that there are very few possible exceptions to the conjecture, and under reasonable hypotheses either approach would show that there are no exceptions. However, it seems difficult to show unconditionally that there are no exceptions using just one of these approaches. Fortunately, as we shall explain below, the two approaches are complementary, and the few rare cases that are untreated by one method fall easily to the other method. Both approaches use in an essential way that the Hecke eigenvalues of a holomorphic eigencuspform satisfy the Ramanujan conjecture which was established by Deligne. Deligne's theorem tells us that the eigenvalues of the Hecke operator T_p (for a prime p) are bounded in magnitude by 2. The Ramanujan conjecture remains open for Maass forms, and this is the (only) barrier to using our methods in the non-holomorphic setting.

While the holomorphic analog of QUE does not have an interpretation in terms of quantum versus classical dynamics, it does imply a striking corollary on the distribution of zeros of modular forms. A cusp form of weight k has about $k/12$ zeros inside a fundamental domain. How are these zeros distributed? If we take a large power of Ramanujan's Δ function, then there is only one zero of multiplicity $k/12$ at the cusp at ∞. However, if the L^2-mass of f is equidistributed on the fundamental domain, then Rudnick [40] (see also the work of Schiffman and Zelditch [45], and Nonnenmacher and Voros [39] in related contexts) showed that the zeros are also equidistributed (with the measure $\frac{3}{\pi}\frac{dx\ dy}{y^2}$). In particular, Theorem 2 implies the following corollary.

Corollary 1. *The zeros of a Hecke eigencusp form of large weight k are equidistributed inside a fundamental domain with respect to the measure $\frac{3}{\pi}\frac{dx\ dy}{y^2}$.*

2. Spectral Expansions, and Expansions into Incomplete Eisenstein and Poincare Series

Let h denote a smooth bounded function on $X = SL_2(\mathbb{Z})\backslash\mathbb{H}$. Considering h as fixed, the holomorphic Rudnick-Sarnak conjecture asserts that for a Hecke eigencuspform f of weight k (normalized to have L^2-norm 1) we have

$$\int_X y^k|f(z)|^2 h(z)\frac{dx\ dy}{y^2} \to \frac{3}{\pi}\int_X h(z)\frac{dx\ dy}{y^2}, \tag{1}$$

as $k \to \infty$ with the rate of convergence depending on the function h. To attack the conjecture (1), it is convenient to decompose the function h in terms of a basis of smooth functions on X. There are two natural ways of doing this, and both decompositions play important roles in the proof of the Rudnick-Sarnak conjecture.

First, we could use the spectral decomposition of a smooth function on X in terms of eigenfunctions of the Laplacian. The spectral expansion will involve (i) the constant function $\sqrt{3/\pi}$, (ii) Maass cusp forms ϕ that are also eigenfunctions of all the Hecke operators, and (iii) Eisenstein series on the $\frac{1}{2}$ line. Recall that the Eisenstein series is defined for $\mathrm{Re}(s) > 1$ by

$$E(z,s) = \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \mathrm{Im}(\gamma z)^s,$$

where $\Gamma = SL_2(\mathbb{Z})$ and Γ_∞ denotes the stabilizer group of the cusp at infinity (namely the set of all translations by integers). The Eisenstein series $E(z,s)$ admits a meromorphic continuation, with a simple pole at $s = 1$, and is analytic for s on the line $\mathrm{Re}(s) = \frac{1}{2}$. For more on the spectral expansion see Iwaniec's book [25].

Note that (1) is trivial when h is the constant eigenfunction. To establish (1) using the spectral decomposition, we would need to show that for a fixed

Maass eigencuspform ϕ, and for a fixed real number t that

$$\left|\int_X y^k|f(z)|^2\phi(z)\frac{dx\ dy}{y^2}\right|, \qquad \text{and} \qquad \left|\int_X y^k|f(z)|^2E(z,\tfrac{1}{2}+it)\frac{dx\ dy}{y^2}\right|$$

tend to 0 as $k\to\infty$. The above statement should be thought of as an analog of Weyl's equidistribution criterion. The two inner products above may be related to special values of certain L-functions, and we shall discuss this connection in the next section.

Alternatively, one could expand the function h in terms of incomplete Poincare and Eisenstein series. Let ψ denote a smooth function, compactly supported in $(0,\infty)$. For an integer m the incomplete Poincare series is defined by

$$P_m(z\mid\psi)=\sum_{\gamma\in\Gamma_\infty\backslash\Gamma} e(m\gamma z)\psi(\mathrm{Im}(\gamma z)).$$

In the special case $m=0$ we obtain incomplete Eisenstein series $E(z\mid\psi)=P_0(z\mid\psi)$. By taking a Fourier expansion of $h(x+iy)$ for each fixed value of y we may approximate h using incomplete Poincare and Eisenstein series; for details see Luo and Sarnak [34]. Now conjecture (1) can be reformulated (again analogously to Weyl's equidistribution criterion) as saying that as $k\to\infty$

$$\left|\int_X y^k|f(z)|^2P_m(z\mid\psi)\frac{dx\ dy}{y^2}\right|\to 0,$$

for $m\neq 0$ (considered to be fixed), and any given smooth function ψ. In the case $m=0$ we want that

$$\int_X y^k|f(z)|^2E(z\mid\psi)\frac{dx\ dy}{y^2}\to\frac{3}{\pi}\int_X E(z,\psi)\frac{dx\ dy}{y^2},$$

for any fixed ψ and as $k\to\infty$. The Rankin-Selberg unfolding method can be used to handle these inner products. For example the inner product with Poincare series (for $m\neq 0$) was related by Luo and Sarnak [33] to the problem of estimating the shifted convolution sums (for m fixed, and as $k\to\infty$)

$$\sum_{n\asymp k}\lambda_f(n)\lambda_f(n+m),$$

where the sum is over n of size k, and $\lambda_f(n)$ denotes the n-th Hecke eigenvalue of f. We will discuss the inner products with these incomplete Poincare and Eisenstein series in more detail in section 4.

3. Relation to L-functions and the Subconvexity Problem

In the approach to the Rudnick-Sarnak conjecture via a spectral expansion, we need to estimate the inner products of $y^k|f(z)|^2$ with fixed Hecke-Maass form

ϕ, and Eisenstein series $E(\cdot, \frac{1}{2}+it)$ with t fixed. Both these inner products are related to L-functions; the latter being a classical result of Rankin and Selberg, and the former a recent result given explicitly by Watson [51].

Let $\lambda_f(n)$ denote the n-th Hecke eigenvalue of the cusp form f. For a prime number p we may write the Hecke eigenvalue $\lambda_f(p)$ as $\alpha_f(p)+\beta_f(p)$ where $\alpha_f(p)$ and $\beta_f(p)$ are complex numbers satisfying $\alpha_f(p)\beta_f(p) = 1$ and, by Deligne's theorem, $|\alpha_f(p)| = |\beta_f(p)| = 1$. The L-function associated to f is then

$$L(s,f) = \sum_{n=1}^{\infty} \frac{\lambda_f(n)}{n^s} = \prod_p \left(1 - \frac{\alpha_p}{p^s}\right)^{-1} \left(1 - \frac{\beta_p}{p^s}\right)^{-1},$$

where the series and product above are absolutely convergent in $\sigma > 1$, and $L(s,f)$ extends analytically to $\mathbb{C}$ with a functional equation connecting the values at s and $1-s$. Of greater importance for us is the related symmetric square L-function which is given by

$$L(s, \mathrm{sym}^2 f) = \sum_{n=1}^{\infty} \frac{\lambda_f^{(2)}(n)}{n^s} = \prod_p \left(1 - \frac{\alpha_p^2}{p^s}\right)^{-1} \left(1 - \frac{1}{p^s}\right)^{-1} \left(1 - \frac{\beta_p^2}{p^s}\right)^{-1}.$$

The series and product above converge absolutely in $\mathrm{Re}(s) > 1$, and by the work of Shimura [46], we know that $L(s, \mathrm{sym}^2 f)$ extends analytically to the entire complex plane, and satisfies the functional equation

$$\begin{aligned} \Lambda(s, \mathrm{sym}^2 f) &= \Gamma_{\mathbb{R}}(s+1)\Gamma_{\mathbb{R}}(s+k-1)\Gamma_{\mathbb{R}}(s+k)L(s, \mathrm{sym}^2 f) \\ &= \Lambda(1-s, \mathrm{sym}^2 f), \end{aligned}$$

where $\Gamma_{\mathbb{R}}(s) = \pi^{-s/2}\Gamma(s/2)$. The symmetric square L-function appears naturally when we normalize the cusp form f to have L^2-norm 1. Recall that f has a Fourier expansion

$$f(z) = C \sum_{n=1}^{\infty} \lambda_f(n) n^{\frac{k-1}{2}} e(nz),$$

where C is a positive constant which is chosen so as to make the L^2 norm of f equal to 1. This constant C is then related to the symmetric square L-function by

$$|C|^2 = \frac{(4\pi)^{k-1}}{\Gamma(k)} \frac{2\pi^2}{L(1, \mathrm{sym}^2 f)}.$$

Now we return to the inner products of $y^k|f(z)|^2$ with Eisenstein series and Maass forms. In the former case of the classical "unfolding method" of Rankin and Selberg (starting with $E(z,s)$ in the domain of absolute convergence, and extending to $s = 1/2 + it$ by analytic continuation) leads to

$$\left| \int_{\Gamma\backslash\mathbb{H}} y^k |f(z)|^2 E(z, \tfrac{1}{2}+it) \frac{dx\, dy}{y^2} \right| = \left| \pi^{\frac{3}{2}} \frac{\zeta(\frac{1}{2}+it) L(\frac{1}{2}+it, \mathrm{sym}^2 f)}{\zeta(1+2it) L(1, \mathrm{sym}^2 f)} \frac{\Gamma(k-\frac{1}{2}+it)}{\Gamma(k)} \right|.$$

Since $|\Gamma(k-\frac{1}{2}+it)| \le \Gamma(k-\frac{1}{2})$, $|\zeta(\frac{1}{2}+it)| \ll (1+|t|)^{\frac{1}{4}}$, and $|\zeta(1+2it)| \gg 1/\log(1+|t|)$, using Stirling's formula it follows that

$$\left|\int_{\Gamma\backslash\mathbb{H}} y^k |f(z)|^2 E(z,\tfrac{1}{2}+it)\frac{dx\ dy}{y^2}\right| \ll \frac{(1+|t|)^2 |L(\frac{1}{2}+it,\mathrm{sym}^2 f)|}{k^{\frac{1}{2}} L(1,\mathrm{sym}^2 f)}.$$

The works of Hoffstein and Lockhart [21] and Goldfeld, Hoffstein and Lieman [8] establish a classical zero free region for the symmetric square L-function. Further, their work shows that the denominator above, $L(1,\mathrm{sym}^2 f)$, is $\gg 1/(\log k)$. Hence the inner product with Eisenstein series tends to zero provided we can establish an upper bound for $|L(\frac{1}{2}+it,\mathrm{sym}^2 f)|$ which is better than $k^{\frac{1}{2}}/\log k$.

This is a special case of the general problem of bounding L-functions on the critical line. This problem has a long history, going back to work of Weyl, Hardy and Littlewood in the case of the Riemann zeta-function. In general one has a bound for L-functions of the form $\ll C^{\frac{1}{4}}$, where C is an object called the *analytic conductor* which measures the complexity of the L-function. Such a bound is called the convexity bound; usually the convexity bound is stated as $\ll C^{\frac{1}{4}+\epsilon}$, and the refined bound we have stated is a recent observation of Heath-Brown [18]. For example, for the zeta-function the convexity bound states that $|\zeta(\frac{1}{2}+it)| \ll |t|^{\frac{1}{4}}$ and the work of Weyl-Hardy-Littlewood furnished improvements over this, leading for example to $|\zeta(\frac{1}{2}+it)| \ll |t|^{\frac{1}{6}}$. Here the truth is expected to be the Lindelöf bound $|\zeta(\frac{1}{2}+it)| \ll |t|^{\epsilon}$, and this bound is a consequence of the Riemann hypothesis. The problem of obtaining a bound for L-values of the shape $C^{\frac{1}{4}-\delta}$ for some $\delta > 0$ is known as the *subconvexity problem*, and is an important outstanding problem in number theory. The subconvexity problem is now resolved for L-functions arising from $GL(1)$ or $GL(2)$, and a handful of other cases, but in general the problem is wide open. One of the most striking applications of subconvexity is to the problem of representing integers by ternary quadratic forms (see [3]). We refer the reader to [28], [36], and [37] for comprehensive accounts on the subconvexity problem.

Returning to the case at hand, we need a bound for $|L(\frac{1}{2}+it,\mathrm{sym}^2 f)|$ and the analytic conductor for this L-function is about $(1+|t|)^3 k^2$. The convexity bound gives $|L(\frac{1}{2}+it,\mathrm{sym}^2 f)| \ll k^{\frac{1}{2}}(1+|t|)^{\frac{3}{4}}$. Using this in our inner product with Eisenstein series, we realize that this is barely insufficient to show that decay of this inner product, and any subconvexity bound would be sufficient. The Generalized Riemann Hypothesis implies such a bound (in fact that the L-value is $\ll k^{\epsilon}(1+|t|)^{\epsilon}$), but unconditionally subconvexity for symmetric square L-functions is not known. Recently, X. Li [30] obtained a subconvexity bound for k fixed and $t \to \infty$, but for our application we want the opposite case of t fixed and $k \to \infty$. For a general class of L-functions, I established recently a *weak subconvexity* bound (described in §5) which implies that

$$|L(\tfrac{1}{2}+it,\mathrm{sym}^2 f)| \ll \frac{k^{\frac{1}{2}}(1+|t|)^{\frac{3}{4}}}{(\log k)^{1-\epsilon}}. \tag{2}$$

Since we only know that $L(1, \mathrm{sym}^2 f) \gg 1/\log k$ we see that weak subconvexity also fails (now only by $(\log k)^\epsilon$) to show the decay of inner products with Eisenstein series. However one can show that $L(1, \mathrm{sym}^2 f)$ is very rarely less than $(\log k)^{-\delta}$ for any $\delta > 0$ (there are at most K^ϵ exceptional Hecke eigenforms with weight below K), and so for the vast majority of cases weak subconvexity suffices. On GRH we also know that $L(1, \mathrm{sym}^2 f) \gg 1/\log\log k$, but improving lower bounds for L-functions on the 1-line unconditionally is a very difficult problem connected with widening the zero-free region for that L-function (and so quite likely harder than subconvexity!).

Now let us turn to the inner product with a fixed Hecke-Maass cusp form. Let ϕ denote a fixed Hecke-Maass cusp form with Laplace eigenvalue $\lambda_\phi = \frac{1}{4} + t_\phi^2$, and normalized to have L^2 norm 1. In exact analogy with the Eisenstein series case, a deep and beautiful formula of Tom Watson (see Theorem 3 of [51]) shows that

$$\left| \int_{\Gamma\backslash\mathbb{H}} y^k |f(z)|^2 \phi(z) \frac{dx\ dy}{y^2} \right|^2 = \frac{1}{8} \frac{L_\infty(\frac{1}{2}, f \times f \times \phi) L(\frac{1}{2}, f \times f \times \phi)}{\Lambda(1, \mathrm{sym}^2 f)^2 \Lambda(1, \mathrm{sym}^2 \phi)}$$

where $L(s, f \times f \times \phi)$ is the triple product L-function and L_∞ denotes its Gamma factors, whose definitions we give below. Also, in the formula above we have set

$$\Lambda(s, \mathrm{sym}^2 f) = \Gamma_{\mathbb{R}}(s+1)\Gamma_{\mathbb{R}}(s+k-1)\Gamma_{\mathbb{R}}(s+k) L(s, \mathrm{sym}^2 f),$$

and

$$\Lambda(s, \mathrm{sym}^2 \phi) = \Gamma_{\mathbb{R}}(s)\Gamma_{\mathbb{R}}(s+2it_\phi)\Gamma_{\mathbb{R}}(s-2it_\phi) L(s, \mathrm{sym}^2 \phi).$$

Recall that we wrote the p-th Hecke eigenvalue of f as $\alpha_f(p) + \beta_f(p)$ where $\alpha_f(p)\beta_f(p) = 1$ and $|\alpha_f(p)| = |\beta_f(p)| = 1$. Similarly write the p-th Hecke eigenvalue of ϕ as $\alpha_\phi(p) + \beta_\phi(p)$ where $\alpha_\phi(p)\beta_\phi(p) = 1$, but we do not know here the Ramanujan conjecture that these are both of size 1. The triple product L-function $L(s, f \times f \times \phi)$ is then defined by means of the Euler product of degree 8 (absolutely convergent in $\mathrm{Re}(s) > 1$)

$$\begin{aligned} \prod_p \quad & \left(1 - \frac{\alpha_f(p)^2 \alpha_\phi(p)}{p^s}\right)^{-1} \left(1 - \frac{\alpha_\phi(p)}{p^s}\right)^{-2} \left(1 - \frac{\beta_f(p)^2 \alpha_\phi(p)}{p^s}\right)^{-1} \\ \times \quad & \left(1 - \frac{\alpha_f(p)^2 \beta_\phi(p)}{p^s}\right)^{-1} \left(1 - \frac{\beta_\phi(p)}{p^s}\right)^{-2} \left(1 - \frac{\beta_f(p)^2 \beta_\phi(p)}{p^s}\right)^{-1}. \end{aligned}$$

This L-function is not primitive and factors as $L(s, \phi) L(s, \mathrm{sym}^2 f \times \phi)$. The archimedean factor $L_\infty(s, f \times f \times \phi)$ is defined as the product of eight Γ-factors

$$\prod_{\pm} \Gamma_{\mathbb{R}}(s+k-1 \pm it_\phi)\Gamma_{\mathbb{R}}(s+k \pm it_\phi)\Gamma_{\mathbb{R}}(s \pm it_\phi)\Gamma_{\mathbb{R}}(s+1 \pm it_\phi).$$

The work of Garrett [6] shows that the *completed L-function* $L(s, f \times f \times \phi)L_\infty(s, f \times f \times \phi)$ is an entire function in $\mathbb{C}$, and its value at s equals its value at $1-s$.

Using Stirling's formula we deduce that

$$\left| \int_{\Gamma\backslash\mathbb{H}} y^k |f(z)|^2 \phi(z) \frac{dx\ dy}{y^2} \right|^2 \ll_\phi \frac{L(\frac{1}{2}, f \times f \times \phi)}{kL(1, \mathrm{sym}^2 f)^2}.$$

Now the analytic conductor of $L(\frac{1}{2}, f \times f \times \phi)$ is about k^4, and again we see that the convexity bound ($\ll_\phi k$) is insufficient to show that the triple product above tends to zero, but any subconvexity bound would suffice. In particular GRH again gives that these triple products tend to zero as $k \to \infty$, and the Rudnick-Sarnak conjecture is thus implied by GRH. In this case also we have a weak subconvexity bound

$$L\left(\tfrac{1}{2}, f \times f \times \phi\right) \ll_{\phi,\epsilon} \frac{k}{(\log k)^{1-\epsilon}} \tag{3}$$

and so if $L(1, \mathrm{sym}^2 f) \geq (\log k)^{-\frac{1}{2}+\delta}$ for some $\delta > 0$ then we would be done. Such a bound holds in all but a very small number of exceptional cases, but establishing such a lower bound in all cases seems extremely difficult: even for the zeta-function we only know that $|\zeta(1+it)| \gg (\log|t|)^{-\frac{2}{3}-\epsilon}$, and the methods of Vinogradov that achieve this are unavailable for general L-functions.

4. Inner Products with Poincare Series and the Shifted Convolution Problem

Now we turn to the approach to the Rudnick-Sarnak conjecture via incomplete Eisenstein and Poincaré series. First let us consider the inner product of $y^k|f(z)|^2$ with the Poincaré series $P_m(z \mid \psi)$ where $m \neq 0$ is fixed. This inner product can be evaluated by the Rankin-Selberg unfolding method, and this was carried out by Luo and Sarnak [33]. We have (recall $X = SL_2(\mathbb{Z})\backslash\mathbb{H}$)

$$\begin{aligned} \int_X y^k |f(z)|^2 P_m(z \mid \psi) \frac{dx\ dy}{y^2} &= \int_X y^k |f(z)|^2 \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} e(m\gamma z)\psi(\mathrm{Im}(\gamma z)) \frac{dx\ dy}{y^2} \\ &= \int_0^1 \int_0^\infty y^k |f(z)|^2 \psi(y) e(mz) \frac{dxdy}{y^2} \end{aligned}$$

and by Parseval this equals

$$C^2 \sum_{r=1}^\infty \lambda_f(r)\lambda_f(r+m)(r(m+r))^{\frac{k-1}{2}} \int_0^\infty y^{k-1}\psi(y) e^{-4\pi(r+m)y} \frac{dy}{y},$$

where we set the Hecke eigenvalues at negative integers to be zero.

Now it is easy to analyze the integral over y above. The term $y^{k-1}e^{-4\pi(r+m)y}$ attains its maximum for $y = (k-1)/(4\pi(r+m))$, and is sharply peaked at that maximum. Note also that $\int_0^\infty y^{k-1}e^{-4\pi(r+m)y}\frac{dy}{y} = (4\pi(r+m))^{-(k-1)}\Gamma(k-1)$. From these remarks, and using from §3 the formula for $|C|^2$, we obtain that the inner product with $P_m(z \mid \psi)$ is

$$\sim \frac{2\pi^2}{(k-1)L(1,\mathrm{sym}^2 f)} \sum_{r\geq 1} \left(\frac{r}{r+m}\right)^{\frac{k-1}{2}} \lambda_f(r)\lambda_f(r+m)\psi\left(\frac{k-1}{4\pi(r+m)}\right).$$

Since ψ is a fixed smooth function compactly supported in $(0,\infty)$ we may think of the above sum as essentially being

$$\frac{1}{kL(1,\mathrm{sym}^2 f)} \sum_{r\asymp k} \lambda_f(r)\lambda_f(r+m),$$

where r runs over a range of values of size k. Finding cancellation in such sums is known as the *shifted convolution problem*. If $m \neq 0$ then we expect that the terms $\lambda_f(r)$ and $\lambda_f(r+m)$ behave independently and cancel out on average. If that were so, then we would reach the desired conclusion that the triple product with Poincare series tends to zero. For fixed m and k, and as $x \to \infty$ it is known that there is cancellation in $\sum_{r\leq x} \lambda_f(r)\lambda_f(r+m)$, however in our case we are interested in the delicate range where x is of size k, and such cancellation remains unknown. Holowinsky's ingenious idea is to forego cancellation in shifted convolution sums, and instead just bound $\sum_{r\asymp k} |\lambda_f(r)\lambda_f(r+m)|$. The insight is that the Hecke eigenvalues tend to be small in size, and we will explain this in more detail in §6.

One can also carry out the above argument for $m = 0$ when we have the incomplete Eisenstein series $E(z \mid \psi)$. The only difference is that here we have a main term to deal with. Here we want to show that

$$\int_{\Gamma\backslash\mathbb{H}} y^k |f(z)|^2 E(z \mid \psi)\frac{dx\ dy}{y^2} \to \frac{3}{\pi}\int_{\Gamma\backslash\mathbb{H}} E(z \mid \psi)\frac{dx\ dy}{y^2} = \frac{3}{\pi}\int_0^\infty \psi(y)\frac{dy}{y^2},$$

where the equality above follows by unfolding. Arguing as above we find that the LHS above is

$$\sim \frac{2\pi^2}{(k-1)L(1,\mathrm{sym}^2 f)} \sum_{r=1}^{\infty} |\lambda_f(r)|^2 \psi\left(\frac{k-1}{4\pi r}\right),$$

and so the problem here is to show that

$$\frac{2\pi^2}{(k-1)L(1,\mathrm{sym}^2 f)} \sum_{r=1}^{\infty} |\lambda_f(r)|^2 \psi\left(\frac{k-1}{4\pi r}\right) \sim \frac{3}{\pi}\int_0^\infty \psi(y)\frac{dy}{y^2}. \tag{4}$$

In this context we recall that by Rankin-Selberg theory we have

$$\sum_{n\leq x} |\lambda_f(n)|^2 \sim L(1,\mathrm{sym}^2 f)x,$$

for $x \geq k^{1+\epsilon}$. This makes our asymptotic above plausible, but just out of reach. A subconvexity bound for the symmetric square L-function would give our desired asymptotic, but as noted earlier this remains unknown.

Here Holowinsky introduces an important refinement of evaluating the above inner product. The idea is to use the Siegel domain $\{0 \leq x \leq 1, y > 1/Y\}$ for some parameter Y. This Siegel domain contains essentially $3Y/\pi$ copies of the fundamental domain for $SL_2(\mathbb{Z})\backslash\mathbb{H}$, and further it is relatively easy to compute inner products on this Siegel domain. In this manner we can reduce the problem of establishing (4) to proving asymptotics for $\sum_{n\leq x} |\lambda_f(n)|^2$ for x of size kY. In this argument Y will be chosen to be a power of $(\log k)$, and this small extra flexibility allows the use of weak subconvexity to resolve this problem.

5. Mean Values of Multiplicative Functions and Weak Subconvexity

We saw in §3 how the Rudnick-Sarnak conjecture is related to obtaining subconvex bounds for values of certain L-functions on the critical line. We now give some of the ideas behind the weak subconvexity result described there. We shall confine ourselves to the case of the triple product L-function $L(s, f \times f \times \phi)$, and refer the reader to [50] for the general result.

The class of L-functions covered by weak subconvexity satisfy the standard properties of having a Dirichlet series, an Euler product, and a functional equation. In addition to these, we require an assumption on the size of the Dirichlet series coefficients of our L-function, which we call a weak Ramanujan condition. From §3 recall that the triple product L-function has an Euler product and functional equation. To explain the weak Ramanujan condition in this context, write

$$-\frac{L'}{L}(s, f \times f \times \phi) = \sum_{n=1}^{\infty} \frac{\lambda(n)\Lambda(n)}{n^s},$$

where $\Lambda(n)$ is the von Mangoldt function, and if $n = p^k$ then $\lambda(n) = (\alpha_f(p)^k + \beta_f(p)^k))^2(\alpha_\phi(p)^k + \beta_\phi(p)^k)$ in the notation of §3. The Ramanujan conjecture for Maass forms gives that $|\lambda(p^k)| \leq 8$ for all primes p, but this is not known and we only have $|\lambda(p^k)| \leq 4|\alpha_\phi(p)^k + \beta_\phi(p)^k|$, using Deligne's theorem for the holomorphic form f. The Rankin-Selberg theory for ϕ now tells us that there is a constant A_ϕ such that for all $x \geq 1$ we have

$$\sum_{x<n\leq ex} \frac{|\lambda(n)|^2}{n}\Lambda(n) \leq A_\phi^2. \tag{5}$$

It is this average form of the Ramanujan conjecture that we call a weak Ramanujan condition.

Write $L(s, f \times f \times \phi) = \sum_{n=1}^{\infty} a(n)n^{-s}$. A straightforward argument using the convexity bound shows that

$$\sum_{n \le x} a(n) \ll \frac{x}{\log x}, \tag{6}$$

provided $x \ge k^2(\log k)^B$ for some positive constant B; recall that the analytic conductor of the triple product L-function was about k^4, and more generally one would have such cancellation for x a little larger than the square-root of the analytic conductor. Our main idea is to show that similar cancellation holds even when $x = k^2(\log k)^{-B}$ for any constant B: For any $\epsilon > 0$, any positive constant B, and all $x \ge k^2(\log k)^{-B}$ we have

$$\sum_{n \le x} a(n) \ll \frac{x}{(\log x)^{1-\epsilon}}. \tag{7}$$

The implied constant above may depend on ϕ, B and ϵ.

Once (7) is established, (3) will follow from a standard partial summation argument using an approximate functional equation for $L(\frac{1}{2}, f \times f \times \phi)$. In (7) and (3), by keeping track of the various parameters involved, it would be possible to quantify ϵ. However, the limit of our method would be to obtain a bound $k^{\frac{1}{2}}/\log k$ in (3), and $x/\log x$ in (7).

Why does the extrapolation (7) hold? At the heart of its proof is the fact that mean values of multiplicative functions vary slowly. Knowing (6) in the range $x \ge k^2(\log k)^B$, this fact will enable us to extrapolate (6) to the range $x \ge k^2(\log k)^{-B}$.

The possibility of obtaining such extrapolations was first considered by Hildebrand [19], [20]. If g is a multiplicative function, we shall denote by $S(x) = S(x; g)$ the partial sum $\sum_{n \le x} g(n)$. Hildebrand [20] showed that if $-1 \le g(n) \le 1$ is a real valued multiplicative function then for $1 \le w \le \sqrt{x}$

$$\frac{1}{x} \sum_{n \le x} g(n) = \frac{w}{x} \sum_{n \le x/w} g(n) + O\left(\left(\log \frac{\log x}{\log 2w} \right)^{-\frac{1}{2}} \right). \tag{8}$$

In other words, the mean value of g at x does not change very much from the mean-value at x/w. Hildebrand [19] used this idea to show that from knowing Burgess's character sum estimates for $x \ge q^{\frac{1}{4}+\epsilon}$ one may obtain some non-trivial cancellation even in the range $x \ge q^{\frac{1}{4}-\epsilon}$ (we assume for simplicity that q is cube-free).

Elliott [4] generalized Hildebrand's work to cover complex valued multiplicative functions with $|g(n)| \le 1$, and also strengthened the error term in (8). Notice that a direct extension of (8) for complex valued functions is false. Consider $g(n) = n^{i\tau}$ for some real number $\tau \ne 0$. Then $S(x; g) = x^{1+i\tau}/(1 + i\tau) + O(1)$, and $S(x/w; g) = (x/w)^{1+i\tau}/(1 + i\tau) + O(1)$. Therefore (8) is false, and instead

we have that $S(x)/x$ is close to $w^{i\tau}S(x/w)/(x/w)$. Building on the pioneering work of Halasz [13], [14] on mean-values of multiplicative functions, Elliott showed that for a multiplicative function g with $|g(n)| \le 1$, there exists a real number $\tau = \tau(x)$ with $|\tau| \le \log x$ such that for $1 \le w \le \sqrt{x}$

$$S(x) = w^{1+i\tau}S(x/w) + O\left(x\left(\frac{\log 2w}{\log x}\right)^{\frac{1}{19}}\right). \tag{9}$$

In [10], Granville and Soundararajan give variants and stronger versions of (9), with $\frac{1}{19}$ replaced by $1 - 2/\pi - \epsilon$.

In order to establish (7), we require similar results when the multiplicative function is no longer constrained to the unit disc. The situation here is considerably more complicated, and instead of showing that a suitable linear combination of $S(x)/x$ and $S(x/w)/(x/w)$ is small, we will need to consider linear combinations involving several terms $S(x/w^j)/(x/w^j)$ with $j = 0, \ldots, J$. In order to motivate our main result, it is helpful to consider two illustrative examples.

Example 1. Let k be a natural number, and take $g(n) = d_k(n)$, the k-th divisor function. Then, it is easy to show that $S(x) = xP_k(\log x) + O(x^{1-1/k+\epsilon})$ where P_k is a polynomial of degree $k-1$. If $k \ge 2$, it follows that $S(x)/x - S(x/w)/(x/w)$ is of size $(\log w)(\log x)^{k-2}$, which is not $o(1)$. However, if $1 \le w \le x^{1/2k}$, the linear combination

$$\begin{aligned}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\frac{S(x/w^j)}{x/w^j} &= \sum_{j=0}^{k}(-1)^j\binom{k}{j}P_k(\log x/w^j) + O(x^{-\frac{1}{2k}}) \\ &= O(x^{-\frac{1}{2k}})\end{aligned}$$

is very small.

Example 2. Let $\tau_1, \ldots, \tau_R$ be distinct real numbers, and let $k_1, \ldots, k_R$ be natural numbers. Let g be the multiplicative function defined by $G(s) = \sum_{n=1}^{\infty} f(n)n^{-s} = \prod_{j=1}^{R}\zeta(s - i\tau_j)^{k_j}$. Consider here the linear combination (for $1 \le w \le x^{1/(2(k_1+\ldots+k_R))}$)

$$\begin{aligned}\frac{1}{x}\sum_{j_1=0}^{k_1}\cdots\sum_{j_R=0}^{k_R} \quad & (-1)^{j_1+\ldots+j_R}\binom{k_1}{j_1}\cdots\binom{k_R}{j_R}w^{j_1(1+i\tau_1)+\ldots+j_R(1+i\tau_R)} \\ \times \quad & S\left(\frac{x}{w^{j_1+\ldots+j_R}}\right).\end{aligned}$$

By Perron's formula we may express this as, for $c > 1$,

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\prod_{j=1}^{R}\zeta(s - i\tau_j)^{k_j}(1 - w^{1+i\tau_j-s})^{k_j}x^{s-1}\frac{ds}{s}.$$

Notice that the poles of the zeta-functions at $1 + i\tau_j$ have been cancelled by the factors $(1 - w^{1+i\tau_j-s})^{k_j}$. Thus the integrand has a pole only at $s = 0$, and

a standard contour shift argument shows that this integral is $\ll x^{-\delta}$ for some $\delta > 0$.

Fortunately, it turns out that Example 2 captures the behavior of mean-values of the multiplicative functions of interest to us. In order to state our result, we require some notation. For a multiplicative function g we write $G(s) = \sum_{n=1}^{\infty} g(n)n^{-s}$ and we suppose that this series converges absolutely in $\mathrm{Re}(s) > 1$. Moreover we write

$$-\frac{G'}{G}(s) = \sum_{n=1}^{\infty} \frac{\lambda_g(n)\Lambda(n)}{n^s} = \sum_{n=1}^{\infty} \frac{\Lambda_g(n)}{n^s},$$

where $\lambda_g(n) = \Lambda_g(n) = 0$ unless n is the power of a prime p. In analogy with the weak Ramanujan hypothesis (5), we suppose that there exists a constant A such that for all $x \geq 1$ we have

$$\sum_{x<n\leq ex} \frac{|\lambda_f(n)|^2\Lambda(n)}{n} \leq A^2. \tag{10}$$

Let R be a natural number, and let $\tau_1, \ldots, \tau_R$ denote R real numbers. Let $\underline{\ell} = (\ell_1, \ldots, \ell_R)$ and $\underline{j} = (j_1, \ldots, j_R)$ denote vectors of non-negative integers, with the notation $\underline{j} \leq \underline{\ell}$ indicating that $0 \leq j_1 \leq \ell_1, \ldots, 0 \leq j_R \leq \ell_R$. Define

$$\binom{\underline{\ell}}{\underline{j}} = \binom{\ell_1}{j_1} \cdots \binom{\ell_R}{j_R}.$$

Finally, we define a measure of the oscillation of the mean-values of g by setting

$$\begin{aligned} \mathcal{O}_{\underline{\ell}}(x,w) &= \mathcal{O}_{\underline{\ell}}(x,w;\tau_1,\ldots,\tau_R) \\ &= \sum_{\underline{j}\leq\underline{\ell}} (-1)^{j_1+\ldots+j_R} \binom{\underline{\ell}}{\underline{j}} w^{j_1(1+i\tau_1)+\ldots+j_R(1+i\tau_R)} S\left(\frac{x}{w^{j_1+\ldots+j_R}}\right). \end{aligned}$$

With the above notations, the estimate (7) follows from the following result: Let $X \geq 10$ and $1 \geq \epsilon > 0$ be given. Let $R = [10A^2/\epsilon^2]+1$ and put $L = [10AR]$, and $\underline{L} = (L, \ldots, L)$. Let w be such that $0 \leq \log w \leq (\log X)^{\frac{1}{3R}}$. There exist real numbers $\tau_1, \ldots, \tau_R$ with $|\tau_j| \leq \exp((\log\log X)^2)$ such that for all $2 \leq x \leq X$ we have

$$|\mathcal{O}_{\underline{L}}(x,w;\tau_1,\ldots,\tau_R)| \ll \frac{x}{\log x}(\log X)^{\epsilon}. \tag{11}$$

The implied constant above depends on A and ϵ. In other words, as in Example 2, we can find a linear combination of mean values of g that is guaranteed to be small.

Before expanding on the result (11), we indicate how (7) follows from it. Let $x_0 = k^2(\log k)^B$ be such that the convexity bound gives cancellation in $\sum_{n\leq x} a(n)$ for $x \geq x_0$ as mentioned in(6). Let $x_0 \geq x \geq k^2/(\log k)^B$, and take $w = x_0/x$ and $X = xw^{LR}$. Applying (11) to the multiplicative function $a(n)$

(note that (5) gives the assumption (10)) we find that for an appropriate choice of $\tau_1, \ldots, \tau_R$ that

$$|\mathcal{O}_{\underline{L}}(X,w)| \ll \frac{X}{(\log X)^{1-\epsilon}}. \tag{12}$$

But, by definition, the LHS above is

$$w^{LR}\left|\sum_{n\le X/w^{LR}} a(n)\right| + O\left(\sum_{j=0}^{LR-1} w^j \left|\sum_{n\le X/w^j} a(n)\right|\right). \tag{13}$$

Now $X/w^{LR} = x$, and for $0 \le j \le LR-1$ we have $X/w^j \ge xw = x_0$ so that the bound of (6) applies. Therefore (13) equals

$$w^{LR}\left|\sum_{n\le x} a(n)\right| + O\left(\frac{X}{\log X}\right),$$

From (12) we conclude that

$$\left|\sum_{n\le x} a(n)\right| \ll w^{-LR}\frac{X}{(\log X)^{1-\epsilon}} \ll \frac{x}{(\log x)^{1-\epsilon}},$$

which proves (7).

For a general multiplicative function, we cannot hope for any better bound for the oscillation than $x/\log x$. To see this, suppose $w \ge 2$, and consider the multiplicative function g with $g(n) = 0$ for $n \le x/2$ and $g(p) = 1$ for primes $x/2 < p \le x$. Then $S(x) \gg x/\log x$ whereas $S(x/w^j) = 1$ for all $j \ge 1$, and therefore for any choice of the numbers $\tau_1, \ldots, \tau_R$ we would have $\mathcal{O}_{\underline{L}}(x,w) \gg x/\log x$.

Our proof of (11) builds both on the techniques of Halasz (as developed in [4] and [10]), and also the idea of *pretentious* multiplicative functions developed by Granville and Soundararajan (see [11] and [12]). We describe just a couple of the main ideas used: how the numbers in τ_j in (11) are defined, and more generally what is special about mean-values of multiplicative functions?

We start by describing the numbers τ_j appearing in (11). As suggested by Example 2 these points correspond to large values of the generating function $G(1 + 1/\log X + it)$. A precise description is as follows. Write $T = \exp((\log\log X)^2)$, and define τ_1 to be that point t in the compact set $\mathcal{C}_1 = [-T,T]$ where the maximum of $|G(1 + 1/\log X + it)|$ is attained. Now remove the interval $(\tau_1 - (\log X)^{-\frac{1}{R}}, \tau_1 + (\log X)^{-\frac{1}{R}})$ from $\mathcal{C}_1 = [-T,T]$, and let $\mathcal{C}_2$ denote the remaining compact set. We define τ_2 to be that point t in $\mathcal{C}_2$ where the maximum of $|G(1 + 1/\log X + it)|$ is attained. Next remove the interval $(\tau_2 - (\log X)^{-\frac{1}{R}}, \tau_2 + (\log X)^{-\frac{1}{R}})$ from $\mathcal{C}_2$ leaving behind the compact set $\mathcal{C}_3$. Define τ_3 to be the point where the maximum of $|G(1+1/\log X+it)|$ for $t \in \mathcal{C}_3$

is attained. We proceed in this manner, defining the successive maxima $\tau_1, \ldots, \tau_R$, and the nested compact sets $\mathcal{C}_1 \supset \mathcal{C}_2 \supset \ldots \supset \mathcal{C}_R$. Notice that all the points $\tau_1, \ldots, \tau_R$ lie in $[-T, T]$, and moreover are well-spaced: $|\tau_j - \tau_k| \geq (\log X)^{-\frac{1}{R}}$ for $j \neq k$.

From the assumption (10) we see that $|G(1 + 1/\log X + it)| \ll (\log X)^A$. For $t \in [-T, T]$ a much better bound holds for $|G(1 + 1/\log X + it)|$ unless t happens to be near one of the points $\tau_1, \ldots, \tau_R$. Precisely, if $1 \leq j \leq R$ and t is a point in $\mathcal{C}_j$, then

$$|F(1 + 1/\log X + it)| \ll (\log X)^{A\sqrt{1/j+(j-1)/(jR)}}. \tag{14}$$

In particular if $t \in \mathcal{C}_R$ we have $|F(1+1/\log X+it)| \ll (\log X)^{\epsilon/2}$. The estimate (14) is inspired by the ideas in [11] and [12], and a proof may be found in [50].

Finally we indicate very briefly why we may expect mean values of multiplicative functions to behave nicely. For simplicity consider a completely multiplicative function g with $|g(n)| \leq 1$. We start with

$$\log x \sum_{n \leq x} g(n) = \sum_{n \leq x} g(n) \log n + O\left(\sum_{n \leq x} \log x/n\right) = \sum_{n \leq x} g(n) \log n + O(x).$$

Writing $\log n = \sum_{d|n} \Lambda(d)$ we obtain that

$$\sum_{n \leq x} g(n) \log n = \sum_{d \leq x} g(d)\Lambda(d) \sum_{m \leq x/d} g(m),$$

so that we deduce

$$|S(x)| \log x \leq \sum_{d \leq x} \Lambda(d)|S(x/d)| + O(x),$$

and by a "partial summation" argument (this needs some elaboration and is not obvious) we find that this is

$$\ll x + \int_1^x |S(x/t)|dt = x \int_1^x |S(t)| \frac{dt}{t^2}.$$

We conclude that

$$|S(x)| \ll \frac{x}{\log x} + \frac{x}{\log x} \int_1^x |S(t)| \frac{dt}{t^2}. \tag{15}$$

The relation above is crucial, and it shows how the mean value of a multiplicative function is dominated by an average of such mean values. This forces a smoother structure of these mean values than one would have expected. Wirsing's pioneering result [52] (on mean-values of real valued multiplicative functions) and Halasz's work, [13] and [14], on complex valued multiplicative

functions both exploit this feature very nicely. See also the work of Granville and Soundararajan [9] on the "spectrum of multiplicative functions" where the analogy with integral equations is made precisely.

We have mentioned several times Halasz's theorem without stating it properly. We now describe the result in general terms. If g is real valued multiplicative function with $|g(n)| \leq 1$ then Wirsing, proving a conjecture of Erdos and Wintner, showed that $\lim_{x\to\infty} \frac{1}{x}\sum_{n\leq x} g(n)$ exists. Moreover the limit is non-zero if and only if $\sum_{p\leq x}(1-g(p))/p$ converges; that is g looks like the function that is 1 always. To see that this result is non-trivial, just consider $g(n) = \mu(n)$. Halasz generalized Wirsing's result to complex valued multiplicative functions with $|g(n)| \leq 1$. If we consider the example $g(n) = n^{i\alpha}$ where $\sum_{n\leq x} g(n) \sim \frac{x^{1+i\alpha}}{1+i\alpha}$ we see that the limiting mean-value need no longer exist. Halasz realized that this example is the only obstruction, and the limiting mean-value tends to zero (and he quantified this nicely) unless it happens that $\sum_p (1-\mathrm{Re}\ g(p)p^{-i\alpha})/p$ converges for some α; that is, g is pretending to be the function $n^{i\alpha}$. When g is no longer restricted to the unit circle, matters are more complicated. But, extending Halasz's insight we may look for functions of the form $n^{-i\alpha_j}$ which g correlates with (or pretends to be). This is the motivation for the successive maxima that we identified earlier, and the oscillation result shows that we can handle the effect of those bounded number of functions that g can pretend to be.

6. Sieve Methods and Holowinsky's Work

Here we describe Holowinsky's approach to bounding the shifted convolution sums that arose in §4. We only deal with the inner products with Poincare series $P_m(z \mid \psi)$ with $m \neq 0$; as discussed in §4, the case $m = 0$ requires more care. We begin by explaining why we might expect the size of Hecke eigenvalues to be small on average; such a result goes back to work of Elliott, Moreno and Shahidi [5] in the context of Ramanujan's τ-function.

Consider, for simplicity, a completely multiplicative function g which is non-negative, and bounded by 1. Then

$$\log x \sum_{n\leq x} g(n) = \sum_{n\leq x} g(n)\log n + O\left(\sum_{n\leq x}\log(x/n)\right) = \sum_{n\leq x} g(n)\log n + O(x).$$

Writing $\log n = \sum_{d|n}\Lambda(d)$ we obtain that

$$\log x \sum_{n\leq x} g(n) = \sum_{m\leq x} g(m) \sum_{d\leq x/m} g(d)\Lambda(d) + O(x) \leq (1+o(1))x \sum_{m\leq x}\frac{g(m)}{m} + O(x),$$

using the prime number theorem. We conclude that

$$\sum_{n\le x} g(m) \ll \frac{x}{\log x}\sum_{n\le x}\frac{g(n)}{n} \ll \frac{x}{\log x}\exp\left(\sum_{p\le x}\frac{g(p)}{p}\right). \tag{16}$$

In fact the estimate above can be established for a larger class of non-negative multiplicative functions whose values on the primes are bounded on average and satisfying some mild conditions on the values at prime powers. Such results were first explored by Hall [16], and see also Halberstam and Richert [15]

The estimate (16) while simple, is nevertheless extremely useful. Take $g(n) = \tau(n)n^{-11/2}$ where τ denotes Ramanujan's function $\sum_{n=1}^{\infty}\tau(n)q^n = q\prod_{n=1}^{\infty}(1-q^n)^{24}$. By Deligne's theorem $|g(n)| \le d(n)$, and the estimate (16) applies to the non-negative multiplicative function $|g(n)|$. We obtain that

$$\sum_{n\le x}|g(n)| \ll \frac{x}{\log x}\sum_{n\le x}\frac{|g(n)|}{n} \ll \frac{x}{\log x}\exp\left(\sum_{p\le x}\frac{|\tau(p)p^{-11/2}|}{p}\right).$$

By Rankin-Selberg theory we know that

$$\sum_{p\le x}\frac{g(p)^2}{p} \sim \log\log x.$$

Using Rankin-Selberg for the $GL(3)$ automorphic form associated to $g(p^2) = g(p)^2 - 1$ (see [7]) we obtain that

$$\sum_{p\le x}\frac{(g(p)^2-1)^2}{p} \sim \log\log x.$$

But $(g(p)^2-1)^2 \le 9(|g(p)|-1)^2$ so that

$$\sum_{p\le x}\frac{(|g(p)|-1)^2}{p} \ge \frac{1}{9}\log\log x + O(1),$$

and we deduce that

$$\sum_{p\le x}\frac{|g(p)|}{p} \le \frac{17}{18}\log\log x + O(1).$$

Consequently

$$\sum_{n\le x}|\tau(n)n^{-\frac{11}{2}}| \ll x(\log x)^{-\frac{1}{18}},$$

which shows that on average the values of $|f(n)|$ are somewhat small.

Here is an explanation of why we might expect $|f(n)|$ to be small. By Rankin-Selberg we know that $\sum_{n\le x} g(n)^2 \sim cx$, for a positive constant c

(which is related to the symmetric square L-function of Δ evaluated at 1). So by Cauchy-Schwarz we know that $\sum_{n\le x}|g(n)|\ll x$. For this estimate to be tight, one would need that the $g(n)$ should be of constant size, and since g is multiplicative this means that most $|g(p)|$ should be close to 1. However the distribution of $g(p)$ is governed by the Sato-Tate law (now known thanks to the work of Taylor and others), and so there is considerable fluctuation in the sizes of $|g(n)|$. The mean square is dominated by the large values of $|g(n)|$, and so naturally we would expect the average of $|g(n)|$ to be small. Our argument above uses information about the first four symmetric powers of Δ, which was known for a while, whereas Sato-Tate amounts to using information about all symmetric powers.

In Holowinsky's work, roughly speaking we need an estimate for the shifted convolution sums $\sum_{n\le k}|\lambda_f(n)\lambda_f(n+m)|$ where $m\neq 0$. We want an analog of (16) for these shifted convolution sums. There is a lovely result of Mohan Nair [38] which establishes such an analog for general classes of multiplicative functions evaluated on polynomials. Nair's work extends work of Peter Shiu [48] who had considered such estimates for multiplicative functions in short intervals and arithmetic progressions.

We do not describe Nair's result in full generality, but restrict ourselves to the special case at hand. The basic point is that if m is a fixed non-zero integer then the multiplicative structure of the integers n and $n+m$ should have very little in common (e.g. if $m=1$ the two numbers are coprime), and hence the values $|\lambda_f(n)|$ and $|\lambda_f(n+m)|$ should behave independently of each other. In other words we may expect the average of $|\lambda_f(n)\lambda_f(n+m)|$ to behave like the product of the average of $|\lambda_f(n)|$ and the average of $|\lambda_f(n+m)|$; i.e. like the square of the average of $|\lambda_f(n)|$. Such an analog of (16) is guaranteed by Nair's theorem: we have for $m\neq 0$

$$\sum_{n\le k}|\lambda_f(n)\lambda_f(n+m)|\ll_m k\exp\left(\sum_{p\le k}\frac{2|\lambda_f(p)|-2}{p}\right). \tag{17}$$

Holowinsky [23] gives an independent proof of a slightly weaker result using a simple Selberg sieve argument.

Using the bound (17) in our work in §4 we find that

$$\left|\int_{\Gamma\backslash\mathbb{H}} y^k|f(z)|^2P_m(z\mid\psi)\frac{dx\,dy}{y^2}\right|\ll_{m,\psi}\frac{1}{L(1,\mathrm{sym}^2 f)}\exp\left(\sum_{p\le k}\frac{2|\lambda_f(p)|-2}{p}\right). \tag{18}$$

In the next section we show how this estimate complements the weak subconvexity bounds of §3, and together the two approaches give a proof of the Rudnick-Sarnak conjecture.

7. Proof of Mass Equidistribution

We begin by observing that

$$L(1, \mathrm{sym}^2 f) \gg \exp\left(\sum_{p \le k} \frac{\lambda_f(p^2)}{p}\right). \tag{19}$$

In fact $L(1, \mathrm{sym}^2 f)$ is of the same size as the RHS above, and the RHS is essentially the Euler product defining $L(s, \mathrm{sym}^2 f)$. This can be established generally for any L-function at the edge of the critical strip, provided there are no Siegel zeros. In the symmetric square case, we already noted that the work of Hoffstein-Lockhart and Goldfeld-Hoffstein-Lieman rules out the existence of Siegel zeros. A slightly weaker version of this bound is described in Lemma 2 of [24].

Using this bound, and noting that $\lambda_f(p^2) = \lambda_f(p)^2 - 1$, in (18) we deduce that

$$\left| \int_{\Gamma \backslash \mathbb{H}} y^k |f(z)|^2 P_m(z \mid \psi) \frac{dx\ dy}{y^2} \right| \ll_{m,\psi} \exp\left(-\sum_{p \le k} \frac{(|\lambda_f(p)| - 1)^2}{p} \right).$$

Thus Holowinsky's argument would give the decay of inner products with Poincare series unless it so happened that

$$\sum_{p \le k} \frac{(|\lambda_f(p)| - 1)^2}{p} \ll 1.$$

But if the above holds then

$$\sum_{p \le k} \frac{\lambda_f(p^2)}{p} = \sum_{p \le k} \frac{(\lambda_f(p) + 1)(\lambda_f(p) - 1)}{p} \ge -3 \sum_{p \le k} \frac{||\lambda_f(p)| - 1|}{p},$$

and using Cauchy-Schwarz we have

$$\sum_{p \le k} \frac{||\lambda_f(p)| - 1|}{p} \ll \sqrt{\log \log k}.$$

Inserting this in (19) we find that in the case when Holowinsky's argument fails we must have $L(1, \mathrm{sym}^2 f) \gg (\log k)^{-\epsilon}$. But recall from §3 that the argument via weak subconvexity succeeds if $L(1, \mathrm{sym}^2 f) \gg (\log k)^{-\frac{1}{2}+\delta}$. In other words, if Holowinsky's method fails then weak subconvexity succeeds! A variant of this argument is described in [24], together with the more delicate arguments needed for the incomplete Eisenstein series case that we have ignored here.

8. The Escape of Mass Argument

In this last section we give a description of the argument in [49] which eliminates the possibility of escape of mass for Hecke-Maass cusp forms, and thus completes Lindenstrauss's proof of QUE for $SL_2(\mathbb{Z})\backslash\mathbb{H}$.

As remarked in the introduction, Lindenstrauss has shown that any weak-$*$ limit of the micro-local lifts of Hecke-Maass forms is a constant c (in $[0,1]$) times the normalized volume measure on $Y = SL_2(\mathbb{R})\backslash SL_2(\mathbb{Z})$. Projecting these measures down to the modular surface X, we see that any weak-$*$ limit of the measures μ_ϕ associated to Hecke-Maass forms is of the shape $c\frac{3}{\pi}\frac{dx\ dy}{y^2}$. Our aim is to show that in fact $c = 1$, and there is no escape of mass. If on the contrary $c < 1$ for some weak-$*$ limit, then we have a sequence of Hecke-Maass forms ϕ_j with eigenvalues λ_j tending to infinity such that for any fixed $T \geq 1$ and as $j \to \infty$

$$\int_{\substack{z\in\mathcal{F}\\ y\leq T}} |\phi_j(z)|^2 \frac{dx\ dy}{y^2} = (c+o(1))\frac{3}{\pi}\int_{\substack{z\in\mathcal{F}\\ y\leq T}} \frac{dx\ dy}{y^2} = (c+o(1))\left(1-\frac{3}{\pi T}\right);$$

here $\mathcal{F} = \{z = x+iy : |z| \geq 1,\ -1/2 \leq x \leq 1/2, y > 0\}$ denotes the usual fundamental domain for $SL_2(\mathbb{Z})\backslash\mathbb{H}$. It follows that as $j \to \infty$

$$\int_{\substack{|x|\leq\frac{1}{2}\\ y\geq T}} |\phi_j(z)|^2 \frac{dx\ dy}{y^2} = 1 - c + \frac{3}{\pi T}c + o(1). \tag{20}$$

Now uniformly for any Hecke-Maass form of eigenvalue $\lambda = \frac{1}{4} + r^2$ (and normalized to have Petersson norm 1) we may show that

$$\int_{\substack{|x|\leq\frac{1}{2}\\ y\geq T}} |\phi(z)|^2 \frac{dx\ dy}{y^2} \ll \frac{\log(eT)}{\sqrt{T}}. \tag{21}$$

Clearly (21) contradicts (20) if $c < 1$ for suitably large T, and this establishes that $c = 1$.

Now let us explain why (21) holds. Letting $\lambda(n)$ denote the n-th Hecke eigenvalue of the form ϕ, we recall that ϕ has a Fourier expansion of the form

$$\phi(z) = C\sqrt{y}\sum_{n=1}^{\infty}\lambda(n)K_{ir}(2\pi ny)\cos(2\pi nx),$$

or

$$\phi(z) = C\sqrt{y}\sum_{n=1}^{\infty}\lambda(n)K_{ir}(2\pi ny)\sin(2\pi nx),$$

where C is a constant (normalizing the L^2 norm), K denotes the usual K-Bessel function, and we have cos or sin depending on whether the form is even or odd.

Using Parseval we find that

$$\int_{\substack{|x|\leq\frac{1}{2}\\ y\geq T}} |\phi(x+iy)|^2 \frac{dx\ dy}{y^2} = \frac{C^2}{2}\int_T^\infty \sum_{n=1}^{\infty} |\lambda(n)|^2 |K_{ir}(2\pi ny)|^2 \frac{dy}{y}.$$

By a change of variables we may write this as

$$\frac{C^2}{2}\sum_{n=1}^{\infty}|\lambda(n)|^2\int_{nT}^{\infty}|K_{ir}(2\pi t)|^2\frac{dt}{t}=\frac{C^2}{2}\int_1^{\infty}|K_{ir}(2\pi t)|^2\sum_{n\le t/T}|\lambda(n)|^2\frac{dt}{t}.$$

Now for $t\ge 1$ if we know that

$$\sum_{n\le t/T}|\lambda(n)|^2\le 10^8\frac{\log eT}{\sqrt{T}}\sum_{n\le t}|\lambda(n)|^2, \tag{22}$$

then the above is

$$\begin{aligned}\ll\quad & \frac{\log eT}{\sqrt{T}}\frac{C^2}{2}\int_1^{\infty}|K_{ir}(2\pi t)|^2\sum_{n\le t}|\lambda(n)|^2\frac{dt}{t}\\ =\quad & \frac{\log eT}{\sqrt{T}}\int_{\substack{|x|\le\frac12\\ y\ge 1}}|\phi(x+iy)|^2\frac{dx\ dy}{y^2}\ll\frac{\log eT}{\sqrt{T}},\end{aligned}$$

since the region $|x|\le\frac12$, $y\ge 1$ is contained inside a fundamental domain for $SL_2(\mathbb{Z})\backslash\mathbb{H}$. This would prove (21).

Lastly it remains to justify (22). In fact this statement is a general fact about a large class of multiplicative functions that we will call *Hecke-multiplicative*. We say that a function f is Hecke-multiplicative if it satisfies the Hecke relation

$$f(m)f(n)=\sum_{d|(m,n)}f(mn/d^2),$$

and $f(1)=1$. If f is Hecke-multiplicative, then for all $1\le y\le x$ we have

$$\sum_{n\le x/y}|f(n)|^2\le 10^8\left(\frac{1+\log y}{\sqrt{y}}\right)\sum_{n\le x}|f(n)|^2. \tag{23}$$

Clearly this statement proves (22).

We won't go into the proof of (23), but just mention that it is based on elementary analytic and combinatorial arguments. It is noteworthy that (23) makes no assumptions on the size of the function f. Hecke-multiplicative functions satisfy $f(p^2)=f(p)^2-1$, so that at least one of $|f(p)|$ or $|f(p^2)|$ must be bounded away from zero; this observation plays a crucial role in our proof. We also remark that apart from the $\log y$ factor, (23) is best possible: Consider the Hecke-multiplicative function f defined by $f(p)=0$ for all primes p. The Hecke relation then mandates that $f(p^{2k+1})=0$ and $f(p^{2k})=(-1)^k$. Therefore, in this example, $\sum_{n\le x}|f(n)|^2=\sqrt{x}+O(1)$ and $\sum_{n\le x/y}|f(n)|^2=\sqrt{x/y}+O(1)$.

References

[1] N. Anantharaman *Entropy and localization of eigenfunctions* Ann. of Math. **168** (2008) 435–475.

[2] Y. Colin de Verdiere *Ergodicite et fonctions propres du laplacien* Comm. Math. Phys. **102** (1985), 497–502.

[3] W. Duke and R. Schulze-Pillot *Representation of integers by positive ternary quadratic forms and the equidistribution of lattice points on ellipsoids* Invent. Math. **99** (1990) 49–57.

[4] P.D.T.A. Elliott, *Extrapolating the mean-values of multiplicative functions*, Indag. Math., **51** (1989), 409–420.

[5] P.D.T.A. Elliott, C. Moreno, and F. Shahidi, *On the absolute value of Ramanujan's τ-function*, Math. Ann. **266** (1984) 507–511.

[6] P. Garrett, *Decomposition of Eisenstein series: Rankin triple products*, Ann. of Math. **125**, (1987) 209–235.

[7] S. Gelbart and H. Jacquet *A relation between automorphic representations of* GL(2) *and* GL(3) Ann. Sci. Ecole Norm. Sup. **11** (1978), 471–542.

[8] D. Goldfeld, J. Hoffstein and D. Lieman *Appendix to the paper by Hoffstein and Lockhart* Ann. of Math. **140** (1994) 161–181.

[9] A. Granville and K. Soundararajan *The spectrum of multiplicative functions* Ann. of Math. **153**, (2001), 407–470.

[10] A. Granville and K. Soundararajan *Decay of mean-values of multiplicative functions*, Can. J. Math. **55** (2003) 1191-1230.

[11] A. Granville and K. Soundararajan *Pretentious multiplicative functions and an inequality for the zeta-function*, CRM Proceedings and Lecture Notes, **46** (2008) 191–197.

[12] A. Granville and K. Soundararajan *Large character sums: Pretentious characters and the Polya-Vinogradov theorem* J. Amer. Math. Soc. **20** (2007), 357–384.

[13] G. Halasz, *On the distribution of additive and mean-values of multiplicative functions* Studia Sci. Math. Hungar. **6** (1971) 211–233.

[14] G. Halasz *On the distribution of additive arithmetic functions* Acta Arith. **27** (1975) 143–152.

[15] H. Halberstam and H. Richert *On a result of R. R. Hall* J. Number Theory **11** (1979), 76–89.

[16] R. R. Hall *Halving an estimate obtained from Selberg's upper bound method* Acta Arith. **25** (1973/74) 347–351.

[17] A. Hassell *Ergodic billiards that are not quantum unique ergodic* Ann. of Math. **171** (2010) 605–618.

[18] D. R. Heath-Brown *Convexity bounds for L-functions* Acta Arith. **136** (2009) 391–395.

[19] A. J. Hildebrand *A note on Burgess' character sum estimate* C. R. Math. Rep. Acad. Sci. Canada **8** (1986) 35–37.

[20] A. J. Hildebrand *On Wirsing's mean value theorem for multiplicative functions* Bull. London Math. Soc. **18** (1986) 147–152.

[21] J. Hoffstein and P. Lockhart *Coefficients of Maass forms and the Siegel zero* Ann. of Math. **140** (1994) 161–181.

[22] J. Hoffstein and D. Ramakrishnan *Siegel zeros and cusp forms* IMRN (1995) 279–308.

[23] R. Holowinsky *Sieving for mass equidistribution* Ann of Math. to appear, preprint, available as `arxiv.org:math/0809.1640`.

[24] R. Holowinsky and K. Soundararajan *Mass equidistribution of Hecke eigenforms* Ann. of Math., to appear, preprint, available as `arxiv.org:math/0809.1636`.

[25] H. Iwaniec *Spectral methods of automorphic forms*, AMS Grad. Studies in Math. **53** (2002).

[26] H. Iwaniec *Topics in classical automorphic forms*. Grad. Studies in Math. AMS **17**.

[27] H. Iwaniec and E. Kowalski *Analytic number theory* AMS Coll. Publ. **53** (2004).

[28] H. Iwaniec and P. Sarnak *Perspectives on the analytic theory of L-functions*, Geom. Funct. Analysis Special Volume (2000) 705–741.

[29] D. Jakobson *Quantum unique ergodicity for Eisenstein series on* $PSL_2(\mathbb{Z})\backslash PSL_2(\mathbb{R})$ Ann. Inst. Fourier **44** (1994) 1477–1504.

[30] X. Li *Bounds for* $GL(3) \times GL(2)$ *L-functions and* $GL(3)$ *L-functions*, Ann. of Math., to appear.

[31] E. Lindenstrauss *Invariant measures and arithmetic quantum unique ergodicity*, Ann. of Math. **163** (2006) 165–219.

[32] E. Lindenstrauss *Adelic dynamics and arithmetic quantum unique ergodicity* Curr. Developments in Math. (2004) 111-139.

[33] W. Luo and P. Sarnak *Mass equidistriution for Hecke eigenforms* Comm. Pure Appl. Math. **56** (2003) 874–891.

[34] W. Luo and P. Sarnak *Quantum ergodicity for* $SL_2(\mathbb{Z})/\mathbb{H}^2$ Inst. Haute Etudes Sci. Publ. Math. **81** (1995) 207–237.

[35] J. Marklof *Arithmetic quantum chaos* Encyclopedia of Math. Phys. (Eds: J.-P. Francoise, G.L. Naber and Tsou S.T.) Elsevier (2006) **1** 212–220.

[36] P. Michel *Analytic number theory and families of automorphic L-functions* Automorphic forms and applications 181–295. IAS/Park City Math. Ser. 12, Amer. Math. Soc., Providence, RI (2007)

[37] P. Michel and A. Venkatesh *The subconvexity problem for* $GL(2)$ preprint, available on arxiv (2009).

[38] M. Nair *Multiplicative functions of polynomial values in short intervals* Acta Arith. **62** (1992) 257–269.

[39] S. Nonnenmacher and A. Voros *Chaotic eigenfunctions in phase space* J. Statist. Phys. **92** (1998) 431–518.

[40] Z. Rudnick *On the asymptotic distribution of zeros of modular forms* Int. Math. Res. Not. (2005) 2059–2074.

[41] Z. Rudnick and P. Sarnak *The behaviour of eigenstates of arithmetic hyperbolic manifolds* Comm. Math. Phys. **161** (1994) 195–213.

[42] P. Sarnak *Estimates for Rankin-Selberg L-functions and Quantum Unique Ergodicity* J. Funct. Anal. **184** (2001) 419–453.

[43] P. Sarnak *Recent progress on QUE* preprint available on his website.

[44] P. Sarnak *Arithmetic quantum chaos* Israel Math. Conf. Proc., Bar-Ilan Univ., Ramat Gan, **8** (1995) 183–236.

[45] B. Shiffman and S. Zelditch *Distribution of zeros of random and quantum chaotic sections of positive line bundles* Comm. Math. Phys. **200** (1999), 661–683.

[46] G. Shimura *On the holomorphy of certain Dirichlet series* Proc. London Math. Soc. **31** (1975) 79–98.

[47] A. Shnirelman *Ergodic properties of eigenfunctions* Uspehi Mat. Nauk. **29** (1974) 181-182.

[48] P. Shiu *A Brun-Titchmarsh theorem for multiplicative functions*, J. Reine Angew. Math. **313** (1980) 161–170.

[49] K. Soundararajan *Quantum unique ergodicity for* $SL_2(\mathbb{Z})\backslash\mathbb{H}$, Ann. of Math., to appear, available as `arxiv.org/abs/0901.4060`

[50] K. Soundararajan *Weak subconvexity for central values of L-functions* Ann. of Math., to appear, preprint available at `http://arxiv.org//abs/0809.1635`.

[51] T. Watson *Rankin triple products and quantum chaos* Ph. D. Thesis, Princeton University (eprint available at: `http://www.math.princeton.edu/~tcwatson/watson_thesis_final.pdf`) (2001).

[52] E. Wirsing *Das asymptotische Verhalten von Summen über multiplikative Funktionen* Acta. Math. Acad. Sci. Hungar. **18** (1967) 411–467.

[53] S. Zelditch *Uniform distribution of eigenfunctions on compact hyperbolic surfaces* Duke Math. J. **55** (1987) 919–941.

[54] S. Zelditch *Mean Lindelöf hypothesis and equidistribution of cusp forms* J. Funct. Anal. **97** (1991) 1–49.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Statistics of Number Fields and Function Fields

Akshay Venkatesh* and Jordan S. Ellenberg†

Abstract

We discuss some problems of arithmetic distribution, including conjectures of Cohen-Lenstra, Malle, and Bhargava; we explain how such conjectures can be heuristically understood for function fields over finite fields, and discuss a general approach to their proof in the function field context based on the topology of Hurwitz spaces. This approach also suggests that the Schur multiplier plays a role in such questions over number fields.

Mathematics Subject Classification (2010). 11R47.

1. Arithmetic Counting Problems

We begin with a concrete example, which has been well-understood for many years.

Let $\mathcal{S}_X$ denote the set of squarefree integers in $[0, X]$ that are congruent to 1 modulo 4; let $\mathcal{C}_X$ denote the set of isomorphism classes of cubic field extensions $K/\mathbf{Q}$ whose discriminant belongs to $\mathcal{S}_X$. Davenport and Heilbronn proved [6] that

$$\frac{|\mathcal{C}_X|}{|\mathcal{S}_X|} \longrightarrow \frac{1}{6}, \quad \text{as } X \to \infty. \tag{1}$$

Our goal is to understand *why limits like that of* (1) *should exist, why they should be rational numbers, and what the rational numbers represent.*

More precisely, we will study several variants on (1) – replacing cubic fields by extensions with prescribed Galois group, and "squarefree discriminant" by other forms of prescribed ramification. We make a heuristic argument as to what the corresponding limits should be when $\mathbf{Q}$ is replaced by the function field of a curve over a finite field, and lay out a program for a proof in certain cases. This program has been partially implemented by us in certain settings, leading to a weak form of the Cohen–Lenstra heuristics (see §4.2) over a rational function

*Akshay Venkatesh, Stanford University. E-mail: akshay@math.stanford.edu.
†Jordan Ellenberg, University of Wisconsin. E-mail: ellenber@math.wisc.edu.

field. In the number field case we have no new theorems; however, the study of the function field case suggests interesting refinements of known heuristics, related to the size of Schur multipliers.

Let us describe – briefly and approximately – how the $\frac{1}{6}$ makes an appearance over a function field. Let k be a finite field, and $\bar{k}$ an algebraic closure of k; we consider cubic extensions L of $k(t)$ with squarefree discriminant and totally split at ∞. By a *marking* of L we shall mean an ordering of the three places above ∞. "Marked" cubic extensions can be descended: they are identified with fixed points of a Frobenius acting on marked cubic extensions of $\bar{k}(t)$. Recall that *the average number of fixed points of a random permutation on a finite set is* 1; thus, if the Frobenius behaves like a random permutation, we expect there to be on average *one* marked cover per squarefree discriminant. Since there are six markings for each cubic field that is totally split at ∞, we recover $\frac{1}{6}$.

The rest of this paper will discuss methods for trying to make this heuristic into a proof, and how it suggests corrections to our view of number fields. In the function field case, results such as (1) are related to the group-theoretic structure of étale π_1; we may speculate that results such as (1) are reflections of some (as yet, not understood) group-theoretic features of the absolute Galois group of $\mathbf{Q}$.

1.1. Context. There has been a great deal of work on the topics discussed here. We note in particular that related topics have been discussed [1, 3, 7] in the last three ICMs. Indeed, [1] contains an overview of Bhargava's results for quartic and quintic fields, and [3] discusses both theoretical and numerical evidence for conjectures of the type described in the present paper.

Our point of view is influenced very much by the study of the function field case; in turn, our study of that case was influenced by both Cohen and Lenstra's work and the more recent paper [8] of Dunfield and Thurston on finite covers of hyperbolic 3-manifolds.

The present paper has three sections; although related, they are also to a large extent independent, and can be considered as "variations on the theme of (1)."

- §2 discusses the conjectures of Bhargava-Malle about distribution of number fields, generalizing (1). We also discuss the role that Schur multipliers may play in formulating sharp versions of such conjectures (§2.4). The reader may wish to first look at Section 3.4, which provides the geometric motivation guiding the computations in Sections 2.4 and 2.5.
- §3 discusses the function field setting and its connection with the geometry of Hurwitz spaces; in particular, how purely topological results on the stable homology of Hurwitz spaces would imply function field versions of Bhargava-Malle conjectures.
- §4 discusses the special case of the Cohen–Lenstra heuristics, and our proof (with Westerland) of a weak version in the function field setting.

This proof suggests more general connections between analytic number theory and stable topology.

1.2. Notation. By a G-extension algebra (resp. field) of a number field K, we shall mean a conjugacy class of homomorphisms (resp. surjective homomorphisms) from the Galois group $G_K := \mathrm{Gal}(\overline{K}/K)$ to G. In other words, G-extension fields are in correspondence with isomorphism classes of pairs $(L \supset K, G \xrightarrow{\sim} \mathrm{Gal}(L/K))$, where an isomorphism of pairs (L, f) and (L', f') is simply a K-isomorphism $\phi : L \to L'$ which commutes with the induced G-actions.

A pair (G, c) of a group G and a conjugacy class $c \subset G$ will be called *admissible* – for short, we say that c is an admissible conjugacy class – if

1. c is a rational conjugacy class, i.e., $g \in c \implies g^n \in c$ whenever n is prime to the order of g;
2. c generates the group G.

Given a tamely ramified G-extension and an admissible conjugacy class, we say that *all ramification is of type* c if the image of every inertia group is either trivial or a cyclic subgroup generated by some $g \in c$.

Acknowledgements. We thank Craig Westerland, our collaborator on the work described here, for many years of advice and ideas about the topological side of the subject. We have also greatly benefited from conversations with Manjul Bhargava, Nigel Boston, Ralph Cohen, Henri Cohen, David Roberts, and Melanie Wood.

2. Number Fields

In this section, we discuss Bhargava's heuristics for discriminants of S_n-extensions of $\mathbf{Q}$, and propose that for extensions with certain Galois groups G these heuristics should be modified by a term related to the Schur multiplier of G.

Before proceeding, however, we warn the reader of the alarming gap between theory and experiment. For example, the statement "there are $\frac{1}{6}$ totally real cubic fields per odd squarefree discriminant" is indeed only asymptotically valid; for instance, the smallest squarefree discriminants of real cubic fields are $229, 257$ and 321, and in fact (1) looks quite inaccurate for small X. However, there is convincing numerical evidence [22] that the ratio of (1) converges from below, with a secondary term decreasing proportionally to $D^{-1/6}$. This unpleasant situation – very slow numerical convergence to the expected limit – persists in all the examples we shall discuss in this paper, making it very difficult to test ideas except in somewhat indirect ways. For more discussion of this point, see §2.6.

2.1. Malle's conjecture. For definiteness, we work over the base field $\mathbf{Q}$ for the moment; the ideas generalize in a straightforward fashion, and we will anyway pass to the case of a general number field in §2.4.

Suppose G is provided with an embedding into S_n. In that case, there is a well-defined "discriminant" of any G-extension algebra or field L, since the map $G \hookrightarrow S_n$ associates to L an étale $\mathbf{Q}$-algebra of degree n which has a discriminant in the usual sense. (In what follows, then, the "discriminant" of an S_n extension field refers to the discriminant of the associated degree n field, and not to the discriminant of its Galois closure.)

In this case, Malle has conjectured [18, 19] that

$$\lim_{X\to\infty} \frac{\#\ G\text{-extension fields of discriminant less than } X}{X^{1/a}(\log X)^b} \tag{2}$$

exists and is nonzero, for certain integers a and b depending on G. For instance, $n-a$ is the maximal number of orbits of any nontrivial $g \in G \subset S_n$.[1] This statement has some consequences which are surprising at first glance: for instance, a positive fraction of quartic fields (ordered by discriminant) contain $\mathbf{Q}(\sqrt{-1})$.

Malle's conjecture has been proved by Davenport–Heilbronn in the case $G = S_3$ (prior to Malle's general formulation!) and by Bhargava in the case S_4, S_5, in each instance with a precise description of the limit in (2) as an Euler product.

2.2. The asymptotic constant. As originally stated, Malle's conjecture gives no information about the asymptotic constant, i.e. the limit of (2) as $X \to \infty$.

In fact, we do not regard the limiting value of (2) as the object of primary interest. This is because it conflates several independent issues; in particular, it mingles together fields with many different types of ramification, and it is also strongly influenced by the notion of "discriminant" (if we change the embedding $G \hookrightarrow S_n$, the limit will change).

Instead, we shall study the asymptotic constant only after "controlling" for these effects. For example: if we prescribe a set of primes and the ramification type at each prime, what is the expected number of global extensions realizing this "ramification data"? The word "expected" implies a suitable average; we usually mean to average over all sets S of primes with $\prod_{p\in S} p \le X$, and then let $X \to \infty$.

In the rest of this paper, we shall discuss the case where we *fix a conjugacy class* $c \subset G$ and study G-extensions where *all ramification is of type* c. (See §1.2 for the notation.) For instance, (1) corresponds to the case of $G = S_3$ and c the class of transpositions; in the next section, we shall discuss totally real

[1] The value of b in Malle's original conjecture is now known to be incorrect in some cases when the extension fields being counted can contain extra roots of unity: see [17],[25].

S_n-extensions of odd squarefree discriminant, which corresponds to the case where $G = S_n$ and c is the class of transpositions.

The ideas that we describe can be generalized to multiple ramification types, and one can eventually return to the setting of (2) by putting this information together.

2.3. The asymptotic constant: Bhargava's heuristic. On the basis of his results for $G = S_4$ and $G = S_5$, Bhargava has formulated a general and very beautiful conjecture [2, Conjecture 1.2] for the constant in the case $G = S_n$. We quote from his paper [2, page 10] and then explain by example:

> *The expected (weighted) number of global S_n-number fields of discriminant D is simply the product of the (weighted) number of local extensions of $\mathbf{Q}_v$ that are discriminant-compatible with D, where v ranges over all places of $\mathbf{Q}$ (finite and infinite).*

By a local extension of $\mathbf{Q}_v$, we mean simply a degree n étale algebra E over $\mathbf{Q}_v$; by discriminant–compatible, we mean (in the non-archimedean case) that the valuation of the discriminant of E coincides with the valuation of D and (in the archimedean case) that the signs match. Bhargava conjectures further that the expected number of S_n-extensions of discriminant D with a specified local behavior at v is obtained by the appropriate modification of the local factor at v in the above product.

Let us consider, for instance, what this means for *totally real fields* of odd squarefree discriminant D, i.e. totally real S_n-extensions all of whose ramification is of "transposition" type. To compute the expected number, one takes the product of local factors weight(v), where

$$\text{weight}(v) = \sum \frac{1}{|\text{Aut}(E/\mathbf{Q}_v)|},$$

the sum being taken over all degree n étale algebras $E/\mathbf{Q}_v$ that are:

- unramified, if v is a place not dividing D;
- have discriminant of valuation 1, if $v(D) = 1$;
- totally real, if v is infinite.

The weights in these cases are computed to be $1, 1$ and $\frac{1}{n!}$ respectively; so Bhargava's heuristic suggests that

> There are, on average, $\frac{1}{n!}$ totally real S_n-extensions of $\mathbf{Q}$ per odd squarefree discriminant,

where this is to be interpreted in a fashion analogous to (1) – in particular, we again restrict to discriminants that are congruent to 1 modulo 4, a necessary

condition by Stickelberger's theorem. This statement is compatible with the limit $\frac{1}{6}$ that appears in (1).

One of the remarkable features of Bhargava's heuristic, as well as of the known results in degree ≤ 5, is that there is no restriction to tamely ramified extensions. Our knowledge in more general situations is sadly limited, and we shall unfortunately have to restrict to tame ramification.

2.4. General groups; the role of the Schur multiplier. Let us now consider the case of a general group G. In this case, we shall propose that in many cases a version of Bhargava's heuristic applies: but that the heuristic as stated above often gives too few extensions, and must be modified by a term related to the size of the Schur multiplier of G. The reader will find motivation for this modification in Section 3.4.

Particularly in the number field case, what follows is speculative: We have, in the function field case, theoretical evidence for this modification, described in §3. But in the number field case we do not yet have serious numerical or theoretical evidence.

Let K, then, be a global field – either a number field, or a function field of a curve over a finite field; allowing this generality now allows ease of comparison later. To isolate as far as possible the particular phenomenon we wish to describe, we consider G-extensions L/K with the following properties:

1. G is center-free and has trivial abelianization;
2. $c \subset G$ is an admissible conjugacy class;
3. If K is a function field, we suppose that the characteristic of K does not divide $|G|$;
4. All ramification in L/K is tame of type c;
5. Fixing a set of places S_∞ of K containing all archimedean places, we consider only extensions L/K that are totally split at S_∞.

In what follows, we regard G, c, S_∞, K as fixed, subject to restrictions 1, 2, 3, and will count extensions L satisfying 4, 5.

The direct analogue of Bhargava's S_n heuristic would suggest that the average number of G-extensions L, for each set of ramified primes compatible with conditions 4 and 5, is $|G|^{-|S_\infty|}$. More precisely, let V be S_∞ together with all places whose residue characteristic divides the order of an element of c; if we denote by $\mathcal{S}_X$ the collection $\{S$ a subset of finite places of $K : S \cap V = \emptyset, \prod_{v \in S} q_v \leq X\}$, and by $\mathcal{F}_X$ the set of G-extensions L satisfying conditions 4,5 and which are ramified precisely at some $S \in \mathcal{S}_X$, then

$$\frac{|\mathcal{F}_X|}{|\mathcal{S}_X|} \longrightarrow |G|^{-|S_\infty|}, \text{ as } X \to \infty.$$

Based on our results in the function field case, we do not think this is right in general, and we speculate instead that

$$\frac{|\mathcal{F}_X|}{|\mathcal{S}_X|} \longrightarrow \frac{h(G,c,K)}{|G|^{|S_\infty|}}, \text{ as } X \to \infty, \tag{3}$$

for some rational number $h(G,c,K)$ related to the order of the Schur multiplier of G. We make precise predictions for the value of $h(G,c,K)$ in some special cases below.

Let $Q = Q_c \subset H_2(G,\mathbf{Z})$ be the subgroup generated by $\phi_*(H_2(\mathbf{Z}\times\mathbf{Z},\mathbf{Z}))$, as ϕ ranges over homomorphisms $\mathbf{Z}\times\mathbf{Z} \longrightarrow G$ taking $(0,1)$ to an element of c. We put

$$H_2(G,c;\mathbf{Z}) := H_2(G,\mathbf{Z})/Q_c. \tag{4}$$

The following interpretation in terms of covering groups will be useful: Fix a universal central covering $H_2(G,\mathbf{Z}) \to \tilde{G} \to G$ (such exists because G is assumed perfect). Fix $g \in c$ and a lift $\tilde{g} \in \tilde{G}$. Then any element $h \in G$ centralizing g has the property that $\tilde{h}\tilde{g}\tilde{h}^{-1} \in \tilde{g}Q_c$. Consequently, the natural projection induces a bijection from the conjugacy class of $\tilde{g}Q_c$ in $\tilde{G}/Q_c$ to the conjugacy class c. Write $\tilde{G}_c$ for the quotient of $\tilde{G}$ by Q_c; it is the "largest covering to which the conjugacy class of c lifts bijectively." Then $H_2(G,c;\mathbf{Z})$ is precisely the kernel of $\tilde{G}_c \to G$.

If the ground field K contains sufficiently many roots of unity (i.e., if $\mu_N \subset K$ where N depends only on (G,c)) and S_∞ is large enough, we believe that

$$h(G,c,K) = \#H_2(G,c;\mathbf{Z}). \tag{5}$$

In the general case, we anticipate $h(G,c,K)$ will be a rational number between 0 and $\#H_2(G,c;\mathbf{Z})$ that depends on the number of roots of unity in K.

For instance, if the order of elements of c are relatively prime to $\#H_2(G,c;\mathbf{Z})$, then we believe that $h(G,c,K) = \#H_2(G,c;\mathbf{Z})$ as soon as the number m of roots of unity in K annihilates $H_2(G,c;\mathbf{Z})$ and S_∞ contains all the primes dividing m.

In fact, in the next section, we shall associate (under these conditions) a fundamental class $\mathfrak{z}(\rho) \in H_2(G,c;\mathbf{Z})$ to any G-extension ρ, and we suggest even the following refinement of (3): for any $\alpha \in H_2(G,c;\mathbf{Z})$,

$$\frac{|\mathcal{F}_X^\alpha|}{|\mathcal{S}_X|} \longrightarrow |G|^{-|S_\infty|}, \text{ as } X \to \infty, \tag{6}$$

where $\mathcal{F}_X^\alpha$ is now restricted to those G-extensions with fundamental class α.

Remark. Heuristic (3) is <u>definitely not valid</u> as stated for general (G,c) with no hypotheses on the extension. The case $G = D_4$ is one whose difficulties have been much studied. There are *no* quartic extensions of $\mathbf{Q}$ with Galois group D_4 and squarefree discriminant, although there are no local obstructions to this; indeed, squarefree discriminant implies that the Galois group is S_n. When one

counts quartic dihedral extensions with $|\mathrm{disc}| \leq X$, one gets a positive multiple of X, by a result of Cohen, Diaz y Diaz, and Olivier [5]; however, the constant is not equal to that predicted by the heuristics discussed here. The point is that the conjugacy class of "transpositions" in D_4 is not admissible – it fails to generate D_4.

2.5. Lifting invariants over global fields. Motivated by the considerations of the prior section, we shall now associate to a homomorphism $\rho : G_K \to G$, satisfying certain local conditions, a "fundamental class" $\mathfrak{z}(\rho)$ in $H_2(G, c; \mathbf{Z})$. (This association depends on the choice of a generator for the roots of unity in K.) This fundamental class is an invariant of the conjugacy class of ρ, meant to analogize the Fried-Serre "lifting invariant" of branched G-covers of the projective line [11, 24].

In addition to the group-theoretic conditions from §2.4 (namely, G center-free with trivial abelianization, $c \subset G$ admissible) we impose the following restrictions:

1. Let μ_n be the group of roots of unity in K. Then n annihilates $H_2(G, c; \mathbf{Z})$.
2. The order e of any element of c is relatively prime to $\# H_2(G, c; \mathbf{Z})$.

These conditions are satisfied, for instance, when $G = A_5$ and c is the class of 3-cycles. If these conditions fail, one may still obtain an invariant by passing to a sufficiently large cyclotomic extension, but we have not yet studied the resulting construction in sufficient detail to be confident about its properties.

Let $\tilde{G}_c$ be the covering of G constructed after (4). Condition (2) of the prior paragraph implies that there is a *unique* conjugacy class $\tilde{c}$ of $\tilde{G}_c$ which projects bijectively onto c, and whose elements have order e. Moreover, if x is an element of c, there exists a unique lifting of the cyclic subgroup $\langle x \rangle \subset G$ to a cyclic subgroup of $\tilde{G}_c$ of order e.

For brevity, we denote $H_2(G, c; \mathbf{Z})$ by A. We fix an algebraic closure $\bar{K}$ of K and let $G_K = \mathrm{Gal}(\bar{K}/K)$ be the absolute Galois group; for each place v of K, we let $G_v \subset G_K$ be a decomposition group and (for v finite) $I_v \subset G_v$ an inertia group.

Lemma. *Let S_∞ be a finite set of places of K, containing archimedean places. Let $\rho : G_K \to G$ be a homomorphism satisfying the following local properties:*

a. ρ is trivial on G_v for $v \in S_\infty$.

b. ρ is tamely ramified;

c. If v is a ramified place, $\rho(I_v)$ is a cyclic subgroup of G generated by an element of c.

Then ρ lifts to a representation $\tilde{\rho} : G_K \to \tilde{G}_c$. Moreover $\tilde{\rho}$ can be chosen so that it has properties (a), (b), i.e. it is everywhere tame, and trivial on G_v for $v \in S_\infty$.

Proof. The obstruction to such a lift lies in $H^2(G_K, A)$; it suffices to compute the obstruction locally, since the map

$$H^2(G_K, A) \longrightarrow \bigoplus_v H^2(G_v, A)$$

is injective, by virtue of the assumption that $\mu_n \subset K$.

For v infinite it is clear that $\rho|G_v$ can be lifted to a homomorphism $G_v \to \tilde{G}_c$. For v finite, $\rho|G_v$ factors through the maximal tame quotient of G_v. Fixing a generator τ_v for tame inertia as well as a Frobenius element Fr_v, we can specify $\rho|G_v$ by means of a pair $(t = \rho(\tau_v), F = \rho(\mathrm{Fr}_v)) \in G \times G$ satisfying

$$FtF^{-1} = t^q. \tag{7}$$

where $q = q_v$ is the cardinality of the residue field at v.

Let $\tilde{t}$ be the unique preimage of t with exact order e, and $\tilde{F}$ an arbitrary lift of F. By (7), t^q has order e, and so q is relatively prime to e; thus

$$\tilde{F}\tilde{t}\tilde{F}^{-1} = \tilde{t}^q, \tag{8}$$

since both sides are lifts of t^q with order e.

Thus $\rho|G_v$ lifts to $\tilde{G}_c$ for all v; thus ρ also lifts to a representation $\tilde{\rho} : G_K \to \tilde{G}_c$.

It remains to check that $\tilde{\rho}$ can be chosen to be tame at all finite places and trivial at $v \in S_\infty$. We have already constructed a tamely ramified lift of $\rho|G_v$ for each finite v. It follows that there exists, for every $v \notin S_\infty$ for which $\tilde{\rho}|G_v$ is wild, a character $\chi : G_v \to A$ so that $\chi_v\tilde{\rho}$ is tame at v. Similarly, for $v \in S_\infty$, there exists a character $\chi : G_v \to A$ so that $\chi_v\tilde{\rho}$ is trivial on G_v.

We now twist $\tilde{\rho}$ by any character $\chi : G_K \to A$ which extends χ_v at S_∞ and all other places wildly ramified in $\tilde{\rho}$, and which is tame at all other places; one checks that such a χ exists by using weak approximation. □

We now take S_∞ to be the set of archimedean places, together with all places dividing n. We shall associate an invariant $\mathfrak{z}(\rho) \in H_2(G, c; \mathbf{Z})$ to any $\rho : G_K \to G$ that satisfies the condition of the Lemma. Fix a lifting $\tilde{\rho}$ as in the Lemma.

For each place $v \notin S_\infty$ consider the sequence

$$I_v \to W_v^{\mathrm{ab}} \xrightarrow{\sim} K_v^\times,$$

where W_v is the local Weil group and the latter isomorphism is class field theory. This induces a map $I_v^{\mathrm{tame}} \to k_v^\times$, where k_v is the residue field at v.

Fix a generator g for $\mu_n \subset K$; regarding it as an element of $k_v^\times$, let g_v be any preimage of g inside I_v^{tame} with the property that the image of g_v inside the tame quotient I_v^{tame} generates a subgroup of index $\frac{q_v-1}{n}$.

Such g_v exist and any two choices g_v, g'_v satisfy $g'_v = g_v^\lambda, g_v = {g'_v}^{\lambda^{-1}}$ where λ lies in the kernel of the reduction $\widehat{\mathbf{Z}}^\times \twoheadrightarrow (\mathbf{Z}/n\mathbf{Z})^\times$.

Recall that the image $\rho(I_v)$ is either trivial or a cyclic subgroup generated by some element of c and, in either case, admits a unique lift $x \mapsto x^*$ to a cyclic subgroup of the same order in $\tilde{G}_c$. We define the lifting invariant of the homomorphism $\rho : G_K \to G$ as

$$\mathfrak{z}(\rho) = \prod_{v \notin S_\infty} z_v, \quad z_v = \tilde{\rho}(g_v)\,(\rho(g_v)^*)^{-1} \in A.$$

1. Independence of $\tilde{\rho}$: Any other lift of ρ as in the Lemma is necessarily of the form $\tilde{\rho}\psi$, for some character $\psi : G_K \to A$ that is everywhere tame, and trivial on all G_v ($v \in S_\infty$). Independence now follows from the reciprocity law of class field theory.

2. Independence of g_v (while fixing g): let $g'_v = g_v^\lambda$, with $\lambda \in \ker(\widehat{\mathbf{Z}}^\times \to (\mathbf{Z}/n\mathbf{Z})^\times)$. Then

$$\begin{aligned} \tilde{\rho}(g'_v) = (\tilde{\rho}(g_v))^\lambda &= (z_v \rho(g_v)^*)^\lambda \\ &= z_v(\rho(g_v)^*)^\lambda = z_v \rho(g_v^\lambda)^* = z_v \rho(g'_v)^*. \end{aligned}$$

3. Independence on the choice of I_v: the inertia subgroups of G_K are defined up to conjugacy, and it is clear that replacing each g_v by a conjugate does not affect z_v, and thus leaves $\mathfrak{z}(\rho)$ unchanged.

The invariant *does* depend on g; replacing g with g^α for $\alpha \in (\mathbf{Z}/n\mathbf{Z})^*$ has the effect of replacing $\mathfrak{z}$ with $\mathfrak{z}^\alpha$.

Example. Take $G = A_5$ and $K = \mathbf{Q}$. For the conjugacy class c of 3-cycles, we have $\#H_2(G, c; \mathbf{Z}) = 2$; on the other hand, for the conjugacy class c' of products of two commuting transpositions we have $\#H_2(G, c; \mathbf{Z}) = 1$. Thus, we expect there to be *twice as many* tamely ramified totally real A_5-extensions, all of whose ramification is of type c, than those all of whose ramification is of type c'.

In this case, the lifting invariant is defined as follows: The universal cover of A_5 is just $\mathrm{SL}_2(\mathbf{F}_5) \to \mathrm{PSL}_2(\mathbf{F}_5) \cong A_5$. Using the lemma, lift the given homomorphism $\rho : G_{\mathbf{Q}} \to A_5$ to $\tilde{\rho} : G_{\mathbf{Q}} \to \mathrm{SL}_2(\mathbf{F}_5)$. Let S be the set of primes p congruent to 3 mod 4 for which $\tilde{\rho}(I_p)$ has even order (cf. [24]).

Then the invariant $\mathfrak{z}(\rho)$ is determined by the parity of $|S|$. This is independent of our choice of $\tilde{\rho}$, since, for any homomorphism $\chi : G_{\mathbf{Q}} \to \{\pm 1\}$ that is tame and trivial at ∞ – i.e., the character associated to a quadratic field of positive odd discriminant – the set of $p \equiv 3 \bmod 4$ for which $\chi(I_p) \neq \{1\}$ has even cardinality.

Numerical inspection of John Jones' number field tables [16] indeed shows, amongst totally real, tame A_5-fields of small discriminant, a preponderance of inertial types of order 3, although, as we discuss in the next section, the data is very scarce and one should be very cautious about treating it as evidence.

Remark. Relaxing the condition that $|A|$ and e are coprime leads to more subtle behavior than that discussed here. For instance, take $G = \mathrm{PSL}_2(\mathbf{F}_7)$ and c the unique conjugacy class of order 4, so that $\tilde{G}_c = \mathrm{SL}_2(\mathbf{F}_7)$ and $A = \mathbf{Z}/2\mathbf{Z}$. One can check that c does *not* lift to any conjugacy class of order-4 elements of $\tilde{G}_c$, and $\rho : G_K \to G$ need not lift locally to $\tilde{G}_c$ even though $\mu_2 \subset K$. In this case different modifications to Bhargava's heuristics are needed.

2.6. Numerics. The difficulty of investigating conjectures of this kind increases very rapidly with the degree, not only because of slow convergence but because of the scarcity of examples. For instance, Jones's database of number fields shows that there are just eight totally real quintics with Galois group S_5 and discriminant less than 10^5, while Bhargava's asymptotic (which is provably correct as the discriminant goes to infinity!) would predict around 600.

One can think of this scarcity as following, in part, from analytic lower bounds for the discriminant [21]. Alternatively, one might imagine that the number of S_n-extensions of discriminant in $[0..X]$ has a secondary main term with negative coefficient. In the S_3 case, Roberts has given convincing evidence [22] that the number of totally real cubics of discriminant $\leq X$ admits an asymptotic formula

$$aX - bX^{5/6} + O(X^{1/2+\varepsilon}),$$

for certain explicit constants $a, b > 0$. This modified heuristic, which arises naturally from the pole structure of the pertinent Shintani zeta function, fits numerical data *far better* than does the Davenport-Heilbronn asymptotic aX.

Question. *Describe the lower order terms in the counting function for S_n-discriminants, the main term of which is provided by the conjectures of Malle and Bhargava.*

It seems quite likely that the phenomenon of lower-order terms only slightly smaller than the main term is rather general; thus (barring a sudden increase in the range where number fields can be counted exhaustively) a principled answer to the Question above is likely necessary for any serious numerical investigation of the conjectures, even insofar as the main term is concerned.

3. Function Fields and Hurwitz Spaces

We now discuss features of the topology of Hurwitz spaces that are responsible for the truth of theorems such as (1) over the function fields of finite fields.

We begin by discussing Hurwitz spaces in a purely topological setting (§3.1, §3.2); then, in §3.3, we discuss Hurwitz *schemes* over finite fields and their relevance to function field analogues of Bhargava-type heuristics; finally in §3.4 we discuss motivation for the "Schur correction" from §2.4.

We fix throughout an admissible pair (G, c) of a finite group G and a conjugacy class $c \subset G$; for simplicity we suppose that G is center-free.

3.1. Hurwitz spaces. In this section, $\mathbf{P}^1_{\mathbf{C}}$ denotes the complex points of the projective line; however, in §3.1 and §3.2 we are only interested in its topology, and the reader could replace it by a two-sphere without changing the meaning.

We will consider the Hurwitz space $\mathsf{Hur}_{G,c}(n)$ that, informally speaking, parameterizes G-covers of $\mathbf{P}^1_{\mathbf{C}}$, branched at n points distinct from ∞, with the monodromy around each branch point lying in c, and with a "marking" of the fiber above ∞, i.e. a G-equivariant identification of this fiber with G. For brevity, we shall regard (G, c) as fixed and write simply $\mathsf{Hur}(n)$ in place of $\mathsf{Hur}_{G,c}(n)$.

Here is the precise definition of $\mathsf{Hur}(n)$: Let $\mathsf{Conf}(n)$ be the configuration space of n points in the complex plane $\mathbf{C}$. It is a $K(\pi, 1)$ whose fundamental group, the *Artin braid group*, is generated by elements $\{\sigma_i : \quad 1 \leq i \leq n-1\}$ which pull one point in front of the next. The generators σ_i and σ_j commute when $|i-j| \neq 1$, and σ_i and σ_{i+1} satisfy the braiding relation $\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}$; these relations give a presentation of the braid group. Let $\mathsf{Hur}(n)$ be the covering space of $\mathsf{Conf}(n)$ whose fiber above a configuration $D \in \mathsf{Conf}(n)$ is the set of homomorphisms

$$\pi_1(\mathbf{P}^1_{\mathbf{C}} - D, \infty) \longrightarrow G,$$

sending a loop around each puncture to an element of c.

Equivalently, the action of the fundamental group of Conf_n on the fiber of $\mathsf{Hur}(n) \to \mathsf{Conf}(n)$ is equivalent to the standard action of the braid group on $\{(g_1, \ldots, g_n) \in c^n : g_1 g_2 \ldots g_n = 1\}$, given by the rule

$$\sigma_i : (g_1, \ldots, g_i, g_{i+1}, \ldots, g_n) \longrightarrow (g_1, \ldots, g_{i+1}, g_{i+1}^{-1} g_i g_{i+1}, \ldots, g_n). \tag{9}$$

3.2. Stable homology. The first thing to note is that the Hurwitz space $\mathsf{Hur}(n)$ need not be connected: Let $\mathsf{CHur}(n) \subset \mathsf{Hur}(n)$ be the subspace of $\mathsf{Hur}(n)$ corresponding to *surjective* homomorphisms $\pi_1(\mathbf{P}^1_{\mathbf{C}} - D, \infty) \longrightarrow G$; then $\mathsf{CHur}(n)$ parametrizes *connected* covers of $\mathbf{P}^1_{\mathbf{C}}$. Then $\mathsf{CHur}(n)$ is open and closed in $\mathsf{Hur}(n)$.

It was proved in the nineteenth century by Clebsch, Lüroth, and Hurwitz that $\mathsf{CHur}(n)$ has only *one* component when G is the symmetric group and c is the conjugacy class of transpositions. However, in general, even $\mathsf{CHur}(n)$ may

not be connected, an issue we study further in §3.4. These questions can be reduced to problems in combinatorial group theory: by (9), the components of $\mathsf{Hur}(n)$ are in bijection with the orbits of the braid group on $\{(g_1, \ldots, g_n) \in c^n : g_1 g_2 \ldots g_n = 1\}$; the components of $\mathsf{CHur}(n)$ are the orbits consisting of n-tuples which generate G.

More generally (that is, for arbitrary G, c) it is a pleasant exercise to check that the connected components *stabilize*: that is, there exists an integer E so that $\mathsf{CHur}(n)$ and $\mathsf{CHur}(n+E)$ have the same number of connected components whenever n is large enough relative to G, c.

One can think of this as a stabilization statement for the degree-zero homology group $H_0(\mathsf{CHur}(n))$. What about the higher homology?

We say that (G, c) satisfies the stability (resp. vanishing) condition if:

1. Stability condition: There exists $A > 0, E \in \mathbf{Z}$ so that

$$\dim H_j(\mathsf{CHur}(n), \mathbf{Q}) = \dim H_j(\mathsf{CHur}(n+E), \mathbf{Q}), \quad j < An - 1.$$

2. Vanishing condition: There exists $A > 0$ so that the map $\mathsf{CHur}(n) \longrightarrow \mathsf{Conf}(n)$ induces an isomorphism on rational homology

$$H_j(X, \mathbf{Q}) \xrightarrow{\sim} H_j(\mathsf{Conf}(n), \mathbf{Q})$$

in degrees $j < An$, for each connected component X (if any) of $\mathsf{CHur}(n)$.

Note that $H_j(\mathsf{Conf}(n), \mathbf{Q})$ is vanishing for $j > 1$ and one-dimensional for $j = 1$, thus the name "vanishing condition." It is very interesting to ask to what extent the regularities above might be satisfied with integral coefficients, or $\mathbf{Z}[\frac{1}{|G|}]$-coefficients. In these settings, $\mathsf{Conf}(n)$ has nontrivial cohomology in many degrees.

In [9], we prove a first theorem in this direction.

Theorem. *(E., V., Westerland). Let A be an abelian group of odd order, and $D(A)$ the generalized dihedral group $A \rtimes \mathbf{Z}/2\mathbf{Z}$, where the $\mathbf{Z}/2\mathbf{Z}$ acts on A by $a \mapsto -a$. Then the pair* $(D(A), \text{involutions})$ *satisfies the stability condition.*

The proof follows, in the large, the same lines as Harer's proof [14] of homological stability for $\mathcal{M}_g$. As in his argument, an essential element is the high connectivity of a combinatorially defined complex – in this case, the "arc complex" studied by Hatcher and Wahl [15] – on which the braid group acts. In the Hurwitz space case, a key role is played by the stable H_0 discussed above:

$$R = \oplus_{n \geq 0} H_0(\mathsf{Hur}(n), \mathbf{Q}) \tag{10}$$

As the notation suggests, R is a *ring*, with product given by concatenation of n-tuples. The animating principle of our argument is that, under the conditions

of the theorem, the homological algebra of the category of R-modules is "approximately" the same as that of the category of $\mathbf{Q}[t]$-modules. (Warning: this ring R is not exactly the same as that used in [9], where we consider covers of $\mathbf{A}^1_{\mathbf{C}}$ rather than $\mathbf{P}^1_{\mathbf{C}}$.)

We believe that far more than the Theorem above is true: not only the stability but also the vanishing condition will hold for a wide range – perhaps all – admissible pairs (G, c). For safety, we formulate this as a conjecture only in the case where we feel most secure:

Conjecture. $(S_n, \text{transpositions})$ *satisfies the vanishing condition.*

The significance for arithmetic lies in the fact that the vanishing condition essentially implies the function field version of Malle-Bhargava heuristics (including Schur corrections as in §2.4). For instance, the stated conjecture implies

> There are, on average, $\frac{1}{n!}$ totally real S_n-extensions of $\mathbf{F}_q(T)$ per squarefree discriminant,

where in this context "totally real" means that the extension is totally split at ∞; also – analogous to restricting to discriminants congruent to 1 mod 4 in (1) – we restrict to discriminants of even degree, since there are no such extensions if the discriminant degree is odd.

Similarly, the Theorem implies a (somewhat weaker) form of Malle's conjecture in the case of dihedral groups; since this particular case is usually formulated in terms of the "Cohen–Lenstra heuristics," we return to it separately in §4.

More generally, it seems that many questions of analytic number theory, when considered over a function field, are related to topological phenomena of homology stabilization, a topic that is discussed in [9, §1.7], and which we intend to take up elsewhere.

We now turn to explaining the relation between homological stability conditions, as discussed above, and counting extensions of function fields. The crucial tool that allows us to pass from topology of complex moduli spaces to enumerative questions over finite fields is, as might be expected, the Grothendieck-Lefschetz trace formula.

3.3. Hurwitz schemes.

We now explain how the homology of the Hurwitz space is related to function-field analogues of Malle's conjecture. In what follows, all schemes are over Spec $\mathbf{Z}[\frac{1}{|G|}]$.

Let $\mathscr{C}(n)$ be the scheme parameterizing configurations of n unordered distinct points on $\mathbf{A}^1$. This can be identified with the complement of the discriminant divisor inside the affine space $\mathrm{Sym}^n \mathbf{A}^1$ of degree n monic polynomials; it is an algebraic version of $\mathsf{Conf}(n)$.

It is also possible to define an algebraic version of the space $\mathsf{CHur}(n)$, i.e., a Hurwitz *scheme* $\mathscr{CH}(n)$ over $\mathbf{Z}[\frac{1}{|G|}]$; it is an étale cover of $\mathscr{C}(n)$ parameterizing

branched G-covers of $\mathbf{P}^1$ with n branch points, all of whose ramification is tame of type c, and which are endowed with an extra structure called "marking at ∞." The complex points of $\mathscr{C}\mathscr{H}(n)$ are naturally identified with the topological space $\mathsf{CHur}(n)$ of the previous section. For an exposition of the construction of $\mathscr{C}\mathscr{H}(n)$ we refer to [23].

In particular, if k is a finite field, the size of $\mathscr{C}\mathscr{H}_n(k)$ is equal to the number of isomorphism classes of G-extensions of $k(t)$, totally split at ∞ and all ramification of type c, together with a "marking at ∞."

The Grothendieck–Lefschetz fixed point formula gives a relation between the number of $\mathbf{F}_q$-points of an algebraic variety and its étale cohomology. It is possible (cf. [9, §7]) to compare the singular homology of the Hurwitz space $\mathsf{CHur}(n)$, and the étale cohomology of the Hurwitz scheme $\mathscr{C}\mathscr{H}(n)$ over $\overline{\mathbf{F}_q}$. In this way, we obtain a relation between the singular homology of the Hurwitz space and Malle's conjecture.

The stability condition for (G, c) alone, together with relatively elementary bounds, shows that the étale cohomology of the Hurwitz scheme is "not too large;" although the only *a priori* control on the Frobenius action comes from the Weil conjectures, this already suffices for interesting upper and lower bounds for $|\mathscr{C}\mathscr{H}(\mathbf{F}_q)|$. An example of a result thus obtained is the theorem given in §4.2; see also [9, pp. 5–6] for further discussion of this technique.

However, if the Hurwitz scheme and the configuration scheme have *the same* rational homology in some range – as the vanishing conjecture predicts, in cases where $\mathsf{CHur}(n)$ is connected – then $|\mathscr{C}\mathscr{H}(n)(\mathbf{F}_q)|$ and $|\mathscr{C}(n)(\mathbf{F}_q)|$ will be approximately equal, as long as q is sufficiently large relative to (G, c). Equivalently, as $n \to \infty$ with q fixed there will be an average of *one* marked G-extension per discriminant. (The $1/n!$ term in the conjecture stated in the previous section comes from the existence of $n!$ different markings on each extension that is totally split at ∞.)

3.4. Stable components and the Schur correction. The vanishing condition of §3.2 gives very strong control of the homology of each component of $\mathsf{CHur}(n)$; but $\mathsf{CHur}(n)$ is not connected in general, and indeed, the description of the set of connected components is somewhat subtle, especially when the Galois action is taken into account.

Remarkably, it is possible to completely understand the connected components in the large n limit, owing to a beautiful theorem of Conway–Parker and Fried-Völklein. Only a proof of a special case is available in print: [12, Appendix], but it contains all the necessary ideas. We also do not attempt to formulate it in the most general case, restricting to G perfect. For the definition and basic properties of $\tilde{G}_c$ used in the definition below, we refer the reader to the discussion after (4).

Theorem. *(Conway-Parker-Fried-Völklein). Suppose G is perfect; let $g \mapsto g^*$ be a conjugacy-equivariant bijection from $c \subset G$ to a conjugacy class of $\tilde{G}_c$*

lifting c. Then, for sufficiently large n, the map

$$(g_1, \ldots, g_n) \in c^n \longrightarrow g_1^* \ldots g_n^* \in \tilde{G}_c$$

induces a bijection from the stable component group to $H_2(G, c; \mathbf{Z})$, as defined in (4)*. In particular, the number of components of* $\mathsf{CHur}(n)$ *equals $\#H_2(G, c; \mathbf{Z})$ for n sufficiently large.*

The source of the "Schur correction" in the function field case is now apparent: it arises from the fact that the Hurwitz scheme $\mathscr{CH}_n$ may have multiple components, and therefore more points than expected; the number of components is related to the size of a Schur multiplier. More precisely, the group $H_2(G, c; \mathbf{Z})$ bijects onto the set of *geometric* components of $\mathscr{CH}_n/\mathbf{F}_q$; if some of these components are not defined over $\mathbf{F}_q$, there may be *fewer* G-covers than expected. This is what happens in the situation of the Remark, page 393; and this is the reason why we have postulated "enough roots of unity" in (5).

Let us make precise the relation to §2.4. Let k be a finite field of order q and $K = k(t)$. We maintain the hypotheses that (G, c) is admissible, that q is relatively prime to $|G|$, that G is center-free and that G^{ab} is trivial; we take the set S_∞ of §2.4 to consist of the place corresponding to the point at ∞.

The methods of [9] establish the following: if k contains sufficiently many roots of unity (that is, if $q-1$ is sufficiently divisible), and q is sufficiently large – both notions depending on (G, c) – then, with $h = \#H_2(G, c; \mathbf{Z})$ the number of connected components of $\mathsf{CHur}(n)$,

> The vanishing condition for (G, c) implies the function field case of (3): $\frac{|\mathcal{F}_{q^m}|}{|\mathcal{S}_{q^m}|} \sim \frac{h}{|G|}$, as $m \to \infty$.

and the weakened version:

> The stability condition for (G, c) implies that $\left| \frac{|\mathcal{F}_{q^m}|}{|\mathcal{S}_{q^m}|} - \frac{h}{|G|} \right| \leq Aq^{-1/2}$, where the constant $A = A(G, c)$ is independent of q, m.

Concerning the notion of "enough" roots of unity: It is very likely[2] that the assumptions of §2.5 – i.e. $\mu_m \subset K$ and $(e, m) = 1$ where e is the order of an element of c and m annihilates $\#H_2(G, c; \mathbf{Z})$ – are sufficient to ensure the validity of the above results, and moreover that, in this case, the refined statement (6) is valid.

Remark. It is particularly interesting to examine from this point of view the phenomenon of "lower order terms" discussed in §2.6. It is natural to suppose that such lower order terms correspond to natural families of cohomology classes on the Hurwitz schemes (albeit classes in degrees which increase with

[2]To verify this amounts to checking compatibility between certain definitions in characteristic zero and positive characteristic; we have not done so carefully.

the dimension of the scheme). Is there any explicit description of these unstable cohomology classes?

4. Special Case: The Cohen–Lenstra Heuristics

The Cohen-Lenstra heuristics – as originally formulated in [4] – are concerned with the average behavior of class groups of quadratic fields; in particular, they try to explain numerical observations such as

$$\mathbf{Z}/9\mathbf{Z} \text{ occurs in class groups more often than } \mathbf{Z}/3\mathbf{Z} \times \mathbf{Z}/3\mathbf{Z}. \tag{11}$$

As we explain, they can be considered a special case of the Bhargava–Malle type conjectures formulated earlier, in the case of dihedral groups. Indeed, (1) establishes one of the few known cases of the Cohen-Lenstra heuristics.

These heuristics are of particular importance because they are readily formulated and relatively easy to investigate numerically.

4.1. Number fields. For any global field L, denote by C_L the class group. Let $\mathcal{Q}_X$ be the set of imaginary quadratic extensions of $\mathbf{Q}$ with discriminant in $[-X, 0]$. Let ℓ be an odd prime. The Cohen-Lenstra conjecture asserts that, for any finite abelian ℓ-group A,

$$\lim_{X\to\infty} \frac{\sum_{L\in\mathcal{Q}_X} |\mathrm{Epi}(C_L, A)|}{|\mathcal{Q}_X|} = 1. \tag{12}$$

where $\mathrm{Epi}(C_L, A)$ denotes the set of surjective homomorphisms from C_L to A. This implies that[3], for any ℓ-group B, the fraction of $L \in \mathcal{Q}_X$ with $C_L[\ell^\infty] \cong B$ is asymptotically $\frac{\prod_{i=1}^{\infty}(1-\ell^{-i})}{|\mathrm{Aut}(B)|}$. This makes manifest why (11) should be true. But the formulation (12) emphasizes the "rationality" feature of the answer.

Before returning to function fields, let us describe a heuristic for (12) in the spirit of Cohen and Lenstra's original work. The class group of C_L is the quotient of all ideals by principal ideals. If we fix a sufficiently large set of finite places V, C_L will be isomorphic to the quotient of the free group $\mathbf{Z}[V]$ by the image U of the $V \cup \{\infty\}$-units. There are $A^{|V|}$ homomorphisms from $\mathbf{Z}[V]$ to A; the chance that a homomorphism is trivial on U is $|A|^{-\mathrm{rank}\ U}$; since $\mathrm{rank}(U) = |V|$, this suggests (12).

The Cohen-Lenstra heuristics can be viewed as a special case of Malle's conjecture: let $D(A)$ be the group $A \rtimes \mathbf{Z}/2\mathbf{Z}$, where $\mathbf{Z}/2\mathbf{Z}$ acts on A via $a \mapsto -a$, and let c be the set of elements which project to the nontrivial element of $\mathbf{Z}/2\mathbf{Z}$. Then there is a bijection between $D(A)$-extensions of $\mathbf{Q}$, all of whose ramification is of type c, and pairs $(L, f : C_L \twoheadrightarrow A)$, where f is defined only up to ± 1. Thus (12) becomes equivalent to a question of the type considered in §2.

[3]For the implication, see [10] for the case of ℓ-torsion, [9, Corollary 8.2] in general.

4.2. Function fields. Now let k be a finite field of odd cardinality q and let $\mathcal{Q}'_X$ be the set of imaginary[4] quadratic extensions of $k(t)$ with discriminant less than X. The following counting result then follows from the homological stability proved in [9]:

Theorem. *(E. , V., Westerland). Suppose $q \not\equiv 0, 1$ modulo ℓ.*

$$\limsup_{X\to\infty} \frac{\sum_{L\in\mathcal{Q}'_X} |\mathrm{Epi}(C_L, A)|}{|\mathcal{Q}'_X|} = 1 + O(q^{-1/2}), \tag{13}$$

for all q sufficiently large (this notion depending only on A). The same is true for $\liminf$.

There is a similar statement [9, Theorem 1.2] concerning the fraction of L for which $C_L[\ell^\infty]$ lies in a specific isomorphism class.

The proof of this theorem is based on the "program" outlined in §3.3 and in particular is a corollary to Theorem of §3.2. As discussed in §3.3, a proof of the *vanishing* conjecture would remove the factor $O(q^{-1/2})$ and show the existence of the limit, as long as q is sufficiently large relative to A.

Remark. The restriction $q \not\equiv 1$ modulo ℓ ensures that k does not contain μ_ℓ. If k contains μ_ℓ, the methods of [9] give a corresponding theorem, but the limit changes, because the pertinent Hurwitz scheme acquires more $\mathbf{F}_q$-rational connected components. A corresponding phenomenon in the number field case has been predicted by Malle [20]. We regard this as an *example* of a Schur correction phenomenon, even though we did not formulate §2.4 in sufficient generality to include dihedral groups. Forthcoming work of Garton [13] will provide an explanation of the phenomena discovered by Malle from the viewpoint of function-field analogies.

References

[1] Manjul Bhargava. Higher composition laws and applications. In *International Congress of Mathematicians. Vol. II*, pages 271–294. Eur. Math. Soc., Zürich, 2006.

[2] Manjul Bhargava. Mass formulae for extensions of local fields, and conjectures on the density of number field discriminants. *Int. Math. Res. Not. IMRN*, (17):Art. ID rnm052, 20, 2007.

[3] H. Cohen. Constructing and counting number fields. In *Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002)*, pages 129–138, Beijing, 2002. Higher Ed. Press.

[4] H. Cohen and H. W. Lenstra, Jr. Heuristics on class groups. In *Number theory (New York, 1982)*, volume 1052 of *Lecture Notes in Math.*, pages 26–36. Springer, Berlin, 1984.

[4] That is to say: ramified at ∞.

[5] H. Cohen, F.D. Y Diaz, and M. Olivier. Enumerating Quartic Dihedral Extensions of $\mathbb{Q}$. *Compositio Mathematica*, 133(01):65–93, 2002.

[6] H. Davenport and H. Heilbronn. On the density of discriminants of cubic fields. II. *Proc. Roy. Soc. London Ser. A*, 322(1551):405–420, 1971.

[7] W. Duke. Bounds for arithmetic multiplicities. In *Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998)*, number Extra Vol. II, pages 163–172 (electronic), 1998.

[8] Nathan M. Dunfield and William P. Thurston. Finite covers of random 3-manifolds. *Invent. Math.*, 166(3):457–521, 2006.

[9] Jordan Ellenberg, Akshay Venkatesh, and Craig Westerland. Homological stability for Hurwitz spaces and the Cohen-Lenstra conjecture over function fields. `http://arxiv.org/abs/0912.0325`.

[10] Étienne Fouvry and Jürgen Klüners. On the 4-rank of class groups of quadratic number fields. *Invent. Math.*, 167(3):455–513, 2007.

[11] M. Fried. Enhanced review: Serre's Topics in Galois theory. *Proceedings of the Recent developments in the Inverse Galois Problem conference, AMS Cont. Math*, 186:15–32, 1995.

[12] Michael D. Fried and Helmut Völklein. The inverse Galois problem and rational points on moduli spaces. *Math. Ann.*, 290(4):771–800, 1991.

[13] Derek Garton. Random matrices and the Cohen-Lenstra statistics for global fields with roots of unity. UW-Madison Ph.D. thesis, in progress, 2010.

[14] John L. Harer. Stability of the homology of the mapping class groups of orientable surfaces. *Ann. of Math. (2)*, 121(2):215–249, 1985.

[15] Allen Hatcher and Nathalie Wahl. Stabilization for mapping class groups of 3-manifolds. *preprint;* `arXiv:math/0601310`.

[16] John Jones. Table of number fields. `http://hobbes.la.asu.edu/NFDB/`.

[17] Jürgen Klüners. A counterexample to Malle's conjecture on the asymptotics of discriminants. *C. R. Math. Acad. Sci. Paris*, 340(6):411–414, 2005.

[18] Gunter Malle. On the distribution of Galois groups. *J. Number Theory*, 92(2):315–329, 2002.

[19] Gunter Malle. On the distribution of Galois groups. II. *Experiment. Math.*, 13(2):129–135, 2004.

[20] Gunter Malle. Cohen-Lenstra heuristic and roots of unity. *J. Number Theory*, 128(10):2823–2835, 2008.

[21] A. M. Odlyzko. Bounds for discriminants and related estimates for class numbers, regulators and zeros of zeta functions: a survey of recent results. *Sém. Théor. Nombres Bordeaux (2)*, 2(1):119–141, 1990.

[22] David P. Roberts. Density of cubic field discriminants. *Math. Comp.*, 70(236):1699–1705 (electronic), 2001.

[23] Matthieu Romagny and Stefan Wewers. Hurwitz spaces. In *Groupes de Galois arithmétiques et différentiels*, volume 13 of *Sémin. Congr.*, pages 313–341. Soc. Math. France, Paris, 2006.

[24] Jean-Pierre Serre. Revêtements à ramification impaire et thêta-caractéristiques. *C. R. Acad. Sci. Paris Sér. I Math.*, 311(9):547–552, 1990.

[25] Seyfi Türkelli. Connected components of Hurwitz schemes and Malle's conjecture. `http://arxiv.org/abs/0809.0951`, 2008.

Section 4

Algebraic and Complex Geometry

Prakash Belkale
The Tangent Space to an Enumerative Problem 405

Christopher D. Hacon* and James McKernan*
Boundedness Results in Birational Geometry 427

Daniel Huybrechts
Hyperkähler Manifolds and Sheaves 450

D. Kaledin
Motivic Structures in Non-commutative Geometry 461

Chiu-Chu Melissa Liu
Gromov-Witten Theory of Calabi-Yau 3-folds 497

Christopher D. Hacon* and James McKernan*
Flips and Flops 513

Mihai Păun
Quantitative Extensions of Twisted Pluricanonical Forms and Non-vanishing 540

Shuji Saito
Cohomological Hasse Principle and Motivic Cohomology of Arithmetic Schemes 558

Frank-Olaf Schreyer* and David Eisenbud
Betti Numbers of Syzygies and Cohomology of Coherent Sheaves 586

Vasudevan Srinivas
Algebraic Cycles on Singular Varieties 603

Richard P. Thomas
An Exercise in Mirror Symmetry 624

Jean-Yves Welschinger
Invariants Entiers en Géométrie Énumérative Réelle 652

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

The Tangent Space to an Enumerative Problem

Prakash Belkale*

Abstract

We will discuss recent work on the relations between the intersection theory of homogeneous spaces (and their quantum, and higher genus generalizations), invariant theory, and non-abelian theta functions. The main theme is that the analysis of transversality in enumerative problems can be viewed as a bridge from intersection theory to representation theory. Some of the new results proved using these ideas are reviewed: multiplicative generalizations of the Horn and saturation conjectures, generalizations of Fulton's conjecture, the deformation of cohomology of homogeneous spaces, and the strange duality conjecture in the theory of vector bundles on algebraic curves.

Mathematics Subject Classification (2010). Primary 14M17, 14N15, 14D20; Secondary 14L24, 14N15.

Keywords. Intersection theory, homogeneous spaces, theta functions, invariant theory, Horn conjecture, saturation conjecture, strange duality.

1. Introduction

An enumerative problem is a problem of counting the number of points in a space that satisfy certain geometrically defined conditions. For example, one may consider the classical problem of intersecting Schubert varieties (in general position) in a Grassmannian, or the more recent problem in quantum cohomology of counting maps from the projective line to a homogeneous space satisfying certain incidence conditions, or of counting subbundles of a fixed (general) vector bundle of a given degree and rank on an algebraic curve.

Take an enumerative problem such as one of the above. Under suitable conditions the enumerative problem counts the number of points of a certain

*Partially supported by NSF grants DMS-0901249 and (FRG) DMS-0554247.
Department of Mathematics, UNC-Chapel Hill, CB #3250, Phillips Hall, Chapel Hill NC 27599. E-mail: belkale@email.unc.edu.

scheme X (which is frequently an intersection of several smooth subvarieties of a smooth variety). In many cases X is reduced as a scheme, and of dimension zero. Such enumerative problems will be said to have the transversality property.

The general theme of this report is to review recent work analyzing the implications of transversality in such situations. The following questions will be considered.

(a) Can we obtain (hopefully "simpler") consequences of a (non-empty) transversal intersection which are equivalent to the enumerative problem having a non-empty solution?

(b) Transversality can be immediately translated as the non-zeroness of suitable determinants (see Section 5 for an example). This leads one to sections ("the theta sections") of line bundles over suitable moduli-spaces which have representation theoretic significance. How effective is this link between intersection theory and representation theory?

Note that the assumptions on the scheme X (that it is smooth and of the expected dimension) is often the consequence of powerful theorems in algebraic geometry. It usually comes about by combining Kleiman's transversality theorem and Grothendieck's computation of tangent spaces of quot schemes.

The following enumerative problems will be considered in this report. In each case we will try to analyze the implications of transversality. Quite surprisingly, the transversality properties often link up to invariant theory (and generalizations).

- *The classical Schubert calculus in a Grassmannian* $\mathrm{Gr}(r,n)$: In this case a tangent space analysis can be made to illuminate the known relation of the Schubert structure constants to invariant theory of $\mathrm{GL}(r)$ and $\mathrm{GL}(n-r)$. Analysis of tangent spaces can be used to geometrically prove many (previously known) results in the area: Geometric proofs of Horn and Saturation conjectures; Fulton's conjecture.

- *The (small) quantum cohomology of a Grassmannian:* In this case the transversality analysis illuminates the known relation of structure constants in the small quantum cohomology to fusion rings. It also allows for a generalization of the Horn and Saturation conjectures to this setting.

- *The classical Schubert calculus in arbitrary homogeneous spaces G/P's:* The relationship between intersection theory and invariant theory is more complicated here. The analysis of tangent spaces reveals a deformation of the product in the singular cohomology $H^*(G/P,\mathbb{Z})$. This deformed product has links to invariant theory and is closely tied in with the Hermitian eigenvalue problem. In particular an analogue of Fulton's conjecture holds here.

- *Higher genus generalizations:* There are many "quantum cohomology" type enumerative problems in higher genus. The structure coefficients in

the fusion rings above are dimensions of the spaces of parabolic theta functions. In higher genus, there are several flavors of these spaces (because of the non triviality of the Jacobian). But the transversality techniques manage to illuminate the picture here, leading to the proof of the strange duality conjecture.

There are a few other enumerative problems that have not yet been analyzed in the above fashion. I want to single out the big quantum cohomology rings, and the enumerative problems of mapping curves into non-homogeneous spaces. Transversality may not hold in general in the last problem, but recent work in enumerative geometry has found a way around this difficulty (virtual fundamental classes etc). Could one hope for a variation of the above theme of analyzing tangent spaces to shed further light on these problems?

I have tried to make the sections independent of each other. Section 2 provides the background to the context in which the transversality techniques were developed (this section can be skipped by a more knowledgeable reader).

The focus for this report is the relation between enumerative questions and invariant theory. We will not discuss relations to geometric invariant theory (a recent highlight here is the work of Ressayre [52]), or questions in representation theory ("saturation conjectures", a highlight here is the work of Kapovich and Millson [32]), or relations to the eigenvalue problem. We refer the reader to [26, 30, 38] for many related aspects of these questions.

There are other ways of relating the number of points of intersections of Schubert varieties with invariant theory [42, 56]. The approach of [42] incorporates an asymptotic study of solutions to the Kniznik-Zamolodchikov (KZ) equations. It can be hoped that this will somehow link to the discussion on non-abelian theta functions in Section 6, which in turn incorporates the study of the Hitchin connection, a higher genus generalization of the KZ equations.

We will work over the field of complex numbers. The representations considered are complex representations of complex algebraic groups. In fact, the methods considered here allow one to prove transversality in some enumerative problems in characteristic p (see e.g. [12]), but we will not consider these applications here.

I would like to thank V. Srinivas for his comments on a preliminary version of this report and L. Matusevich for some references in Section 4.1.

2. The Context for Many of the Problems

The following questions have been the source of some recent developments in geometry and representation theory. Many of the tangent space techniques were developed while trying to understand the geometry behind these problems.

Question 2.1. Given the eigenvalues of two hermitian matrices, what are the possible eigenvalues of their sum?

Question 2.2. Given irreducible representations V_λ and V_μ of $\mathrm{GL}(n)$, which irreducible representations of $\mathrm{GL}(n)$ appear in the tensor product $V_\lambda \otimes V_\mu$?

Question 2.3. Given the eigenvalues of two unitary matrices, what are the possible eigenvalues of their product?

Question 2.4. ("The Schubert calculus") Given cycle classes of Schubert varieties $[\omega_I]$ and $[\omega_J] \in H^*(\mathrm{Gr}(r,n))$ in the cohomology of a Grassmannian, which cycle classes $[\omega_K]$ appear with non-zero coefficient in the cup product of $[\omega_I]$ and $[\omega_J]$?

Question 2.1 has an interesting history going back to the work of Weyl (see the survey article of Fulton [26]). Question 2.3 can be seen as a Riemann-Hilbert problem by restating it as: "What are the possible eigenvalues of unitary matrices A, B and C which satisfy $ABC = 1$?", and recognizing the equation $ABC = 1$ as the fundamental relation in the fundamental group of $\mathbb{P}^1 - \{0, 1, \infty\}$. It has a very illustrious history (the unitary group can be replaced by any group, for example the general linear group). Question 2.1 can be considered to be the Lie algebra version of Question 2.3. Questions 2.1 and 2.3 are examples of "eigenvalue problems".

In 1962, Horn [29] gave a conjectural solution to Question 2.1, by a recursively determined system of inequalities. This conjecture was free of cohomology. Klyachko [35] gave another solution to Question 2.1; in terms of a list of inequalities parameterized by non-vanishing structure constants in the Schubert calculus of Grassmannians $\mathrm{Gr}(r,n)$ (Question 2.4).

In [36], Knutson and Tao proved the saturation theorem which says that for irreducible representations V_λ, V_μ and V_ν of $\mathrm{Gl}(n)$ given by Young diagrams λ, μ and ν, V_ν appears in $V_\lambda \otimes V_\mu$ if and only if for some positive integer N, $V_{N\nu}$ appears in $V_{N\lambda} \otimes V_{N\mu}$.

It turns out that the problem: "Does there exist N so that $V_{N\nu}$ appears in the tensor product of $V_{N\lambda}$ and $V_{N\mu}$?" is equivalent to the Hermitian eigenvalue problem for $n \times n$ matrices (for eigenvalues of the summands given by λ and μ, and that of the sum by ν); and the Schubert calculus problem is equivalent to an instance of Question 2.2 for the smaller group $\mathrm{GL}(r)$. Note that Schubert cycle classes for the Grassmannians $\mathrm{Gr}(r,n)$ are also parameterized by Young diagrams, see Section 3.

These works taken together implied Horn's original conjecture. They also implied that the non-vanishing question for a product of Schubert cohomology classes (Question 2.4) in a given Grassmannian has an inductive solution: It is characterized by a series of inequalities coming from knowing the answer to the same question for smaller Grassmannians (see Theorem 3.3).

This set of questions, solved in the late nineties, forms a point of confluence of combinatorics, algebraic geometry, representation theory and symplectic geometry (see Fulton's survey [26]). For the generalizations considered below, the combinatorial apparatus is largely missing. The numerical relationship between the invariant theory of $\mathrm{GL}(r)$ and the Schubert calculus of Grassmannians

$\mathrm{Gr}(r,n)$, which at first glance seems to be a phenomenon restricted to the general linear groups, was used rather crucially at several places in the first known proofs of many of these results.

2.1. Generalizations. There are three basic objects in this picture. The first one is a topological one concerning representations of fundamental groups (or the Lie algebra version). The second one is representation theoretic (tensor product problem), and the third is intersection theoretic (cohomology of Grassmannians).

Viewed in this way, there are many potential generalizations. We may replace the group $\mathrm{GL}(n)$ by an arbitrary reductive group. Actually, we will find it convenient to state results for semi-simple rather than reductive groups.

We may also consider multiplicative eigenvalue problems (see Question 2.3), replace invariant theory by the fusion ring, and the cohomology by quantum cohomology of Grassmannians (one can change the group too).

Another generalization is to replace $\mathbb{P}^1$ by an arbitrary curve (with or without punctures), the fusion ring by the non-abelian theta functions and quantum cohomology by some higher genus generalization.

There are also potential generalizations to the subgroup embedding problems pioneered by Berenstein and Sjamaar [20].

3. Intersection Theory in an Ordinary Grassmannian

Let I be a subset of $[n]=\{1,\dots,n\}$ of cardinality r. Write $I=\{i_1<i_2<\cdots<i_r\}$. Let

$$E_\bullet : 0=E_0\subsetneq E_1\subsetneq\cdots\subsetneq E_n=\mathbb{C}^n$$

be a complete flag of subspaces of $\mathbb{C}^n$. Define the Schubert variety

$$\Omega_I(E_\bullet) \;=\; \{V\in \mathrm{Gr}(r,W)\mid \mathrm{rk}(V\cap E_{i_a})\geq a,\ \text{for } 1\leq a\leq r\}$$

Denote the cycle class in the cohomology of this subvariety by ω_I. The codimension of $\Omega_I(E_\bullet)$ in $\mathrm{Gr}(r,n)$ is $\mathrm{codim}(\omega_I)=\sum_{a=1}^r(n-r+a-i_a)$.

Now suppose that we are given three subsets I, J and K of $[n]$ of cardinality r each such that

$$\mathrm{codim}(\omega_I)+\mathrm{codim}(\omega_J)+\mathrm{codim}(\omega_K)=\dim\mathrm{Gr}(r,n).$$

The prototypical enumerative problem is the one of counting the number of points in the intersection

$$\Omega_I(E_\bullet)\cap\Omega_J(F_\bullet)\cap\Omega_K(H_\bullet) \tag{1}$$

when the triple of complete flags $(E_\bullet,F_\bullet,G_\bullet)$ is in general position. This number also equals the number m such that

$$\omega_I\cdot\omega_J\cdot\omega_K=m[\text{class of a point}] \tag{2}$$

3.1. The work of Klyachko and Knutson-Tao. Irreducible polynomial representations of the $\mathrm{GL}(r)$, (or equivalently, the unitary group $\mathrm{U}(r)$) are parameterized by weakly decreasing sequences of non-negative integers $\lambda = (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r)$. These restrict to irreducible representations $\bar{\lambda}$ of $\mathrm{SL}(r)$ (equivalently the special unitary group $\mathrm{SU}(r)$). Sequences λ and μ restrict to give the same irreducible representation of $\mathrm{SU}(r)$ if and only if, the difference $\lambda_a - \mu_a = c$ for some constant c and all $a \in [r]$. The congruence class $|\bar{\mu}| = \sum_{a=1}^{r} \mu_a \pmod r \in \mathbb{Z}/r\mathbb{Z}$ is therefore well defined.

In [36], Knutson and Tao and proved the saturation conjecture for $\mathrm{SL}(r)$:

Theorem 3.1. *Consider representations V_λ, V_μ and V_ν of $\mathrm{SL}(r)$, such that*

$$|\bar{\lambda}| + |\bar{\mu}| + |\bar{\nu}| \equiv 0 \pmod r \tag{3}$$

then the following are equivalent

1. $(V(\lambda) \otimes V(\nu) \otimes V(\nu))^{\mathrm{SL}(r)} \neq 0$
2. *For some positive integer N,* $(V(N\lambda) \otimes V(N\mu) \otimes V(N\nu))^{\mathrm{SL}(r)} \neq 0$.

By the work of Klyachko [35], the second property in Theorem 3.1 is characterized by a system of inequalities which is parameterized by non vanishing structure constants in smaller Grassmannians $\mathrm{Gr}(\tilde{r}, r)$ where $1 \leq \tilde{r} < r$. A proof of the saturation conjecture using the quiver theory was later given by Derken and Weyman [25].

3.2. Numerical relations between intersection numbers and invariant theory. Intersection theory of Grassmannians and invariant theory of the special linear group $\mathrm{GL}(r)$ are related, and this has been known for a long time. To describe this, note that sequences I, J and K as above also parameterize some irreducible (polynomial) representations of $\mathrm{GL}(r)$. The association takes

$$I \mapsto \lambda_I = (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r),\ \lambda_a = n - r + a - i_a,\ 1 \leq a \leq r.$$

Denote the corresponding representation of $\mathrm{GL}(r)$ by $V(\lambda_I)$. Then, the intersection multiplicity m (from (2)) equals the dimension of the space of invariants

$$\big(V(\lambda_I) \otimes V(\lambda_J) \otimes V(\lambda_K)\big)^{\mathrm{SL}(r)} \tag{4}$$

(See [26] for some history, and for the proof of this assertion.) Note that since $\mathrm{Gr}(r, n)$ and $\mathrm{Gr}(n - r, n)$ are isomorphic, these sets up a "numerical strange duality" between the invariant theories of $\mathrm{GL}(r)$ and $\mathrm{GL}(n - r)$.

A geometric "reason" was given in [9], and [8]: Each point of intersection V of the Schubert varieties (1) produces an non-zero invariant θ_V in the dual of the vector space (4).

It is easy to describe θ_V, using the Borel-Weil theorem. It is known that ${V_\lambda}^* = H^0(\mathrm{Fl}(r), \mathcal{L}_\lambda)$ where $\mathrm{Fl}(r)$ is the variety of complete flags in $\mathbb{C}^r$, and $\mathcal{L}_\lambda$

is a suitable line bundle with a $\mathrm{SL}(r)$-action. The dual of the space of invariants (4) is therefore a subspace of $H^0(\mathrm{Fl}(r)^3, \mathcal{L}_\lambda \boxtimes \mathcal{L}_\mu \boxtimes \mathcal{L}_\nu)$.

Therefore, to describe θ_V up to scalars, we may as well describe its zero scheme. Fix a point V as above, let $Q = \mathbb{C}^n/V$. The given flags on $\mathbb{C}^n$ induce flags $E'_\bullet$, $F'_\bullet$ and $G'_\bullet$ on V and also $E''_\bullet$, $F''_\bullet$ and $G''_\bullet$ on Q. The zero locus of θ_V is the set of points $(R_\bullet, S_\bullet, T_\bullet) \in \mathrm{Fl}(r)^3$ such there exists a non-zero homomorphism (of vector spaces) $\phi : \mathbb{C}^r \to Q$ so that for every $a \in [r]$, $\phi(R_a) \subset E''_{i_a - a}$, $\phi(S_a) \subset F''_{j_a - a}$ and $\phi(T_a) \subset G''_{k_a - a}$. These conditions were suggested by a computation of tangent spaces of Schubert varieties. They can be readily converted into a determinantal condition [8] and are reminiscent of the theta divisor from the theory of vector bundles (the parabolic version).

It is easy to see that each point V as above also gives a point V in $\mathrm{Fl}(r)^3$ (well defined upto the diagonal $\mathrm{SL}(r)$ action). We can show that if V_i and V_j are in the transversal intersection (1), then θ_{V_i} vanishes at V_j if and only if $i \neq j$. These claims for $i \neq j$ hold because there is a natural non-zero map $V_j \to \mathbb{C}^n/V_i$ (inclusion into $\mathbb{C}^n$ followed by projection) satisfying the above conditions. For $i = j$, we use the fact that the intersection (1) is transverse at V_i.

The linear independence of the sections θ_{V_i} follows immediately. Together with the known agreement of m with the dimension of (4), we get [9, 8]:

Theorem 3.2. *Let $V_1, \ldots, V_m$ be the points in* (1)*. Then, $\theta_{V_1}, \ldots, \theta_{V_m}$ form a basis for the space of invariants* (4)*.*

3.3. The Geometric Horn Property. The work of Klyachko-Knutson-Tao and the numerical relation between intersection numbers and invariant theory implies the following theorem which says that we can decide whether $m \neq 0$ by writing down a series of inequalities coming from knowing the answer to the same question for smaller Grassmannians. Here m is the intersection number from (1).

Theorem 3.3. *Let $\lambda = \lambda_I$, $\mu = \lambda_J$ and $\nu = \lambda_K$. The following are equivalent*

1. *$m \neq 0$.*
2. *For every $1 \leq \tilde{r} < r$ and choice of subsets A, B, C of $[r]$ each of cardinality $\tilde{r}$ so that $\omega_A \cdot \omega_B \cdot \omega_C \neq 0 \in H^*(\mathrm{Gr}(\tilde{r}, r))$ the following inequality holds*

$$\sum_{a \in A} \lambda_a + \sum_{b \in B} \mu_b + \sum_{c \in C} \nu_c \leq \tilde{r}(n - r)$$

Fulton proposed the challenge of finding a geometric proof of Theorem 3.3. This was achieved in [10]. It was based on the following idea: If general Schubert varieties intersect at a point, then by Kleiman's transversality, they intersect transversally there. Conversely, one can detect if general Schubert varieties intersect, by a tangent space calculation (also see [49] where similar ideas were

pursued independently). The tangent space of the Grassmannian $\mathrm{Gr}(r, W)$ at point V is isomorphic to $\mathrm{Hom}(V, W/V)$. One may use the action of $\mathrm{GL}(V) \times \mathrm{GL}(W/V)$ on $\mathrm{Hom}(V, W/V)$ (this action has only finitely many orbits) to study the reasons for a non-transverse intersection.

It was shown in [10] that Theorem 3.3 immediately implies the Knutson-Tao saturation theorem. In fact using Theorem 3.2, we can construct geometrically "explicit" elements of the space of invariants (4). The explicitness of these sections should be of use, when analyzing functoriality issues [18].

3.4. Fulton's conjecture. Fulton conjectured a statement which runs parallel to the saturation conjecture:

If $|\lambda| + |\mu| + \nu| \equiv 0 \pmod r$ then, $\mathrm{rk}(V(\lambda) \otimes V(\mu) \otimes V(\nu))^{\mathrm{SL}(r)} = 1$ implies that for all positive integers N, $\mathrm{rk}(V(N\lambda) \otimes V(N\mu) \otimes V(N\nu))^{\mathrm{SL}(r)} = 1$ (the opposite implication is also true, but that follows from the saturation conjecture)

As in Section 3.2, one may view the representations $V(\lambda)$ via the Borel-Weil theorem, as the space of global sections of certain line bundles over complete flag varieties. Using GIT, one can view the space of invariants in a tensor product as the space of global sections of a certain line bundle over a suitable moduli space $\mathcal{M}$ (of "semi-stable" parabolic vector spaces of rank r). The property $\mathrm{rk}(V(N\lambda) \otimes V(N\mu) \otimes V(N\nu))^{\mathrm{SL}(r)} = 1$, for all positive integers N, implies the rigidity statement that $\mathcal{M}$ is a point.

Fulton's conjecture was first proved by Knutson, Tao and Woodward [37] using combinatorial methods. In [13], a geometric proof was given. The proof used the connections to intersection theory and the special theta sections θ_V from Section 3.2.

Remark 1. The saturation theorem of Knutson-Tao also takes a nice form in terms of $\mathcal{M}$:

$$\mathcal{M} \neq \emptyset \implies h^0(\mathcal{M}, \mathcal{L}) \neq 0$$

Here $\mathcal{L}$ on $\mathcal{M}$ is a natural line bundle obtained through descent (see e.g. [13] for more details).

4. Quantum Cohomology

The intersection theoretic problem considered here is the (small) quantum cohomology of Grassmannians. For simplicity we restrict to three punctures, although the results are valid for any number of punctures. Fix three points $0, 1$ and ∞ on $\mathbb{P}^1$. Let I, J and K be subsets of $[n]$ each of cardinality r. Let d be a non-negative integer such that

$$\mathrm{codim}(\omega_I) + \mathrm{codim}(\omega_J) + \mathrm{codim}(\omega_K) = \dim \mathrm{Gr}(r, n) + dn$$

Define the Gromov-Witten number $\langle \omega_I, \omega_J, \omega_K \rangle$ to be the number of maps $\mathbb{P}^1 \to \mathrm{Gr}(r, n)$ of degree d such that $f(0) \in \Omega_I(E_\bullet), f(1) \in \Omega_J(F_\bullet), f(\infty) \in$

$\Omega_K(G_\bullet)$. The triple of flags $(E_\bullet, F_\bullet, G_\bullet)$ is assumed to be in general position. The (small) quantum cohomology ring, a generalization of ordinary cohomology of Grassmannians encapsulates the Gromov-Witten numbers as structure coefficients [39, 27].

It is easy to see that sub-bundles of the trivial rank n bundle $\mathcal{O}_{\mathbb{P}^1}^{\oplus n}$, of degree rank r and degree $-d$ correspond bijectively to maps $\mathbb{P}^1 \to \mathrm{Gr}(r,n)$ of degree d. The conditions at 0, 1 and ∞ above, can be interpreted in terms of the corresponding sub-bundle.

The analogue of the Hermitian eigenvalue problem is the problem of characterizing possible eigenvalues of a product of unitary matrices. This problem can be reinterpreted as the problem of determining the possible local monodromies in a unitary representation of the fundamental group of $\mathbb{P}^1 - \{0, 1, \infty\}$.

The analogue of the ring of invariants is the fusion ring (see e.g. [5]) of the special unitary group $\mathrm{SU}(r)$, which incorporates an additional parameter of a nonnegative "level". The structure coefficients in the fusion ring are the dimensions of spaces of sections of suitable line bundles on moduli stacks of parabolic bundles on $\mathbb{P}^1$ [46]. Therefore, the quantum generalization replaces GIT quotients of products of flag varieties with the moduli spaces of parabolic bundles.

A theorem of Witten [59] relates the (small) quantum cohomology of Grassmannians to the fusion rings of unitary groups.

The theorems of Sections 3 generalize to the quantum setting. The generalizations of Klyachko's theorem are known [22, 3, 7]; these works are based on the theorem of Mehta and Seshadri [41] which relates unitary representations of the fundamental group of a punctured curve and semi-stable parabolic bundles. According to this generalization, the multiplicative eigenvalue problem is controlled by a system of inequalities, parameterized by non-vanishing Gromov-Witten numbers.

The analogue of theorem 3.3 is the following theorem, proved in [12]. The proof is by an examination of the transversality in the enumerative problem. We will use notation from Section 3.

Theorem 4.1. *Let I, J and K be subsets of $[n]$, each of cardinality r and let d be a non-negative integer such that*

$$\mathrm{codim}(\omega_I) + \mathrm{codim}(\omega_J) + \mathrm{codim}(\omega_K) = dn + r(n-r),$$

Write $d = qr + h$ with $0 \le h < r$ and $(q,h) \in \mathbb{Z}^2$. Let

$$L = \{x \in [n] \mid \exists\, y \in I, x \equiv y - i_h \pmod{n}\}$$

(where $i_h = 0$ if $h = 0$). Let $(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r) = \lambda_L$, $\mu = \lambda_J$ and $\nu = \lambda_K$. The following are equivalent:

(a) $\langle \omega_I, \omega_J, \omega_K \rangle_d \neq 0$.

(b) *For any integers $\tilde{d}$ and $\tilde{r}$ with $0 < \tilde{r} < r$, $\tilde{d} \geq 0$, and A, B, C subsets of $[r]$ each of cardinality $\tilde{r}$, such that $\langle \omega_A, \omega_B, \omega_C \rangle_{\tilde{d}} = 1$, the following inequality holds:*

$$\sum_{\alpha \in A} \tilde{\lambda}_a + \sum_{b \in B} \mu_b + \sum_{c \in C} \nu_c \leq \tilde{d}(n-r) + \tilde{r}(qn + i_h) + \tilde{r}(n-r).$$

The following version of the above theorem links non-vanishing Gromov-Witten numbers and the multiplicative eigenvalue problem for a smaller group:

Recall that conjugacy classes in $\mathrm{SU}(r)$ are in one to one correspondence with points in the $(n-1)$-simplex

$$\Delta(r) = \left\{ \alpha = (\alpha_1, \ldots, \alpha_r) \mid \alpha_1 \geq \cdots \geq \alpha_r \geq \alpha_1 - 1, \ \sum_{t=1}^{r} \alpha_t = 0 \right\} \subseteq \mathbb{R}^r$$

where, to $(\alpha_1, \ldots, \alpha_r)$, we associate the conjugacy class of the diagonal matrix with entries $\exp(2\pi\sqrt{-1}\alpha_t)$ for $t = 1, \ldots, r$.

For a subset of $[n]$ of cardinality r, with associated sequence $\lambda_I = (\lambda_1 \geq \cdots \geq \lambda_r)$, define a conjugacy class $\beta(I) = (\beta_1, \ldots, \beta_r)$ for $\mathrm{SU}(r)$ as follows:

$$\beta(I) = \frac{1}{n-r}(\lambda_1, \ldots, \lambda_r) - \frac{|\lambda(I)|}{r(n-r)}(1, \ldots, 1)$$

where $|\lambda(I)| = \sum_{a=1}^{r} \lambda_a$.

The center of $\mathrm{SU}(r)$ acts on the conjugacy classes of elements in $\mathrm{SU}(r)$. Set $\zeta_r = \exp(\frac{2\pi\sqrt{-1}}{r}) \in \mathbb{C}$. Given a conjugacy class α for $\mathrm{SU}(r)$ we can multiply α by ζ_r and obtain a new conjugacy class $\zeta_r \alpha$.

Theorem 4.2. *[12] Under the conditions of Theorem 4.1, the assertions (a), (b) there are equivalent to each of the following*

1. *There exist $U, V, W \in \mathrm{SU}(r)$ satisfying*
   - *$UVW = I$.*
   - *U, V and W are in the conjugacy classes corresponding to $\zeta_r^d \beta(I)$, $\beta(J)$, and $\beta(K)$ respectively.*
2. *There exists a $\mathrm{SU}(n)$-local system $\mathcal{L}$ on $\mathbb{P}^1 - \{0, 1, \infty\}$ such that the local monodromies of $\mathcal{L}$ at 0, 1 and ∞ are $\zeta_r^d \beta(I)$, $\beta(J)$ and $\beta(K)$ respectively.*

4.1. Fulton's conjecture. It can be shown that, using the techniques of [13] we have the following generalization of Fulton's conjecture (unpublished). In the setting of Theorem 4.2, let $\mathcal{M}$ be the moduli-space of unitary local systems $\mathcal{L}$ on $\mathbb{P}^1 - \{0, 1, \infty\}$ such that the local monodromies of $\mathcal{L}$ at 0, 1 and ∞ are $\zeta_r^d \beta(I)$, $\beta(J)$ and $\beta(K)$ respectively.

Theorem 4.3. *The following are equivalent.*

1. $\langle \omega_I, \omega_J, \omega_K \rangle_d = 1$.
2. $\mathcal{M}$ *is a point.*

In the above setting, if $\langle \omega_I, \omega_J, \omega_K \rangle_d = 1$, and the corresponding local system $\mathcal{L}$ is irreducible, then one gets examples of rigid local systems, in the sense of [33], with unitary monodromy. Could this relation of unitary rigid local systems to quantum cohomology shed light on the problem of classifying (and constructing) rigid local systems with finite global monodromy? (see [7, 21, 28]). We note that Theorem 4.3 has an extension to an arbitrary number of punctures.

4.2. Fusion rings, and saturation. A theorem of Witten [59] relates the (small) quantum cohomology of Grassmannians to the fusion rings of unitary groups. We can use this numerical information to obtain a geometrically defined basis of the space of global sections of the corresponding line bundle on a moduli stack of parabolic bundles on $\mathbb{P}^1$, in exactly the same manner as in Section 3.2.

Theorem 4.1 and Witten's theorem can be used to obtain a generalization of the Knutson-Tao saturation theorem to fusion rings (see [12]).

5. Intersection Theory in an Arbitrary Homogeneous Space

Let G be a connected semi-simple complex algebraic group. We choose a Borel subgroup B and a maximal torus $H \subset B$. Let $P \supseteq B$ be a (standard) parabolic subgroup of G. The enumerative problem that we want to consider is the problem of intersecting Schubert varieties in G/P. The representation theoretic problem is the invariant theory of G. Let L be the Levi subgroup of P, and $L^{ss} = [L, L]$ its semi-simple part.

Let K be a maximal compact in G, and let $\mathfrak{k}$ be the Lie algebra of K. K acts on $\mathfrak{k}$ via the adjoint representation. The orbits of this conjugation (or adjoint) action are parameterized by the positive Weyl chamber in $\mathfrak{h}$, the Lie algebra of H. The analogue of the Hermitian eigenvalue problem is the problem of characterizing the conjugacy class of a sum $C = A + B$ where $A, B \in \mathfrak{k}$, given the conjugacy classes of A and B. Recall that Hermitian matrices of trace zero form the Lie algebra of $\mathrm{SU}(n)$.

There are some genuine surprises in the case of arbitrary groups. Many of the known properties for the case $G = \mathrm{SL}(n)$ fail (without suitable modifications). For example, the saturation conjecture has to be amended [32] and the generalized Klyachko inequalities (as in [20, 31]) describing the solution to the Hermitian eigenvalue problem turn out to be redundant. In joint work with S. Kumar [17], the author discovered a deformation of $H^*(G/P)$ which seems to

be more finely tuned to the cohomological/intersection theoretic issues and to eigenvalue problems.

Let us first recall the definition and parameterizations of Schubert varieties in G/P. Let W be the Weyl group of G, W_P the Weyl group of P, and let W^P be the set of minimal length coset representatives in W/W_P. For any $w \in W^P$, define the (shifted) Schubert cell

$$\Lambda_w^P = w^{-1}BwP/P \subset G/P$$

Let $[\bar{\Lambda}_w^P] \in H^{2\dim(G/P)-2\ell(w)}(G/P, \mathbb{Z})$ denote the cycle class of the closure $\bar{\Lambda}_w^P$ of Λ_w^P. Here, $\dim(G/P) - \ell(w)$ is the codimension of the Schubert variety Λ_w^P in G/P.

5.1. Intersection of Schubert varieties in a G/P. Now assume that we are given a triple $(u, v, w) \in W^P \times W^P \times W^P$ (we will state results for three factors, but these results are valid for any number of factors) such that

$$\operatorname{codim}(\Lambda_u^P) + \operatorname{codim}(\Lambda_v^P) + \operatorname{codim}(\Lambda_w^P) = \dim(G/P) \tag{5}$$

In this case, by Kleiman's theorem, for general $(g, h, k) \in G^3$, $g\Lambda_u^P$, $h\Lambda_v^P$ and $k\Lambda_u^P$ meet transversally in a finite set of points. The number of points of intersection

$$m = |g\Lambda_u^P \cap h\Lambda_v^P \cap k\Lambda_w^P|$$

can be calculated using cohomology of G/P. More precisely,

$$[\bar{\Lambda}_u^P] \cdot [\bar{\Lambda}_v^P] \cdot [\bar{\Lambda}_w^P] = m[\bar{\Lambda}_e^P]$$

where $e \in W^P$ is the identity element.

The enumerative problem is the one of calculating m. It is easy to see that $m \neq 0$ if (and only if) we can get $g\Lambda_u^P$, $h\Lambda_v^P$ and $k\Lambda_u^P$ to intersect transversally at some point. We may assume by translations that this point is $\dot{e} \in G/P$. Note that

- $g\Lambda_u^P$ passes through $\dot{e} \in G/P$ if and only if $g\Lambda_u^P = p\Lambda_u^P$ for some $p \in P$.
- The Borel B_L of the Levi L of P does not move the Schubert varieties $B_L\Lambda_u^P = \Lambda_u^P$. Therefore, P/B_L is a complete parameter space of Schubert varieties (having fixed $u \in W^P$) passing through $\dot{e}$.

Lemma 5.1. *[17] The following are equivalent*

1. $|g\Lambda_u^P \cap h\Lambda_v^P \cap k\Lambda_w^P| \neq 0$ *for general* $(g, h, k) \in G^3$.
2. *For general* $p_1, p_2, p_3 \in P$, *the following map between vector spaces of the same dimension is an isomorphism*

$$T_{\dot{e}}(G/P) \to \frac{T_{\dot{e}}(G/P)}{p_1T(G/P)_{\dot{e}}} \oplus \frac{T_{\dot{e}}(G/P)}{p_2T(G/P)_{\dot{e}}} \oplus \frac{T_{\dot{e}}(G/P)}{p_3T(G/P)_{\dot{e}}}$$

By taking determinants in (2), we can view (2) as the non-vanishing of a natural section θ of a line bundle $\mathcal{L}$ on a product $(P/B_L)^3$ which is invariant for the diagonal P-action. This line bundle $\mathcal{L}$ can easily be identified as a product of line bundles from the factors. By the Borel-Weil theory, global sections of positive line bundles on G/B, give irreducible representations. Therefore one is tempted to view θ in a tensor product of irreducible representations of P.

But P is not reductive. This problem is not serious when P is a cominiscule maximal parabolic, where the unipotent radical of P acts as zero on the tangent space $T(G/P)_{\dot{e}}$, so θ is really a invariant section of a line bundle over $(L/B_L)^3$ (in the case of Grassmannians, the author had shown this before, in [8]).

5.2. A deformation of cohomology. We are led to the following definition (see [17]).

Definition 1. We call a triple (u, v, w) satisfying (5) *Levi-movable* if, for generic $(l_1, l_2, l_3) \in L^3$, the intersection $l_1\Lambda_u^P \cap l_2\Lambda_v^P \cap l_3\Lambda_w^P$ is a transverse intersection at $\dot{e}$ (Hence the number $m \neq 0$).

It turns out that Levi-movable triples with $m = 1$, and P maximal parabolic are exactly the ones required in eigenvalue problems (all others are redundant) [17]. By Ressayre's work [52], the reduced set of inequalities corresponding to the Levi-movable triples form an irredundant set of inequalities (this generalizes a theorem of Knutson-Tao-Woodward [37] in the case of $G = \mathrm{SL}(n)$).

In the case of the ordinary Grassmannians, and more generally for cominiscule flag varieties, the above condition is vacuous. In general, the condition of Levi-movability is equivalent to having $m \neq 0$ and a system of linear equalities (see [17], Theorem 15). A triple (u, v, w) satisfying these numerical equalities will be called numerically Levi-movable in this report.

Definition 2. Suppose we write the structure coefficients in the usual cup product in $H^*(G/P)$ by

$$[\bar{\Lambda}_u^P] \cdot [\bar{\Lambda}_v^P] = \sum c_{u,v}^w [\bar{\Lambda}_w^P]$$

Define the deformed product $\odot_0$ by the following rule

$$[\bar{\Lambda}_u^P] \odot_0 [\bar{\Lambda}_v^P] = \sum{}' c_{u,v}^w [\bar{\Lambda}_w^P]$$

where the sum is restricted to w so that the triple $(u, v, w_o w w_o^P)$ is Levi-movable. Here w_o (resp. w_o^P) is the longest element of W (resp. W_P).

Remark 2. 1. It is an open problem to find combinatorial "manifestly non-negative" rules for the structure coefficients in the usual cup product on $H^*(G/P)$ (in the Schubert basis). However such rules are known, for cominiscule (maximal) parabolics. In these cases, the deformation of cohomology is trivial. One is therefore tempted to ask if there are combinatorial rules for the structure coefficients in the product $\odot_0$ (for arbitrary G and P).

2. Are there formulas analogous to the Pieri rules, which give a description of $\odot_0$ in terms of generators and relations? Is there a description analogous to the classical Giambelli formula, of cycle classes of Schubert varieties in terms of the generators?

5.3. The deformed product and invariant theory. A preliminary connection to the representation theory of the Levi subgroup is established in [17]. For every $w \in W^P$, a line bundle $L(\chi_w)$ on P/B_L was constructed where χ_w is a corresponding character in $\mathfrak{h}^*$ (see [17] for the definitions). It was shown there that if (u, v, w) are L-movable, then $H^0((L/B_L)^3, L(\chi_u) \boxtimes L(\chi_v) \boxtimes L(\chi_w))^{L^{ss}} \neq 0$. Note that $H^0(L/B_L, L(\chi_u)) = V(\chi_u)$ is an irreducible representation of L and we have the following result [17]:

Proposition 5.2. *If* $[\bar{\Lambda}_u^P] \odot_0 [\bar{\Lambda}_v^P] \odot_0 [\bar{\Lambda}_v^P] = m[\bar{\Lambda}_e^P] \in H^*(G/P), m \neq 0$, *then* $I = \mathrm{rk}(V(\chi_u) \otimes V(\chi_v) \otimes V(\chi_w))^{L^{ss}} \neq 0$.

The corresponding construction in the case of $G = \mathrm{GL}(n)$ and $G/P = \mathrm{Gr}(r, n)$ was previously carried out in [8]. In this case, $L^{ss} = SL(r) \times SL(n-r)$. Starting from the Schubert cell parameterized by λ, the corresponding representation of L^{ss} coincides with $V(\lambda)^* \otimes V(\tilde{\lambda})$, where $V(\lambda)$ is the corresponding irreducible representation of $\mathrm{SL}(r)$ and $\tilde{\lambda}$ is the conjugate partition giving rise to the irreducible representation $V(\tilde{\lambda})$ of $\mathrm{SL}(n-r)$. We therefore have the stronger relation $I = m^2$.

In general, however there are no known numerical relations between m and I. In fact we know that m is not given by a formula in I (or the other way). The following question is expected (at least by this author!) to have a positive answer. Suppose (u, v, w) is numerically Levi-movable.

Question 5.3. For every semi-simple group G and every maximal parabolic P, does $I \neq 0$ imply that $m \neq 0$ (the opposite implication is true by the above discussion).

Also note that if we set $I_N = \mathrm{rk}(V(N\chi_u) \otimes V(N\chi_v) \otimes V(N\chi_w))^{L^{ss}}$, then there is a finite set of inequalities characterizing the stable tensor product problem: $I_N \neq 0$ for some N (see [20]). It is not known if the following is true:

Question 5.4. Does $I_N \neq 0$ for some $N \geq 1$ imply that $I \neq 0$.

A positive answer to Question 5.4 will constitute saturation for some special representations of L^{ss} ("of intersection theoretic origin"). If Questions 5.3 and 5.4 have positive answers, then we would have a generalization of the geometric version of the Horn conjecture (Theorem 3.3). In the case of miniscule parabolics [50], and in the case of maximal parabolics in symplectic and odd orthogonal groups [18], there are known versions of the geometric form of the Horn conjecture. These are different from the generalization that would follow from the truth of Questions 5.3 and 5.4.

The analogue of Fulton's conjecture holds [19]:

Theorem 5.5. *If $m = 1$, then for every positive integer N, $I_N = 1$.*

Let $\mathcal{M}$ be the GIT quotient by the diagonal action of L^{ss} of the space $(L/B_L)^3$ linearized by $L(\chi_u) \otimes L(\chi_v) \otimes L(\chi_w)$. The conclusion of the theorem is equivalent to the rigidity statement that $\mathcal{M} =$ point. Therefore, multiplicity one in intersection theory leads to rigidity in representation theory.

The converse to Theorem 5.5 is not true as stated. There are examples where $I_N = 1$ for all N, but $m > 1$ (see [19] for an example).

The condition $m = 1$ in Theorem 5.5 can be translated into the statement that a certain map of parameter spaces $X \to Y = (G/B)^3$ appearing in Kleiman's theorem is birational. Here X is the "universal intersection" of closed Schubert varieties. If X were smooth, we could argue as follows: Let $R \subset X$ be the ramification divisor. Since $X \to Y$ is birational, no multiple of R can move (even infinitesimally). We may therefore conclude that $h^0(X, \mathcal{O}(NR)) = 1$ for every positive integer N. In the case at hand, X is not smooth, and $H^0(X, \mathcal{O}(NR))$ needs to be connected to invariant theory.

Remark 3. Let $G = \mathrm{SL}(n)$, and P a maximal parabolic. It would be very interesting to obtain the numerical relation $I = m^2$ using (only) the geometry of the map $X \to (G/B)^3$. Let X^0 be the universal intersection of the smooth parts of the (closed) Schubert varieties. $X - X^0$ has codimension ≥ 2 in X. It is shown in [19] that $(V(N\chi_u) \otimes V(N\chi_v) \otimes V(N\chi_w))^{L^{ss}}$ is dual to $H^0(X^0, \mathcal{O}(R))^G$. The intersection number m is the degree of the generically finite map $X \to (G/B)^3$. Perhaps a clever application of a (suitable) equivariant-Riemann-Roch theorem will achieve this end?

6. Non-abelian Theta Functions

We have seen that invariant theory of a group G generalizes to the fusion ring. The structure coefficients in the fusion rings are the dimensions of the spaces of certain line bundles over suitable moduli spaces of parabolic bundles on $\mathbb{P}^1$. We will now consider higher genus generalizations. For simplicity, we will not consider parabolic structures. There are different flavors of moduli spaces of vector bundles on curves.

6.1. Moduli spaces and theta functions. To define these objects, let X be a connected smooth projective algebraic curve X of genus $g \geq 1$ over $\mathbb{C}$. Let $\mathrm{SU}_X(r)$ be the moduli space of semi-stable vector bundles of rank r with trivial determinant over X. For any line bundle L of degree $g - 1$ on X define $\Theta_L = \{E \in \mathrm{SU}_X(r), h^0(E \otimes L) \geq 1\}$ which is a Cartier divisor on $\mathrm{SU}_X(r)$ and let $\mathcal{L} = \mathcal{O}(\Theta_L)$ (which is independent of L). The spaces $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ should be considered as a non-abelian generalization of invariant theory.

Because of the non-triviality of the Jacobian of X, we can consider a slightly different space as well. Let $U_X^*(k)$ be the moduli space of semi-stable rank k and degree $k(g-1)$ bundles on X. Recall that on $U_X^*(k)$ there is a canonical non-zero theta (Cartier) divisor Θ_k whose underlying set is $\{F \in U_X^*(k), h^0(X,F) \neq 0\}$. Put $\mathcal{M} = \mathcal{O}(\Theta_k)$. The spaces $H^0(U_X^*(k), \mathcal{M}^r)$ may also be considered to be higher genus generalizations of invariant theory (in this way, Riemann-Jacobi theta functions are generalizations of invariant theory). In fact, these spaces can be considered as a twisted version of the spaces from the previous paragraph.

The analogue of the Hermitian eigenvalue problem is the problem of the moduli of special unitary representations of the fundamental group of X. By the classical Narasimhan-Seshadri theorem [45], this topological moduli-space is homeomorphic to the algebraic variety $\mathrm{SU}_X(r)$. A significant departure from the genus zero situation is that unitary local systems with given local monodromies at a set of punctures always exist on curves of genus ≥ 1. Motivated by Remark 1, we asked if the spaces of parabolic theta functions (with an arbitrary number of punctures), satisfying conditions analogous to (3) are always non-zero in genus $g \geq 1$ (for $G = \mathrm{SL}(n)$). This was proved by Boysal [23].

6.2. Strange duality. The intersection theoretic generalization in the higher genus setting is not immediately clear. In analogy with Section 3.2, we may expect that having a corresponding enumerative problem would lead to spanning sets for the spaces $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$. The strange duality conjecture predicts a good spanning set for the space $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ which we will now describe (also see [48, 47] for expository accounts, especially for the history of this problem).

Consider the natural map $\tau_{k,r} : \mathrm{SU}_X(r) \times U_X^*(k) \to U_X^*(kr)$ given by tensor product. From the theorem of the square, it follows that $\tau_{k,r}^*\mathcal{M}$ is isomorphic to $\mathcal{L}^k \boxtimes \mathcal{M}^r$. The canonical element $\Theta_{kr} \in H^0(U_X^*(kr), \mathcal{M})$ and the Kunneth theorem gives a map well defined up to scalars:

$$H^0(U_X^*(k), \mathcal{M}^r)^* \to H^0(\mathrm{SU}_X(r), \mathcal{L}^k). \tag{6}$$

The strange duality conjecture asserts that (6) is an isomorphism. Consider the restriction of Θ_{kr} to $\mathrm{SU}_X \times \{F\}$ for each $F \in U_X^*(k)$. This gives rise to a section $\theta_F \in H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ well defined up to scalars. The strange duality conjecture is equivalent to the statement that the sections θ_F span $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$. Notice that this situation is strikingly analogous to the discussion in Section 3. To have the analogy lead to the strange duality, we need an enumerative problem with the same number of solutions as the dimension of the space $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ (we would then need to analyze the consequences of transversality).

6.2.1. The enumerative problem. This led to the work in [14]. The enumerative problem was not on X but on a nodal degeneration of X. Let T be a general vector bundle of degree $k(g-1)$ and rank $r+k$ on $\mathbb{P}^1$. Let $p_1, , \ldots p_g$ and $q_1, \ldots, q_g$ be distinct points on $\mathbb{P}^1$. For $i = 1, \ldots, g$, consider vector space

homomorphisms $\eta_i : T_{p_i} \to T_{q_i}$ with kernel of rank 1 (but general up to this condition). The enumerative problem is the following: Count the set of subbundles E of T of degree zero and rank r such that the following two conditions are satisfied:

1. $\eta_i(E_{p_i}) \subset E_{q_i}$
2. E_{p_i} contains the kernel of η_i.

Say we get bundles $E_1, \ldots, E_m$ above. The enumerative number m above equals $h^0(SU_X(r), L^k)$, where X as before, is an arbitrary connected, smooth and projective curve of genus g. This follows from the degeneration formulas of [57], the known agreement of the structure coefficients of the fusion and quantum cohomology rings (in genus 0), and a study of the cohomology class of the diagonal in a (self) product of a Grassmannian.

Now consider the nodal curve X' obtained by gluing p_i to q_i, for $i = 1, \ldots, g$. It can be shown that η_i descend to give isomorphisms $(T/E_i)_{p_i} \to (T/E_i)_{q_i}$. These can then be used to glue, and therefore one obtains vector bundles $F'_1, \ldots, F'_m$ of rank k and degree $k(g-1)$ on X'. Consider the deformations $F_1, \ldots, F_m$ of $F'_1, \ldots, F'_m$ to a general smooth X. The strange duality on the general curve X was proved in [14] in the following form:

Theorem 6.1. *The theta sections $\theta_{F_1}, \ldots, \theta_{F_m}$ form a basis of $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$.*

Notice that this is formally analogous to Theorem 3.2 (In Theorem 3.2, the θ_V are defined through the quotient $Q = \mathbb{C}^n/V$).

In fact there is a very natural enumerative problem on the curve (not necessarily general) X itself. Let T be a general vector bundle of degree $k(g-1)$ and rank $r+k$ on X. The enumerative problem is the problem of counting subbundles of T of degree zero and rank $k(g-1)$. It is easy to see that transversality holds in this enumerative problem (using the first order consequence of the fact that the sub-bundle deforms with deformations of T). There are finitely many such bundles $(E_1, \ldots, E_{m'})$. It is also immediate that $\mathrm{Hom}(E_i, T/E_j)$ is non-zero if and only if $i \neq j$ (such a calculation occurred first in [8], variants of this calculation occur in both [14] and [40]). Now if E_i had trivial determinants we could use the T/E_j as candidates for the spanning set of F's predicted by the strange duality conjecture. However this may not be the case, and m' may not equal $m = h^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ (it is in fact larger). In [14] a part of m' was identified (in a degeneration) which corresponds to $m = h^0(\mathrm{SU}_X(r), \mathcal{L}^k)$

In subsequent work, Marian and Oprea built an enumerative problem on an arbitrary curve that corresponds to a variant of $H^0(SU_X(r), \mathcal{L}^k)$ and succeeded in proving strange duality for all curves [40]. This work introduced an interesting variant of the original strange duality map which is symmetric in both sides.

6.3. Other perspectives. Conformal field theory introduces a new perspective in the study of the spaces $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$. The starting point is an

observation (made rigorous by many authors) that since vector bundles with trivial determinant on X minus a point p are trivial, the moduli stack of rank r vector bundles with trivialized determinant is a double quotient (where z is a formal coordinate at p)

$$\mathrm{SL}_r(\mathcal{O}(X-p))\backslash \mathrm{SL}_r(\mathbb{C}((z)))/\,\mathrm{SL}_r(\mathbb{C}[[z]])$$

Let $\mathcal{Q} = \mathrm{SL}_r(\mathbb{C}((z)))/\,\mathrm{SL}_r(\mathbb{C}[[z]])$ and $\mathcal{L}'$ the pull back of $\mathcal{L}^k$ under the natural map from $\mathcal{Q}$ to the double quotient. The space $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ is the subspace of sections of $H^0(\mathcal{Q}, \mathcal{L}')$ invariant under the action of $\mathrm{SL}_r(\mathbb{C}[[z]])$. However the bundle $\mathcal{L}'$ is linearized not for the action of $\mathrm{SL}_r(\mathbb{C}((z)))$, but rather, for a central extension of it, and the space $H^0(\mathcal{Q}, \mathcal{L}')$ is the dual V_k^* of an irreducible representation of this central extension (a result of Kumar and Mathieu). Therefore, there is a nice Borel-Weil picture of $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ as a subspace of V_k^*. This point of view makes contact with the theory of conformal blocks [57] and the representation theory of Kac-Moody algebras (for more details, see Sorger's Bourbaki report [55]).

The intuition from physics gave rise to many surprises in the theory of non-abelian theta functions. One of the surprises is the existence of a flat projective connection (Hitchin's connection) on the spaces $H^0(\mathrm{SU}_X(r), \mathcal{L}^k)$ as X varies in a family. In [15], I pointed out that the strange duality map (6) is projectively flat for Hitchin's connection (and hence the conjecture for general curves implies it for all curves). The flatness of the strange duality in the genus zero case lies at the very heart of the physics expectation on strange duality (see [43, 44]), and follows easily from the theory of conformal embeddings.

The work of Abe ([1, 2], also see [44, 16]) in proving Beauville's symplectic strange duality conjecture [6] shows that monodromy arguments may play an essential role in proving strange duality type theorems in their most general context. An interesting aspect of Beauville's symplectic duality is that it seems to have no classical analogue. The author is not aware of any numerical relationships between the invariant theories of different symplectic groups. There are no known relations to enumerative geometry either. Similarly, the work of Boysal-Pauly [24] on the strange duality for the exceptional groups does not seem to have classical analogues, or relations to enumerative geometry!

6.3.1. Motives and strange duality:. The KZ/Hitchin connection in genus zero with insertions (and the choice of representations of $\mathrm{SL}(n)$ associated to the insertion points) is known to be motivic (see [58] and the references therein) i.e., the non-abelian theta functions can be realized as a subspace of the cohomology group of a smooth variety (which varies with the pointed curve), consistent with the KZ/Hitchin and Gauss-Manin connections (also see [51]).

Question 6.2. Is the strange duality also of a motivic origin?

One may ask a similar question in higher genus as well. It is not known whether the Hitchin connection in genus ≥ 1 is motivic.

References

[1] T. Abe, *Degeneration of the strange duality map for symplectic bundles*, J. Reine Angew. Math. **631** (2009), 181–220.

[2] T. Abe *Strange duality for parabolic symplectic bundles on a pointed projective line,* Int. Math. Res. Not. Vol. **2008** : Art.ID rnn121.

[3] S. Agnihotri and C. Woodward, *Eigenvalues of products of Unitary matrices and Quantum Schubert calculus,* Math. Res. Lett. **5** (1998), 817–836.

[4] A. Beauville, *Vector bundles on curves and generalized theta functions: recent results and open problems,* Current topics in complex algebraic geometry (Berkeley,CA, 1992/93), 17–33. Math. Sci. Res. Publ., **28**. Cambridge university Press, Campridge, 1995.

[5] A. Beauville, *Conformal blocks, fusion rules and the Verlinde formula,* Proceedings of the Hirzebruch 65 Conference on Algebraic Geometry, 75–96, Israel Math. Conf. Proc., **9**, Bar-Ilan Univ., Ramat Gan, 1996.

[6] A. Beauville, *Orthogonal bundles on curves and theta functions,* Ann. Inst. Fourier (Grenoble) **56** (2006), no. 5, 1405–1418.

[7] P. Belkale, *Local systems on $\mathbb{P}^1 - S$ for S a finite set*, Compositio Math. **129** (2001) no.1., p. 67–86.

[8] P. Belkale, *Invariant theory of* $\mathrm{GL}(n)$ *and intersection theory of Grassmannians,* Int. Math. Res. Not. Vol **2004**, no.49, p. 2655–2670.

[9] P. Belkale, *Geometric Proofs of Horn and Saturation conjectures,* November 2002, arXiv:math/0208107v2.

[10] P. Belkale, *Geometric Proofs of Horn and Saturation conjectures,* J. Algebraic Geom. **15** (2006), no. 1, 133–173.

[11] P. Belkale, *Extremal unitary local systems on $\mathbb{P} - \{p_1, \ldots, p_s\}$,* Algebraic groups and homogeneous spaces, 37–64, Tata Inst. Fund. Res. Stud. Math., Tata Inst. Fund. Res., Mumbai, 2007.

[12] P. Belkale, *Quantum generalization of the Horn conjecture,* J. Amer. Math. Soc. **21** (2008), 365–408.

[13] P. Belkale, *Geometric proof of a conjecture of Fulton,* Adv. Math. **216** (2007), 346–357.

[14] P. Belkale, *The strange duality conjecture for generic curves,* J. Amer. Math. Soc. **21** (2008), 235–258.

[15] P. Belkale, *Strange duality and the Hitchin/WZW connection*, J. Diff. Geom. **82** (2009) 445–465.

[16] P. Belkale, *Orthogonal bundles, theta characteristics and the symplectic strange duality*, Preprint, arXiv:0808.0863.

[17] P. Belkale and S. Kumar, *Eigenvalue problem and a new product in cohomology of flag varieties,* Invent. Math. **166**, 185–228 (2006).

[18] P. Belkale and S. Kumar, Eigencone, saturation and Horn problems for symplectic and odd orthogonal groups, J. Algebraic Geom. **19** (2010), 199–242.

[19] P. Belkale, S. Kumar and N. Ressayre, A generalization of Fulton's conjecture for arbitrary groups, preprint.

[20] A. Berenstein and R. Sjamaar, *Projections of coadjoint orbits, moment polytopes, and the Hilbert-Mumford criterion,* J. Amer. Math. Soc. **13** (2000), 433–466.

[21] F. Beukers and G. Heckman, Monodromy for the hypergeometric function ${}_nF_{n-1}$, Invent. Math. **95** (1989), 325–354.

[22] I. Biswas, *A criterion for the existence of a parabolic stable bundle of rank two over the projective line,* Internat. J. Math. (1998) no. **5**, 523–533.

[23] A. Boysal, *Nonabelian theta functions of positive genus,* Proc. Amer. Math. Soc. **136** (2008), 4201–4209.

[24] A. Boysal and C. Pauly, *Strange duality for Verlinde spaces of exceptional groups at level one,* Int. Math. Res. Not. **2009**; doi:10.1093/imrn/rnp151.

[25] H. Derksen and J. Weyman, *Semi-invariants of quivers and saturation for Littlewood-Richardson coefficients,* J. Amer. Math. Soc. **13** (2000), 467–479.

[26] W. Fulton, *Eigenvalues, invariant factors, highest weights, and Schubert calculus,* Bull. Amer. Math. Soc. **37** (2000), 209–249.

[27] W. Fulton and R. Pandharipande *Notes on stable maps and quantum cohomology. Algebraic geometry—Santa Cruz 1995,* 45–96, Proc. Sympos. Pure Math., **62**, Part 2, Amer. Math. Soc., Providence, RI, 1997.

[28] Y. Haraoka, *Finite monodromy of Pochhammer equation,* Ann. Inst. Fourier (Grenoble) **44** (1994), no. 3, 767–810.

[29] A. Horn, *Eigenvalues of sums of Hermitian matrices,* Pacific J. Math. **12** (1962), p. 225–241.

[30] M. Kapovich, *Generalized triangle inequalities and their applications.* International Congress of Mathematicians. Vol. II, 719–741, Eur. Math. Soc., Zürich, 2006.

[31] M. Kapovich, B. Leeb and J. Millson, *The generalized triangle inequalities in symmetric spaces and buildings with applications to algebra,* Memoirs of AMS, **192** (2008).

[32] M. Kapovich and J. Millson, *A path model for geodesics in Euclidean buildings and its applications to representation theory*, Groups, Geometry and Dynamics, vol **2**, 2008, p. 405–480.

[33] N. Katz, *Rigid Local systems*, Annals of Mathematics Studies, vol. 139, Princeton University Press, Princeton, NJ, 1996.

[34] S.L. Kleiman, *Transversality of the general translate*, Compos. Math. **28** (1973), 287-297.

[35] A. Klyachko, *Stable bundles, representation theory and Hermitian operators*, Selecta Mathematica **4** (1998), p. 419–445.

[36] A. Knutson and T. Tao, *The honeycomb model of* $\mathrm{GL}_n(\mathbb{C})$ *tensor products. I. Proof of the saturation conjecture,* J. Amer. Math. Soc., vol. **12** (1999), no. 4, p. 1055–1090.

[37] A. Knutson, T. Tao and C. Woodward, *The honeycomb model of* $\mathrm{GL}_n(\mathbb{C})$ *tensor products. II. Puzzles determine the facets of the Littelwood-Richardson cone,* J. Amer. Math. Soc. **17** (2004), p. 19–48.

[38] A. Knutson, *The symplectic and algebraic geometry of Horn's problem* Special Issue: Workshop on Geometric and Combinatorial Methods in the Hermitian Sum Spectral Problem (Coimbra, 1999). Linear Algebra Appl. **319** (2000), no. 1-3, 61–81.

[39] M. Kontsevich and Yu. Manin, *Gromov-Witten classes, quantum cohomology and enumerative geometry,* Comm. Math. Phys. **164(3)**:525–562, 1994.

[40] A. Marian and D. Oprea, *The level rank duality for non-abelian theta functions*, Invent. Math., **168** (2007), 225–247.

[41] V. Mehta and C.S. Seshadri, *Moduli of vector bundles on curves with parabolic structures*, Math. Ann. **248** (1980), no 3, 205–239.

[42] E. Mukhin, V. Tarasov and A. Varchenko, *Schubert calculus and representations of the general linear group*, J. Amer. Math. Soc. **22** (2009) 909–940.

[43] S. Naculich and H. Schnitzer. Duality relations between $\mathrm{SU}(N)_k$ and $\mathrm{SU}(k)_N$ WZW models and their braid matrices, Phys. Lett. B **244** (1990), no. 2, 235–240.

[44] T. Nakanishi and A. Tsuchiya, *Level-rank duality of WZW models in conformal field theory,* Comm. Math. Phys. **144** (1992), no. 2, 351–372.

[45] M. S. Narasimhan and C. S. Seshadri, *Stable and unitary vector bundles on a compact Riemann surface,* Ann. of Math., **82** (1965) 540–64.

[46] C. Pauly, *Espaces de modules de fibrés paraboliques et blocs conformes*, Duke Math. J. **84** (1996), no. 1, 217–235.

[47] C. Pauly La Dualité Étrange, Séminaire Bourbaki, **994**, June 2008.

[48] M. Popa, *Generalized theta linear series on moduli spaces of vector bundles on curves* Notes from the Cologne Summer School, August 2006, to appear in the Handbook of Moduli.

[49] K. Purbhoo, *A Vanishing and a nonvanishing condition for Schubert Calculus on* G/B, Int. Math. Res. Not. **2006**, Art. ID 24590, 38 pages.

[50] K. Purbhoo and F. Sottile, *The recursive nature of cominuscule Schubert calculus, with Frank Sottile*, Adv. Math., **217** (2008), 1962-2004.

[51] T. R. Ramadas, *The Harder-Narasimhan trace and unitarity of the KZ/Hitchin connection: genus 0*, Ann. of Math. Vol. **169** (2009), No. 1, 1–39.

[52] N. Ressayre, *Geometric Invariant Theory and the Generalized Eigenvalue Problem,* Invent. Math., to appear.

[53] E. Richmond, *A partial Horn recursion in the cohomology of flag varieties*, J. of Alg. Comb (2009) **30**: 1–17.

[54] A. N. Schellekens and N. P. Warner, *Conformal subalgebras of Kac-Moody algebras*, Phys. Rev. D (3) **34** (1986) 3092–3096.

[55] C. Sorger, *La formule de Verlinde,* Séminaire Bourbaki **794** (1994).

[56] H. Tamvakis, *The connection between representation theory and Schubert calculus*, Enseign. Math. (2) **50** (2004), no. 3–4, 267–286.

[57] A. Tsuchiya, K. Ueno and Y. Yamada, *Conformal field theory on universal family of stable curves with gauge symmetries*, Integrable systems in quantum field theory and statistical mechanics, 459–566, Adv. Stud. Pure Math., **19**, Academic Press, Boston, MA, 1989.

[58] A. Varchenko, Special functions, KZ type equations, and representation theory, CBMS Regional Conference Series in Mathematics, **98**, Providence, RI, 2003. viii+118 pp.

[59] E. Witten, *The Verlinde algebra and the cohomology of the Grassmannian*, In: "Geometry, topology and physics," p. 357–422, Conf. Proc. Lecture Notes Geom. Topology, IV, Internat. Press, Cambridge, MA, 1995.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Boundedness Results in Birational Geometry

Christopher D. Hacon* and James McKernan†

Abstract

We survey results related to pluricanonical maps of complex projective varieties of general type.

Mathematics Subject Classification (2010). Primary 14E05; Secondary 14J40

Keywords. Pluricanonical map, boundedness, minimal model program.

1. Introduction

Let X be a smooth complex projective variety of dimension n. In order to study the geometry of X one would like to choose a natural embedding $X \subset \mathbb{P}^N_{\mathbb{C}}$. This is equivalent to the choice of a very ample line bundle $\mathcal{L}$ on X i.e. of a line bundle $\mathcal{L}$ such that its sections define an embedding

$$\phi_{\mathcal{L}} : X \hookrightarrow \mathbb{P}H^0(X, \mathcal{L}).$$

(If $s_0, \ldots, s_N$ is a basis of $H^0(X, \mathcal{L})$, then we let $\phi_{\mathcal{L}}(x) = [s_0(x) : \ldots : s_N(x)]$.) Conversely, given an embedding $\phi : X \hookrightarrow \mathbb{P}^N_{\mathbb{C}}$, we have that $\mathcal{L} := \phi^*\mathcal{O}_{\mathbb{P}^N}(1)$ is a very ample line bundle on X. Since any projective variety X may have many different embeddings in $\mathbb{P}^N$ it is important to find a "natural" choice of this embedding (or equivalently a natural choice of a very ample line bundle).

*Department of Mathematics, University of Utah, 155 South 1400 East, JWB 233, Salt Lake City, UT 84112, USA. E-mail: hacon@math.utah.edu.

†Department of Mathematics, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. E-mail: mckernan@math.mit.edu.

[0]Christopher Hacon was partially supported by NSF Grant 0757897. James McKernan was partially supported by NSF Grant 0701101 and by the Clay Mathematics Institute. The authors would like to thank C. Xu for many useful discussions and comments.

The only known natural choice is the canonical bundle $\omega_X = \wedge^n T_X^\vee$ and its tensor powers $\omega_X^{\otimes m}$ for $m \in \mathbb{Z}$.

When $\dim X = 1$, we have that ω_X is a line bundle of degree $\deg \omega_X = 2g-2$ where g denotes the genus of X so that $\deg \omega_X > 0$ if and only if $g \geq 2$. There exist curves with genus $g \geq 2$ such that ω_X is not very ample, however we have the following classical result.

Theorem 1.1. *If X is a curve of genus $g \geq 2$, then $\omega_X^{\otimes m}$ is very ample for any integer $m \geq 3$.*

Proof. Let $\phi_m = \phi_{\omega_X^{\otimes m}}$. In order to show the theorem, we must show that ϕ_m is a morphism and separates points and tangent directions. This is equivalent to showing (cf. [Hartshorne77, II.7.3, IV.3.1]) that

1. $h^0(X, \omega_X^{\otimes m}(-P)) = h^0(X, \omega_X^{\otimes m}) - 1$ for any $P \in X$, and
2. $h^0(X, \omega_X^{\otimes m}(-P-Q)) = h^0(X, \omega_X^{\otimes m}) - 2$ for any points P and Q on X.

Considering the short exact sequence of coherent sheaves on X

$$0 \to \omega_X^{\otimes m}(-P) \to \omega_X^{\otimes m} \to \mathbb{C}_P \to 0$$

(where the last homomorphism is given by evaluating sections at P) we obtain a short exact sequence of vector spaces over $\mathbb{C}$

$$0 \to H^0(X, \omega_X^{\otimes m}(-P)) \to H^0(X, \omega_X^{\otimes m}) \to \mathbb{C} \to H^1(X, \omega_X^{\otimes m}(-P)) \dots$$

Since $\deg \omega_X^{\otimes(1-m)}(P) = (1-m)(2g-2) - 1 < 0$, we have that

$$H^1(X, \omega_X^{\otimes m}(-P)) \cong H^0(X, \omega_X^{\otimes(1-m)}(P))^\vee = 0,$$

and so the homomorphism $H^0(X, \omega_X^{\otimes m}) \to \mathbb{C}$ is surjective (or equivalently $h^0(X, \omega_X^{\otimes m}(-P)) = h^0(X, \omega_X^{\otimes m}) - 1$). Therefore ϕ_m is a morphism.

The proof that ϕ_m separates points and tangent directions is similar. □

Remark 1.2. *Note that:*

1. *If $\mathcal{L}$ is a line bundle, then $\mathcal{L}(-P)$ denotes the coherent sheaf of sections of $\mathcal{L}$ vanishing at P. Since $\dim X = 1$ this is also a line bundle.*
2. *$h^0(X, \mathcal{L})$ denotes the dimension of the $\mathbb{C}$-vector space $H^0(X, \mathcal{L})$.*
3. *The isomorphism $H^1(X, \omega_X^{\otimes m}(-P)) \cong H^0(X, \omega_X^{\otimes(1-m)}(P))^\vee$ is implied by Serre Duality: If X is a smooth projective variety of dimension n and F is locally free, then $H^i(X, F) \cong H^{n-i}(X, \omega_X \otimes F^\vee)^\vee$.*
4. *The vanishing $H^1(X, \omega_X^{\otimes m}(-P))$ is also implied by Kodaira vanishing which says that if X is a smooth projective variety and $\mathcal{L}$ is an ample line bundle, then $H^i(X, \omega_X \otimes \mathcal{L}) = 0$ for all $i > 0$.*

It follows that we can hope to use some *multiple* of ω_X to study the geometry of *most* varieties. In dimension ≥ 2 the situation is further complicated by the fact that there exist (smooth) birational varieties which are not isomorphic. For example if $P \in X$ is a point on a smooth projective surface (i.e. $\dim X = 2$), then one can construct a new surface $X' = \mathrm{Bl}_P X$ the *blow up* of X at P such that there is a morphism $f : X' \to X$ which is an isomorphism over the complement of P and whose fiber over P is a curve $E \subset X'$ which is isomorphic to $\mathbb{P}^1 \cong \mathbb{P}(T_x X)$. It is easy to see that $E \cdot E = -1$ and (since $\omega_{X'} = f^*\omega_X \otimes \mathcal{O}_X(E)$) that $\omega_{X'} \cdot E = -1$. Therefore, E is known as a -1-*curve.*

Consider now the example of a quintic surface $X \subset \mathbb{P}^3$ which is a smooth surface defined by a homogeneous polynomial of degree 5 in $\mathbb{C}[x_0, \ldots, x_3]$. By adjunction, one has $\omega_X \cong \omega_{\mathbb{P}^3}(X) \otimes \mathcal{O}_X \cong \mathcal{O}_{\mathbb{P}^3}(1)|_X$ so that ω_X is very ample (and $\phi_1 : X \hookrightarrow \mathbb{P}^3$ coincides with the given embedding). However, if $f : X' \to X$ is the blow up of X at a point $P \in X$ and E is the exceptional curve, then $\omega_{X'} \cdot E = -1$. Therefore, for any $m > 0$, sections of $H^0(X', \omega_{X'}^{\otimes m})$ must vanish along E and so ϕ_m is not a morphism along points of E. If we remove the singularities of ϕ_m (or more precisely we subtract the fixed divisor mE of $\omega_{X'}^{\otimes m}$), we obtain a morphism $\phi_{\omega_{X'}^{\otimes m}(-mE)} : X' \to \mathbb{P}^3$ whose image is X.

Therefore in dimension ≥ 2, we can not expect that, for most varieties, multiples of ω_X define an embedding in projective space. We can only hope that for most varieties, multiples of ω_X define a birational map (i.e. there is an open subset of X on which the given map is an embedding).

We have the following definition.

Definition 1.3. *Let X be a smooth projective variety, then X is of general type if the sections of $\omega_X^{\otimes m}$ define a birational map for some $m > 0$.*

It is known that if X is of general type, then in fact the sections of $\omega_X^{\otimes m}$ define a birational map for *all* sufficiently big integers $m > 0$.

When $\dim X = 2$ (and X is of general type), it is known by a result of Bombieri (cf. [Bombieri70]) that:

Theorem 1.4. *If X is a surface of general type, then ϕ_m is birational for all $m \geq 5$.*

In fact we have that (after subtracting the fixed divisor) $\phi_m : X \to \mathbb{P}^N$ is a morphism whose image X_{can} is uniquely determined by $X_{\mathrm{can}} \cong \mathrm{Proj}\, R(\omega_X)$ where $R(\omega_X) = \oplus_{m \geq 0} H^0(X, \omega_X^{\otimes m})$ is the canonical ring. Note that X_{can} has rational double point singularities so that $\omega_{X_{\mathrm{can}}}$ is a line bundle. We have $\omega_X = \phi_m^* \omega_{X_{\mathrm{can}}} \otimes \mathcal{O}_X(E)$ for some effective exceptional divisor E or equivalently $\omega_X^{\otimes m} \otimes \mathcal{O}_X(-mE) \cong \phi_m^* \mathcal{O}_{\mathbb{P}^N}(1)$.

Since X_{can} may be singular, it is convenient to consider the minimal desingularization $X_{\min} \to X_{\mathrm{can}}$. For surfaces of general type, the minimal model is uniquely determined. It can also be obtained from X by contracting all -1 curves. Therefore there is a morphism $X \to X_{\min}$. It is known that $\omega_{X_{\min}}^{\otimes m}$ is

base point free for all $m \geq 5$ (in fact ϕ_m defines the morphism $X_{\min} \to X_{\text{can}}$ and $\phi_m^* \omega_{X_{\text{can}}} = \omega_{X_{\min}}$).

By Riemann-Roch and (a generalization of) Kodaira vanishing, we have that for all $m \geq 2$

$$h^0(\omega_X^{\otimes m}) = h^0(\omega_{X_{\min}}^{\otimes m}) = \frac{m(m-1)}{2} K_{X_{\min}}^2 + \chi(\mathcal{O}_{X_{\min}})$$

where $K_{X_{\min}}^2 \in \mathbb{Z}_{>0}$ is the self intersection of the canonical divisor $K_{X_{\min}}$ (a divisor corresponding to the zeroes of a section of $\omega_{X_{\min}}$). Note that as X is of general type

$$\chi(\mathcal{O}_X) = \chi(\mathcal{O}_{X_{\min}}) = \sum (-1)^i h^i(\mathcal{O}_{X_{\min}}) > 0.$$

In particular we have that for all $m \geq 2$

$$P_m(X) := h^0(\omega_X^{\otimes m}) > \frac{m(m-1)}{2} K_{X_{\min}}^2.$$

One important consequence of the above results is that X_{can} is a subvariety of $\mathbb{P}^{10K_{X_{\min}}^2 + \chi(\mathcal{O}_{X_{\min}}) - 1}$ of degree $25K_{X_{\min}}^2$. It follows by a Hilbert scheme type argument that there exists a parameter space for canonical (and hence also for minimal) surfaces of general type:

Theorem 1.5. *Let $M \in \mathbb{Z}_{>0}$. There exists a morphism $\mathcal{X} \to S$ such that for any $s \in S$, the fiber $\mathcal{X}_s$ is a canonical surface of general type and for any canonical surface of general type X such that $K_X^2 \leq M$, there is a point $s \in S$ and an isomorphism $X \cong \mathcal{X}_s$.*

Remark 1.6. *The moduli space for minimal complex projective surfaces of general type was constructed in [Gieseker77].*

It is then important to generalize (1.4) to higher dimensions. Even though many of the features of the classification of surfaces of general type were shown to hold for threefolds in the 80's (cf. [Kolláretal92]), the generalization of (1.4) turned out to be more difficult than expected and was only completed in [Tsuji07], [HM06] and [Takayama06]. One of the difficulties encountered, is that in dimension ≥ 3 even though minimal models $X_{\min}$ are known to exist (but are not uniquely determined cf. [BCHM09]), they have mild (terminal) singularities and so $K_{X_{\min}}^{\dim X}$ is a positive rational number. In fact the threefold X_{46} given by a degree 46 hypersurface in weighted projective space $\mathbb{P}(4,5,6,7,23)$, satisfies $K_{X_{\min}}^3 = 1/420$ and ϕ_m is birational if and only if $m = 23$ or $m \geq 27$ (cf. [Iano-Fletcher00]). A further complication is given by the fact that we have little control over other terms of the Riemann-Roch formula for multiples of the canonical bundle (however see Section 2.1 for the 3-fold case). In particular we do not control $\chi(\mathcal{O}_X)$. (This should be contrasted with the above mentioned results for surfaces: $K_{X_{\min}}^2 \geq 1$ and $\chi(\mathcal{O}_X) \geq 1$.)

Using ideas of Tsuji, the following result was proven in [HM06], [Takayama06] and [Tsuji07].

Theorem 1.7. *For any positive integer n, there exists an integer r_n such that if X is a smooth variety of general type and dimension n, then $\phi_r : X \dashrightarrow \mathbb{P}(H^0(X, \omega_X^{\otimes r}))$ is birational for all $r \geq r_n$.*

In fact it turns out that proving the above result is equivalent to showing that the volume

$$\mathrm{vol}(\omega_X) := \lim_{m\to\infty} \frac{n! h^0(\omega_X^{\otimes m})}{m^n}$$

is bounded from below by a positive constant v_n depending only on the dimension $n = \dim X$. We will discuss the ideas behind the proof of this result in Section 2.

Remark 1.8. *Notice that in characteristic $p > 0$ Theorem 1.7 is only known to hold in dimension ≤ 2.*

It should be observed that the proof (1.7) is not effective so that we are unable to compute r_n the minimum integer such that ϕ_r is birational for all n-dimensional varieties of general type and for all $r \geq r_n$.

Recently, effective results were proven for 3-folds of general type. In [Todorov07], it is shown that if $\mathrm{vol}(\omega_X)$ is sufficiently big, then ϕ_m is birational for all $m \geq 5$ (see [DiBiagio10] for related results in dimension 4). In [CC08], J. A. Chen and M. Chen show the following almost optimal result.

Theorem 1.9. *Let X be a smooth projective 3-fold of general type, then ϕ_r is birational for all $r \geq 77$.*

Their proof is based on a detailed analysis of Reid's exact plurigenera formula for threefolds (see also [CC08b], [Zhu09a], [Zhu09b] for related results). In higher dimensions the situation is more complicated and effective results are not known.

Naturally, one may ask whether similar results are known for varieties not of general type. Recall that by definition the Kodaira dimension of a complex projective variety X is given by

$$\kappa(X) = \max\{\dim \phi_m(X) | m \in \mathbb{Z}_{>0}\}.$$

Here we make the convention that if $h^0(\omega_X^{\otimes m}) = 0$ for all $m \in \mathbb{Z}_{>0}$, then $\kappa(X) = -1$ so that $\kappa(X) \in \{-1, 0, 1, \ldots, \dim X\}$. Note that in this case some authors define $\kappa(X) = -\infty$ (instead of $\kappa(X) = -1$) and some others simply say $\kappa(X) < 0$. With our convention we have $\kappa(X) = \mathrm{tr.deg.}_{\mathbb{C}} R(\omega_X) - 1$. (Note that by [BCHM09], the graded ring $R(\omega_X)$ is finitely generated.) Another equivalent definition is $\kappa(X) = \dim \mathrm{Proj}\, R(\omega_X)$. In fact, ϕ_r is birational to the Iitaka fibration and its image is birational to $\mathrm{Proj}\, R(\omega_X)$ for all sufficiently divisible integers $r > 0$. The natural conjecture is then:

Conjecture 1.10. *Fix $n \in \mathbb{Z}_{>0}$ and $\kappa \in \mathbb{Z}_{\geq 0}$. Then there exist a positive integer k_n depending only on n and κ such that for all smooth complex projective varieties of dimension $\dim X = n$ and Kodaira dimension $\kappa(X) = \kappa$, the image of ϕ_r is birational to $\mathrm{Proj}\, R(\omega_X)$ for all integers $r > 0$ divisible by k_n.*

By work of Fujino and Mori cf. [FM00], it is known that there exist positive integers m_1 and m_2 such that

$$R(K_X)^{(m_1)} \cong R(K_Z + B)^{(m_2)}$$

where (Z, B) is a klt pair of general type birational to $\operatorname{Proj} R(\omega_X)$ and for any positive integer m, $R^{(m)} = \oplus_{t \geq 0} R_{mt}$ is the m-th truncation of the graded ring $R = \oplus_{t \geq 0} R_t$. Therefore, this problem is closely related to the natural problem of studying pluricanonical maps for varieties of log general type. These issues will be discussed in Section 3.

Pluricanonical maps for varieties of log general type also arise when studying the automorphism groups of varieties of general type. We now illustrate this in dimension 1.

Theorem 1.11 (Hurwitz). *Let X be a curve of genus $g \geq 2$ with automorphism group G. Then $|G| \leq 84(g-1)$.*

Proof. Let $f : X \to Y = X/G$ be the induced morphism, then

$$K_X = f^* \left(K_Y + \sum \left(1 - \frac{1}{n_i} \right) P_i \right)$$

where n_i is the order of ramification of f over P_i. We have

$$2(g-1) = \deg K_X = |G| \cdot \deg \left(K_Y + \sum \left(1 - \frac{1}{n_i} \right) P_i \right).$$

Therefore, the theorem follows since by (1.12), we have

$$\deg \left(K_Y + \sum \left(1 - \frac{1}{n_i} \right) P_i \right) \geq \frac{1}{42}. \qquad \square$$

Theorem 1.12. *Let $\mathcal{A} \subset [0,1]$ be a DCC set (so that any non-increasing sequence $a_i \in \mathcal{A}$ is eventually constant). Then*

$$\mathcal{V} := \{2g - 2 + \sum d_i | g \in \mathbb{Z}_{\geq 0},\ d_i \in \mathcal{A}\} \cap (0,1]$$

is a DCC set and in particular there is a minimal element $v_0 \in \mathcal{V}$.

If $\mathcal{A} = \{1 - \frac{1}{m} | m \in \mathbb{Z}_{>0}\}$, then $v_0 = \frac{1}{42}$.

The proof is elementary, but we recall it for the convenience of the reader.

Proof. We may assume that $g \in \{0,1\}$. It is easy to see that the set $\mathcal{A}_+ = \{\sum a_i | a_i \in \mathcal{A}\} \cap [0,1]$ is also a DCC set and hence so is $\mathcal{V}$.

If $\mathcal{A} = \{1 - \frac{1}{m} | m \in \mathbb{Z}_{>0}\}$ then $v_0 = \sum a_i + 2g - 2$ where $g \in \{0,1\}$ and $a_i = 1 - \frac{1}{n_i}$ for some $n_i \in \mathbb{Z}_{>0}$. If $g = 0$, $a_1 = 1 - \frac{1}{2}$, $a_2 = 1 - \frac{1}{3}$ and $a_3 = 1 - \frac{1}{7}$, then $v_0 = \frac{1}{42}$.

If $g = 1$, then $\sum a_i \geq \frac{1}{2}$. Therefore, we may assume that $g = 0$. In this case $v_0 = \sum_{i=1}^{t} a_i - 2$. Since $1 \geq a_i = 1 - \frac{1}{n_i} \geq \frac{1}{2}$, we may assume $t \in \{3, 4\}$. Let $2 \leq n_1 \leq n_2 \leq \ldots$. If $t = 4$, then as $v_0 > 0$, we have $n_4 \geq 3$ and hence $v_0 = 2 - \sum \frac{1}{n_i} \geq \frac{1}{6}$. If $t = 3$, then $v_0 = 1 - \sum \frac{1}{n_i}$. If $n_1 > 3$, then $v_0 \geq \frac{1}{4}$. If $n_1 = 3$, as $v_0 > 0$, we have $n_3 \geq 4$ and hence $v_0 \geq \frac{1}{12}$. If $n_1 = 2$ and $n_2 \geq 5$, then $v_0 \geq \frac{1}{10}$. If $n_1 = 2$ and $n_2 = 4$, then as $v_0 > 0$, $n_3 \geq 5$ and so $v_0 \geq \frac{1}{20}$. If $n_1 = 2$ and $n_2 = 3$, then as $v_0 > 0$, $n_3 \geq 7$ and so $v_0 \geq \frac{1}{42}$. Finally, if $n_1 = n_2 = 2$, then $v_0 < 0$. □

One expects results similar to (1.11) to hold for automorphism groups of varieties of general type (regardless of their dimension). Results in this direction will be discussed in Section 3.1.

Another reason to be interested in pluricanonical maps for varieties of log general type is that they naturally arise when studying moduli spaces of canonically polarized varieties of general type cf. Section 3.3 and open varieties cf. Section 3.4.

At the opposite end of the spectrum, we have varieties with $\kappa(X) < 0$. From the point of view of the minimal model program, the typical representatives of this class of varieties are Fano varieties. For these varieties we have that $\omega_X^\vee$ is ample. Therefore, we consider the maps induced by sections of $\omega_X^{\otimes m}$ for $m \in \mathbb{Z}_{<0}$. The geometry of Fano varieties is briefly discussed in Section 3.5.

2. Varieties of General Type

In this section we will explain the main ideas behind the proof of (1.7). Our goal is to show that if X is an n-dimensional projective variety of general type, then ϕ_r is birational for all $r \gg r_n$. To this end, it suffices then to show that there exists a subset $X^0 \subset X$ given by the complement of countably many closed subsets of X such that ϕ_r is defined at points of X^0 and ϕ_r separates any two distinct points $x, y \in X^0$. The first major reduction in the proof of (1.7) is to show the following.

Proposition 2.1. *In order to prove* (1.7) *it suffices to show that there exist positive constants A and B (depending only on n) such that for any integer*

$$r \geq \frac{A}{\mathrm{vol}(\omega_X)^{1/n}} + B,$$

the rational map ϕ_r is birational.

Proof. If $\mathrm{vol}(\omega_X) \geq 1$, then the assertion is clear as ϕ_r is birational for all $r \geq A + B$. We may therefore assume that $\mathrm{vol}(\omega_X) < 1$. Let r_0 be the smallest

integer such that ϕ_{r_0} is birational, then

$$1 \leq \deg \overline{\phi_{r_0}(X)} \leq \mathrm{vol}(\omega_X^{\otimes r_0}) = r_0^n \mathrm{vol}(\omega_X) \leq$$

$$\left(\frac{A}{\mathrm{vol}(\omega_X)^{1/n}} + B + 1\right)^n \mathrm{vol}(\omega_X) < (A+B+1)^n.$$

It follows that the degree of the closure of $\phi_{r_0}(X)$ is bounded. Therefore, by a Hilbert scheme type argument, there is a projective morphism of quasi-projective varieties $f : \mathcal{X} \to S$ such that if X is any smooth n-dimensional complex projective variety with $0 < \mathrm{vol}(\omega_X) < 1$, then there exists a point $s \in S$ such that X is birational to the fiber $\mathcal{X}_s$. By Noetherian induction, possibly replacing S by a union of locally closed subsets, we may assume that f is smooth and S is irreducible. Let $\eta = \mathrm{Spec}(K)$ be the generic point of S and $\mathcal{X}_K$ be the generic fiber. Then there exists r_η such that $\phi_{\omega_{\mathcal{X}_K}^{\otimes r}}$ is birational for all $r_\eta \leq r \leq 2r_\eta$ (and hence for all $r \geq r_\eta$). It then follows that there exists an open subset S^0 of S such that $\phi_{\omega_{\mathcal{X}_t}^{\otimes r}}$ is birational for all $t \in S^0$ and all $r_\eta \leq r \leq 2r_\eta$ (and hence for all $r \geq r_\eta$).

By Noetherian induction, there is an integer r_S such that $\phi_{\omega_{\mathcal{X}_t}^{\otimes r}}$ is birational for all $t \in S$ and all $r \geq r_S$. □

Remark 2.2. *By the above discussion,* (1.7) *implies that for any* $n \in \mathbb{Z}_{>0}$*, there exist a positive constant* $v_n > 0$ *such that if* X *is a projective variety of general type and* $\dim X = n$*, then* $\mathrm{vol}(\omega_X) \geq v_n$*.*

In order to show that a rational map ϕ_r is birational, we would like to imitate the proof of the curve case of this theorem cf. (1.1) and show that the evaluation map

$$H^0(X, \omega_X^{\otimes r}) \to \mathbb{C}_x \oplus \mathbb{C}_y$$

at very general points $x, y \in X$ is surjective. The problem is that in higher dimensions it is very hard to ensure that cohomology groups of the form $H^1(X, \omega_X^{\otimes r} \otimes m_x \otimes m_y)$ vanish (here m_x denotes the maximal ideal of $x \in X$). In order to achieve this, the usual strategy is to use a far reaching generalization of Kodaira vanishing known as Kawamata-Viehweg vanishing or Nadel vanishing. Recall the following:

Theorem 2.3 (Nadel vanishing). *Let* X *be a smooth complex projective variety,* $\mathcal{L}$ *a line bundle on* X *and* D *a* $\mathbb{Q}$*-divisor such that* $\mathcal{L}(-D)$ *is nef and big. Then* $H^i(X, \omega_X \otimes \mathcal{L} \otimes \mathcal{J}(D)) = 0$ *for all* $i > 0$*.*

Remark 2.4. *Recall that a line bundle is nef if* $\deg(\mathcal{L}|_C) \geq 0$ *for any curve* $C \subset X$*. In this case,* $\mathcal{L}$ *is big if and only if* $\mathcal{L}^{\dim X} > 0$*. These definitions readily extend to* $\mathbb{Q}$*-divisors.*

Remark 2.5. *Recall that if* $D \subset X$ *is a* $\mathbb{Q}$*-divisor, then the multiplier ideal* $\mathcal{J}(D) \subset \mathcal{O}_X$ *is defined as follows. Let* $f : Y \to X$ *be a log resolution so that* f

is a projective birational morphism, Y is smooth, $\mathrm{Exc}(f)$ *and* $\mathrm{Exc}(f) \cup f_*^{-1}D$ *are divisors with simple normal crossings support. Then*

$$\mathcal{J}(D) := f_*\mathcal{O}_Y(K_{Y/X} - \llcorner f^*D \lrcorner).$$

It is well known that $\mathcal{J}(D)$ is trivial at points $x \in X$ where $\mathrm{mult}_x(D) < 1$ *and $m_x \subset \mathcal{J}(D)$ if* $\mathrm{mult}_x(D) \geq \dim X$. *The interested reader can consult [Lazarsfeld05] for a clear and comprehensive treatment of the properties of multiplier ideal sheaves.*

Using Nadel vanishing we obtain the following.

Proposition 2.6. *In order to prove* (1.7) *it suffices to show that there exists positive constants A and B (depending only on n) such that for any two distinct very general points $x, y \in X$ there is a $\mathbb{Q}$-divisor $D_{x,y}$ such that*

1. $D_{x,y} \sim \lambda K_X$ *where* $\lambda < \frac{A}{\mathrm{vol}(\omega_X)^{1/n}} + B - 1$;
2. *x is an isolated point of the co-support of $\mathcal{J}(D_{x,y})$ and y is contained in the co-support of $\mathcal{J}(D_{x,y})$.*

Proof. Let $r \geq \frac{A}{\mathrm{vol}(\omega_X)^{1/n}} + B$ be any integer. By (2.1), it suffices to show that ϕ_r is birational.

Since ω_X is big, there exists an integer $m > 0$, an ample divisor H and an effective divisor $G \geq 0$ such that $mK_X \sim G + H$. We may assume that x, y are not contained in the support of G. We let $D'_{x,y} = D_{x,y} + \frac{r-1-\lambda}{m}G$. Then $(r-1)K_X - D'_{x,y} \sim_{\mathbb{Q}} \frac{r-1-\lambda}{m}H$ is ample so that by (2.3) $H^1(X, \omega_X^{\otimes r} \otimes \mathcal{J}(D'_{x,y})) = 0$.

Consider the short exact sequence of coherent sheaves on X

$$0 \to \omega_X^{\otimes r} \otimes \mathcal{J}(D'_{x,y}) \to \omega_X^{\otimes r} \to \mathcal{Q} \to 0$$

where $\mathcal{Q}$ denotes the corresponding quotient. Since, as observed above, $H^1(X, \omega_X^{\otimes r} \otimes \mathcal{J}(D'_{x,y})) = 0$, the homomorphism

$$H^0(X, \omega_X^{\otimes r}) \to H^0(X, \mathcal{Q})$$

is surjective. Since x is an isolated point in the co-support of $\mathcal{J}(D'_{x,y})$, $\mathbb{C}_x$ is a summand of $\mathcal{Q}$. Since y is also contained in the support of $\mathcal{Q}$, we may find a section $s \in H^0(X, \omega_X^{\otimes r})$ vanishing at y but not at x. Since x and y are very general points on X, by symmetry we may also find a section $t \in H^0(X, \omega_X^{\otimes r})$ vanishing at x but not at y. It follows that the evaluation map

$$H^0(X, \omega_X^{\otimes r}) \to \mathbb{C}_x \oplus \mathbb{C}_y$$

is surjective and hence ϕ_r is birational. $\square$

Proof of Theorem 1.7. By (2.6), it suffices to show that there exists positive constants A and B (depending only on n) such that for any two distinct very general points $x, y \in X$ there is a $\mathbb{Q}$-divisor $D_{x,y} \sim_{\mathbb{Q}} \lambda K_X$ where $\lambda < \frac{A}{\mathrm{vol}(\omega_X)^{1/n}} + B - 1$ such that x is an isolated point of the co-support of $\mathcal{J}(D_{x,y})$ and y is contained in the co-support of $\mathcal{J}(D_{x,y})$.

For ease of exposition, we will however just show that there is a $\mathbb{Q}$-divisor $D_x \sim_{\mathbb{Q}} \lambda K_X$ where $\lambda < \frac{A}{\mathrm{vol}(\omega_X)^{1/n}} + B - 1$ such that x is an isolated point of the co-support of $\mathcal{J}(D_x)$. The interested reader can consult [Tsuji07] or [Takayama06] for the remaining details or [HM06] for an alternative argument.

We will also assume that ω_X is ample. This can be achieved replacing X by its canonical model. Of course X is no longer smooth, but it has mild (canonical) singularities and the proof goes through with minor changes.

We will proceed by induction on the dimension and hence we may assume that (1.7) holds for varieties of dimension $\leq n-1$. Note that by (1.1), the theorem holds when $n = 1$. We will not keep careful track of the various constants and so we will say that $\lambda = O(\mathrm{vol}(\omega_X)^{-1/n})$ (instead of $\lambda < \frac{A}{\mathrm{vol}(\omega_X)^{1/n}} + B - 1$).

Since

$$h^0(\mathcal{O}_X(mK_X)) = \frac{\mathrm{vol}(\omega_X)}{n!} m^n + O(m^{n-1})$$

and since vanishing to order k at a smooth point $x \in X$ imposes at most $k^n/n! + O(k^{n-1})$ conditions, by an easy calculation it follows that for any smooth point $x \in X$, we may find $m \gg 0$ and a $\mathbb{Q}$-divisor $D_x^m \sim mK_X$ such that $\mathrm{mult}_x(D_x^m) > \frac{m}{2}\mathrm{vol}(\omega_X)^{1/n}$. Note that if we assume that $x \in X$ is a very general point, then we can assume that the integer m is independent of the point x. Let τ be defined by

$$\tau = \sup\{t \geq 0 | m_x \subset \mathcal{J}(X, tD_x^m)\}.$$

By (2.5), $\tau < \frac{2n}{m \cdot \mathrm{vol}(\omega_X)^{1/n}}$. Note that if $D_x := \tau D_x^m$, then $m_x \subset \mathcal{J}(X, D_x)$ and $D_x \sim \lambda K_X$ where $\lambda \leq \frac{2n}{\mathrm{vol}(\omega_X)^{1/n}}$ so that $\lambda = O(\mathrm{vol}(\omega_X)^{-1/n})$.

By a standard perturbation technique, we may assume that on a neighborhood of $x \in X$ there is a unique irreducible subvariety V_x contained in the co-support of $\mathcal{J}(D_x)$. (More precisely, if $f : Y \to X$ is a log resolution of (X, D_x), we may assume that there is a unique divisor $E \subset Y$ such that $\mathrm{mult}_E(K_{Y/X} - f^*D_x) = -1$ and $E \cap f^{-1}(x) \neq \emptyset$. V_x is then the center of E on X.) The problem is that we may have $\dim V_x > 0$. The idea is to then use the techniques of [AS95] to "cut down" the cosupport of $\mathcal{J}(D_x)$ i.e. to reduce to the case when $\dim V_x = 0$. We will use the following result:

Proposition 2.7. *Let V_x and (X, D_x) be as above. If for a general point $x' \in V_x$ there exists a divisor $F_{x'}$ on X whose support does not contain V_x such that $\mathrm{mult}_{x'}(F_{x'}|_{V_x}) > \dim V_x$, then there exist rational numbers $0 < \alpha, \beta < 1$ such that $m_{x'} \subset \mathcal{J}(\alpha D_x + \beta F_{x'})$ and in a neighborhood of x', every component of the co-support of $\mathcal{J}(\alpha D_x + \beta F_{x'})$ has dimension less than $\dim V_x$.*

The established strategy to produce the $\mathbb{Q}$-divisor $F_{x'}$ is as follows:

1. produce a divisor $E_{x'}$ on V_x such that $\mathrm{mult}_{x'}(E_{x'}) > \dim V_x$, and then
2. lift this divisor to X, that is find a $\mathbb{Q}$-divisor $F_{x'} \sim_{\mathbb{Q}} \lambda' K_X$ such that $F_{x'}|_{V_x} = E_{x'}$ and $\lambda' = O(\mathrm{vol}(\omega_X)^{-1/n})$.

In order to complete the first step, we need to bound the volume of $\omega_X|_{V_x}$ from below. This is achieved by comparing $K_X + D_x$ with K_{V_x} via a result of Kawamata (cf. [Kawamata98]):

Theorem 2.8. *Let V_x and (X, D_x) be as above, and let A be an ample divisor. If $\nu : V_x^\nu \to V_x$ is the normalization, then for any rational number $\epsilon > 0$, there exists a $\mathbb{Q}$-divisor $\Delta_\epsilon \geq 0$ such that*

$$\nu^*(K_X + D_x + \epsilon A) \sim_{\mathbb{Q}} K_{V_x^\nu} + \Delta_\epsilon.$$

Remark 2.9. *Kawamata's Subadjunction Theorem says that if moreover V_x is a minimal non-klt center at a point $y \in V_x$, then (on a neighborhood of y) V is normal and we may assume that (V_x, Δ_ϵ) is klt.*

Since X is of general type and $x \in X$ is a very general point, it follows that V_x is also of general type. Let $n' = \dim V_x$ and $\mu : \tilde{V}_x \to V_x^\nu$ be a resolution of singularities. Assume for simplicity that V_x is normal. By our inductive hypothesis, for general $x' \in \tilde{V}_x$ there is a $\mathbb{Q}$-divisor $E_{x'} \sim_{\mathbb{Q}} \gamma K_{\tilde{V}_x}$ on $\tilde{V}_x$ with $\mathrm{mult}_{x'}(E_{x'}) > n'$ and $0 < \gamma < n/v_{n'}$ so that $\gamma = O(1)$ (for the definition of $v_{n'}$ see (2.2)). Fix a rational number $0 < \epsilon \ll 1$ and let $A = K_X$. Pushing forward, we obtain a $\mathbb{Q}$-divisor

$$\nu_*(\mu_* E_{x'} + \gamma\Delta_\epsilon) \sim_{\mathbb{Q}} \gamma\nu_*(K_{V_x^\nu} + \Delta_\epsilon) \sim_{\mathbb{Q}} \gamma(1 + \lambda + \epsilon)K_X|_{V_x}$$

on V_x with $\mathrm{mult}_{x'}\nu_*(\mu_* E_{x'} + \gamma\Delta_\epsilon) > n'$.

Since we have assumed that K_X is ample, by Serre vanishing, the homomorphism

$$H^0(X, \mathcal{O}_X(mK_X)) \to H^0(X, \mathcal{O}_{V_x}(mK_X))$$

is surjective for all $m \gg 0$ and so there exists a $\mathbb{Q}$-divisor $F_{x'} \sim_{\mathbb{Q}} \gamma(1+\lambda+\epsilon)K_X$ such that $F_{x'}|_{V_x} = \nu_*(\mu_* E_{x'} + \gamma\Delta_\epsilon)$.

By (2.7), we then have that for some $0 < \alpha, \beta < 1$

1. $m_{x'} \subset \mathcal{J}(\alpha D_x + \beta F_{x'})$,
2. in a neighborhood of x', every component of the co-support of $\mathcal{J}(\alpha D_x + \beta F_{x'})$ has dimension $< \dim V_x$, and
3. $\alpha D_x + \beta F_{x'} \sim_{\mathbb{Q}} \lambda' K_X$ where $\lambda' = O(\mathrm{vol}(\omega_X)^{-1/n})$.

Repeating this procedure at most $n-1$ times, we may assume that for any very general point $x^* \in X$, there is a $\mathbb{Q}$-divisor $D^*_{x^*} \sim_{\mathbb{Q}} \lambda^* K_X$ such that x^* is an isolated point in the co-support of $\mathcal{J}(D^*_{x^*})$ and $\lambda^* = O(\mathrm{vol}(\omega_X)^{-1/n})$. $\square$

2.1. Reid's 3-fold exact plurigenera formula. In dimension 3, an almost optimal version of (1.7) can be obtained using Reid's 3-fold exact plurigenera formula.

Theorem 2.10. *Let X be a minimal 3-fold with terminal singularities, then*

$$\chi(\mathcal{O}_X(mK_X)) = \frac{1}{12}m(m-1)(2m-1)K_X^3 - (2m-1)\chi(\mathcal{O}_X) + l(m),$$

where the correction term $l(m)$ depends only on the (finitely many) singularities of X. More precisely, there is a finite set (basket) of pairs of integers $\mathcal{B}(X) = \{(b_i, r_i)\}$ where $0 < b_i < r_i$ are uniquely determined by the singularities of X such that

$$l(m) := \sum_{Q_i \in \mathcal{B}(X)} l_{Q_i}(m) := \sum_{Q_i \in \mathcal{B}(X)} \sum_{j=1}^{m-1} \frac{\overline{jb_i}(r_i - \overline{jb_i})}{2r_i},$$

where $\overline{x}$ denotes the smallest non-negative residue modulo r_i, so that, $\overline{x} := x - r_i\lfloor \frac{x}{r_i} \rfloor$.

When X is of general type, K_X is nef and big so that by Kawamata-Viehweg vanishing we have

$$P_m(X) := h^0(\mathcal{O}_X(mK_X)) = \chi(\mathcal{O}_X(mK_X)) \qquad \text{for all } m \geq 2.$$

One can therefore hope to use (2.10) to find values of m such that $P_m(X) \geq 1$ or $P_m(X) \geq 2$. If, for example $\chi(\mathcal{O}_X) \leq 0$, then since $l(m) \geq 0$ and $K_X^3 > 0$, we have $P_m(X) \geq 1$ for all $m \geq 2$.

More generally, it is not hard to see that if $P_m(X) = 0$ for some $m \geq 2$ and if $-\chi(\mathcal{O}_X)$ is bounded from below, then there are only finitely many possible baskets of singularities $\mathcal{B}(X)$. This implies that the index r of K_X (i.e. the smallest integer $r > 0$ such that rK_X is Cartier) is bounded from above. In turn this means that $K_X^3 \geq \frac{1}{r^3}$ and hence one obtains an integer m_0 such that $P_m(X) \geq 1$ for all $m \geq m_0$.

By a detailed study of the above Riemann-Roch formula, J.-A. Chen and M. Chen prove the following (cf. [CC08]).

Theorem 2.11. *Let X be a non-singular 3-fold of general type then*

1. $\mathrm{vol}(\omega_X) \geq \frac{1}{2660}$,
2. $P_{12}(X) \geq 1$,
3. $P_{24}(X) \geq 2$, *and*
4. ϕ_r *is birational for all* $r \geq 77$.

Remark 2.12. *The second and third inequalities are optimal.*

There exist examples with $\mathrm{vol}(\omega_X) = \frac{1}{420}$ *(cf. [Iano-Fletcher00]) and hence the first inequality is "almost optimal". By [CC08b], it is known that if* $\chi(\mathcal{O}_X) \leq 0$*, then* $\mathrm{vol}(\omega_X) \geq \frac{1}{30}$*. This inequality is optimal as shown by the example of a canonical hypersurface of degree* 28 *in the weighted projective space* $\mathbb{P}(1, 3, 4, 5, 14)$*.*

When $\chi(\mathcal{O}_X) = 1$*, it is known that* $\mathrm{vol}(\omega_X) \geq \frac{1}{420}$ *(cf. [Zhu09b]) and that* ϕ_r *is birational for all* $r \geq 46$ *(cf. [Zhu09a]).*

As mentioned in the introduction, there are examples where ϕ_{26} *is not birational and so the fourth inequality is also "almost optimal".*

Remark 2.13. *Using similar methods, in [CC08c] it is shown that if* X *is a terminal weak* $\mathbb{Q}$*-Fano* 3*-fold (so that* $-K_X$ *is nef and big), then* $h^0(\mathcal{O}_X(-6K_X)) > 0$*,* $h^0(\mathcal{O}_X(-8K_X)) > 1$ *and* $-K_X^3 \geq \frac{1}{330}$ *(which is the optimal possible lower bound).*

As mentioned above, the idea of using Reid's exact plurigenera formula in this context is not new (see for example [Iano-Fletcher00]). The main new insight of [CC08] is to use (2.10) for various values of m to prove the following inequality:

$$2P_5 + 3P_6 + P_8 + P_{10} + P_{12} \geq \chi(\mathcal{O}_X) + 10P_2 + 4P_3 + P_7 + P_{11} + P_{13}.$$

It follows that if $P_m = 0$ for $m \leq 12$, then $\chi(\mathcal{O}_X) \leq 0$ which as observed above is the well understood case.

The precise results obtained in [CC08] are then a consequence of a detailed study of the terms appearing in Reid's exact plurigenera formula.

3. Varieties of Log General Type

One would like to generalize Theorem 1.7 to the case of log canonical pairs. This is a natural question in its own right, but it is also motivated by the desire to study the geometry of open varieties, of varieties of intermediate Kodaira dimension, of the moduli spaces of varieties of general type, of the automorphism groups of varieties of general type and other related questions.

We start by considering the case of curves. Let (X, D) be a pair consisting of a smooth curve X and a $\mathbb{Q}$-divisor $D = \sum d_i D_i$ such that $K_X + D$ has general type. We ask the following:

Question 3.1. *Is there a lower bound for the volume of* $K_X + D$*?*

The answer in this case is simple:

$$\mathrm{vol}(K_X + D) = \deg(K_X + D) = 2g - 2 + \sum d_i > 0$$

where g denotes the genus of the curve X. If $g \geq 2$, then $\mathrm{vol}(K_X + D) \geq 2$, but if $g \leq 1$, one sees immediately that no such bound exists unless we impose

some restrictions on the possible values that d_i are allowed to take. The most natural answer was given in (1.12): *If $\mathcal{A} \subset [0,1]$ is a DCC set, then there exists a constant $v_0 > 0$ such that $2g - 2 + \sum d_i \geq v_0$ for any $g \in \mathbb{Z}_{\geq 0}$ and $d_i \in \mathcal{A}$.*

The most optimistic generalizations of Theorem 1.12 are the following two conjectures.

Conjecture 3.2. *Let $\mathcal{A} \subset (0,1]$ be a DCC set, $n \in \mathbb{Z}_{>0}$ and*

$$\mathcal{V} = \{\mathrm{vol}(K_X + D) | (X, D) \text{ is lc}, \ \dim X = n, \ D \in \mathcal{A}\}.$$

Then $\mathcal{V}$ is a DCC set.

Conjecture 3.3. *Let $\mathcal{A} \subset (0,1]$ be a DCC set and $n \in \mathbb{Z}_{>0}$. Then there exists a positive integer $N > 0$ such that if (X, D) is a lc pair of dimension n with $K_X + D$ big and $D \in \mathcal{A}$, then $|\llcorner m(K_X + D) \lrcorner|$ is birational for all $m \geq N$.*

Notice that the above conjectures were proven in dimension 2 by Alexeev and Alexeev-Mori (cf. [Alexeev94] and [AM04]).

At first sight one may hope to apply the techniques used in the proof of Theorem 1.7, however there are several problems that arise:

It is easy to produce a divisor $D_x \sim_{\mathbb{Q}} k(K_X + D)$ such that $m_x \subset \mathcal{J}(D_x)$ for very general $x \in X$ and $k = O(\mathrm{vol}(K_X + D)^{1/n})$. Assume for simplicity that there is an irreducible subvariety $V_x \subset X$ such that $\mathcal{J}(D_x) = \mathcal{I}_{V_x}$ on a neighborhood of $x \in X$. If $\dim V_x > 0$, we must bound $\mathrm{vol}((K_X + D)|_{V_x})$ from below. To this end, one applies Kawamata sub-adjunction

$$\nu^*(K_X + D_x + \epsilon A) = K_{V_x^\nu} + \Delta_\epsilon$$

where $\nu : V_x^\nu \to V_x$ is the normalization morphism, A is an ample line bundle and $0 < \epsilon \ll 1$.

In order to proceed by induction on the dimension, we must show that $K_{V_x^\nu} + \Delta_\epsilon$ satisfies the inductive hypothesis. This is problematic. Even if we ignore the dependence on ϵ (which is at least conjecturally a reasonable assumption), in order to control the coefficients of Δ_ϵ, we must control the coefficients of D_x. In higher dimension, there is no known strategy to accomplish this.

3.1. Automorphism groups of varieties of general type.

Let X be a variety of general type with automorphism group G, then it is known that G is finite. It is a natural question to find effective bounds on the order of G.

Naturally, one would hope to generalize Hurwitz's Theorem cf. (1.11) to higher dimensions. The natural conjecture is:

Conjecture 3.4. *For any $n \in \mathbb{Z}_{>0}$, there exists a constant $C > 0$ (depending only on n) such that if X is an n-dimensional variety of general type with automorphism group G, then*

$$|G| \leq C \cdot \mathrm{vol}(\omega_X).$$

Over the years there has been much interest in results related to the above conjecture; see for example [Andreotti50], [Corti91], [HS91], [Xiao94], [Xiao95], [Xiao96], [Szabo96], [CS96], [Ballico93] and [Cai00].

One would hope to use the ideas in the proof of (1.11) to attack Conjecture 3.4 in higher dimensions. We can still write

$$K_X = f^*\left(K_Y + \sum(1-\frac{1}{n_i})P_i\right)$$

where $f : X \to Y = X/G$ is the induced morphism and n_i is the order of ramification of f over P_i. We also have

$$\mathrm{vol}(\omega_X) = |G| \cdot \mathrm{vol}\left(K_Y + \sum(1-\frac{1}{n_i})P_i\right).$$

Therefore, a positive answer to Conjecture 3.2, would imply a positive answer to Conjecture 3.4.

Remark 3.5. *It is likely that proving that* $\mathrm{vol}(K_Y + \sum(1-\frac{1}{n_i})P_i)$ *is bounded from below, is substantially easier than Conjecture 3.2, and that this problem is even more accessible when* $(Y, \sum(1-\frac{1}{n_i})P_i)$ *arises as the quotient of a variety of general type by its automorphism group.*

3.2. Varieties of intermediate Kodaira dimension. Let X be a smooth projective variety of Kodaira dimension $0 \le \kappa(X) < \dim X$, then it is known that for all $m > 0$ sufficiently divisible $\phi_m : X \to Z$ defines a map birational to the Iitaka fibration so that $\dim Z = \kappa(X)$ and $\kappa(F) = 0$ where F is a general fiber of ϕ_m. (In fact Z is birational to $\mathrm{Proj}R(K_X)$.) When $\kappa(X) = 0$, $Z = \mathrm{Spec}(k)$ and there is an integer $N > 0$ such that $P_m(X) > 0$ if and only if m is divisible by N.

It is natural to conjecture the following:

Conjecture 3.6. *Fix positive integers* $0 \le \kappa < n$. *Then there exists an integer* $N > 0$ *(depending only on* κ *and* n*) such that if* X *is a smooth projective variety of dimension* n *and Kodaira dimension* κ*, and* $m > 0$ *is an integer divisible by* N*, then* ϕ_N *is birational to the Iitaka fibration.*

For surfaces, this conjecture is known to be true. In fact we have the following:

1. If $\kappa(X) = 0$ then $P_{12}(X) > 0$, and
2. if $\kappa(X) = 1$, then $P_{12}(X) > 0$ and $P_m(X) > 1$ for some $m \le 42$.

In dimension 3 the following results are known:

1. If $\kappa(X) = 0$ then by [Kawamata86] and [Morrison86]

$$P_{2^5\cdot 3^3\cdot 5^2\cdot 7\cdot 11\cdot 13\cdot 17\cdot 19}(X) > 0;$$

2. if $\kappa(X) = 1$, then by [FM00] there exists an explicit constant $N > 0$ (presumably far from optimal) such that ϕ_m is birational to the Iitaka fibration for all $m > 0$ divisible by N; and

3. if $\kappa(X) = 2$, then by [VZ09] and [Ringler07], there exists an explicit constant $N > 0$ such that ϕ_m is birational to the Iitaka fibration for all $m > 0$ divisible by N (in fact $m \geq 48$ and divisible by 12 suffices).

We now outline a strategy for proving Conjecture (3.6). Let X be a smooth projective variety of dimension n and Kodaira dimension $\kappa \geq 0$.

Step 1. By the minimal model program, it is expected that there is a minimal model $\phi : X \dashrightarrow X'$ (given by a finite sequence of flips and divisorial contractions) such that $R(K_X) \cong R(K_{X'})$, X' has terminal singularities and $K_{X'}$ is semiample. This means that for some $m_0 > 0$, the linear series $|m_0 K_{X'}|$ is base point free and it defines a morphism $f' : X' \to Z'$ which is birational to the Iitaka fibration of X. In particular $\dim Z = \kappa$ and $\kappa(F') = 0$ where F' is a very general fiber of f'. In fact we have $K_{F'} \sim_{\mathbb{Q}} 0$. Note this step requires the abundance conjecture.

Step 2. Using the ideas of Fujino and Mori [FM00], we write the "canonical bundle formula"

$$K_{X'} \sim_{\mathbb{Q}} f'^*(K_Z + B + M)$$

where the "boundary" part B is determined by the singularities of the morphism f' and the "moduli" part M is determined by the variation in moduli of the general fiber F'.

When f' is an elliptic fibration, then $M = \frac{1}{12} j^* \mathcal{O}_{\mathbb{P}^1}(1)$ where $j : Z \to \mathbb{P}^1$ is the j-function. In general, one expects M to be the pull-back of a big semiample $\mathbb{Q}$-divisor on a moduli scheme.

In order to make use of Fujino-Mori's canonical bundle formula, it is important to bound the denominators of the $\mathbb{Q}$-divisors B and M.

By [FM00, 3.1], there exists a positive integer $k = k(b, B_m) > 0$ such that kM is a divisor, where b is the smallest positive integer such that $P_b(F') > 0$, $m = n - \kappa$ and B_m is the m-th Betti number of a desingularization of the $\mathbb{Z}_m$-cover $E \to F'$ determined by the divisor in $|bK_{F'}|$. In fact we have $k = \operatorname{lcm}\{y \in \mathbb{Z}_{>0} | \phi(y) \leq B_m\}$ where ϕ is Euler's function.

The boundary part, B is defined as follows: Let P be a codimension 1 point on Z and let b_P be the supremum of $b \geq 0$ such that (X, bf^*P) is log canonical over the general point of P. We then set $B = \sum (1 - b_P)P$. Note that $b_P = 1$ for all but finitely many codimension 1 points P on Z. An interesting feature is that the coefficients of b_p are of the form $b - \frac{v}{ku}$ where $0 < v \leq bk$ cf. [FM00, 2.8].

The upshot is that if we control the invariants b and B_m of the general fiber F', then we can bound the denominators of B and M.

Step 3. Apply Conjecture (3.3) to conclude.

Remark 3.7. *Note that Step 2 depends on bounding k independently of the general fiber F'. If $m = 1$ then F' is a curve of genus 1 and hence $b = 1$ and $B_1 = 2$. If $m = 2$, then F is either an abelian, a K3, and Enriques or a bielliptic surface. As mentioned above, we have $b \leq 12$. Let $E \to F'$ be the corresponding cover. We again have $\kappa(E) = 0$ and hence $B_2(E) \leq 24$. If $m = 3$, by Kawamata's result mentioned above, there is a known bound for b, but there is no known bound for B_3. In higher dimensions, these questions are completely open.*

Remark 3.8. *One may make the analogous conjecture for log pairs. The case when $\dim X \leq 3$ and $\kappa(K_X + \Delta) = \dim X - 1$ is treated in [Todorov08]. The case where $\dim X \leq 4$ and $\kappa(K_X + \Delta) = \dim X - 2$ is treated in [TX08].*

3.3. Moduli spaces of varieties of general type. As we have remarked above, boundedness of varieties of general type is an essential ingredient in the proof of the existence of a moduli space for canonically polarized varieties of general type.

Recall that if X is a projective variety of general type, then its canonical model X_{can} is defined by $X_{\text{can}} := \text{Proj}R(K_X)$. X_{can} has canonical singularities (in particular $K_{X_{\text{can}}}$ is $\mathbb{Q}$-Cartier and $R(K_{X_{\text{can}}}) \cong R(K_X)$) and $K_{X_{\text{can}}}$ is ample. As a consequence of (1.7), we have:

Theorem 3.9. *For every $n, v \in \mathbb{Z}_{>0}$ then there exists a projective morphism of normal quasi-projective varieties $f : \mathcal{X} \to B$ such that any fiber $\mathcal{X}_b$ is a canonically polarized variety of general type with canonical singularities, and if X is a canonically polarized variety of general type with canonical singularities and $\text{vol}(\omega_X) \leq v$, then there exists $b \in B$ such that $X \cong \mathcal{X}_b$.*

Idea of the proof. By (1.7) and its proof, there exists a projective morphism of normal quasi-projective varieties $f : \mathcal{Z} \to B$ such that for any X as above, there exists $b \in B$ such that $\mathcal{Z}_b$ is birational to X. Note that

$$X \cong \text{Proj}R(K_X) \cong \text{Proj}R(K_{Y_b})$$

where $Y_b \to \mathcal{Z}_b$ is any log resolution.

We may assume that B is irreducible. Let $\eta = \text{Spec}(K)$ be its generic point and $\mathcal{Y}_\eta \to \mathcal{Z}_\eta$ be a log resolution. By [BCHM09], it follows that $R(K_{\mathcal{Y}_\eta})$ is finitely generated. We may therefore pick an integer $N > 0$ such that $R(NK_{\mathcal{Y}_\eta})$ is generated in degree 1. There is an open subset $B^0 \subset B$ such that $R(NK_{\mathcal{Y}^0})$ is generated over B^0 in degree 1 where $\mathcal{Y}^0 = \mathcal{Y} \times_B B^0$. By Noetherian induction, we may assume that $R(NK_{\mathcal{Y}})$ is generated over B in degree 1. Replacing $\mathcal{Y}$ by an appropriate resolution, we may assume that $|NK_{\mathcal{Y}}|$ defines a morphism $\phi_N : \mathcal{Y} \to \mathcal{X} \cong \text{Proj}_B R(K_{\mathcal{Y}})$. We then have

$$X \cong \text{Proj}R(K_{\mathcal{Y}_b}) \cong \text{Proj}R(K_X). \qquad \square$$

Ideally, one would like to construct **proper** moduli spaces for varieties of general type. In order to do this, it is necessary to allow certain degenerations of these varieties. For example in dimension 1 it is necessary to consider stable curves and in higher dimensions we must consider semi-log canonical varieties i.e. varieties X such that

1. X is reduced and S_2,
2. K_X is $\mathbb{Q}$-Cartier, and
3. if $f : \tilde{X} \to X$ is a semiresolution of singularities, then $K_{\tilde{X}} \equiv f^*K_X + \sum a_i E_i$ where $a_i \geq -1$.

This is the generalization of log canonical singularities to the non-normal situation.

If we let $\nu : X^\nu \to X$ be the normalization, then $X^\nu = \coprod_{i=1,\dots,m} X_i$ and we may write $K_{X_i} + \Delta_i = (\nu^* K_X)|_{X_i}$ where (X_i, Δ_i) is a log canonical pair and Δ_i is a reduced divisor.

If we are to construct proper moduli spaces, it is therefore important to prove the boundedness of n-dimensional canonically polarized semi log canonical varieties X with fixed volume $K_X^n = M$.

The first step is provided by an affirmative answer to Conjecture 3.2: Since $K_X^n = \sum_{i=1,\dots,m}(K_{X_i} + \Delta_i)^n$ and since by (3.2) the numbers $(K_{X_i} + \Delta_i)^n$ belong to a DCC set $\mathcal{V}$, then there exists a positive constant $v > 0$ such that $(K_{X_i} + \Delta_i)^n \geq v$ for all i. In particular there is an upper bound for the number of irreducible components of X i.e. $m \leq M/v$. Moreover, by (3.3) and arguing as in the proof of (3.9), one expects that the pairs (X_i, Δ_i) (and hence the variety X) belong to a bounded family.

3.4. Open varieties. Let X be a smooth quasi-projective variety, and consider $\bar{X}$ a smooth projective variety such that $X = \bar{X} - F$ where F is a simple normal crossing divisor on $\bar{X}$.

The geometry of X is then studied in terms of the rational maps defined by $H^0(\omega_{\bar{X}}^{\otimes m}(mF))$ for $m > 0$. Note these maps are independent of the chosen compactification $\bar{X}$ of X. Conjectures 3.2 and 3.3 would allow us to generalize (1.7) to this context.

3.5. Fano varieties. A terminal Fano variety X is a normal projective variety with terminal singularities such that $-K_X$ is ample. (We have similar definitions for canonical singularities, log terminal singularities, etc.) These varieties naturally arise in the context of the minimal model program, which predicts that if Y is a variety with $\kappa(Y) < 0$, then there is a finite sequence of flips and divisorial contractions $Y \dashrightarrow Y'$ and a projective morphism $f : Y' \to Z$ whose general fiber is a terminal Fano variety (of dimension > 0). Therefore, one can think of terminal Fano varieties with Picard number one

as the building blocks for smooth projective varieties with negative Kodaira dimension $\kappa(Y) < 0$.

If $\dim X = 1$, there is only one terminal Fano variety: the projective line $\mathbb{P}^1$. If $\dim X = 2$, terminal Fano varieties are known as Del Pezzo surfaces (a terminal surface is necessarily smooth). There are ten families of such surfaces. In higher dimensions, one expects a similar result to hold. The following fundamental result (cf. [Nadel90], [Nadel91], [Campana91], [Campana92], [KMM92a], [KMM92b]) shows that at least for smooth Fano varieties, this is the case:

Theorem 3.10. *Let $n \in \mathbb{Z}_{>0}$. Then there are only finitely many families of n-dimensional smooth projective Fano varieties.*

The proof of this Theorem is based on the study of the properties of rational curves on these manifolds. When X has singularities, then the behavior of rational curves on X is more subtle. Nevertheless we have the following conjecture known as the BAB (or Borisov-Alexeev-Borisov) Conjecture.

Conjecture 3.11. *Let $n \in \mathbb{Z}_{>0}$. Then there are only finitely many families of canonical $\mathbb{Q}$-factorial Fano varieties.*

Remark 3.12. *The above conjecture is already interesting for Fano varieties of Picard number* 1.

One also expects a similar conjecture for ϵ-log terminal Fano varieties (not necessarily $\mathbb{Q}$-factorial with arbitrary Picard number). Recall that if $\epsilon > 0$, then X is ϵ-log terminal if for any log resolution $f : X' \to X$, we have $K_{X'} = f^*K_X + \sum a_i E_i$ where $a_i > \epsilon - 1$. The example of cones over a rational curve of degree n show that the ϵ-log terminal condition is indeed necessary.

Conjecture 3.11 is known for canonical Fano varieties of dimension ≤ 3 (in characteristic zero) of arbitrary Picard number cf. [Kawamata92] and [KMMT00]; for toric varieties [BB92] and for spherical varieties [AB04].

A positive answer to Conjecture 3.11 would have profound implications on the birational geometry of higher dimensional projective varieties. In particular (3.11) is related to the famous conjectures on the ACC for mld's, the ACC for log canonical thresholds and the termination of flips.

The techniques for the study of varieties of positive Kodaira dimension (that we have described above) do not readily apply to this context. However we would like to mention [McKernan03] for a related approach and [HM10b] for one possible connection showing that it is possible that results for varieties of log general type may be useful in the study of log-Fano varieties.

References

[Alexeev94] V. Alexeev, *Boundedness and K^2 for log surfaces.* Internat. J. Math. **5** (1994), no. 6, 779–810.

[AB04] V. Alexeev and M. Brion, *Boundedness of spherical Fano varieties.* The Fano Conference, 69–80, Univ. Torino, Turin, 2004.

[AM04] V. Alexeev and S. Mori, *Bounding singular surfaces of general type.* Algebra, arithmetic and geometry with applications (West Lafayette, IN, 2000), 143–174, Springer, Berlin, 2004.

[Andreotti50] A. Andreotti, *Sopra le superficie algebriche che posseggono trasformazioni birazionali in se.* (Italian) Univ. Roma Ist. Naz. Alta Mat. Rend. Mat. e Appl. (5) 9, (1950). 255–279.

[AS95] U. Angehrn and Y.-T. Siu, *Effective freeness and point separation for adjoint bundles.* Invent. Math. **122** (1995), no. 2, 291–308.

[Ballico93] E. Ballico, *On the automorphisms of surfaces of general type in positive characteristic,* Rend. Mat. Acc. Lincei (9) **4** (1993) 121–129.

[Benveniste86] X. Benveniste, *Sur les applications pluricanoniques des variétés de type tres général en dimension 3.* Am. J. Math. 108, 433–449 (1986)

[BCHM09] C. Birkar, P. Cascini, C. Hacon, J. McKernan, *Existence of minimal models for varieties of log general type.* J. Amer. Math. Soc. **23** (2010), 405–468.

[Bombieri70] E. Bombieri, *The pluricanonical map of a complex surface.* In: Several Complex Variables, I (Proc. Conf., Univ. of Maryland, College Park, MD 1970), pp. 35–87. Berlin: Springer 1970

[BB92] A. A. Borisov and L. A. Borisov, *Singular toric Fano three-folds.* (Russian) Mat. Sb. 183 (1992), no. 2, 134–141; translation in Russian Acad. Sci. Sb. Math. 75 (1993), no. 1, 277–283

[Cai00] J.-X. Cai, *Bounds of automorphisms of surfaces of general type in positive characteristic.* J. Pure Appl. Algebra **149** (2000), no. 3, 241–250.

[Campana91] F. Campana, *Une version géométrique généralizée du théorème du produit de Nadel.* Bull. Soc. Math. France **119** (1991), 479–493.

[Campana92] F. Campana, *Connexité rationnelle des variétés de Fano.* Ann. Sci. École Norm. Sup. (4) 25 (1992), no. 5, 539–545.

[CS96] F. Catanese and M. Schneider, *Polynomial bounds for abelian groups of automorphisms.* Special issue in honour of Frans Oort. Compositio Math. **97** (1995), no. 1–2, 1–15.

[Chen01] M. Chen, *The relative pluricanonical stability for 3-folds of general type.* Proc. Am. Math. Soc. 129, 1927–1937 (2001) (electronic)

[CC08] J. A. Chen, M. Chen, *On projective threefolds of general type.* Electron. Res. Announc. Math. Sci. 14 (2007), 69–73 (electronic).

[CC08b] J. A. Chen, M. Chen, *The canonical volume of 3-folds of general type with* $\chi \leq 0$. J. Lond. Math. Soc. (2) 78 (2008), no. 3, 693–706.

[CC08c] J. A. Chen, M. Chen, *An optimal boundedness on weak* $\mathbb{Q}$*-Fano 3-folds.* Adv. Math. **219** (2008), no. 6, 2086–2104.

[Corti91] A. Corti, *Polynomial bounds for the number of automorphisms of a surface of general type.* Ann. Sci. École Norm. Sup. (4) **24** (1991), no. 1, 113–137.

[DiBiagio10] L. Di Biagio, *Pluricanonical systems for 3-folds and 4-folds of general type.* arXiv:1001:3340

[FM00] O. Fujino, S. Mori, *A canonical bundle formula.* J. Differential Geom. 56 (2000), no. 1, 167–188.

[Gieseker77] D. Gieseker, *Global moduli for surfaces of general type.* Invent. Math. **43** (1977), no. 3, 233–282.

[HM06] C. Hacon, J. McKernan, *Boundedness of pluricanonical maps of varieties of general type.* Invent. Math. **166** (2006), no. 1, 1–25.

[HM07] C. Hacon, J. McKernan, *Extension Theorems and the existence of Flips.* To appear in: Flips for 3-folds and 4-folds, A. Corti editor. Oxford Lecture Series in Mathematics and Its Applications, **35**.

[HM10a] C. Hacon, J. McKernan, *Existence of minimal models for varieties of log general type II* J. Amer. Math. Soc. **23** (2010), 469–490.

[HM10b] C. Hacon, J. McKernan, *Flips and flops.* Preprint (2010).

[Hanamura85] M. Hanamura, *Pluricanonical maps of minimal 3-folds.* Proc. Japan Acad., Ser. A. Math. Sci. 61, 116–118 (1985)

[Hartshorne77] R. Hartshorne, *Algebraic geometry.* Graduate Texts in Mathematics, No. **52**. Springer-Verlag, New York-Heidelberg, 1977. xvi+496 pp.

[HS91] A. T. Huckleberry, M. Sauer, *On the order of the automorphism group of a surface of general type.* Math. Z. **205** (1990), no. 2, 321–329.

[Iano-Fletcher00] A.R. Iano-Fletcher, *Inverting Reid's exact plurigenera formula.* Math. Ann. 284 (1989), no. 4, 617–629.

[Iano-Fletcher00] A.R. Iano-Fletcher, *Working with weighted complete intersections.* In: Explicit Birational Geometry of 3-folds. Lond. Math. Soc. Lect. Note Ser., vol. 281, pp. 101–173. Cambridge: Cambridge Univ. Press (2000)

[Kawamata86] Y. Kawamata, *On the plurigenera of minimal algebraic* 3*-folds with* $K \equiv 0$. Math. Ann. **275** (1986), no. 4, 539–546.

[Kawamata92] Y. Kawamata, *Boundedness of* $\mathbb{Q}$*-Fano threefolds. Proceedings of the International Conference on Algebra, Part 3 (Novosibirsk, 1989),* 439–445, Contemp. Math., 131, Part 3, Amer. Math. Soc., Providence, RI, 1992.

[Kawamata98] Y. Kawamata, *Subadjunction of log canonical divisors. II.* MR1646046 (2000d:14020) Kawamata, Yujiro Subadjunction of log canonical divisors. II. Amer. J. Math. 120 (1998), no. 5, 893–899.

[Kolláretal92] J. Kollár et. al. *Flips and abundance for algebraic threefolds*, Asterisque **211** (1992).

[KM98] J. Kollár, S. Mori, *Birational geometry of algebraic varieties.* With the collaboration of C. H. Clemens and A. Corti. Translated from the 1998 Japanese original. Cambridge Tracts in Mathematics, **134**. Cambridge University Press, Cambridge, 1998. viii+254 pp.

[Kollár95] J. Kollár, *Singularities of pairs.* Algebraic geometry—Santa Cruz 1995, 221–287, Proc. Sympos. Pure Math., **62**, Part 1, Amer. Math. Soc., Providence, RI, 1997.

[Kollár07] J. Kollár, *Kodaira's canonical bundle formula and adjunction.* Flips for 3-folds and 4-folds, 134–162, Oxford Lecture Ser. Math. Appl., **35**, Oxford Univ. Press, Oxford, 2007.

[Kollár08] J. Kollár, *Is there a topological Bogomolov-Miyaoka-Yau inequality?*, Pure Appl. Math. Q., 2008, **4** no. 2, 203–236

[KMM92a] J. Kollár, Y. Miyaoka and S. Mori, *Rational curves on Fano varieties.* In Proc. Alg. Geom. Conf. Trento, Lecture Notes in Mathematics, vol. 1515. Springer, 1992, pp. 100–105.

[KMM92b] J. Kollár, Y. Miyaoka and S. Mori, *Rational connectedness and boundedness of Fano manifolds.* J. Differential Geom. 36 (1992), no. 3, 765–779.

[KMMT00] J. Kollár, Y. Miyaoka, S. Mori, H. Takagi, *Boundedness of canonical $\mathbb{Q}$-Fano 3-folds.* Proc. Japan Acad. Ser. A Math. Sci. 76 (2000), no. 5, 73–77.

[Lazarsfeld05] R. Lazarsfeld, *Positivity in algebraic geometry. II. Positivity for vector bundles, and multiplier ideals.* Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], **49**. Springer-Verlag, Berlin, 2004. xviii+385 pp.

[McKernan03] J. McKernan, *Boundedness of log terminal Fano pairs of bounded index* arXiv:math/0205214.

[Morrison86] D. Morrison, *A remark on: "On the plurigenera of minimal algebraic 3-folds with $K \equiv 0$.* Math. Ann. **275** (1986), no. 4, 547–553.

[Luo94] T. Luo, *Plurigenera of regular threefolds.* Math. Z. 217, 37–46 (1994)

[Luo00] T. Luo, *Global holomorphic 2-forms and pluricanonical systems on threefolds.* Math. Ann. 318, 707–730 (2000)

[Nadel90] A. M. Nadel, *A finiteness theorem for Fano 4-folds.* Unpublished.

[Nadel91] A. M. Nadel, *The boundedness of degree of Fano varieties with Picard number one.* J. Amer. Math. Soc. **4** (1991), 681–692.

[Pacienza09] G. Pacienza, *On the uniformity of the Iitaka fibration.* Math. Res. Lett. 16 (2009), no. 4, 663–681.

[Ringler07] A. Ringler, *On a conjecture of Hacon and McKernan in dimension three.* arXiv:0708.3662v2

[Szabo96] E, Szabó, *Bounding automorphism groups.* Math. Ann. **304** (1996), no. 4, 801–811.

[Takayama06] S. Takayama, *Pluricanonical systems on algebraic varieties of general type.* Invent. Math. **165** (2006), no. 3, 551–587.

[Todorov07] G. Todorov, *Pluricanonical maps for threefolds of general type.* Ann. Inst. Fourier (Grenoble) 57 (2007), no. 4, 1315–1330.

[Todorov08] G. Todorov, *Effective log Iitaka fibrations for surfaces and threefolds.* arXiv:0805.3494

[TX08] G. Todorov, C. Xu, *On Effective log Iitaka Fibration for 3-folds and 4-folds.* arXiv:0811.3998

[Tsuji00] H. Tsuji, *Bound of automorphisms of projective varieties of general type.*, arXiv:math/0004138

[Tsuji07] H. Tsuji, *Pluricanonical systems of projective varieties of general type. II.* Osaka J. Math. **44** (2007), no. 3, 723–764.

[VZ09] E. Viehweg, D.-Q. Zhang, *Effective Iitaka fibrations.* J. Algebraic Geom. 18 (2009), no. 4, 711–730.

[Xiao94] G. Xiao, *Bound of automorphisms of surfaces of general type. I.* Ann. of Math. (2) **139** (1994), no. 1, 51–77.

[Xiao95] G. Xiao, *Bound of automorphisms of surfaces of general type. II.* J. Algebraic Geom. **4** (1995), no. 4, 701–793.

[Xiao96] G. Xiao, *Linear bound for abelian automorphisms of varieties of general type.* J. Reine Angew. Math. **476** (1996), 201–207.

[Zhu09a] L. Zhu, *On threefolds of general type with* $\chi = 1$. Manuscripta Math. 129 (2009), no. 1, 99–126.

[Zhu09b] L. Zhu, *The sharp lower bound for the volume of threefolds of general type with* $\chi(\mathcal{O}_X) = 1$. Math. Z. 261 (2009), no. 1, 123–141.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Hyperkähler Manifolds and Sheaves

Daniel Huybrechts*

Abstract

Moduli spaces of hyperkähler manifolds or of sheaves on them are often non-separated. We will discuss results where this phenomenon reflects interesting geometric aspects, e.g. deformation equivalence of birational hyperkähler manifolds or cohomological properties of derived autoequivalences. In these considerations the Ricci-flat structure often plays a crucial role via the associated twistor space providing global deformations of manifolds and bundles.

Mathematics Subject Classification (2010). Primary 14F05, 53C26; Secondary 18E30,14J28.

Keywords. Hyperkähler manifolds, moduli spaces, derived categories, holomorphic symplectic manifolds.

K3 surfaces and (over the last ten years or so) also their higher dimensional analogues, compact hyperkähler manifolds, have been studied intensively from various angles. In the case of abelian varieties, the interplay between algebraic, arithmetic, and complex geometric techniques makes the study of this particular class of varieties interesting and rewarding. In many respects, K3 surfaces and hyperkähler manifolds behave very much like abelian varieties, one can even pass from one to the other via the Kuga–Satake construction. There are however two features that are new: Non-separation (of various moduli spaces) and twistor spaces (associated to Ricci-flat metrics).

In a way, it is the group structure that prevents both issues from playing any role for abelian varieties. For example, Ricci-flat metrics on complex tori are actually flat and hence without much geometric significance. As for the non-separation, we will discuss birational hyperkähler manifolds giving rise to non-separated points in the moduli space of varieties, whereas the group structure allows one to extend any birational correspondence between abelian varieties to an isomorphism right away.

*Mathematisches Institut, Universität Bonn, Germany.
E-mail: huybrech@math.uni-bonn.de.

The first is the more intriguing of the two features. Usually, non-Hausdorff phenomena, e.g. for an algebraic geometer non-separated schemes, are considered unpleasant and better avoided. As it turns out, the occurrence of non-separated points, e.g. in the moduli space of manifolds or of (complexes of) sheaves, can be turned into a useful technique applicable to various problems. This general idea seems to work best when combined with the existence of twistor spaces. The latter also allows one to go back and forth between algebraic and non-algebraic complex geometry. For purists this technique might be a weakness of the theory, but we shall try to convince the reader that it is indeed very powerful.

The aim of this note is to review a few scattered results for which non-separation phenomena and twistor spaces play a decisive role. We will touch upon questions concerning the birational geometry of hyperkähler manifolds, derived categories of coherent sheaves on K3 surfaces and their autoequivalences, Brauer classes, hyperholomorphic bundles, Chow groups, etc. There is no attempt at completeness and I apologize for not covering the material in a more concise form. I believe that some of the techniques can be pushed further to treat other interesting open problems in the area, some of which will be mentioned at the end.

Acknowledgement. I wish to thank Emanuele Macrì, Paolo Stellari, and Richard Thomas for the pleasant and stimulating collaboration over the years.

1. Introduction

To get an idea what kind of non-Hausdorff phenomena we have in mind let us recall the following two classical examples.

– The bundles E_t on the projective line $\mathbb{P}^1$ (say over a field k) parametrized by classes $t \in \mathrm{Ext}^1(\mathcal{O}(1), \mathcal{O}(-1)) \cong H^1(\mathbb{P}^1, \mathcal{O}(-2)) \cong k$ are isomorphic to $\mathcal{O} \oplus \mathcal{O}$ for $t \neq 0$ and to $\mathcal{O}(1) \oplus \mathcal{O}(-1)$ for $t = 0$. In other words, there exists a vector bundle E on $\mathbb{P}^1 \times \mathbb{A}^1$ such that on all fibres of the projection $\mathbb{P}^1 \times \mathbb{A}^1 \to \mathbb{A}^1$ with the exception of the fibre over the origin the bundle is the trivial bundle of rank two. Equivalently, there exist two bundles E and E' on $\mathbb{P}^1 \times \mathbb{A}^1 \to \mathbb{A}^1$ which are isomorphic on the open set $\mathbb{P}^1 \times \mathbb{A}^1 \setminus \{0\}$ but with different restrictions $E_0 \cong \mathcal{O}(1) \oplus \mathcal{O}(-1)$ respectively $E'_0 \cong \mathcal{O} \oplus \mathcal{O}$ to the special fibre. This classical observation can easily be translated into a more geometric non-separation phenomenon for Hirzebruch surfaces: $\mathbb{F}_2 = \mathbb{P}(\mathcal{O} \oplus \mathcal{O}(2))$ and $\mathbb{F}_0 = \mathbb{P}^1 \times \mathbb{P}^1$ define non-separated points in the moduli space of varieties

– The Atiyah flop describes two crepant resolutions $Z \to Z_0 \leftarrow Z'$ of the three-dimensional rational double point $Z : xy - zw = 0$ both replacing the singular point by a $\mathbb{P}^1$. Equivalently, the blow-up $\tilde{Z} \to Z_0$ of the singular point admits two projections $Z \leftarrow \tilde{Z} \to Z'$ extending the two projections of the exceptional divisor $\mathbb{P}^1 \times \mathbb{P}^1$. Put in a more global context this observation can be used to construct two non-isomorphic families of K3 surfaces $\mathcal{X} \to D \leftarrow \mathcal{X}'$

over a disk D isomorphic over the punctured disk D^*, i.e. $\mathcal{X}|_{D^*} \cong \mathcal{X}'|_{D^*}$. In particular, all fibres $\mathcal{X}_t, \mathcal{X}'_t$, $t \neq 0$, are isomorphic in a way compatible with the projection to D, but these isomorphisms do not converge to an isomorphism of the special fibres $\mathcal{X}_0$ and $\mathcal{X}'_0$. In fact, the graph Γ_t of the fibrewise isomorphism for $t \neq 0$ degenerates to a cycle $\Gamma_0 + \mathbb{P}^1 \times \mathbb{P}^1 \subset \mathcal{X}_0 \times \mathcal{X}'_0$ with Γ_0 itself being, somewhat accidentally due to the small dimension, the graph of an isomorphism.

A compact *hyperkähler manifold* or irreducible holomorphic *symplectic manifold* is, by the definition we adopt here, a simply-connected compact complex Kähler manifold X such that $H^0(X, \Omega^2_X)$ is spanned by a nowhere degenerate two-form σ. Since all our manifolds will be compact, we simply call them hyperkähler. The definition can be adapted to projective varieties over other fields, but most of the existing theory is concerned with complex manifolds. K3 surfaces are the two-dimensional hyperkähler manifolds and for them there is also a rich theory over number fields and in finite characteristic.

What makes the complex case special is the Calabi–Yau theorem proving the existence of a unique Ricci-flat Kähler metric in each Kähler class on X. In fact, Ricci-flat Kähler metrics exist on the larger class of Calabi–Yau manifolds, but for hyperkähler manifolds they lead to a global complex geometric structure, the *twistor space*.

To be more precise, let $\mathcal{K}_X \subset H^2(X, \mathbb{R}) \cap H^{1,1}(X)$ denote the open cone of Kähler classes (among them all ample classes if X is projective). With any $\alpha \in \mathcal{K}_X$ there is associated a complex manifold $\mathcal{X}(\alpha)$ together with a smooth proper holomorphic map $\pi : \mathcal{X}(\alpha) \to \mathbb{P}^1$. One of the fibres, say $\mathcal{X}_0$ is actually isomorphic to X, but most of the other fibres are not.

Note that by construction the twistor space as a differentiable manifold is simply $X \times \mathbb{P}^1$ and π is the second projection. Moreover, the natural (twistor) sections $\{x\} \times \mathbb{P}^1$ of π are holomorphic with normal bundle $\mathcal{O}(1) \oplus \ldots \oplus \mathcal{O}(1)$.

2. Non-separation for Hyperkähler Manifolds

Birational K3 surfaces are always isomorphic, e.g. because the minimal model of a surface of non-negative Kodaira dimension is unique. In fact, any birational correspondence extends to an isomorphism. (Note that by abuse of language we will speak about birational maps etc. even when the manifolds are not algebraic and one should more accurately say bimeromorphic.)

In higher dimension the situation changes. The easiest example of a non-trivial birational correspondence between, in general non-isomorphic, hyperkähler manifolds has been constructed already in [21] and is called the Mukai flop. Any hyperkähler manifold containing a half-dimensional projective space can be flopped replacing the projective space $\mathbb{P}$ by its dual $\mathbb{P}^*$. The new manifold is holomorphic symplectic, but not always Kähler and hence not hyperkähler (see [29] for an example that starts with a projective moduli space of sheaves).

In general and in particular in dim > 4, birational correspondences between hyperkähler manifolds will be more complicated than simple Mukai flops. But as it turns out, any birational correspondence between hyperkähler manifolds can be obtained as the limit of isomorphisms (see [10, 11]):

2.1. *Any two birational hyperkähler manifolds X and X' define non-separated points in the moduli space of varieties. Equivalently, there exist two smooth proper families $\mathcal{X} \to D \leftarrow \mathcal{X}'$ over a disk D with central fibres $\mathcal{X}_0 \cong X$ respectively $\mathcal{X}'_0 \cong X'$ and such that the two families are isomorphic over the punctured disk D^*, i.e. $\mathcal{X}|_{D^*} \cong \mathcal{X}'|_{D^*}$.*

This result had first been proved for projective hyperkähler manifolds and under an additional assumption on the codimension of the exceptional locus by projective techniques which are valid over arbitrary fields. Later, twistor spaces have been used instead to prove the result in the above form. Note that even for X and X' projective, the nearby fibres in the families in (2.1) are usually non-projective.

The result is intimately related to the description of the Kähler cone and its birational variant. For K3 surfaces the Kähler cone is determined by smooth rational curves and a less explicit version of this holds true also in higher dimensions. In particular, for generic hyperkähler manifolds, which do not admit any curves, the Kähler cone is maximal, i.e. coincides with the positive cone. For the general theory see the survey [7] and references therein. A detailed investigation of the shape of the ample cone in the known examples has been initiated by Hassett and Tschinkel, see e.g. [8].

Let us state explicitly the following immediate consequence of (2.1):

2.2. *Two birational hyperkähler manifolds are deformation equivalent. In particular, their Hodge, Betti, and Chern numbers coincide.*

The result was used to show that most of the known examples, with the exception of O'Grady's exceptional examples in dimension six and ten, are deformations of the two standard series provided by Hilbert schemes of points on K3 surfaces and generalized Kummer varieties.

Note that deformation equivalence does not hold for birational Calabi–Yau manifolds in general, which need not even be homeomorphic and might even have different Chern numbers (see [2, 26]). For general Calabi–Yau manifolds the result that comes close to (2.1) is due to Batyrev and Kontsevich and proves equality of Hogde and Betti numbers. Motivic integration originated by Kontsevich for this purpose has been developed to a beautiful general theory by Denef and Loeser (see [20]). Applied to birational Calabi–Yau manifolds X and X' it shows that the (infinite-dimensional) spaces of formal arcs $J(X)$ respectively $J(X')$ differ only by insignificant bits. For birational hyperkähler manifolds and the non-separating families $\mathcal{X}, \mathcal{X}'$ as in (2.1) one can consider the spaces $J_0(\mathcal{X})$ and $J_0(\mathcal{X}')$ of formal arcs with support in the central fibre. The twistor sections provide a canonical section of the projection $J_0(\mathcal{X}) \to X$

which should lead to a stratified isomorphism of X and X' (non-holomorphic on the exceptional locus). It would be interesting to incorporate the Ricci-flat metric in a stronger way into birational geometry of hyperkähler manifolds and also to extend some of it to general Calabi–Yau manifolds.

The graph Γ_t of the isomorphism of the general fibres $\mathcal{X}_t \cong \mathcal{X}'_t$ in (2.1) degenerates to a cycle $Z + \sum Y_i \subset X \times X'$ where Z is the original birational correspondence. This is reminiscent of the Atiyah flop. The additional components Y_i do not dominate the factors but are in general difficult to describe explicitly. E.g. in the case of a Mukai flop there is only one additional component which is simply $\mathbb{P} \times \mathbb{P}^*$. So, more in the spirit of our philosophy here, (2.1) says that up to adding non-dominating components any birational correspondence $X \leftarrow Z \rightarrow X'$ between hyperkähler manifolds can be deformed to an isomorphism of generic deformations of X respectively X'. Derived versions will be discussed later, see (4.2) and (5.1).

3. Twistor Spaces

Deformation theory is a technical but well developed subject. The standard techniques deal with finite order or formal deformations. Convergence or algebraicity is usually more difficult. Global deformations of a variety X, i.e. a flat family $\mathcal{X} \rightarrow B$ with $\mathcal{X}_0 \cong X$ over a proper base B of positive dimension are hardly ever constructed explicitly. This makes twistor spaces stand apart. The twistor space $\mathcal{X} = \mathcal{X}(\alpha) \rightarrow \mathbb{P}^1$ associated with a Kähler class α on a hyperkähler manifold X connects X with other, possibly far away, hyperkähler manifolds $\mathcal{X}_t$. The price one has to pay is the loss of algebraicity. In fact, the total space $\mathcal{X}$ is not even Kähler and only countable many fibres $\mathcal{X}_t$ are projective. Nevertheless, it seems that essential information about the geometry of a projective hyperkähler manifold X is preserved along the twistor space deformation to other projective fibres.

Twistor spaces or almost equivalently hyperkähler metrics play a central role already in the standard theory of K3 surfaces and, partially due to the absence of a proper analogue of the Global Torelli theorem, even more so in higher dimensions (see [7, 11]). We will not go into the details of the general theory of hyperkähler manifolds, but let us mention that twistor spaces are crucial e.g. for the proof of the surjectivity of the period map and the description of the (birational) Kähler cone.

To underpin the global nature of twistor spaces let us just mention that for any polarized K3 surface (X, L), e.g. $X \subset \mathbb{P}^3$ a quartic and L the restriction of $\mathcal{O}(1)$, and any Kähler class on X, e.g. the one given by $\mathrm{c}_1(L)$, the associated twistor space will also parametrize polarized K3 surfaces (X', L') of other degrees, e.g. a double cover of the plane. In dimension four the twistor space can be used to connect e.g. the Hilbert scheme $\mathrm{Hilb}^2(S)$ of a K3 surface S with the Fano variety of lines on a cubic fourfold. The reason behind this observation is that the base of the twistor space yields a curve in the moduli space of marked

hyperkähler manifolds whereas the other loci are of codimension one, which therefore are expected to intersect.

By construction, twistor spaces are associated to hyperkähler metrics. A similar relation exists, due to the work of Donaldson, Hitchin and others, between stable vector bundles and Hermite–Einstein metrics. A combination of both leads to the following result of Verbitsky [27] which applies to stable vector bundles with trivial first Chern class on K3 surfaces.

3.1. *Let X be a hyperkähler manifold and E a holomorphic bundle on X which is stable with respect to a Kähler class α. If the first and second Chern classes of E stay algebraic (i.e. of type $(1,1)$ resp. $(2,2)$) on the fibres of the associated twistor space $\mathcal{X}(\alpha) \to \mathbb{P}^1$, then E is hyperholomorphic, i.e. extends naturally to a holomorphic vector bundle on $\mathcal{X}(\alpha)$.*

The idea of the proof is to show that the curvature of the Hermite–Einstein connection on E is of type $(1,1)$ with respect to all complex structures associated to the Ricci-flat structure given by α. That this is controlled by the first two Chern classes is reminiscent of Simpson's observation that the vanishing of the second Chern character of a stable bundle implies its (projective) flatness. Then on each fibre $\mathcal{X}_t$ the $(0,1)$-part of the natural Hermite–Einstein connection defines the $\bar{\partial}$-operator for E on this fibre.

The result can be applied to cases where the first Chern class is not trivial or not even orthogonal to the Kähler class α, but then it is only $\mathbb{P}(E)$ that deforms and not the bundle E itself. In [12] this was used to prove that cohomological and geometric Brauer group coincide for K3 surfaces, a result well known for algebraic surfaces. Roughly, the idea is to follow a given cohomological Brauer class along a twistor space and show that it becomes trivial somewhere. (Picard and hence Brauer group jump in a countable and dense subset.) When the class is trivial one represents it by a stable vector bundle which deforms back to the original K3 surface as a projective bundle that represents the chosen Brauer class.

Verbitsky used his result to deduce that very general (and hence non-projective) K3 surfaces have equivalent abelian categories $\mathrm{Coh}(X)$. This is in contrast to Gabriel's result (see [6]) that the abelian category $\mathrm{Coh}(X)$ of an algebraic variety, or more generally any scheme, determines X, but confirms the belief that for non-algebraic manifolds the abelian category of coherent sheaves is too small. Note that even for a very general $K3$ surface the category of coherent sheaves is very rich due to the many stable bundles that continue to exist.

Another point of view on Verbitsky's result, already studied by Itoh and others, is that the moduli space of stable vector bundles on a K3 surface inherits a natural hyperkähler structure. Equivalently, the relative moduli space of stable bundles on the fibres of $\mathcal{X}(\alpha) \to \mathbb{P}^1$ is nothing but the twistor space of the moduli space on one fibre. Note however that this does not extend to the boundary, i.e. to the moduli space of (semi-)stable sheaves and hence does not

allow one to construct the hyperkähler structure on the Hilbert scheme or on the moduli space of stable sheaves.

4. Non-separation for Sheaves and Complexes

That there are bundles that define, like $\mathcal{O}(1) \oplus \mathcal{O}(-1)$ and $\mathcal{O} \oplus \mathcal{O}$ on $\mathbb{P}^1$, non-separated points in the moduli space of bundles is a common feature and not special to $\mathbb{P}^1$. Also the existence of non-trivial homomorphisms between the bundles (in both directions) is frequently observed even for simple bundles (see [24]). The moduli space of simple bundles on a variety, an algebraic space, is in general not expected to be separated. Only stability prevents sheaves of the same slope (or normalized Hilbert polynomial, or phase, etc.) to have non-trivial homomorphisms between each other and this leads to separated and in fact quasi-projective moduli spaces.

The situation seems easier for simple sheaves not allowing any deformation, they do define isolated and hence separated points in their moduli space. Recall that a sheaf F has no infinitesimal deformations if and only if $\mathrm{Ext}^1(F,F) = 0$. Simple sheaves with this property on a K3 surface X are called *spherical*, i.e. they satisfy $\mathrm{Ext}^*(F,F) = H^*(S^2,k)$. So in particular, two spherical sheaves F and F' on X will always define separated points in the moduli space of sheaves on X, but this changes if also deformations of X are allowed. For the rest of this section X will be a projective K3 surface.

4.1. *Suppose F and F' are spherical sheaves with the same numerical invariants on a K3 surface X. Then there exists a deformation $\mathcal{X} \to D$ of X over a disk D and two D-flat sheaves $\mathcal{F}$ and $\mathcal{F}'$ on $\mathcal{X}$ with isomorphic restrictions to $\mathcal{X}^* := \mathcal{X}|_{D^*}$ and special fibres F respectively F'.*

In fact, X can be deformed together with F and F' such that simple implies stable with respect to any Kähler class or polarization. A beautiful observation going back to Mukai says that moduli spaces of stable sheaves with fixed numerical invariants are irreducible (see e.g. [9] or the original [22]). This allows one to conclude that in particular the generic deformations of F and F' are isomorphic.

A rather straightforward consequence of this is that numerically equivalent spherical bundles can also not be distinguished by any other continuous invariants, e.g. they are also rationally equivalent, i.e. their Chern characters in $\mathrm{CH}^*(X)$ coincide. The result also holds for spherical objects in the derived category, see below.

The result can be generalized to sheaves on products of K3 surfaces. This is central for the proof of a conjecture of Szendrői [25] as we shall explain shortly. Let X be an algebraic K3 surface and let $\Phi := \Phi_{\mathcal{E}_0} : \mathrm{D^b}(X) \xrightarrow{\sim} \mathrm{D^b}(X)$ be a linear exact autoequivalence of the derived category $\mathrm{D^b}(X) := \mathrm{D^b}(\mathrm{Coh}(X))$ given as a Fourier–Mukai transform $F \mapsto \mathrm{pr}_{2*}(\mathcal{E}_0 \otimes \mathrm{pr}_1^* F)$ with $\mathcal{E}_0 \in \mathrm{D^b}(X \times X)$.

Then $\mathcal{E}_0$ is rigid, i.e. $\mathrm{Ext}^1(\mathcal{E}_0, \mathcal{E}_0) = 0$, but $\mathrm{Ext}^2(\mathcal{E}_0, \mathcal{E}_0)$ is of dimension 22. In [15] it was proved that X and Φ (or rather $\mathcal{E}_0$) can be deformed together to a very general K3 surface and an equivalence that can be written as a product of explicitly described autoequivalences (shifts and spherical twists $T_{\mathcal{O}}$) whenever the action of Φ on the cohomology of X allows it. Informally, it can be rephrased by saying that any autoequivalence acting trivially on cohomology is a degeneration of the identity. This should be compared to (2.1). As we will mention later, the group of cohomologically trivial autoequivalences is a very rich group. More in the spirit of this review we state the result as (see [15]):

4.2. *If $\Phi, \Phi' : \mathrm{D^b}(X) \xrightarrow{\sim} \mathrm{D^b}(X)$ are two linear exact autoequivalences inducing the same action on $H^*(X, \mathbb{Z})$, then there exist formal deformations $\mathcal{X} \to \mathrm{Spf}(\mathbb{C}[[t]])$ of X and $(\tilde{\Phi}, \tilde{\Phi}')$ of (Φ, Φ') whose restrictions to the generic fibre $\mathcal{X}_K$ of $\mathcal{X}$ over $K := \mathbb{C}((t))$ are isomorphic Fourier–Mukai transforms up to shift and a power of the simple spherical twist $T_{\mathcal{O}}$.*

The assumption in (4.1) that the two sheaves have the same Chern characters in $H^*(X)$ is here replaced by Φ, Φ' inducing the same action on $H^*(X)$. In fact, any spherical sheaf F induces an autoequivalence T_F, the spherical twist, which on cohomology acts by reflection. In this sense, (4.2) is a generalization of (4.1), but due to the deformation theory involved its proof is rather more technical.

Note also that the deformation $\mathcal{X} \to \mathrm{Spf}(\mathbb{C}[[t]])$ used in [14] is the formal neighbourhood of a very generic twistor space $\mathcal{X}(\alpha) \to \mathbb{P}^1$ and thus highly non-algebraic. In particular, the generic fibre $\mathcal{X}_K$ does not exist as a projective variety. Instead of working with rigid analytic varieties, [14] makes only use of $\mathrm{Coh}(\mathcal{X}_K)$ and its derived category which can both be constructed directly as quotients of $\mathrm{Coh}(\mathcal{X})$ respectively $\mathrm{D^b}(\mathcal{X})$ without ever defining $\mathcal{X}_K$.

The result (4.2) has interesting consequences. Firstly, in [15] it was proved that autoequivalences of K3 surfaces, thought of as mirrors of symplectomorphisms, behave as predicted by mirror symmetry (see [25]):

4.3. *If $\Phi : \mathrm{D^b}(X) \to \mathrm{D^b}(X)$ is a linear exact autoequivalence, then the induced action on $H^*(X, \mathbb{Z})$ preserves the natural orientation of any positive four-space.*

The orthogonal group $\mathrm{O}(H^*(X, \mathbb{R}))$ has four connected components and the result says that derived autoequivalences avoid two of them. The result completes earlier work of Mukai, Orlov, and others and allows one to describe the image of $\mathrm{Aut}(\mathrm{D^b}(X)) \to \mathrm{O}(H^*(X, \mathbb{Z}))$ explicitly as the group of orientation preserving Hodge isometries. This can be seen as a derived version of the Global Torelli theorem for automorphisms of K3 surfaces.

Secondly, since the Fourier–Mukai kernels $\mathcal{E}_0$ and $\mathcal{E}_0'$ of Φ respectively Φ' as in (4.2) cannot be separated in the larger moduli space of complexes on deformations of $X \times X'$, all their usual invariants will be the same. E.g. the action on cohomology determines the action on the much larger (at least over $\mathbb{C}$) Chow groups $\mathrm{CH}^*(X)$. Combined with Lazarsfeld's result that indecomposable curves on K3 surfaces are Brill–Noether general this leads to (see [16]):

4.4. *For $\rho(X) \geq 2$ all spherical complexes $F \in \mathrm{D^b}(X)$ take Chern classes in the Beauville–Voisin ring $R(X) \subset \mathrm{CH}^*(X)$.*

Recall that the Beauville–Voisin ring naturally splits the cycle map $\mathrm{CH}^*(X) \to H^*(X, \mathbb{Z})$ (see [1]) and for X defined over $\bar{\mathbb{Q}}$ it is conjectured (Bloch–Beilinson) to be the Chow ring of X over $\bar{\mathbb{Q}}$ (see [16]). The assumption on the Picard number $\rho(X)$ in (4.4) should be superfluous.

5. Open Problems

5.1. It is generally believed that birational Calabi–Yau varieties are derived equivalent. The conjecture seems more accessible for hyperkähler manifolds. One approach could be to use (2.1) and put the two birational hyperkähler manifolds X and X' as special fibres of the same family and then construct the Fourier–Mukai kernel as a degeneration of the diagonal. How to degenerate the diagonal explicitly is not clear. This has been worked out in a few cases (e.g. [17, 23]) and progress in the non-compact situation has been made in [5]. One could wonder whether the autoequivalence can be produced without actually explicitly giving the Fourier–Mukai kernel $\mathcal{E}$. Again, the degeneration argument could be helpful, but since not a single Fourier–Mukai equivalence has ever been described without also giving its Fourier–Mukai kernel, this seems not obvious.

5.2. As explained, all equivalences in the kernel of the natural representation

$$\rho : \mathrm{Aut}(\mathrm{D^b}(X)) \to \mathrm{O}(H^*(X, \mathbb{Z}))$$

can be obtained as degenerations of the diagonal on deformations of X (up to shift and twist $T_{\mathcal{O}}$). Conjecturally, $\ker(\rho)$ is described by Bridgeland [3] as the fundamental group of a certain period domain depending only on the Hodge structure of $H^2(X, \mathbb{Z})$. In particular, $\ker(\rho)$ is usually a non-residually finite group. Also, it should be viewed as the group of deck-transformations of the space of stability conditions on $\mathrm{D^b}(X)$. How exactly the spaces of stability conditions $\mathrm{Stab}(\mathcal{X}_t)$ on the generic deformation $\mathcal{X}_t$ of $X = \mathcal{X}_0$, which has been shown to be simply connected in [13], fit together and 'degenerate' to $\mathrm{Stab}(X)$ is unclear.

5.3. Can non-separation be avoided for hyperkähler manifolds? I believe it cannot and this should be seen as a good thing. E.g. hyperkähler manifolds giving rise to non-separated points are birational and non-isomorphic birational correspondences produce rational curves. The existence and the counting of rational curves on K3 surfaces is a highly interesting subject, see e.g. [4] and [18]. Clearly, if a hyperkähler manifold contains a rational curve, it cannot be hyperbolic as predicted by the Kobayashi conjecture. In fact, non-separated points should be dense in the moduli space of hyperkähler manifolds which could eventually prove non-hyperbolicity for all hyperkähler manifolds. Note

that non-separation would also imply topological restrictions, e.g. $b_2 > 3$ which is widely expected but proved only in small dimensions.

5.4. Another interesting subject concerns the arithmetic of hyperkähler manifolds and whether certain arithmetic properties, e.g. to be defined over particular fields or to admit (many) rational points, is transferred along the twistor space from one algebraic fibre to another. (Compare the work of Hausel and Rodriguez-Villegas on moduli spaces of bundles on curves and the character variety.)

5.5. In analogy to (3.1) it would be interesting to define hyperholomorphic complexes, i.e. complexes of sheaves that naturally deform to the whole twistor space. The stability condition should be phrased in terms of Bridgeland stability. However, since (3.1) works only for bundles, one would also need to find a derived version for locally freeness. How exactly the Hermite–Einstein metric should come in seems unclear.

5.6. The general fibre of a twistor space is a rigid analytic variety. In [14] its category of sheaves was studied, and somehow identified with the variety. As a geometric object or as a category, it should be viewed as naturally associated to the Ricci-flat metric. However, in the construction only the formal neighbourhod of one twistor fibre was used and it would be interesting to see whether this leads to equivalent notions for all fibres. For an algebraic family it would just be the fibre over the generic point.

References

[1] A. Beauville, C. Voisin *On the Chow ring of a K3 surface*, J. Alg. Geom. 13 (2004), 417–426.

[2] L. Borisov, A. Libgober *Elliptic genera of singular varieties*, Duke Math. J. 116 (2003), 319–351.

[3] T. Bridgeland *Stability conditions on K3 surfaces*, Duke Math. J. 141 (2008), 241–291.

[4] F. Bogomolov, B. Hassett, Y. Tschinkel *Constructing rational curves on K3 surfaces*, arXiv:0907.3527.

[5] S. Cautis, J. Kamnitzer, A. Licata *Derived equivalences for cotangent bundles of Grassmannians via categorical sl(2) actions*, arXiv:0902.1797.

[6] P. Gabriel *Des catégories abéliennes*, Bull. Soc. Math. France 90 (1962), 323–448.

[7] M. Gross, D. Joyce, D. Huybrechts *Calabi–Yau manifolds and related geometries*, Springer (2002).

[8] B. Hassett, Y. Tschinkel *Intersection numbers of extremal rays on holomorphic symplectic varieties*, arXiv:0909.4745.

[9] D. Huybrechts, M. Lehn *The geometry of moduli spaces of sheaves*, 2nd edition. Cambridge University Press (2010).

[10] D. Huybrechts *Birational symplectic manifolds and their deformations*, J. Diff. Geom. 45 (1997), 488–513.

[11] D. Huybrechts *Compact hyperkähler manifolds: Basic results*, Invent. Math. 135 (1999), 63–113. Erratum: Invent. math. 152 (2003), 209–212.

[12] D. Huybrechts, S. Schröer *The Brauer group of analytic K3 surfaces*, IMRN. 50 (2003), 2687–2698.

[13] D. Huybrechts, E. Macrì, P. Stellari, *Stability conditions for generic K3 surfaces*, Comp. Math. **144** (2008), 134–162.

[14] D. Huybrechts, E. Macrì, P. Stellari, *Formal deformations and their categorical general fibre*, arXiv:0809.3201. to appear in Com. Math. Helv.

[15] D. Huybrechts, E. Macrì, P. Stellari *Derived equivalences of K3 surfaces and orientation*, Duke Math. J. 149 (2009), 461–507.

[16] D. Huybrechts *Chow groups of K3 surfaces and spherical objects*, arXiv:0809.2606v2. to appear in J. EMS.

[17] Y. Kawamata *Derived equivalence for stratified Mukai flop on* $G(2,4)$, Mirror symmetry. V, AMS/IP Stud. Adv. Math. 38, AMS (2006), 285–294.

[18] A. Klemm, D. Maulik, R. Pandharipande, E. Scheidegger *Noether–Lefschetz theory and the Yau–Zaslow conjecture*, arXiv:0807.2477.

[19] R. Lazarsfeld *Brill–Noether–Petri without degenerations*, J. Diff. Geom. 23 (1986), 299–307.

[20] E. Looijenga *Motivic measures*, Séminaire Bourbaki, Exp. 874, Astérisque 276 (2002), 267–297.

[21] S. Mukai *Symplectic structures of the moduli space of sheaves on an abelian or K3 surface*, Invent. Math. 77 (1984), 101–116.

[22] S. Mukai, *On the moduli space of bundles on K3 surfaces, I*, In: Vector Bundles on Algebraic Varieties, Oxford University Press, Bombay and London (1987), 341–413.

[23] Y. Namikawa *Mukai flops and derived categories II*, Alg. struct. and moduli spaces, CRM Proc. Lect. Not. 38 AMS (2004), 149–175.

[24] A. Norton *Non-separation in the moduli of complex vector bundles*, Math. Ann. 235 (1978), 1–16.

[25] B. Szendrői, *Diffeomorphisms and families of Fourier–Mukai transforms in mirror symmetry*, Applications of Alg. Geom. to Coding Theory, Phys. and Comp. NATO Science Series. Kluwer (2001), 317–337.

[26] B. Totaro *Chern numbers for singular varieties and elliptic homology*, Annals Math. 151 (2000), 757–791.

[27] M. Verbitsky *Hyperholomorphic bundles over a hyper-Kähler manifold*, J. Alg. Geom. 5 (1996), 633–669.

[28] M. Verbitsky *Coherent sheaves on generic compact tori*, CRM Proc. and Lecture Notices 38 (2004), 229–249.

[29] K. Yoshioka *Moduli spaces of stable sheaves on abelian surfaces*, Math. Ann. 321 (2001), 817–884.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Motivic Structures in Non-commutative Geometry

D. Kaledin*

Abstract

We review recent theorems and conjectures saying that periodic cyclic homology of a smooth non-commutative algebraic variety carries all the additional structures the usual de Rham cohomology has in the commutative case, such as a mixed Hodge structure, and a structure of a filtered Dieudonné module.

Mathematics Subject Classification (2010). 14F05, 14F30 and 14F40.

Keywords. Motivic, non-commutative, cyclic, p-adic, Hodge-de Rham.

1. Generalities on Mixed Motives

The conjectural category $\mathcal{MM}$ of *mixed motives*, as described by Deligne, Beilinson and others, unifies and connects various cohomology theories which appear in modern algebraic geometry. Recall that one expects $\mathcal{MM}$ to be a symmetric tensor abelian category with a distiguished invertible object $\mathbb{Z}(1)$ called the *Tate motive*. One expects that for any smooth projective algebraic variety X defined over $\mathbb{Q}$, there exist a functorial *motivic cohomology complex* $H^\bullet(X) \in \mathcal{D}^b(\mathcal{MM})$ with values in the derived category $\mathcal{D}^b(\mathcal{MM})$, whose cohomology groups

$$H^i(X) \in \mathcal{MM}$$

are called *motivic cohomology groups*. If X is the projective space $\mathbb{P}^n$, $n \geq 1$, then one expects to have

$$H^{2i}(\mathbb{P}^n) \cong \mathbb{Z}(-i)$$

for $0 \leq i \leq n$, and 0 otherwise. For a general X and any integer j, one defines the *absolute cohomology complex* by

$$H^\bullet_{abs}(X, \mathbb{Z}(j)) = \mathrm{RHom}^\bullet_{\mathcal{MM}}(\mathbb{Z}(-j), H^\bullet(X)),$$

*Independent University of Moscow & Steklov Math Institute, Moscow, USSR.
E-mail: kaledin@mi.ras.ru.

with its cohomology groups $H^i_{abs}(X, \mathbb{Z}(j))$ known as *absolute cohomology groups*. It is expected that the absolute cohomology groups are related to the algebraic K-theory groups $K_\bullet(X)$ by means of a functorial *regulator map*

$$r : K_\bullet(X) \to \bigoplus_j H^{2j-\bullet}_{abs}(X, \mathbb{Z}(j)), \tag{1.1}$$

and it is expected that the regulator map is "not far from an isomorphism" (for example, it ought to be an isomorphism modulo torsion).

The above picture, with its many refinements which we will not need, is, unfortunately, still conjectural. In some applications, one can get away with considering the "triangulated category of motives" of Hanamura, Levine and Voevodsky, see e.g. [Le]. In other applications, one has to be content with *categories of realizations*. These follow the same general pattern, but the hypothetical category $\mathcal{MM}$ is replaced with a known category Real whose definition axiomatizes the features of a particular known cohomology theory. The prototype example is that of l-adic cohomology. Recall that for any algebraic variety $X/\mathbb{Q}$, its l-adic étale cohomology groups

$$H^i_{et}(X, \mathbb{Q}_l)$$

are $\mathbb{Q}_l$-vector spaces equipped with an additional structure of an *l-adic representation* of the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. These representations form a tensor symmetric abelian category $\mathrm{Rep}_l(\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}))$ with a distiguished Tate module $\mathbb{Q}_l(1)$, and one can treat l-adic cohomology as taking values in this category. One can them define a double-graded absolute cohomology theory

$$H^\bullet_{abs}(X, \mathbb{Q}_l(j)) = H^\bullet(\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}), H^\bullet_{et}(X, \mathbb{Q}_l(j))),$$

known as *absolute l-adic cohomology*, and construct a regulator map of the form (1.1). Conjecturally, we have an exact tensor "realization functor" $\mathcal{MM} \to \mathrm{Rep}_l(\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}))$, l-adic cohomology is obtained by applying realization to motivic cohomology, and the étale regulator map factors through the motivic one. In practice, one can treat $\mathrm{Rep}_l(\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}))$ as a replacement for $\mathcal{MM}$, and hope that the regulator map still captures essential information about $K_\bullet(X)$.

In this paper, we will be concerned with another family of cohomology theories and realizations which appear as refinements of *de Rham cohomology*. By its very nature, de Rham cohomology of a smooth algebraic variety X has coefficients in the field or ring of definition of X. Thus it is not necessary to require that X is defined over $\mathbb{Q}$, and it is convenient to classify de Rham-type cohomology theories by their rings of definitions. There are two main examples.

(i) The ring of definition is either $\mathbb{R}$ or $\mathbb{C}$; the corresponding category of realizations is Deligne's category of mixed $\mathbb{R}$-Hodge structures, and the absolute cohomology theory is Hodge-Deligne cohomology (with a refinement

by Beilinson). The regulator map is the subject of the famous Beilinson Conjectures [Be].

(ii) The ring of definition is $\mathbb{Z}_p$; the corresponding category of realizations is the category of *filtered Dieudonné modules* of Fontaine-Lafaille [FL], and the absolute cohomology theory is *syntomic cohomology* of Fontaine and Messing [FM].

The goal of this paper is to report on recent discoveries and conjectures which state, roughly speaking, that all these additional "motivic" structures on de Rham cohomology of an algebraic variety should exist in a much more general setting of periodic cyclic homology of properly understood *non-commutative* algebraic varieties. As opposed to the usual commutative setting, the "classical" case (i) is more difficult and largely conjectural; in the p-adic case (ii), most of the statements have been proved. Moreover, the p-adic story shows an unexpected relation to algebraic topology which we will also explain. Before we start, however, we should define exactly what we mean by a "non-commutative algebraic variety", and recall basic facts on cyclic homology.

2. Non-commutative Setting

We start by a brief recollection on cyclic homology; a very good overview can be found in J.-L. Loday's book [Lo], and an old overview [FT] is also quite useful. *Hochschild homology* $HH_\bullet(A/k)$ of an associative unital algebra A flat over a commutative ring k is given by

$$HH_\bullet(A) = HH_\bullet(A/k) = \mathrm{Tor}_\bullet^{A^{opp}\otimes_k A}(A, A),$$

where A^{opp} is A with multiplication written in the opposite direction. It has been discovered by Hochschild, Kostant and Rosenberg [HKR] that if A is commutative and $X = \operatorname{Spec} A$ is a smooth algebraic variety over k, then

$$HH_i(A) \cong H^0(X, \Omega^i(X)),$$

the space of i-forms on X over k. *Cyclic homology* $HC_\bullet(A)$ is a refinement of Hochschild homology discovered independently by A. Connes and B. Tsygan. It is functorial in A, and related to $HH_\bullet(A)$ by the *Connes' long exact sequence*

$$HH_\bullet(A) \longrightarrow HC_\bullet(A) \xrightarrow{u} HC_{\bullet-2}(A) \longrightarrow ,$$

where u is a canonical *periodicity map* of degree 2. Both $HH_\bullet(A)$ and $HC_\bullet(A)$ can be represented by functorial complexes $CH_\bullet(A)$, $CC_\bullet(A)$, and the Connes' exact sequence then becomes a short exact sequence of complexes. The complex

$CC_\bullet(A)$ is the total complex of a bicomplex

$$\begin{tikzcd}
\dots \arrow[r] & A \arrow[r, "\mathsf{id}"] & A \arrow[r, "0"] & A \\
\dots \arrow[r] & A\otimes A \arrow[r, "\mathsf{id}+\tau"] \arrow[u, "b"'] & A\otimes A \arrow[r, "\mathsf{id}-\tau"] \arrow[u, "b'"'] & A\otimes A \arrow[u, "b"'] \\
\dots & \dots \arrow[u, "b"'] & \dots \arrow[u, "b'"'] & \dots \arrow[u, "b"'] \\
\dots \arrow[r] & A^{\otimes n} \arrow[r, "\mathsf{id}+\tau+\dots+\tau^{n-1}"] \arrow[u, "b"'] & A^{\otimes n} \arrow[r, "\mathsf{id}-\tau"] \arrow[u, "b'"'] & A^{\otimes n} \arrow[u, "b"']
\end{tikzcd} \tag{2.1}$$

Here it is understood that the whole thing extends indefinitely to the left, all the even-numbered columns are the same, all odd-numbered columns are the same, and the bicomplex is invariant with respect to the horizontal shift by 2 columns which gives the periodicity map u. The map $\tau : A^{\otimes i} \to \mathcal{A}^{\otimes i}$ is the cyclic permutation of order i multiplied by $(-1)^{i+1}$, and b, b' are certain explicit differentials expressed in terms of the multiplication map $m : A^{\otimes 2} \to A$. The complex $CH_\bullet(A)$ is the rightmost column of (2.1), and also any odd-numbered column when counting from the right; the even-numbered columns are acyclic.

Periodic cyclic homology $HP_\bullet(A)$ is obtained by inverting the map u, namely, $HP_\bullet(A)$ is the homology of the complex

$$CP_\bullet(A) = \lim_{\overset{u}{\leftarrow}} CC_\bullet(A)$$

(explictily, this is the total complex of a bicomplex obtained by extending (2.1) to the right as well as to the left). *Negative cyclic homology* $HC^-_\bullet(A)$ is the homology of the complex $CC^-_\bullet(A)$ obtained as the third term in a short exact sequence

$$0 \longrightarrow CC^-(A) \longrightarrow CP_\bullet(A) \longrightarrow CC_{\bullet-2}(A) \longrightarrow 0$$

(equivalently, one extends (2.1) to the right but not to the left).

The reason cyclic homology is interesting in algebraic geometry is the following comparison theorem. In the situation of the Hochschild-Kostant-Rosenberg Theorem, let d be the dimension of $X = \operatorname{Spec} A$, and assume in addition that $d!$ is invertible in the base ring k. Then there exists a canonical isomorphism

$$HP_\bullet(A) \cong H^\bullet_{DR}(X)((u)), \tag{2.2}$$

where the right-hand side is a shorthand for "formal Laurent power series in one variable u of degree 2 with coefficients in de Rham cohomology $H_{DR}(X)$".

By (2.2), periodic cyclic homology classes can be thought of as non-commutative generalizations of de Rham cohomology classes. Some information is lost in this generalization: because of the presense of u in the right-hand side of (2.2), what we recover from $HP_\bullet(A)$ is not the de Rham cohomology of X but rather, the de Rham cohomology of the product $X \times \mathbb{P}^\infty$ of X and the infinite projective space $\mathbb{P}^\infty$, where we moreover invert the generator $u \in H^2_{DR}(\mathbb{P}^\infty)$. Thus given a category of realizations Real and a Real-valued refinement of de Rham cohomology, the appropriate target for its non-commutative generalization is not the derived category $\mathcal{D}(\mathsf{Real})$ but the *twisted* 2*-periodic derived category* $\mathcal{D}^{per}(\mathsf{Real})$ obtained by inverting quasiisomorphisms in the category of complexes $M_\bullet$ of objects in Real equipped with an isomorphism $u : M_\bullet \cong M_\bullet(1)[2]$, where we denote $M(n) = M \otimes \mathbb{Z}(n)$, $n \in \mathbb{Z}$.

We note, however, that this causes no problem with the regulator map, since the summation in the right-hand side of (1.1) is the same as in the right-hand side of (2.2). Thus for a Real-valued refinement $H^\bullet_{\mathsf{Real}}(-)$ of de Rham cohomology and any smooth affine algebraic variety $X = \operatorname{Spec} A$, the regulator map (1.1) takes the form

$$K_\bullet(A) \to \operatorname{RHom}^\bullet_{\mathcal{D}^{per}(\mathsf{Real})}(k, HP_\bullet(A)) = \operatorname{RHom}^\bullet_{\mathcal{D}^{per}(\mathsf{Real})}(k, H^\bullet_{\mathsf{Real}}(X)((u))),$$

where k in the right-hand side is the unit object of Real.

Somewhat surprisingly, non-affine algebraic varieties can be included in the above picture with very little additional effort. To do it, it is convenient to use the machinery of differential graded (DG) algebras and DG categories. An excellent overview can be found in [Ke2]; for the convenience of the reader, let us summarize the relevant points.

Roughly speaking, a k-linear DG category is a category $C^\bullet$ whose Hom-sets $C^\bullet(-,-)$ are equipped with a structure of complexes of k-modules in such a way that composition maps are k-linear and compatible with the differentials (for precise definitions, see [Ke2, Section 2]). For any small k-linear DG category $C^\bullet$, one defines a triangulated *derived category of DG modules* $\mathcal{D}(C^\bullet)$ ([Ke2, Section 3]). Any k-linear DG functor $\gamma : C_1^\bullet \to C_2^\bullet$ induces a triangulated functor $\gamma^* : \mathcal{D}(C_2^\bullet) \to \mathcal{D}(C_1^\bullet)$. The functor γ ia a *derived Morita equivalence* if the induced functor γ^* is an equivalence of triangulated categories. It turns out – this mostly due to the work of G. Tabuada and B. Toën, see [Ke2, Section 4] and references therein – that there is a closed model structure on the category of small k-linear DG categories whose weak equivalences are exactly derived Morita equivalences. Denote by $\mathsf{Morita}(k)$ the corresponding homotopy category, that is, the category of "small k-linear DG categories up to a derived Morita equivalence".

Any k-algebra A is a k-linear DG category with one object pt and $\operatorname{Hom}(\mathsf{pt}, \mathsf{pt}) = A$ placed in degree 0, so that we have an embedding $\mathsf{Alg}(k) \to \mathsf{Morita}(k)$ from the category $\mathsf{Alg}(k)$ of associative k-algebras to $\mathsf{Morita}(k)$. Then, as explained in [Ke2, Section 5], Hochschild homology, cyclic homology, periodic cyclic homology and negative cyclic homology extend to functors

$$\mathsf{Morita}(k) \to \mathcal{D}(k).$$

Moreover, so does the algebraic K-theory functor $K^\bullet(-)$, and other "additive invariants" in the sense of [Ke2, Section 5].

In general, a DG category with one object pt is the same thing as an associative unital DG algebra $A^\bullet = \mathrm{Hom}^\bullet(\mathsf{pt}, \mathsf{pt})$. The category of DG algebras over k has a natural closed model structure whose weak equivalences are quasi-isomorphisms, and whose fibrations are surjective maps. The corresponding homotopy category $\mathsf{DG\text{-}Alg}(k)$ is the category of DG algebras "up to a quasi-isomorphism". One shows that a quasiisomorphism between DG algebras is in particular a derived Morita equivalence, so that we have a natural functor

$$\mathsf{DG\text{-}Alg}(k) \to \mathsf{Morita}(k). \tag{2.3}$$

It is not diffucult to show that for every cofibrant DG algebra $A^\bullet$, the individual terms of the complex $A^\bullet$ are flat k-modules. In this case, the Hochschild, cyclic etc. homology of $A^\bullet$ are especially simple – they are given by exactly the same bicomplex (2.1) and its versions as in the case of ordinary algebras. This is manifestly invariant under quasiisomorphisms, so that the Hochschild, cyclic etc. homology obviously descend to functors from $\mathsf{DG\text{-}Alg}(k)$ to the derived category $\mathcal{D}(k)$. The DG category approach shows that there is even more invariance: even if two DG algebras $A_1^\bullet$, $A_2^\bullet$ are not quasiisomorphic but only have isomorphic images in $\mathsf{Morita}(k)$, their Hochschild, cyclic etc. homology is naturally identified. This statement is already non-trivial in the case of usual algebras, see [Lo, Section 1.2].

Definition 2.1. A DG category $T_\bullet \in \mathsf{Morita}(k)$ is *derived-affine* if it lies in the essential image of the functor (2.3).

Remark 2.2. A small k-linear DG category $C^\bullet$ with a finite number of objects is automatically derived-Morita equivalent to a DG algebra $A^\bullet$, thus affine. For example, one can take

$$A^\bullet = \bigoplus_{c,c'} C^\bullet(c, c'),$$

where the sum is taken over all pairs of objects in $\mathbb{C}^\bullet$.

Now, it has been proved ([BV] combined with [Ke1]) that for any quasiseparated quasicompact scheme X over k, there exists a DG algebra $A^\bullet/k$ such that the derived category $\mathcal{D}(X)$ of quasicoherent sheaves on X is equivalent to the derived category $\mathcal{D}(A^\bullet)$,

$$\mathcal{D}(X) \cong \mathcal{D}(A^\bullet),$$

and such a DG algebra $A^\bullet$ is unique up to a derived Morita equivalence, so that we have a canonical functor from the category of algebraic varieties over k to the category $\mathsf{Morita}(k)$. Roughly speaking, any algebraic variety is derived Morita-equivalent to a DG algebra, or, in a succint formulation of [BV], "every algebraic variety is derived-affine".

Moreover, it turns out that the properties of X which are relevant for the present paper are reflected in the properties of a Morita-equivalent DG algebra $A^\bullet$. For example, one introduces the following (see e.g. [KKP]).

Definition 2.3. (i) A DG algebra $A^\bullet/k$ is *proper* if $A^\bullet$ is perfect as an object in the derived category $\mathcal{D}(k)$ of complexes of k-modules.

(ii) A DG algebra $A^\bullet/k$ is *smooth* if $A^\bullet$ is perfect as an object in the derived category $\mathcal{D}(A^{\bullet opp} \otimes A^\bullet)$ of $A^\bullet$-bimodules.

Then $A^\bullet$ is proper, resp. smooth if and only if X is proper, resp. smooth (in the affine case $X = \operatorname{Spec} A$, the second claim is the famous Serre regularity criterion). Moreover, the correspondence $X \mapsto A^\bullet$ is compatible with algebraic K-theory, $K_\bullet(X) \cong K_\bullet(A^\bullet)$, and if the variety X/k is smooth of dimension d, and $d!$ is invertible in k, then the Hochschild homology of such a Morita-equivalent DG algebra $A^\bullet$ is canonically isomorphic to

$$HH_i(A^\bullet) \cong \bigoplus_j H^j(X, \Omega^{i+j}(X)),$$

the so-called "Hodge cohomology" of X, while the periodic cyclic homology $HP_\bullet(A)$ is exactly as in (2.2).

Thus as far as homological invariants are concerned, one can treat DG algebras "up to a derived Morita-equivalence" as non-commutative generalizations of algebraic varieties:

- A non-commutative algebraic variety over k is a DG algebra $A^\bullet$ over k considered as an object of the Tabuada-Toën category $\mathsf{Morita}(k)$.

This is the point of view we will adopt.

3. Hodge-to-de Rham Spectral Sequence

A convenient way to pack all the structures related to Hochschild homology $HH_\bullet(A^\bullet)$ of a DG algebra $A^\bullet/k$ is by considering the equivariant derived category $\mathcal{D}_{S^1}(k)$ of S^1-equivariant constructible sheaves of k-modules on the point pt. Then the claim is that the Hochschild homology complex $CH_\bullet(A^\bullet)$, while *a priori* simply a complex of k-modules, in fact underlies a canonical object $\widetilde{CH}_\bullet(A^\bullet) \in \mathcal{D}_{S^1}(k)$ (loosely speaking, "$CH_\bullet(A^\bullet)$ carries a canonical S^1-action"). The negative cyclic homology appears as S^1-equivariant cohomology

$$H^\bullet_{S^1}(\mathsf{pt}, \widetilde{CH}_\bullet(A^\bullet)),$$

the periodicity map u is the generator of $H^\bullet_{S^1}(\mathsf{pt}) \cong H^\bullet(BS^1)$, and $HP_\bullet(A^\bullet)$ is the localization $HC^-_\bullet(A^\bullet)(u^{-1})$.

Another way to pack the same data is by considering the *filtered derived category* $\mathcal{DF}(k)$ of k-modules of [BBD] – that is, the triangulated category

obtained by considering complexes $V_\bullet$ of k-modules equipped with a descreasing filtration $F^\bullet$ numbered by all integers, and inverting those maps which induce quasiisomorphisms on the associated graded quotients gr^F. This has a "twisted 2-periodic" version $\mathcal{DF}^{per}(k)$, obtained from filtered complexes $V_\bullet$ equipped with an isomorphism $V_\bullet \cong V_\bullet[2](1)$, where (1) means renumbering the filtration: $F^iV(1) = F^{\bullet-1}V$.

Lemma 3.1. *We have*

$$\mathcal{D}_{S^1}(k) \cong \mathcal{DF}^{per}(k).$$

Sketch of the proof. Let us just indicate the equivalence: it sends $V_\bullet \in \mathcal{D}_{S^1}(k)$ to the equivariant cohomology complex $C^\bullet_{S^1}(\mathsf{pt}, V_\bullet)(u^{-1})$, with the (generalized) filtration given by

$$F^iH^\bullet_{S^1}(\mathsf{pt}, V_\bullet)(u^{-1}) = u^iC^\bullet_{S^1}(\mathsf{pt}, V_\bullet),$$

where $u \in C^2_{S^1}(k)$ represents the generator of the equivariant cohomology ring $H^\bullet_{S^1}(\mathsf{pt}, k) \cong k[u]$. □

In the case of the Hochschild homology complex $\widetilde{CH}_\bullet(A^\bullet)$, the corresponding periodic filtered complex is $CP_\bullet(A^\bullet)$, with the filtration given by

$$F^iCP_\bullet((A^\bullet) = u^iCC^-_\bullet(A^\bullet) \subset CP_\bullet(A^\bullet).$$

One can treat $\mathcal{DF}^{per}(k)$ as a very crude "category of realization" Real in the sense of Section 1, or rather, of its periodic derived category $\mathcal{D}^{per}(\mathsf{Real})$. The expected regulator map then takes the form

$$K_\bullet(A^\bullet) \to HC^-_\bullet(A^\bullet) = \mathrm{RHom}^\bullet_{\mathcal{DF}^{per}(k)}(k, HP_\bullet(A)). \tag{3.1}$$

Such a map does indeed exist, see [Lo, Chapter 8]. In general, it is very far from being an isomorphism. The only general result is a theorem of T. Goodwillie [Good] which shows that at least the tangent spaces to both sides are the same. Namely, given an alegbra A with an ideal $I \subset A$, one defines the relative K-theory $K_\bullet(A, I)$ spectrum as the cone of the natural map $K_\bullet(A) \to K_\bullet(A/I)$, and analogously for the cyclic homology functors. Then it has been proved in [Good] that if k is a field of characteristic 0 and $I \subset A$ is a nilpotent ideal, then the map

$$K_\bullet(A, I) \to HC^-(A, I)$$

induced by the regulator map (3.1) is a quasiisomorphism. An analogous statement also holds for DG algebras over k.

While filtered complexes are a very crude approximation to mixed motives, already on this level the smoothness and properness of a DG algebra leads to non-trivial consequences. Namely, a filtered complex gives rise to a spectral sequence. In the case of cyclic homology, it takes the form

$$HH_\bullet(A^\bullet)((u)) \Rightarrow HP_\bullet(A^\bullet), \tag{3.2}$$

where we use the same shorthand as in (2.2). When the DG algebra $A^\bullet/k$ is Morita-equivalent to a smooth algebraic variety X/k, the filtration $F^\bullet$ on $HP_\bullet(A^\bullet) \cong H^\bullet_{DR}(X)((u))$ is just the Hodge filtration on de Rham cohomology, and (3.2) is the usual Hodge-to-de Rham spectral sequence

$$H^p(X, \Omega^q(X)) \Rightarrow H^{p+q}_{DR}(X)$$

tensored with $k((u))$. Because of this, (3.2) in general is also called "Hodge-to-de Rham spectral sequence". Then the following is a partial proof of a general conjecture of M. Kontsevich and Ya. Soibelman [KS].

Theorem 3.2 ([Ka1]). *Assume that $A^\bullet$ is a smooth and proper DG algebra over a field k of characteristic* $\mathsf{char}\, k = 0$. *Assume further that $A^i = 0$ for $i < 0$. Then the Hodge-to-de Rham spectral sequence* (3.2) *degenerates.*

The assumption $A^i = 0$, $i < 0$ is technical (note, however, that it can always be achieved for a DG algebra $A^\bullet$ corresponding to a smooth and proper algebraic variety X/k, see e.g. [O, Theorem 4]).

In the usual commutative case, the Hodge-to-de Rham degeneration statement is well-known and has two proofs. Classically, it follows from the general complex-analytic package of Hodge theory and harmonic forms. An alternative proof by Deligne and Illusie [DI] uses reduction to positive characteristic and p-adic methods. So far, it is only the second technique that has been generalized to the non-commutative case. We will now explain this.

4. Review of Filtered Dieudonné Modules

A p-adic analog of the notion of a mixed Hodge structure has been introduced in 1982 by Fontaine and Lafaille [FL]. Here is the definition.

Definition 4.1. Let k be a finite field of characteristic p, with its Frobenius map, and let W be its ring of Witt vectors, with its canonical lifting φ of the Frobenius map. A *filtered Dieudonné module* over W is a finitely generated W-module M equipped with a decreasing filtration $F^\bullet M$, indexed by all integers and such that $\cap F^i M = 0$, $\cup F^i M = M$, and a collection of Frobenius-semilinear maps $\varphi_i : F^i M \to M$, one for each integer i, such that

(i) $\varphi_i|_{F^{i+1}M} = p\varphi_{i+1}$, and

(ii) the map

$$\sum \varphi_i : \bigoplus_i F^i M \to M$$

is surjective.

We will denote by $\mathcal{FDM}(W)$ the category of filtered Dieudonné modules over W. It is an abelian category. A symmetric tensor product in $\mathcal{FDM}(W)$

is defined in the obvious way, and we have the Tate object $W(1)$ given by: $W(1) = W$ as a W-module, $F^1W(1) = W(1)$, $F^2W(1) = 0$, $\varphi_1 : F^1W(1) \to W(1)$ equal to φ. We also have the derived category $\mathcal{D}(\mathcal{FDM}(W))$.

If a filtered Dieudonné module $M \in \mathcal{FDM}(W)$ is annihilated by p, then (i) of Definition 4.1 insures that the map in (ii) factors through a surjective map

$$\widetilde{\varphi} : \mathsf{gr}_F^\bullet M \to M.$$

Since both sides are k-vector spaces of the same dimension, $\widetilde{\varphi}$ must be an isomorphism. For a general filtered W-module $\langle M, F^\bullet\rangle$, one lets $\widetilde{M}$ be the cokernel of the map

$$\bigoplus_i F^i M \xrightarrow{t-p\,\mathsf{id}} \bigoplus_i F^i M, \tag{4.1}$$

where $t : F^{\bullet+1}M \to F^\bullet M$ is the tautological embedding. Then again, (i) insures that the map $\sum_i \varphi_i$ factors through a map

$$\widetilde{\varphi} : \widetilde{M} \to M \tag{4.2}$$

and this map must be an isomorphism if (ii) were to be satisfied. This allows to generalize the definition of a filtered Dieudonné module: instead of a finitely generated filtered W-module, one can consider a filtered W-module $\langle M, F^\bullet\rangle$ such that M is p-adically complete and complete with respect to the topology induced by $F^\bullet$ (these conditions together with the non-degeneracy conditions $\cap F^i M = 0$, $\cup F^i M = M$ insure that the map (4.1) is injective). Then a *unbounded Dieudonné module* structure on M is given by a Frobenius-semilinear isomorphism $\widetilde{\varphi}$ of the form (4.2).

I do not know whether the category of unbounded filtered Dieudonné modules is still abelian. However, complexes of unbounded filtered Dieudonné modules can be defined in the obvious way, and the correspondence $M \mapsto \widetilde{M}$ sends filtered quasiisomorphisms into quasiisomorphisms, so that we obtain a triangulated derived category $\mathcal{DFDM}(W) \supset \mathcal{D}(\mathcal{FDM}(W))$ and its twisted 2-periodic version $\mathcal{DFDM}^{per}$.

Moreover, one can drop the requirement that the map $\widetilde{\varphi}$ is an isomorphism and allow it to be an arbitrary map. Let us call the resulting objects "weak filtered Dieudonné modules". The category of weak filtered Dieudonné modules is definitely not abelian, but the above procedure still applies: we can invert filtered quasiisomorphisms and obtain triangulated categories denoted $\widetilde{\mathcal{DFDM}}(W)$, $\widetilde{\mathcal{DFDM}}^{per}(W)$. We then have a fully faithful inclusions $\mathcal{DFDM}(W) \subset \widetilde{\mathcal{DFDM}}(W)$, $\mathcal{DFDM}^{per}(W) \subset \widetilde{\mathcal{DFDM}}^{per}(W)$, and their essential images consist of those $M_\bullet$ in $\widetilde{\mathcal{DFDM}}(W)$, resp. $\widetilde{\mathcal{DFDM}}^{per}(W)$ for which the map $\widetilde{\varphi}$ of (4.2) is a quasiisomorphism.

Assume given a algebraic variety X smooth over W, of dimension $d < p$. Then de Rham cohomology $H^\bullet_{DR}(X/W)$ equipped with the filtration induced by the stupid filtration on the de Rham complex has the structure of a complex of generalized filtered Dieudonné modules. If X/W is proper, the groups

$H^i_{DR}(X/W)$ are finitely generated, so that they are filtered Dieudonné modules in the strict sense (and the filtration is then the Hodge filtration). This Dieudonné module structure can be seen explicitly under the following strong additional assumption:

- the Frobenius endomorphism Fr of the special fiber $X_k = X \otimes_W k$ of X/W lifts to a Frobenius-semilinear endomorphism $\widetilde{\mathsf{Fr}} : X \to X$.

Then one checks easily that for any $i \geq 0$, the natural map $\widetilde{\mathsf{Fr}}^* : \Omega^i(X/W) \to \Omega^i(X/W)$ is divisible by p^i. The Dieudonné module structure maps φ_i are induced by the corresponding maps $\frac{1}{p^i}\widetilde{\mathsf{Fr}}^*$. We note that in this special case, the map φ_i sends F^i into F^i. In the general case, the construction is due to G. Faltings [F, Theorem 4.1]; roughly speaking, it uses a comparison theorem which gives a quasiisomorphism

$$H^\bullet_{cris}(X_k) \cong H^\bullet_{DR}(X),$$

where in the left-hand side, we have the cristalline cohomology of the special fiber X_k. The Frobenius endomorphism of X_k induces an endomorphism on cristalline cohomology, and this gives the structure map φ_0. By an additional argument, one shows that $\varphi_0|F^i$ is canonically divisible by p^i, and this gives the other structure maps φ_i (in general, they do not preserve the Hodge filtration $F^\bullet$).

In particular, for any smooth X/W, one has the isomorphism (4.2). Its reduction $\mod p$ is an isomorphism

$$\mathsf{gr}^\bullet_F H^\bullet_{DR}(X_k) \cong \bigoplus_i H^{\bullet-i}(X_k, \Omega^i(X_k)) \cong H^\bullet_{DR}(X_k) \tag{4.3}$$

between Hodge and de Rham cohomology of the special fiber X_k.

If X is affine, this is nothing but the inverse to the Cartier isomorphism, discovered by P. Cartier back in the 1950-ies; as such, it depends only on the special fiber X_k and not on the lifting X/W. In the general case, it has been shown by Deligne and Illusie in [DI] that (4.3) depends on the lifting $X \otimes_W W_2(k)$ of X_k to the second Witt vectors ring $W_2(k) = W(k)/p^2$ (but not on the lifting to higher orders, nor even on the existence of such a lifting).

The absolute cohomology theory associated to the $\mathcal{FDM}$-valued refinement of de Rham cohomology is the *syntomic cohomology* of Fontaine and Messing. As it happens, the functors $\mathrm{RHom}^\bullet(W(-j), -)$ in the category $\mathcal{DFDM}$ are easy to compute explicitly — for any complex $M_\bullet \in \mathcal{DFDM}$, $\mathrm{RHom}^\bullet(W(-j), -)$ is the cone of the natural map

$$F^j M_\bullet \xrightarrow{\mathsf{id} - \varphi_j} M_\bullet.$$

When applied to a smooth proper variety X/W, this gives syntomic cohomology groups $H^\bullet_{synt}(X, \mathbb{Z}_p(j))$. The construction can even be localized with respect

to the Zariski topology on X_k, so that the syntomic cohomology is expressed as hypercohomology of X_k with coefficients in certain canonical complexes of Zariski sheaves, as in [FM].

The existence and properties of the regulator map for the syntomic cohomology have been studied by M. Gros [Gr1, Gr2]. In principle, one can construct the regulator by the standard procedure for "twisted cohomology theories" in the sense of [BO], but there is one serious problem: the filtered Dieudonné module structure on $H^\bullet_{DR}(X)$ only exists if $p > \dim X$. Since the standard procedure works by considering infinite projective spaces and Grassmann varieties, this condition is inevitably broken no matter what p we start with. To circumvent this, Gros had to modify (in [Gr2]) the definition of syntomic cohomology by including additional structures such as the rigid analytic space associated to X/W. The resulting picture becomes extremely complex, and at present, it is not clear whether it can be generalized to non-commutative varieties.

5. FDM in the Non-commutative Case

What we do have for non-commutative varieties is the following result.

Definition 5.1. The *Hochschild cohomology* $HH^\bullet(A^\bullet/R)$ of a DG algebra $A^\bullet$ over a ring R is given by

$$HH^\bullet(A^\bullet/R) = \mathrm{RHom}^\bullet_{A^{\bullet opp}\otimes_R A}(A^\bullet, A^\bullet).$$

Theorem 5.2 ([Ka1]). *Assume given an associative DG algebra $A^\bullet$ over a finite field k. Assume that $A^i = 0$ for $i < 0$. Assume also that $A^\bullet$ is smooth, that it can be lifted to a flat DG algebra $\widetilde{A}^\bullet$ over $W_2(k)$, and that $HH^i(A^\bullet) = 0$ for $i \geq 2p-1$. Then there exists a canonical Cartier-type isomorphism*

$$HH_\bullet(A^\bullet)((u)) \cong HP_\bullet(A^\bullet).$$

Remark 5.3. If a DG algebra $A^\bullet$ is derived Morita-equivalent to a smooth algebraic variety X/k, then we have $HH^i(A^\bullet) = 0$ automatically for $i > 2\dim X$, so that the last condition on $A^\bullet$ in Theorem 5.2 reduces to the condition $p > \dim X$ already mentioned in Section 4.

Remark 5.4. Theorem 3.2 easily follows from Theorem 5.2 by the same dimension argument as in the original proof of Deligne and Illusie in [DI]. The only non-trivial additional input is a beautiful recent theorem of B. Toën [To2] which claims that a smooth and proper DG algebra $A^\bullet$ over a field K comes from a smooth and proper DG algebra $A^\bullet_R$ over a finitely generated subring $R \subset K$, $A^\bullet \cong A^\bullet_R \otimes_R K$. This allows one to reduce problems from $\mathsf{char}\,0$ to $\mathsf{char}\,p$.

Let us give a very rough sketch of how Theorem 5.2 is proved (for more details, see [Ka2], and the complete proof in a slightly different language is in

[Ka1]). As in the commutative story, there are two cases for Theorem 5.2: the easy case when one can construct the Cartier map explicitly, and the general case. The easy case is when $A^\bullet = A$ is concentrated in degree 0, and the algebra A admits a so-called *quasi-Frobenius map*.

Lemma 5.5 ([Ka1]). *For any vector space V over the finite field k of characteristic* $\mathsf{char}\, k = p > 0$*, there is a canonical Frobenius-semilinear isomorphism*

$$\check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, V) \cong \check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, V^{\otimes p}),$$

where $\check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, -)$ means the Tate (co)homology of the group $\mathbb{Z}/p\mathbb{Z}$, the action of $\mathbb{Z}/p\mathbb{Z}$ on V is trivial, and the action on $V^{\otimes p}$ is by the longest cycle permutation $\sigma : V^{\otimes p} \to V^{\otimes p}$. □

Definition 5.6. [[Ka1]] A *quasi-Frobenius map* for an algebra A/k is a $\mathbb{Z}/p\mathbb{Z}$-equivariant algebra map

$$\Phi : A \to A^{\otimes p}$$

which induces the standard isomorphism of Lemma 5.5 on Tate cohomology $\check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, -)$.

If the algebra A admits a quasi-Frobenius map Φ, then the construction of the Cartier isomorphism proceeds as follows. First, recall that for any algebra B equipped with an action of a group G, the *smash product algebra* $B\#G$ is the group algebra $B[G]$ but with the twisted product given by

$$(b_1 \cdot g_1)(b_2 \cdot g_2) = b_1 b_2^{g_1} \cdot g_1 g_2,$$

and one has a canonical decomposition

$$HP_\bullet(B\#G) = \bigoplus_{\langle g \rangle} HP_\bullet(B\#G)_g \tag{5.1}$$

into components numbered by conjugacy classes of elements in G (these components are sometimes called *twisted sectors*). Next, let G be the cyclic group $\mathbb{Z}/p\mathbb{Z}$, and let $\sigma \in G$ be the generator. Then one can show that if the G-action on B is trivial, then

$$HP_\bullet(B\#G)_\sigma \cong \widetilde{HP_\bullet(B)}, \tag{5.2}$$

where $HP_\bullet(B)$ in the right-hand side is equipped with the Hodge filtration, and $\widetilde{M}$ for a filtered group M means the cokernel of the map (4.1), as in Section 4. One the other hand, if we take the p-th power $B^{\otimes p}$ with σ acting by the longest cycle permutation, then one can show that

$$HP_\bullet(B^{\otimes p}\#G)_\sigma \cong HP_\bullet(B). \tag{5.3}$$

Both the isomorphisms (5.2) and (5.3) are completely general and valid for algebras over any ring. So is the decomposition (5.1), which is moreover functorial

with respect to G-equivariant maps. We now apply this to our algebras A and $A^{\otimes p}$ over k, with the G-action as in Lemma 5.5. The quasi-Frobenius map Φ induces a map

$$\varphi : \widetilde{HP_\bullet(A)} \cong HP_\bullet(A\#G)_\sigma \to HP_\bullet(A^{\otimes p}\#G)_\sigma \cong HP_\bullet(A),$$

and since p annihilates $HP_\bullet(A)$, we have

$$\widetilde{HP_\bullet(A)} \cong \mathsf{gr}_F^\bullet HP_\bullet(A) \cong HH_\bullet(A)((u)).$$

The map φ is the Cartier map of Theorem 5.2. One then shows that it is an isomorphism; this requires one to assume that A is smooth.

The general case of Theorem 5.2 is handled by finding a replacement for a quasi-Frobenius map; as far as the cyclic homology is concerned, the argument stays the same. One first shows that for any unital associative algebra A/k, there exists a completely canonical diagram

$$A \xleftarrow{\alpha} Q_\bullet(A) \xrightarrow{\Phi} P_\bullet(A) \xleftarrow{\beta} A^{\otimes p}$$

of DG algebras equipped with an action of $G = \mathbb{Z}/p\mathbb{Z}$ and G-equivariant maps between them. The G action on A and $Q_\bullet(A)$ is trivial. In addition, if A is smooth, the map

$$HP_\bullet(A^{\otimes p}\#G)_\sigma \to HP_\bullet(P_\bullet(A)\#G)_\sigma$$

induced by the map β is an isomorphism (although in general, this isomorphism does not preserve the Hodge filtration). Thus as before, Φ induces a canonical map

$$\overline{\varphi} : HH_\bullet(Q_\bullet(A))((u)) \cong \widetilde{HP_\bullet(Q_\bullet(A))} \to HP_\bullet(A).$$

To construct the Cartier map for the algebra A, it remains to construct a map

$$HH_\bullet(A) \to HH_\bullet(Q_\bullet(A)).$$

To do this, one applies obstruction theory and shows that the map $\alpha : Q_\bullet(A) \to A$ admits a splitting in the category $\mathsf{DG\text{-}Alg}(k)$. The homology of the DG algebra $Q_\bullet(A)$ is given by

$$\mathcal{H}_i(Q_\bullet(A)) = A \otimes \mathsf{St}_i(k), \tag{5.4}$$

where $\mathsf{St}_\bullet(k)$ is the dual k-Steenrod algebra — that is, the dual to the algebra of k-linear cohomological operations in cohomology with coefficients in k. We have $\mathsf{St}_0(k) \cong \mathsf{St}_1(k) \cong k$, and $\mathsf{St}_i(k) = 0$ for $1 < k < 2p-2$. The map $a : Q_\bullet(A) \to A$ is an isomorphism in degree 0. The splitting is constructed degree-by-degree. In degree 1, the obstruction to splitting is exactly the same as the obstruction to lifting the algebra A/k to the ring $W_2(k)$. In any higher degree $i > 1$, the obstruction lies in the Hochschild cohomology group $HH^{2+i}(A \otimes \mathsf{St}_i(k))$, and

this vanishes in the relevant range of degrees by the assumption $HH^i(A) = 0$, $i \geq 2p - 1$.

In the DG algebra case, the construction breaks down since Lemma 5.5 does not have a DG version. Thus one first has to replace a DG algebra $A^\bullet$ with a cosimplicial algebra $\mathcal{A}$ by the Dold-Kan equivalence, and then apply the above construction to $\mathcal{A}$ "pointwise". It is at this point that one has to require $A^i = 0$ for $i < 0$.

Although [Ka1] only provides a Cartier map for DG algebras defined over a finite field k, the same technology should apply to DG algebras over $W = W(k)$ with very little changes, so that for any smooth DG algebra $A^\bullet/W(k)$ with $HH^i(A^\bullet) = 0$ for $i \geq 2p - 1$, one should be able to construct a canonical isomorphism

$$\widetilde{\varphi} : \widetilde{HP_\bullet(A^\bullet)} \cong HP_\bullet(A^\bullet).$$

Equivalently, $HP_\bullet(A^\bullet)$ should carry a filtered Dieudonné module structure (in other words, underlie a canonical object of the periodic derived category $\mathcal{DFDM}^{per}(W)$). One also should be able to check that if $A^\bullet$ is Morita-equivalent to a smooth variety X/W, the comparison isomorphism (2.2) is compatible with the filtered Dieudonné module structures on both sides. However, at present, none of this has been done.

We note that the problem with the regulator map in the p-adic setting mentioned in the end of Section 4 survives in the non-commutative situation. Namely, the standard technology for constructing the regulator map (3.1) ([Lo, Section 8.4]) involves considering the group algebras $k[G]$ for $G = GL_n(A)$, for all $n \geq 1$. As n goes to infinity, the homological dimension of these group alegbras becomes arbitrarily large, and the conditions of Theorem 5.2 cannot be satified.

6. Generalities on Stable Homotopy

The appearance of the Steenrod algebra in (5.4) suggests that the whole story should be related to algebraic topology. This is indeed so. To explain the relation, we need to recall some standard facts on stable homotopy theory.

6.1. Stable homotopy category and homology. Roughly speaking, the *stable homotopy category* StHom is obtained by inverting the suspension functor Σ in the category Hom of pointed CW complexes and homotopy classes of maps between them. Objects of StHom are called *spectra*. A spectrum consists of a collection of pointed CW complexes X_i, $i \geq 0$, and maps $\Sigma X_i \to X_{i+1}$ for all i (in some treatments, these data are required to satisfy additional technical conditions). For the definitions of maps between spectra and homotopies between such maps, we refer the reader to a number of standard references, for example [Ad]. Any CW complex $X \in \mathsf{Hom}$ defines its *suspension spectrum*

$\Sigma^\infty X \in \mathsf{StHom}$ consisting of the suspensions $\Sigma^i X$. For any two CW complexes X, Y, we have

$$\mathrm{Hom}_{\mathsf{StHom}}(\Sigma^\infty X, \Sigma^\infty Y) = \lim_{\overrightarrow{i}}[\Sigma^i X, \Sigma^i Y],$$

where $[-,-]$ denotes the set of homotopy classes of maps.

Any complex of abelian groups $M_\bullet$ defines a spectrum $\mathsf{EM}(M_\bullet)$ called the *Eilenberg-Maclane spectrum of* $M_\bullet$. This is functorial in $M_\bullet$, so that for any commutative ring R, we have a functor

$$\mathsf{EM} : \mathcal{D}(R) \to \mathcal{D}(\mathrm{Ab}) \to \mathsf{StHom},$$

where $\mathcal{D}(R)$ is the derived category of the category of R-modules. This functor has a left-adjoint $H(R) : \mathsf{StHom} \to \mathcal{D}(R)$, known as *homology with coefficients in* R.

The category StHom is a tensor triangulated category. Both functors EM and $H(R)$ are triangulated. Moreover, the homology functor $H(R)$ is a tensor functor – for any two spectra $X, Y \in \mathsf{StHom}$ with smash-product $X \wedge Y$, there exists a functorial isomorphism

$$H(R)(X) \overset{\mathsf{L}}{\otimes}_R H(R)(Y) \cong H(R)(X \wedge Y).$$

The adjoint Eilenberg-Maclane functor EM is pseudotensor – we have a natural map

$$\mathsf{EM}(V_\bullet) \wedge \mathsf{EM}(W_\bullet) \to \mathsf{EM}(V_\bullet \overset{\mathsf{L}}{\otimes}_R W_\bullet)$$

for any two objects $V_\bullet, W_\bullet \in \mathcal{D}(R)$. Thus for any associative ring object $\mathcal{A}$ in StHom, its homology $H(R)(\mathcal{A})$ is a ring object in $\mathcal{D}(R)$, and conversely, for any associative ring object $A_\bullet \in \mathcal{D}(R)$, the Eilenberg-Maclane spectrum $\mathsf{EM}(A_\bullet)$ is a ring object in StHom.

In the homological setting, we know that the structure of a "ring object in $\mathcal{D}(R)$" is too weak, and the right objects to consider are DG algebras over R. To define an analogous notion for spectra is non-trivial, since the traditional topological interpretation of spectra does not behave too well as far as the products are concerned. Fortunately, new models for StHom have appeared more recently, such as for example *S-modules* of [EKMM], *orthogonal spectra* of [MM], or *symmetric spectra* of [HSS]. All these approaches give equivalent results; to be precise, let us choose for example the last one. As shown in [HSS], symmetric spectra form a symmetric monoidal category; denote it by Sym. Then in this paper, a ring spectrum will denote a monoidal object in Sym, and StAlg will denote the category of ring spectra considered up to a homotopy equivalence (formally, this is defined by putting a closed model structure on the category of ring monoidal objects in Sym whose weak equivalences are homotopy equivalences of the underlying symmetric spectra). The homology functor $H(R)$ and the Eilenberg-Maclane functor EM extend to functors

$$H(R) : \mathsf{StAlg} \to \mathsf{DG\text{-}Alg}(R), \qquad \mathsf{EM} : \mathsf{DG\text{-}Alg}(R) \to \mathsf{StAlg}\,.$$

where as in Section 2, $\mathsf{DG\text{-}Alg}(R)$ is the category of DG algebras over R considered up to a quasiisomorphism.

6.2. Equivariant categories. For any compact group G, a pointed "G-CW complex" is a pointed CW complex X equipped with a continuous action of G such that the fixed-point subset $X^g \subset X$ is a pointed subcomplex for any $g \in G$. We will denote by $\mathsf{Hom}(G)$ the category of pointed G-CW complexes and G-equivariant homotopy classes of G-equivariant maps between them. We note that for any closed subgroup $H \subset G$, sending X to the fixed-point subspace $X^H \subset X$ gives a well-defined functor

$$\mathsf{Hom}(G) \to \mathsf{Hom}\,.$$

This functor is representable in the following sense: for any $X \in \mathsf{Hom}(G)$, we have a homotopy equivalence

$$X^H \cong \mathrm{Maps}_G([G/H]_+, X), \tag{6.1}$$

where $\mathrm{Maps}_G(-,-)$ means the space of G-equivariant maps with its natural topology, and $[G/H]_+$ is the pointed G-CW complex obtained by adding a (disjoint) marked point to the quotient G/H with the induced topology and G-action.

To define a stable version of the category $\mathsf{Hom}(G)$, one could again simply invert the suspension functor. However, there is a more interesting alternative: by definition, n-fold suspension Σ^n is the smash-product with an n-sphere, and in the equivariant setting, one can allow the sphere to carry a non-trivial G-action. The corresponding equivariant stable category has been constructed in [LMS]; it is known as the *genuine G-equivariant stable homotopy category* $\mathsf{StHom}(G)$. To define it, one needs to fix a real representation U of the group G which is equipped with a G-invariant inner product and contains every finite-dimensional inner-product representation countably many times; this is called a "complete G-universe". Then a genuine G-equivariant spectrum is a collection of G-CW complexes $X(V)$, one for each finite-dimensional G-invariant inner-product subspace $V \subset U$, and maps $S^W \wedge X(V) \to X(V \oplus W)$, one for each inner-product G-invariant subspace $V \oplus W \subset U$, where S^V is the one-point compactification of the underlying topological space of the representation V, with its natural G-action. As in the non-equivariant case, $\mathsf{StHom}(G)$ is a tensor triangulated category. We have a natural suspension spectrum functor $\Sigma^\infty : \mathsf{Hom}(G) \to \mathsf{StHom}(G)$, and for any two objects $X, Y \in \mathsf{Hom}(G)$, we have

$$\mathrm{Hom}_{\mathsf{StHom}(G)}(\Sigma^\infty X, \Sigma^\infty Y) = \lim_{V \subsetneq U} [S^V \wedge X, S^V \wedge Y]_G,$$

where $[-,-]_G$ is the set of G-homotopy classes of G-equivariant maps, and the limit is over all the finite-dimensional G-invariant inner-product subspaces $V \subset U$. The category $\mathsf{StHom}(G)$ does depend on U, but this is not too drastic:

all complete G-universes are isomorphic, and for any isomorphism $U \cong U'$ between complete G-universes, there is a "change of universe" functor which is an equivalence between the corresponding versions of $\mathsf{StHom}(G)$.

Forgetting the G-action gives a natural forgetful functor $\mathsf{StHom}(G) \to \mathsf{StHom}$, and equipping a spectrum with a trivial G-action gives an embedding $\mathsf{StHom} \to \mathsf{StHom}(G)$. Thus for any $X \in \mathsf{StHom}$ and $Y \in \mathsf{StHom}(G)$, we have a functorial smash product $X \wedge Y \in \mathsf{StHom}(G)$. This has an adjoint: for any $X, Y \in \mathsf{StHom}(G)$, we have a natural spectrum $\mathrm{Maps}_G(X, Y) \in \mathsf{StHom}$ such that for any $Z \in \mathsf{StHom}$, there is a functorial isomorphism

$$\mathrm{Hom}_{\mathsf{StHom}(G)}(Z \wedge X, Y) \cong \mathrm{Hom}_{\mathsf{StHom}}(Z, \mathrm{Maps}_G(X, Y)).$$

For any closed subgroup $H \subset G$ and any $X \in \mathsf{StHom}(G)$, one can extend (6.1) and define the fixed point spectrum X^H by the same formula,

$$X^H = \mathrm{Maps}_G(\Sigma^\infty [G/H]_+, X). \tag{6.2}$$

However, this does not commute with the suspension spectrum functor Σ^∞. In [LMS], a second fixed-points functor is introduced, called the *geometric fixed points functor* and denoted Φ^H. It does commute with Σ^∞, and also commutes with smash products, so that there are functorial isomorphisms

$$\Phi^H(\Sigma^\infty X) \cong \Sigma^\infty X^H, \qquad \Phi^H(X \wedge Y) \cong \Phi^H(X) \wedge \Phi^H(Y)$$

for any $X, Y \in \mathsf{StHom}(G)$. For any $X \in \mathsf{StHom}(G)$, there exists a canonical map

$$\mathsf{can} : X^H \to \Phi^H(X), \tag{6.3}$$

functorial in X. Moreover, let $N_H \subset G$ be the normalizer of the subgroup $H \subset G$, and let $W_H = N_H/H$ be the quotient. Then Φ^H can be extended to a functor

$$\widehat{\Phi}^H : \mathsf{StHom}(G) \to \mathsf{StHom}(W_H),$$

and the same is true for the usual fixed-points functor $X \mapsto X^H$ of (6.2). The map can of (6.3) then lifts to a map of W_H-equivariant spectra. Here if $\mathsf{StHom}(G)$ is defined on a complete G-universe U, then $\mathsf{StHom}(W_H)$ should be defined on the complete W_H-universe U^H. The functor $\widehat{\Phi}^H$ has a right-adjoint which is a fully faithful embedding $\mathsf{StHom}(W_H) \to \mathsf{StHom}(G)$ (for example, if $H = G$, then this is the trivial embedding $\mathsf{StHom} \to \mathsf{StHom}(G)$).

6.3. Mackey functors. Assume from now on that the compact group G is a finite group with discrete topology. It is not difficult to extend the homology functor $H(R)$ to a functor

$$H(R) : \mathsf{StHom}(G) \to \mathcal{D}(G, R)$$

with values in the derived category of $R[G]$-modules. However, this version of equivariant homology looses a lot of information such as fixed points. A more

natural target for equivariant homology is the category of the so-called *Mackey functors*. To define them, one considers an additive category $\mathcal{B}_G$ whose objects are G-orbits G/H for all subgroups $H \subset G$, and whose Hom-groups are given by

$$\begin{aligned}\mathcal{B}^G([G/H_1],[G/H_2]) &= \mathrm{Hom}_{\mathsf{StHom}(G)}(\Sigma^\infty[G/H_1]_+, \Sigma^\infty[G/H_2]_+) = \\ &= \pi_0(\mathrm{Maps}_G(\Sigma^\infty[G/H_1]_+, \Sigma^\infty[G/H_2]_+)).\end{aligned} \tag{6.4}$$

An *R-valued G-Mackey functor* ([Dr], [Li], [tD], [M1]) is an additive functor from $\mathcal{B}_G$ to the category of R-modules. The category of such functors is an abelian category, denoted $\mathcal{M}(G,R)$.

More explicitly, for any subgroups $H_1, H_2 \subset G$, one can consider the groupoid $\mathcal{Q}([G/H_1],[G/H_2])$ of diagrams $[G/H_1] \leftarrow S \rightarrow [G/H_2]$ of finite sets equipped with a G-action, and isomorphisms between such diagrams. Then disjoint union turns these groupoids into symmetric monoidal categories, the Cartesian product turns the collection $\mathcal{Q}(-,-)$ into a 2-category with objects $[G/H]$, and it seems very likely that the mapping spectra $\mathrm{Maps}_G(\Sigma^\infty[G/H_1]_+, \Sigma^\infty[G/H_2]_+)$ are in fact obtained from the classifying spaces $|\mathcal{Q}([G/H_1],[G/H_2])|$ of symmetric monoidal groupoids $\mathcal{Q}([G/H_1],[G/H_2])$ by group completion. At present, this has not been proved ([M2]); however, the corresponding isomorphism is well-known at the level of π_0: we have

$$\pi_0(\mathrm{Maps}_G(\Sigma^\infty[G/H_1]_+, \Sigma^\infty[G/H_2]_+)) \cong \pi_0(\Omega B|\mathcal{Q}([G/H_1],[G/H_2])|),$$

so that the groups $\mathcal{B}_G(-,-)$ are given by

$$\mathcal{B}^G([G/H_1],[G/H_2]) = \mathbb{Z}[\mathrm{Iso}(\mathcal{Q}([G/H_1],[G/H_2]))]/\{[S_1 \coprod S_2] - [S_1] - [S_2]\}, \tag{6.5}$$

where Iso means the set of isomorphism classes of objects.

For any $X \in \mathsf{StHom}(G)$, individual homology groups $H_i(R)(X)$ can be equipped with a natural structure of a Mackey functor in such a way that $H_i(R)(X)([G/H]) \cong H_i(R)(X^H)$, $H \subset G$ (for more details, see [M1]). To collect these into a single homology functor $H(R)$, one has to work out a natural derived version of the abelian category $\mathcal{M}(G,R)$. This has been done recently in [Ka3]. Roughly speaking, instead of π_0 in (6.4), one should the chain homology complexes $C_\bullet(-,\mathbb{Z})$ of the corresponding spectra, and one should set

$$\mathcal{B}^G_\bullet([G/H_1],[G/H_2]) = C_\bullet(\mathrm{Maps}_G(\Sigma^\infty[G/H_1]_+, \Sigma^\infty[G/H_2]_+), \mathbb{Z}).$$

In practice, one replaces this with complexes which compute the homology of the spectra obtained by group completion from the symmetric monoidal groupoids $\mathcal{Q}([G/H_1],[G/H_2])$. This can be computed explicitly, so that the complexes $\mathcal{B}^G_\bullet(-,-)$ introduced in [Ka3, Section 3] are given by an explicit formula, and spectra are not mentioned at all. One then shows that the collection

$\mathcal{B}^G_\bullet(-,-)$ is an A_∞-category in a natural way, and one defines the triangulated category $\mathcal{DM}(G,R)$ of *derived R-valued G-Mackey functors* as the derived category of A_∞-functors from $\mathcal{B}^G_\bullet$ to the category of complexes of R-modules.

In general, the category $\mathcal{DM}(G,R)$ turns out to be different from the derived category $\mathcal{D}(\mathcal{M}(G,R))$ (although both contain the abelian category $\mathcal{M}(G,R)$ as a full subcategory). On the level of slogans, one can hope that the category $\mathcal{DM}(G,R)$ is the "brave new product" of the category $\mathsf{StHom}(G)$ and the derived category $\mathcal{D}(R)$ of R-modules, taken over the non-equivariant stable homotopy category StHom, so that we have a diagram

$$\begin{array}{ccccc} \mathcal{D}(\mathcal{M}(G,R)) & \longrightarrow & \mathcal{DM}(G,R) & \longrightarrow & \mathsf{StHom}(G) \\ & & \downarrow & & \downarrow \\ & & \mathcal{D}(R) & \longrightarrow & \mathsf{StHom}, \end{array}$$

where the square is Cartesian in some "brave new" sense. On a more mundane level, it is expected that the triangulated category $\mathcal{DM}(G,R)$ reflects the structure of the category $\mathsf{StHom}(G)$ in the following way.

(i) There exists a symmetric tensor product $-\otimes-$ on the triangulated category $\mathcal{DM}(G,R)$, and for any subgroup $H\subset G$, we have natural triangulated fixed-point functors $\Phi^H,\Psi^H:\mathcal{DM}(G,R)\to\mathcal{D}(R)$.

(ii) There exists a natural triangulated equivariant homology functor

$$H_G(R):\mathsf{StHom}(G)\to\mathcal{DM}(G,R)$$

and natural functorial isomorphisms

$$\begin{aligned} \Phi^H(H_G(R)(X)) &\cong H(R)(\Phi^H(X)),\\ \Psi^H(H_G(R)(X)) &\cong H(R)(X^H),\\ H_G(X\wedge Y) &\cong H_G(X)\otimes H_G(Y) \end{aligned}$$

for any $X,Y\in\mathsf{StHom}(G)$, $H\subset G$.

In fact, most of these statements has been proved in [Ka3], although only for the so-called "Spanier-Whitehead category", the full triangulated subcategory in $\mathsf{StHom}(G)$ spanned by the suspension spectra of finite CW complexes (the only thing not proved is the compatibility $\Psi^H(H_G(R)X)\cong H(R)(X^H)$ which requires one to leave the Spanier-Whitehead category). It has been also shown in [Ka3] that as in the case of spectra, the fixed point functor Φ^H extends to a functor

$$\widehat{\Phi}^H:\mathcal{DM}(G,R)\to\mathcal{DM}(W_H,R) \tag{6.6}$$

with a fully faithful right-adjoint. These fixed-points functors allow one to give a very explicit description of the category $\mathcal{DM}(G,R)$. Namely, let $I(G)$ be the set of conjugacy classes of subgroups in G, and for any $c\in I(G)$, let

$$\mathcal{DM}_c(G,R)\subset\mathcal{DM}(G,R)$$

be the full subcategory of such $M \in \mathcal{DM}(G,R)$ that $\Phi^H(M) = 0$ unless $H \subset G$ is in the class c.

Proposition 6.1 ([Ka3]). *For any $c \in I(G)$, $\mathcal{DM}_c(G,R) \subset \mathcal{DM}(G,R)$ is an admissible triangulated subcategory, and for any subgroup $H \subset G$ is a subgroup in the class c, the functor $\widehat{\Phi}^H$ of* (6.6) *induces an equivalence*

$$\widehat{\varphi}^H : \mathcal{DM}_c(G,R) \cong \mathcal{D}(W_H, R).$$

Moreover, equip $I(G)$ with the partial order given by inclusion. Then it has been shown in [Ka3] that unless $c \leq c'$, $\mathcal{DM}_c(G,R)$ is left-orthogonal to $\mathcal{DM}_{c'}(G,R)$, so that $\mathcal{DM}_c(G,R)$, $c \in I(G)$ form a semiorhtogonal decomposition of the triangulated category $\mathcal{DM}(G,R)$ indexed by the partially ordered set $I(G)$ (for generalities on semiorthogonal decompositions, see [BK]). To describe the gluing data between the pieces of this semiorthogonal decomposition, one introduces the following.

Definition 6.2. Assume given a finite group G and a module V over $R[G]$. The *maximal Tate cohomology* $\check{H}^\bullet_{max}(G,V)$ is given by

$$\check{H}^\bullet_{max}(G,V) = \mathrm{RHom}^\bullet_{\mathcal{D}^b(G/R)/\,\mathsf{Ind}}(R,V),$$

where $\mathrm{RHom}^\bullet$ is computed in the quotient $\mathcal{D}^b(G,R)/\,\mathsf{Ind}$ of the bounded derived category $\mathcal{D}^b(G,R)$ by the full saturated triangulated subcategory $\mathsf{Ind} \subset \mathcal{D}^b(G,R)$ spanned by representrations $\mathsf{Ind}_G^H(W)$ induced from a representation W of a subgroup $H \subset G$, $H \neq G$.

Then for any two subgroups $H \subset H' \subset G$ with conjucacy classes $c, c' \in I$, $c \leq c'$, the gluing functor between $\mathcal{DM}_c(G,R)$ and $\mathcal{DM}_{c'}(G,R)$ is expressed in terms of maximal Tate cohomology of the group W_H and its various subgroups.

This description turns out to be very effective because maximal Tate cohomology often vanishes. For example, if the order of the group G is invertible in R, $\check{H}^\bullet_{max}(G,V) = 0$ for any $R[G]$-module V, and the category $\mathcal{DM}(G,R)$ becomes simply the direct sum of the categories $\mathcal{DM}_c(G,R) \cong \mathcal{D}(W_H,R)$ (for the abelian category $\mathcal{M}(G,R)$, a similar decomposition theorem has been proved some time ago by J. Thevenaz [Th]). On the other hand, if R is arbitrary but the group $G = \mathbb{Z}/n\mathbb{Z}$ is cyclic, then $\check{H}^\bullet_{max}(G,V) = 0$ for any V unless $n = p$ is prime, in which case $\check{H}^\bullet_{max}(G,V)$ reduces to the usual Tate cohomology $\check{H}^\bullet(G,V)$.

7. Cyclotomic Traces

Returning to the setting of Theorem 5.2, we can now explain the appearance of the Steenrod algebra in (5.4): up to a quasiisomorphism, the DG algebra $Q_\bullet(A)$ of (5.4) is in fact given by

$$Q_\bullet(A) = H(k)(\mathsf{EM}(A))^{k^*},$$

where the k^*-invariants are taken with respect to the natural action of the multiplicative group k^* of the finite field k induced by its action on k.

In particular, this shows that it is not necessary to use dimension arguments to construct a splitting $A \to Q_\bullet(A)$ of the augmentation map $Q_\bullet(A) \to A$. For example, if we are given a ring spectrum $\mathcal{A}$ with homology DG algebra $A^\bullet = H(k)(\mathcal{A})$, then a canonical map

$$A^\bullet = H(k)(\mathcal{A}) \to H(k)(\mathsf{EM}(H(k)(\mathcal{A}))) \tag{7.1}$$

exists simply by adjunction, and being canonical, it is in particular k^*-invariant. Thus for any DG algebra of the form $A^\bullet = H(k)(\mathcal{A})$, the same procedure as in the proof of Theorem 5.2 allows one to construct a Cartier map. However, in this case one can do much more – namely, one can compare the homological story with the theory of *cyclotomic traces* and *topological cyclic homology* known in algebraic topology. Let us briefly recall the setup (we mostly follow the very clear and concise exposition in [HM]).

7.1. Topological cyclic homology. For any unital associative algebra A over a ring k, the Hochschild homology complex $CH_\bullet(A)$ of Section 2 is in fact the standard complex of a simplicial k-module $A_\# \in \Delta^{opp}k$-mod. *Topological Hochschild homology* is a version of this construction for ring spectra. It was originally introduced by Bökstedt [Bo] long before the invention of symmetric spectra, and used the technology of "functors with a smash product". In the language of symmetric spectra, one starts with a unital associative ring spectrum $\mathcal{A}$, and one defines a simplicial spectrum $\mathcal{A}_\#$ by exactly the same formula as in the algebra case. The terms of $\mathcal{A}_\#$ are the iterated smash products $\mathcal{A} \wedge \cdots \wedge \mathcal{A}$, and the face and degeneracy maps are obtained from the multiplication and the unit map in $\mathcal{A}$. Then one sets

$$\mathrm{THH}(\mathcal{A}) = \mathsf{hocolim}_{\Delta^{opp}} \mathcal{A}_\#.$$

As in the algebra case, this spectrum is equipped with a canonical S^1-action, but in the topological setting this means much more: one shows that $\mathrm{THH}(\mathcal{A})$ actually underlies a canonical S^1-equivariant spectum $\mathrm{THH}(\mathcal{A}) \in \mathsf{StHom}(S^1)$.

However, this is not the end of the story. Note that the finite subgroups in S^1 are the cyclic groups $C_n = \mathbb{Z}/n\mathbb{Z} \subset S^1$ numbered by integers $n \geq 1$, and for every n, we have $S^1/C_n \cong S^1$. Fix a system of such isomorphisms which are compatible with the embeddings $C_n \subset C_{nm} \subset S^1$, $n, m \geq 1$, and fix a compatible system of isomorphisms $U^{C_n} \cong U$, where U is the complete S^1-universe used to define $\mathsf{StHom}(S^1)$. Then the following notion has been introduced in [BM].

Definition 7.1. A *cyclotomic structure* on an S^1-equivariant spectrum T is given by a collection of S^1-equivariant homotopy equivalences

$$r_n : \widehat{\Phi}^{C_n} T \cong T,$$

one for each finite subgroup $C_n \subset S^1$, such that $r_1 = \mathsf{id}$ and $r_n \circ r_m = r_{nm}$ for any two integer $n, m > 1$.

Remark 7.2. Here it is tacitly assumed that one works with specific model of equivariant spectra, so that a spectrum means more than just an object of the triangulated category $\mathsf{StHom}(S^1)$; moreover, the functors $\widehat{\Phi}^{C_n}$ are composed with the change of universe functors so that we can treat them as endofunctors of $\mathsf{StHom}(S^1)$. Please refer to [BM] or [HM] for exact definitions.

Example 7.3. *Assume given a CW complex X, and let $LX = \mathrm{Maps}(S^1, X)$ be its free loop space. Then for any finite subgroup $C \subset S^1$, the isomorphism $S^1 \cong S^1/C$ induces a homeomorphism*

$$\mathrm{Maps}(S^1, X)^C = \mathrm{Maps}(S^1/C, X) \cong \mathrm{Maps}(S^1, X),$$

and these homeomorphism provide a canonical cyclotomic structure on the suspension spectrum $\Sigma^\infty LX$.

For any S^1-equivariant spectrum T and a pair of integers $r, s > 1$, one has a natural non-equivariant map

$$F_{r,s} : T^{C_{rs}} \to T^{C_r}.$$

On the other hand, assume that T is equipped with a cyclotomic structure. Then we have a natural map

$$R_{r,s} : T^{C_{rs}} \cong (T^{C_s})^{C_r} \xrightarrow{\mathsf{can}} (\widehat{\Phi}^{C_s} T)^{C_r} \xrightarrow{r_s} T^{C_r},$$

where can is the canonical map (6.3), and r_s comes from the cyclotomic structure on T. To pack together the maps $F_{r,s}$, $R_{r,s}$, it is convenient to introduce a small category $\mathbb{I}$ whose objects are all integers $n \geq 1$, and whose maps are generated by two maps $F_r, R_r : n \to m$ for each pair m, $n = rm$, $r > 1$, subject to the relations $F_r \circ F_s = F_{rs}$, $R_r \circ R_s = R_{rs}$, $F_r \circ R_s = R_s \circ F_r$. Then the maps $T_{r,s}$, $F_{r,s}$ turn the collection T^{C_n}, $n \geq 1$ into a functor $\widetilde{T}$ from $\mathbb{I}$ to the category of spectra.

Definition 7.4. The *topological cyclic homology* $\mathrm{TC}(T)$ of a cyclotomic spectrum T is given by

$$\mathrm{TC}(T) = \mathsf{holim}_{\mathbb{I}} \widetilde{T}.$$

Given a ring spectrum $\mathcal{A}$, Bökstedt and Madsen equip the S^1-equivariant spectrum $THH_\bullet(\mathcal{A})$ with a canonical cyclotomic structure. *Topological cyclic homology* $\mathrm{TC}(\mathcal{A})$ is then given by

$$\mathrm{TC}(\mathcal{A}) = \mathrm{TC}(\mathrm{THH}(\mathcal{A})).$$

Further, they construct a canonical *cyclotomic trace map*

$$K(\mathcal{A}) \to \mathrm{TC}(\mathcal{A}) \tag{7.2}$$

from the K-theory spectrum $K(\mathcal{A})$ to the topological cyclic homology spectrum.

The topological cyclic homology functor $\mathrm{TC}(\mathcal{A})$ and the cyclotomic trace were actually introduced by Bökstedt, Hsiang and Madsen in [BHM]; the more convenient formulation using cyclotomic spectra appeared slightly later in [BM]. Starting with [BHM], it has been proved in many cases that the cyclotomic trace map becomes a homotopy equivalence after taking profinite completions of both sides of (7.2). Moreover, in [Mc] MacCarthy generalized Goodwillie's Theorem and proved that after pro-p completion at any prime p, the cyclotomic trace gives an equivalence of the relative groups $\widehat{K(\mathcal{A},I)}_p \cong \widehat{\mathrm{TC}(\mathcal{A},I)}_p$, where $I \subset \mathcal{A}$ is a nilpotent ideal.

7.2. Cyclotomic complexes. To define a homological analog of cyclotomic spectra, one needs to replace S^1-equivariant spectra with derived Mackey functors. The machinery of [Ka3] does not apply directly to non-discrete groups, since this would require treating the groupoids $\mathcal{Q}(-,-)$ of Subsection 6.2 as topological groupoids. However, for finite subgroups $C_1, C_2 \subset S^1$, the category $\mathcal{Q}([S^1/C_1],[S^1/C_2])$ is still discrete. Thus one can define a restricted version of derived S^1-Mackey functors by discarding the only infinite closed subgroup in S^1 (which is S^1 itself). This is done in [Ka4]. The category $\mathcal{DM}\Lambda(R)$ of *R-valued cyclic Mackey functors* introduced in that paper has the following features.

(i) For every proper finite subgroup $C = C_n \subset S^1$, $n > 1$, there is a fixed-point functor $\widehat{\Phi}_n : \mathcal{DM}\Lambda(R) \to \mathcal{DM}\Lambda(R)$ whose right-adjoint functor $\widehat{\iota}_n : \mathcal{DM}\Lambda(R) \to \mathcal{DM}\Lambda(R)$ is a full embedding. Moreover, there are canonical isomorphisms $\widehat{\Phi}_n \circ \widehat{\Phi}_m \cong \widehat{\Phi}_{mn}$.

(ii) Let $\mathcal{D}_{S^1}(R)$ be the equivariant derived category of Section 3. Then there is a full embedding $\iota_1 : \mathcal{D}_{S^1}(R) \to \mathcal{DM}\Lambda(R)$ with a left-adjoint $\Phi_1 : \mathcal{DM}\Lambda(R) \to \mathcal{D}_{S^1}(R)$.

(iii) The images $\mathcal{DM}\Lambda_n(R)$ of the full embeddings $\iota_n = \widehat{\iota}_N \circ \iota_1 : \mathcal{D}_{S^1}(R) \to \mathcal{DM}\Lambda(R)$, $n \geq 1$, generate the triangulated category $\mathcal{DM}\Lambda(R)$, and $\mathcal{DM}\Lambda_n(R) \subset \mathcal{DM}\Lambda(R)$ is left-orthogonal to $\mathcal{DM}\Lambda_m(R) \subset \mathcal{DM}\Lambda(R)$ unless $n = mr$ for some integer $r \geq 1$.

Thus as in the finite group case of [Ka3], the subcategories $\mathcal{DM}\Lambda_n(R) \subset \mathcal{DM}\Lambda(R)$ form a semiorthogonal decomposition of the category $\mathcal{DM}\Lambda(R)$. The gluing data between $\mathcal{DM}\Lambda_{mr}(R)$ and $\mathcal{DM}\Lambda_r(R)$ can be expressed in terms of the maximal Tate cohomology $\check{H}^{\bullet}_{max}(C_m,-)$ of the cyclic group $C_m = \mathbb{Z}/m\mathbb{Z}$. For any $n \geq 1$, let $\overline{\Phi}_n : \mathcal{DM}\Lambda(R) \to \mathcal{D}(R)$ be the composition of the left-adjoint $\Phi_n = \Phi_1 \circ \widehat{\Phi}_n$ to ι_n and the forgetful functor $\mathcal{D}_{S^1}(R) \to \mathcal{D}(R)$; then the functors $\overline{\Phi}_n$ play the role of fixed points functors Φ^H. There are also functors $\Psi_n : \mathcal{DM}\Lambda(R) \to \mathcal{D}_{S^1}(R)$ analogous to the functors Ψ^H. The homology functor $H(R)$ extends to a functor

$$H_{S^1}(R) : \mathsf{StHom}(S^1) \to \mathcal{DM}\Lambda(R),$$

and we have functorial isomorphisms

$$\overline{\Phi}_n(H_{S^1}(R)(T)) \cong H(R)(\Phi^{C_n}(T)), \qquad \Psi_n(H_{S^1}(R)(T)) \cong H(R)(T^{C_n})$$

for every $n \geq 1$ and every $T \in \mathsf{StHom}(S^1)$.

Another category defined in [Ka4] is a triangulated category $\mathcal{D}\Lambda\mathrm{R}(R)$ of R-*valued cyclotomic complexes*. Essentially, a cyclotomic complex $M_\bullet \in \mathcal{D}\Lambda\mathrm{R}(R)$ is a cyclic Mackey functor $M_\bullet$ equipped with a system of compatible quasiisomorphisms

$$\widehat{\Phi}_n M_\bullet \cong M_\bullet,$$

as in Definition 7.1 (although as in Remark 7.2, the precise definition is different for technical reasons). The homology functor $H_{S^1}(R) : \mathsf{StHom}(S^1) \to \mathcal{DM}\Lambda(R)$ extends to a functor from the category of cyclotomic spectra to the category $\mathcal{D}\Lambda\mathrm{R}(R)$. Moreover, all the constructions used in the definition of topological cyclic homology make sense for cyclotomic complexes, so that one has a natural functor

$$\mathrm{TC} : \mathcal{D}\Lambda\mathrm{R}(R) \to \mathcal{D}(R)$$

and a functorial isomorphism

$$\mathrm{TC}(H_{S^1}(R)(T)) \cong H(R)(\mathrm{TC}(T)) \tag{7.3}$$

for every cyclotomic spectrum T.

7.3. Comparison theorem. We can now formulate the comparison theorem relating Dieudonné modules and cyclotomic complexes. We introduce the following definition.

Definition 7.5. A *generalized filtered Dieudonné module* M over a commutative ring R is an R-module M equipped with a decreasing filtration $F^\bullet M$ and a collection of maps

$$\varphi^p_{i,j} : F^i M \to M/p^j,$$

one for every integers i, j, $j \geq 1$, and a prime p, such that

$$\varphi^p_{i,j+1} = \varphi^p_{i,j} \mod p^j, \quad \varphi^p_{i,j}|_{F^{i+1}M} = p\varphi^p_{i,j}.$$

For any integer i, we define the generalized filtered Dieudonné module $R(i)$ as R with the filtration $F^i R(i) = R$, $F^{i+1}R(i) = 0$, and $\varphi^p_{i,j} = p^i\,\mathsf{id}$ for any p and j. Generalized filtered Dieudonné modules in the sense of Defintion 7.5 do not form an abelian category; however, by inverting the filtered quasiisomorphisms, we can still construct the derived category $\mathcal{DFDM}_g(R)$ and its twisted 2-periodic version $\mathcal{DFDM}_g^{per}(R)$.

Definition 7.5 generalizes (4.1) in that it collects together the data for all primes p. Note, however, that one can rephrase Definition 7.5 by putting together all the maps $\varphi^p_{i,j}$, $j \geq 1$, into a single map

$$\widehat{\varphi}^p_i : F^i M \to \widehat{(M)}_p$$

into the pro-p completion $\widehat{(M)}_p$ of the module M. Then if $R = \mathbb{Z}_p$ and M is finitely generated over $\mathbb{Z}_p$, we have

$$\widehat{(M)}_p \cong M, \quad \widehat{(M)}_l = 0 \text{ for } l \neq p,$$

so that for such an M, the extra data imposed onto M in Definition 7.5 and in Definition 4.1 are the same. In general, for any prime p, we have a fully faithful embedding

$$\widetilde{\mathcal{DFDM}}(\mathbb{Z}_p) \subset \mathcal{DFDM}_g(\mathbb{Z}),$$

where $\widetilde{\mathcal{DFDM}}(\mathbb{Z}_p)$ is as in Section 4, and similarly for the periodic categories. The essential images of these embeddings are spanned by complexes which are pro-p complete as complexes of abelian groups. Note, however, that what appears here are *weak* filtered Dieudonné modules. The requirement that the map (4.2) is a quasiisomorphism can be additionally imposed at each individual prime p; I do not know whether it is useful to impose it in the universal category $\mathcal{DFDM}_g(\mathbb{Z})$.

Here is then the main comparison theorem of [Ka4].

Theorem 7.6 ([Ka4, Section 5]). *For any commutative ring R, there is a canonical equivalence of categories*

$$\mathcal{DFDM}_g^{per}(R) \cong \mathcal{D}\Lambda\mathrm{R}(R).$$

Thus the category $\mathcal{D}\Lambda\mathrm{R}(R)$ of cyclotomic complexes over R admits an extremely simple linear-algebraic description. Roughly speaking, the reason for this is the vanishing of maximal Tate cohomology $\check{H}^\bullet(\mathbb{Z}/n\mathbb{Z}, -)$ for non-prime n mentioned at the end of Subsection 6.3. Due to this vanishing, the only nontrivial gluing between the pieces $\mathcal{DM}\Lambda_m(R)$, $\mathcal{DM}\Lambda_n(R)$ of the semiorthogonal decomposition of the category $\mathcal{DM}\Lambda(R)$ of cyclic Mackey functors occurs when $n = mp$ for a prime p (and this gluing is described by the Tate cohomology of the group $\mathbb{Z}/p\mathbb{Z}$). The gluing data provide the maps $\widehat{\varphi}_i^p$ in the equivalence of Theorem 7.6; the periodic filtered complex comes from the equivalence $\mathcal{D}_{S^1}(R) \cong \mathcal{DF}^{per}(R)$ of Lemma 3.1. These are the main ideas of the proof.

Moreover, there is a second comparison theorem which expresses topological cyclic homology in terms of generalized Dieudonné modules.

Theorem 7.7 ([Ka4, Section 6]). *Under the equivalence of Theorem* 7.6*, there is functorial isomorphism*

$$\widehat{\mathrm{TC}(M_\bullet)}_f \cong \widehat{\mathrm{RHom}^\bullet(R, M_\bullet)}_f \tag{7.4}$$

for any $M_\bullet \in \mathcal{D}\Lambda\mathrm{R}(R)$, where $R = R(0) \in \mathcal{DFDM}_g^{per}(R)$ is the trivial generalized filtered Dieudonné module, and $\widehat{(-)}_f$ stands for profinite completion.

Remark 7.8. Both $\mathrm{TC}(-)$ and $\mathrm{RHom}^\bullet(R,-)$ commute with profinite completions, so that if $M_\bullet$ itself is profinitely complete, the completions in (7.4) can be dropped. In general, it is better to keep the completion; to obtain an isomorphism in the general case, one should, roughly speaking, replace T with the homotopy fixed points T^{hS^1} in the definition of topological cyclic homology $\mathrm{TC}(T)$.

Remark 7.9. It is not unreasonable to hope that Theorem 7.7 has a topological analog: one can define a triangulated category of cyclotomic spectra which is enriched over StHom, and then for any profinitely complete cyclotomic spectrum T, we have a natural homotopy equivalence

$$\mathrm{TC}(T) \cong \mathrm{Maps}(\mathbb{S}, T),$$

where $\mathrm{Maps}(-,-)$ is the mapping spectrum in the cyclotomic category, and $\mathbb{S} = \Sigma^\infty \mathsf{pt}$ is the sphere spectrum with the trivial cyclotomic structure (obtained as in Example 7.3). This would give a conceptual replacement of the somewhat *ad hoc* definition of the functor TC.

7.4. Back to ring spectra. Return now to our original situation: we have a ring spectrum $\mathcal{A} \in \mathsf{StAlg}$, and the DG algebra $A_\bullet = H(W)(\mathcal{A})$ is obtained as its homology with coefficients in the Witt vector ring $W = W(k)$ of a finite field k. Assume for simplicity that $k = \mathbb{Z}/p\mathbb{Z}$ is a prime field, so that $W = \mathbb{Z}_p$.

Then on one hand, we have the cyclotomic spectrum $\mathrm{THH}(\mathcal{A})$ of [BM], and since the homology functor $H(\mathbb{Z}_p)$ commutes with tensor products, we have a quasiisomorphism

$$H(\mathbb{Z}_p)(\mathrm{THH}(\mathcal{A})) \cong CH_\bullet(A_\bullet).$$

But the left-hand side underlies a cyclotomic complex, and by Theorem 7.6, this is equivalent to saying that it has a structure of a generalized filtered Dieudonné module. And on the other hand, $CH_\bullet(A_\bullet)$ has a Dieudonné module structure induced by the splitting map (7.1). We expect that the two structures coincide (although at present, this has not been checked).

Moreover, the functor TC commutes with $H(\mathbb{Z}_p)$ by (7.3), and Theorem 7.7 shows that we have

$$\begin{aligned} H(\mathbb{Z}_p)(\mathrm{TC}(\mathcal{A})) &\cong H(\mathbb{Z}_p)(\mathrm{TC}(\mathrm{THH}(\mathcal{A}))) \cong \mathrm{TC}(CH_\bullet(A_\bullet)) \\ &\cong \mathrm{RHom}^\bullet(\mathbb{Z}_p, CH_\bullet(A)), \end{aligned}$$

where $\mathrm{RHom}^\bullet(-,-)$ is taken in the category $\mathcal{DFDM}^{per}(\mathbb{Z}_p)$ of filtered Dieudonné modules. In other words:

- The homology functor $H(\mathbb{Z}_p)$ sends topological cyclic homology into syntomic periodic cyclic homology.

This principle can be used to study further the regulator map for syntomic homology. Namely, applying $H(\mathbb{Z}_p)$ to the cyclotomic trace map (7.2), we obtain

a functorial map

$$H(\mathbb{Z}_p)(K_\bullet(\mathcal{A})) \to H(\mathbb{Z}_p)(\mathrm{TC}(\mathcal{A})),$$

and the right-hand side is the target of the desired regulator map for the DG algebra $A_\bullet$. The desired source of this map is $K_\bullet(A_\bullet) \cong K_\bullet(H(\mathbb{Z}_p)(\mathcal{A}))$. Thus the question of existence of the syntomic regulator maps reduces to a problem in algebraic K-theory: describe the relation between the homology of the K-theory of a ring spectrum, and the K-theory of its homology.

To finish the Section, let us explain how things work in a very simple particular case. Assume given a CW complex X, and let $\mathcal{A} = \Sigma^\infty \Omega X$, the suspension spectrum of the based loop space ΩX. Then since ΩX is a topological monoid, $\mathcal{A}$ is a ring spectrum. The DG algebra $A_\bullet = H(\mathbb{Z}_p)(\mathcal{A})$ is given by $A_\bullet = C_\bullet(\Omega X, \mathbb{Z}_p)$, the singular chain complex of the topological space ΩX. It is known that in this case, we have

$$CH_\bullet(A_\bullet) \cong C_\bullet(LX, \mathbb{Z}_p),$$

the singular chain complex of the free loop space LX. Analogously, we have $\mathrm{THH}(\mathcal{A}) \cong \Sigma^\infty LX$. The S^1-action on $\mathrm{THH}(\mathcal{A})$ and $CH_\bullet(A_\bullet)$ is induced by the loop rotation action on LX. The cyclotomic structure on $\mathrm{THH}(\mathcal{A})$ is that of Example 7.3. The corresponding Dieudonné module structure map φ on $CP_\bullet(A_\bullet)$ is induced by the cyclotomic structure map $LX^{\mathbb{Z}/p\mathbb{Z}} \cong LX$ of the free loop space LX. To compare this with the constructions of Section 5, specialize even further and assume that ΩX is discrete, so that $A_\bullet$ is quasiisomorphic to an algebra A concentated in degree 0. In this case $X \cong BG$ for a discrete group G, and $A = \mathbb{Z}_p[G]$ is its group algebra. Then the diagonal map $G \to G^p$ induces a map $A \to A^{\otimes p}$ which is a quasi-Frobenius map in the sense of Section 5, thus induces another Dieudonné module structure on the filtered complex $CP_\bullet(A)$. One checks easily that the two structures coincide. For a general X, the Diedonné module structure on $CP_\bullet(A_\bullet)$ can also be described explicitly in the same way as in Section 5, by using the map

$$A_\bullet \to A_\bullet^{\otimes p}$$

induced by the diagonal map $\Omega X \to (\Omega X)^p$ in place of the quasi-Frobenius map.

8. Hodge Structures

In the archimedian setting of (i) of Section 1, much less is known about periodic cyclic homology than in the non-archimedian setting of (ii). One starts with a smooth proper DG algebra $A^\bullet$ over $\mathbb{C}$ and considers its periodic cyclic homology complex $CP_\bullet(A^\bullet)$ with its Hodge filtration. In order to equip $HP_\bullet(A^\bullet)$ with an $\mathbb{R}$-Hodge structure, one needs to define a weight filtration $W_\bullet CP_\bullet(A^\bullet)$ and a

complex conjugation isomorphism $\overline{} : CP_\bullet(A^\bullet) \to \overline{CP_\bullet(A^\bullet)}$. The gradings in the isomorphism (2.2) suggest that $W_\bullet$ should be simply the canonical filtration of the complex $CP_\bullet(A^\bullet)$. However, the complex conjugation is a complete mystery. There is only one approach known at present, albeit a very indirect and highly conjectural one; the goal of this section is to describe it. I have learned all this material from B. Toën and/or M. Kontsevich – it is only the mistakes here that are mine.

The so-called $\mathcal{D}^-$-*stacks* introduced by B. Toën and G. Vezzosi in [ToVe] generalize both Artin stacks and DG schemes and form the subject of what is now known as "derived algebraic geometry"; a very nice overview is avaiable in [To1]. Very approximately, a $\mathcal{D}^-$-stack over a ring k is a functor

$$\mathcal{M} : \Delta^{opp}\,\mathsf{Comm}(k) \to \Delta^{opp}\,\mathrm{Sets}$$

from the category of simplicial commutative algebras over k to the category of simplicial sets. This functor should satisfy some descent-type conditions, and all such functors are considered up to an appropriately defined homotopy equivalence (made sense of by the technology of closed model structures). This generalizes the Grothendieck approach to schemes which treats a scheme over k as its functor of points – a sheaf of sets on the opposite $\mathsf{Comm}(k)^{opp}$ to the category of commutative algebras over k. The category $\mathsf{Comm}(k)$ is naturally embedded in $\Delta^{opp}\,\mathsf{Comm}(k)$ as the subcategory of constant simplicial objects, and resticting a $\mathcal{D}^-$-stack $\mathcal{M}$ to $\mathsf{Comm}(k) \subset \Delta^{opp}\,\mathsf{Comm}(k)$ gives an ∞-stack in the sense of Simpson [S] (this is called the *truncation* of $\mathcal{M}$).

If k contains $\mathbb{Q}$, one may replace simplicial commutative algebras with commutative DG algebras $R_\bullet$ over k placed in non-negative homological degrees, $R_i = 0$ for $i < 0$. If we denote the category of such DG algebras by $\mathsf{DG\text{-}Comm}^-(k)$, then a $\mathcal{D}^-$-stack is a functor

$$\mathcal{M} : \mathsf{DG\text{-}Comm}^-(k) \to \Delta^{opp}\,\mathrm{Sets},$$

again satisfying some conditions, and considered up to a homotopy equivalence. The category of $\mathcal{D}^-$-stacks over k is denoted $\mathcal{D}^-\mathsf{st}(k)$. For every DG algebra $R_\bullet \in \mathsf{DG\text{-}Comm}^-(k)$, its *derived spectrum* $\mathrm{R\,Spec}(R_\bullet) \in \mathcal{D}^-\mathsf{st}(k)$ sends a DG algebra $R'_\bullet \in \mathsf{DG\text{-}Comm}^-(k)$ to the simplicial set of maps from $R_\bullet$ to $R'_\bullet$, with the simplicial structure induced by the model structure on the category $\mathsf{DG\text{-}Comm}^-(k)$. We thus obtain a Yoneda-type embedding

$$\mathrm{R\,Spec} : \mathsf{DG\text{-}Comm}^-(k)^{opp} \to \mathcal{D}^-\mathsf{st}(k).$$

For any DG algebra $R_\bullet \in \mathsf{DG\text{-}Comm}^-(k)$, its de Rham cohomology comple $\Omega^\bullet(R_\bullet)$ is defined in the obvious way; $\Omega^\bullet(-)$ gives a functor

$$\Omega^\bullet : \mathsf{DG\text{-}Comm}^-(k) \to \mathrm{Spaces}_{\mathbb{Q}}$$

from $\mathsf{DG\text{-}Comm}^-(k)$ to the category $\mathrm{Spaces}_{\mathbb{Q}}$ of rational homotopy types in the

sense of Quillen [Q]. By the standard Kan extension machinery, $\Omega^\bullet$ extentds to a de Rham realization functor

$$\Omega^\bullet : \mathcal{D}^-\mathsf{st}(k) \to \mathrm{Spaces}_{\mathbb{Q}}\,.$$

Alternatively, one can take the 0-th homology algebra $H_0(R_\bullet)$ and consider its cristalline cohomology; this gives a DG algebra quasiisomorphic to $\Omega^\bullet(R_\bullet)$ (the higher homology groups behave as nilpotent extensions and do not contribute to cohomology). This shows that the de Rham realization $\Omega^\bullet(\mathcal{M})$ of a $\mathcal{D}^-$-stack $\mathcal{M} \in \mathcal{D}^-\mathsf{st}(k)$ only depends on its truncation.

Moreover, for $\mathcal{D}^-$-stacks satifying a certain finiteness condition ("locally geometric" and "locally finitely presented" in the sense of [ToVa]), instead of considering de Rham cohomology, one can take the underlying topological spaces $\mathrm{Top}(\mathcal{M}(R))$ of the simplicial complex algebraic varieties $\mathcal{M}(R)$, $R \in \mathsf{Comm}(k)$; by Kan extension, this gives a topological realization functor

$$\mathrm{Top} : \mathcal{D}^-\mathsf{st}(k) \to \mathrm{Spaces}$$

into the category of topological spaces. By the standard comparison theorems, $\mathrm{Top}(\mathcal{M})$ and $\Omega^\bullet(\mathcal{M})$ represent the same rational homotopy type.

Now, it has been proved in [ToVa] that for any associative unital DG algebra $A^\bullet$ over k, there exists a $\mathcal{D}^-$-stack $\mathcal{M}(A^\bullet)$ classifying "finite-dimensional DG modules over $A^\bullet$". By definition, for any commutative DG algebra $R_\bullet \in \mathsf{DG\text{-}Comm}^-(k)$, the simplicial set $\mathcal{M}(A^\bullet)(R_\bullet)$ is given by

- $\mathcal{M}(A^\bullet)(R_\bullet)$ is the nerve of the category $\mathsf{Perf}(A^\bullet, R_\bullet)$ of DG modules over $A^\bullet \otimes R_\bullet$ which are perfect over $R_\bullet$, and quasiisomorphisms between such DG modules.

Toën and Vaquié prove that this indeed defines a $\mathcal{D}^-$-stack. Moreover, they prove that if $A^\bullet$ satisfies certain finiteness conditions, the $\mathcal{D}^-$-stack $\mathcal{M}(A^\bullet)$ is locally geometric and locally finitely presented.

In particular, a smooth and proper DG algebra $A^\bullet \in \mathsf{DG\text{-}Alg}(k)$ satisfies the finiteness conditions needed for [ToVa], so that there exists a locally geometric and locally finitely presented $\mathcal{D}^-$-stack $\mathcal{M}(A^\bullet)$. Consider its de Rham realization $\Omega^\bullet(\mathcal{M}(A^\bullet))$. For any $R_\bullet \in \mathsf{DG\text{-}Comm}^-(k)$, the category $\mathsf{Perf}(A^\bullet, R_\bullet)$ is a symmetric monoidal category with respect to the direct sum, so that the realization $\mathrm{Top}(\mathcal{M}(A^\bullet))$ is automatically an E_∞-space.

Lemma 8.1 (Toën). *The E_∞-space $\mathrm{Top}(\mathcal{M}(A^\bullet))$ is group-like.*

Sketch of a possible proof. One has to show that $\pi_0(\mathrm{Top}(\mathcal{M}(A^\bullet)))$ is not only a commutative monoid but also an abelian group. A point in $\mathrm{Top}(\mathcal{M}(A^\bullet))$ is represented by a DG module $M_\bullet$ over $A^\bullet$ which is perfect over k. One observes that $M_\bullet \oplus M_\bullet[1]$ can be deformed to a acyclic DG module; thus the sum of points represented by $M_\bullet$ and $M_\bullet[1]$ lies in the connected component of 0 in $\mathrm{Top}(\mathcal{M}(A^\bullet))$. $\square$

Thus for any smooth and proper DG algebra $A^\bullet \in \mathsf{DG\text{-}Alg}(k)$, the realization $\mathrm{Top}(\mathcal{M}(A^\bullet))$ is an infinite loop space, that is, the 0-th component of a spectrum.

Definition 8.2. The *semi-topological K-theory* $K^{st}_\bullet(A^\bullet)$ of a smooth and proper DG algebra $A^\bullet$ is given by

$$K^{st}_\bullet(A^\bullet) = \pi_\bullet(\mathrm{Top}(\mathcal{M}(A^\bullet))),$$

the homotopy groups of the infinite loop space $\mathrm{Top}(\mathcal{M}(A^\bullet))$.

If we are only interested in $K^{st}_\bullet(A^\bullet) \otimes k$, we may compute it using the de Rham model $\Omega^\bullet(\mathcal{M}(A^\bullet))$. Then $K^{st}_\bullet(\mathcal{M}(A^\bullet))$ is exactly the complex of primitive elements with respect to the natural cocommutative coalgebra structure on $\mathcal{M}(A^\bullet)$ induces by the direct sum map

$$\mathcal{M}(A^\bullet) \times \mathcal{M}(A^\bullet) \to \mathcal{M}(A^\bullet).$$

Since $\mathbb{Q} \subset k$, and rationally, spectra are the same as complexes of $\mathbb{Q}$-vector spaces, the groups $K^{st}_\bullet(A^\bullet) \otimes k$ are the only rational invariants one can extract from the space $\mathcal{M}(A^\bullet)$.

Assume for the moment that $A^\bullet \in \mathsf{DG\text{-}Alg}(k)$ is derived-Morita equivalent to a smooth and proper algebraic variety X/k. Then one can also consider the ∞-stack $\overline{\mathcal{M}}(X)$ of all coherent sheaves on X; for any noetherian $R \in \mathsf{Comm}(k)$, $\overline{\mathcal{M}}(X)(R)$ is by definition the nerve of the category of coherent sheaves on $M \otimes R$ and isomorphisms between them. The realization $\mathrm{Top}(\overline{\mathcal{M}}(X))$ is again an E_∞-space, no longer group-like. By definition, we have a natural map

$$\overline{\mathcal{M}}(X) \to \mathcal{M}(A^\bullet),$$

and the induced E_∞-map of realizations.

Lemma 8.3 (Toën). *The natural E_∞-map*

$$\mathrm{Top}(\overline{\mathcal{M}}(X)) \to \mathrm{Top}(\mathcal{M}(A^\bullet)) \tag{8.1}$$

induces a homotopy equivalence between $\mathrm{Top}(\mathcal{M}(A^\bullet))$ *and the group completion of the E_∞-space* $\mathrm{Top}(\overline{\mathcal{M}}(X))$.

Sketch of a possible proof. Since $\mathrm{Top}(\mathcal{M}(A^\bullet))$ is group-like by Lemma 8.1, it suffices to prove that the delooping

$$B\,\mathrm{Top}(\overline{\mathcal{M}}(X)) \to B\,\mathrm{Top}(\mathcal{M}(A^\bullet))$$

of the E_∞-map (8.1) is a homotopy equivalence. Delooping obviously commutes with geometric realization, so that $B\,\mathrm{Top}(\mathcal{M}(A^\bullet))$ is the realization of the $\mathcal{D}^-$-stack $B\,\mathcal{M}(A^\bullet)$, and similarly for $B\,\mathrm{Top}(\overline{\mathcal{M}}(X))$. Instead of taking deloopings,

we can apply Waldhausen's S-construction. The resulting map

$$S\overline{\mathcal{M}}(X) \to S\,\mathcal{M}(A^\bullet)$$

is then an equivalence by Waldhausen's devissage theorem, so that it suffices to prove that the natural map

$$\mathrm{Top}(B\,\mathcal{M}(A^\bullet)) \to \mathrm{Top}(S\,\mathcal{M}(A^\bullet))$$

is a homotopy equivalence, and similarly for $\overline{\mathcal{M}}(X)$. For this, one argues as in Lemma 8.1: since every filtered complex can be canonically deformed to its associated graded quotient, the terms $\mathrm{Top}(S_n\,\mathcal{M}(A^\bullet))$ of the S-construction can be retracted to n-fold products $\mathrm{Top}(\mathcal{M}(A^\bullet) \times \cdots \times \mathcal{M}(A^\bullet))$, that is, the terms of the delooping $\mathrm{Top}(B\,\mathcal{M}(A^\bullet))$, and similarly for $\overline{\mathcal{M}}(X)$. □

Corollary 8.4. *The semitopological K-theory $K^{st}_\bullet(\mathbb{Q})$ is given by*

$$K^{st}_\bullet(k) \cong \mathbb{Z}[\beta],$$

the algebra of polynomials in one generator β of degree 2.

Proof. By Lemma 8.3, computing $K^{st}_\bullet(k)$ reduces to studying the group completion of the realization

$$\mathrm{Top}(\overline{\mathcal{M}}(\mathsf{pt})) \cong \coprod_n \mathrm{Top}([\mathsf{pt}/GL_n]) \cong \coprod_n BU_n,$$

where $[\mathsf{pt}/GL_n]$ is the Artin stack obtained as the quotient of the point by the trivial action of the algebraic group GL_n. This group completion is well-known to be homotopy equivalent to the classifying space $\mathbb{Z} \times BU$. □

Remark 8.5. At present, Lemma 8.1 and Lemma 8.3 are unpublished, as well as Corollary 8.4. The above sketches of proofs have been kindly explained to me by B. Toën. Lemma 8.1 is slightly older, and it also appears for example in [To3].

Now, since $k \supset \mathbb{Q}$ by our assumption, we have a well-defined tensor product $M_\bullet \otimes V_\bullet$ for any DG module $M_\bullet$ over $A^\bullet$ and every complex $V_\bullet$ of $\mathbb{Q}$-vector spaces. On the level of the stacks $\mathcal{M}(-)$, this tensor product turns $K^{st}_\bullet(A^\bullet)$ into a module over $K^{st}_\bullet(\mathbb{Q}) = \mathbb{Z}[\beta]$. We can now state the main conjecture.

Conjecture 8.6. *Assume that k is a ring conatining $\mathbb{Q}$, and assume that a DG algebra $A^\bullet$ $\mathsf{DG\text{-}Alg}(k)$ is smooth and proper. Then there exists a map*

$$c : K^{st}_\bullet(A^\bullet) \to HP_\bullet(A^\bullet)$$

such that $c(\beta(\alpha)) = u(c(\alpha))$ for any $\alpha \in K^{st}_\bullet(A^\bullet)$, where u is the periodicity

map. The map c is functorial in $A^\bullet$. Moreover, the induced map

$$K^{st}_\bullet(A^\bullet) \otimes_{\mathbb{Z}[\beta]} k[\beta, \beta^{-1}] \to HP_\bullet(A^\bullet) \tag{8.2}$$

is an isomorphism.

The reason this conjecture is relevant to the present paper is that the tensor product $K^{st}_\bullet(A^\bullet) \otimes k$ by its very definition has all the structures possessed by the de Rham cohomology of an algebraic variety. In particular, if $k = \mathbb{C}$, $K^{st}_\bullet(A^\bullet)$ has a canonical real structure.

Conjecture 8.7. *Assume that $K = \mathbb{C}$, and assume given a smooth and proper DG algebra $A^\bullet/K$ for which Conjecture* 8.6 *holds. Equip $CP_\bullet(A^\bullet)$ with the real structure induced from the canonical real structure on $K^{st}_\bullet(A^\bullet) \otimes K$ by the isomorphism* 8.2. *Then for any integer i, the periodic cyclic homology group $HP_\bullet(A^\bullet)$ this real structure and the standard Hodge filtration $F^\bullet$ is a pure $\mathbb{R}$-Hodge structure of weight i.*

The two conjectures above are a slight refinement and/or reformulation of a conjecture made by B. Toën [To3] with a reference to A. Bondal and A. Neeman, and described by L. Katzarkov, M. Kontsevich and T. Pantev in [KKP, 2.2.6].

Apart from the basic case $A^\bullet = k$ of Corollary 8.4, the only real evidence for Conjecture 8.6 comes from recent work of Fiedlander and Walker [FW], where it has been essentially proved for a DG algebra $A^\bullet$ equivalent to a smooth projective algebraic variety X/k. The definition of semi-topological K-theory used in [FW] is different from Definition 8.2, but it is very close to the homotopy groups of the group completion of the E_∞-space $\mathrm{Top}(\overline{\mathcal{M}}(X))$; Lemma 8.3 should then show that the two things are the same. Friedlander and Walker also show that their constructions are compatible with the complex conjugation, so that Conjecture 8.7 then follows by the usual Hodge theory applied to X.

In the general case, as far as I know, both Conjecture 8.6 and Conjecture 8.7 are completely open. They are now a subject of investigation by B. Toën and A. Blanc.

Acknowledgements. I have benefited a lot from discussing this material with A. Beilinson, R. Bezrukavnikov, A. Bondal, V. Drinfeld, L. Hesselholt, V. Ginzburg, L. Katzarkov, D. Kazhdan, B. Keller, M. Kontsevich, A. Kuznetsov, N. Markarian, J.P. May, G. Merzon, D. Orlov, T. Pantev, S.-R. Park, B. Toën, M. Verbitsky, G. Vezzosi, and V. Vologodsky; discussions with M. Kontsevich, on one hand, and B. Toën and G. Vezzosi, on the other hand, were particularly invaluable. I am grateful to the referee for useful comments.

References

[Ad] J.F. Adams, *Stable Homotopy and Generalized Homology*, Univ. of Chicago Press, 1974.

[Be] A. Beilinson, *Higher regulators and values of L-functions*, (in Russian), VINITI Current problems in mathematics **24**, Moscow, 1984; 181–238.

[BBD] A. Beilinson, J. Bernstein, and P. Deligne, *Faiscaux pervers*, Astérisque **100**, Soc. Math. de France, 1983.

[Bo] Bökstedt, *Topological Hochschild homology*, preprint, Bielefeld, 1985.

[BHM] M. Bökstedt, W.C. Hsiang, and I. Madsen, *The cyclotomic trace and algebraic K-theory of spaces*, Invent. Math. **111** (1993), 465–539.

[BM] M. Bökstedt and I. Madsen, *Topological cyclic homology of the integers*, in *K-theory (Strasbourg, 1992)*, Astérisque **226** (1994), 7–8, 57–143.

[BK] A. Bondal and M. Kapranov, *Representable functors, Serre functors, and reconstructions*, (Russian) Izv. Akad. Nauk SSSR Ser. Mat. **53** (1989), 1183–1205, 1337; translation in Math. USSR-Izv. **35** (1990), 519–541.

[BV] A. Bondal and M. Van den Bergh, *Generators and representability of functors in commutative and noncommutative geometry*, Mosc. Math. J. **3** (2003), 1–36,

[BO] S. Bloch and A. Ogus, *Gersten's conjecture and the homology of schemes*, Ann. Sci. École Norm. Sup. (4) 7 (1974), 181–201.

[DI] P. Deligne and L. Illusie, *Relévements modulo p^2 et décomposition du complexe de de Rham*, Inv. Math. **89** (1987), 247–270.

[tD] T. tom Dieck, *Transformation groups*, De Gruyter, Berlin-New York, 1987.

[Dr] A.W.M. Dress, *Contributions to the theory of induced representations*, in *Algebraic K-Theory II*, (H. Bass, ed.), Lecture Notes in Math. **342**, Springer-Verlag, 1973; pp. 183–240.

[EKMM] A.D. Elmendorf, I. Kriz, M.A. Mandell, and J.P. May, *Rings, modules, and algebras in stable homotopy theory*, Mathematical Surveys and Monographs, **47**, AMS, Providence, RI, 1997.

[F] G. Faltings, *Crystalline cohomology and p-adic Galois-representations*, in *Algebraic analysis, geometry, and number theory (Baltimore, MD, 1988)*, Johns Hopkins Univ. Press, Baltimore, MD, 1989; 25–80.

[FT] B. Feigin and B. Tsygan, *Additive K-Theory*, in Lecture Notes in Math. **1289** (1987), 97–209.

[FL] J.-M. Fontaine and G. Lafaille, *Construction de représentations p-adiques*, Ann. Sci. École Norm. Sup. (4) **15** (1982), 547–608 (1983).

[FM] J.-M. Fontaine and W. Messing, *p-adic periods and p-adic étale cohomology*, in *Current trends in arithmetical algebraic geometry (Arcata, Calif., 1985)*, Contemp. Math. **67**, AMS, Providence, RI, 1987; 179–207.

[FW] E. Friedlander and M. Walker, *Semi-topological K-theory*, in *Handbook of K-theory*, Springer, Berlin, 2005; 877–924.

[Good] T.G. Goodwillie, *Relative algebraic K-theory and cyclic homology*, Ann. of Math. **124** (1986), 347–402.

[Gr1] M. Gros, *Régulateurs syntomiques et valeurs de fonctions L p-adiques, I*, Invent. Math. **99** (1990), 293–320.

[Gr2] M. Gros, *Régulateurs syntomiques et valeurs de fonctions L p-adiques, II*, Invent. Math. **115** (1994), 61–79.

[HM] L. Hesselholt and I. Madsen, *On the K-theory of finite algebras over Witt vectors of perfect fields*, Topology **36** (1997), 29–101.

[HKR] G. Hochschild, B. Kostant, and A. Rosenberg, *Differential forms on regular affine algebras*, Trans. AMS **102** (1962), 383–408.

[HSS] M. Hovey, B. Shipley, and J. Smith, *Symmetric spectra*, J. AMS **13** (2000), 149–208.

[Ka1] D. Kaledin, *Non-commutative Hodge-to-de Rham degeneration via the method of Deligne-Illusie*, Pure Appl. Math. Q. **4** (2008), 785–875.

[Ka2] D. Kaledin, *Cartier isomorphism and Hodge theory in the non-commutative case*, in *Arithmetic geometry*, Clay Math. Proc. **8**, AMS, Providence, RI, 2009; 537–562.

[Ka3] D. Kaledin, *Derived Mackey functors*, `arXiv:0812.2519`, to appear in Moscow Math. J.

[Ka4] D. Kaledin, *Cyclotomic complexes*, `arXiv:1003.2810`.

[KKP] L. Katzarkov, M. Kontsevich, and T. Pantev, *Hodge-theoretic aspects of mirror symmetry*, `arvix.org/0806.0107`.

[KS] M. Kontsevich and Y. Soibelman, *Notes on A-infinity algebras, A-infinity categories and non-commutative geometry, I*, preprint math.RA/0606241.

[Ke1] B. Keller, *Deriving DG categories*, Ann. Sci. Ecole Norm. Sup. (4) **27** (1994), 63–102.

[Ke2] B. Keller, *On differential graded categories*, in *International Congress of Mathematicians*, Vol. II, Eur. Math. Soc., Zürich, 2006; 151–190.

[LMS] L.G. Lewis, J.P. May, and M. Steinberger, *Equivariant stable homotopy theory*, with contributions by J. E. McClure, Lecture Notes in Mathematics, **1213**, Springer-Verlag, Berlin, 1986.

[Le] M. Levine, *Mixed motives*, in *Handbook of K-theory*, Springer, Berlin, 2005; 429–521.

[Li] H. Lindner, *A remark on Mackey functors*, Manuscripta Math. **18** (1976), 273–278.

[Lo] J.-L. Loday, *Cyclic Homology*, second ed., Springer, 1998.

[Mc] R. McCarthy, *Relative algebraic K-theory and topological cyclic homology*, Acta Math. **179** (1997), 197–222.

[MM] M.A. Mandell and J.P. May, *Equivariant orthogonal spectra and S-modules*, Memoirs of the AMS **755** (2002).

[M1] J.P. May, *Equivariant homotopy and cohomology theory*, with contributions by M. Cole, G. Comezana, S. Costenoble, A.D. Elmendorf, J.P.C. Greenlees, L.G. Lewis, Jr., R.J. Piacenza, G. Triantafillou, and S. Waner, CBMS Regional Conference Series in Mathematics, **91**. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1996.

[M2] J.P. May, private communication.

[O] D. Orlov, *Remarks on Generators and Dimensions of Triangulated Categories*, Moscow Math. J. **9** (2009), 513–519.

[Q] D. Quillen, *Rational homotopy theory*, Ann. of Math. (2) **90** 1969, 205–295.

[S] C. Simpson, *Algebraic (geometric) n-stacks*, `arXiv:alg-geom/9609014`.

[Th] J. Thevenaz, *Some remarks on G-functors and the Brauer morphism*, J. Reine Angew. Math. **384** (1988), 24–56.

[To1] B. Toën, *Higher and derived stacks: a global overview*, in *Algebraic geometry—Seattle 2005*, Part 1, 435–487, Proc. Sympos. Pure Math., **80**, Part 1, AMS, Providence, RI, 2009.

[To2] B. Toën, *Anneaux de définition des dg-algèbres propres et lisses*, Bull. Lond. Math. Soc. **40** (2008), 642–650.

[To3] B. Toën, *Saturated dg-categories III*, a talk at Workshop on Homological Mirror Symmetry and Related Topics, January 18-23, 2010, University of Miami; handwritten notes by D. Auroux available at `http://www-math.mit.edu/~auroux/frg/miami10-notes/`.

[ToVa] B. Toën and M. Vaquié, *Moduli of objects in dg-categories*, Ann. Sci. École Norm. Sup. (4) **40** (2007), 387–444.

[ToVe] B. Toën and G. Vezzosi, *Homotopical algebraic geometry, II. Geometric stacks and applications*, Mem. Amer. Math. Soc. **193** (2008).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Gromov-Witten Theory of Calabi-Yau 3-folds

Chiu-Chu Melissa Liu*

Abstract

We describe some recent progress and open problems in Gromov-Witten theory of Calabi-Yau 3-folds, focusing on the quintic 3-fold and toric Calabi-Yau 3-folds.

Mathematics Subject Classification (2010). 14N35

Keywords. Gromov-Witten invariants, Calabi-Yau 3-folds

1. Gromov-Witten Invariants of Calabi-Yau 3-folds

1.1. Moduli spaces of stable maps. Let X be a nonsingular projective variety over $\mathbb{C}$. Gromov-Witten (GW) invariants of X can be viewed as intersection numbers on moduli spaces of (parametrized) complex algebraic curves in X. Let $\mathcal{M}_{g,n}(X,\beta)$ be the moduli space of morphisms $f : (C, x_1, \ldots, x_n) \to X$, where C is a smooth complex algebraic curve of genus g, $x_1, \ldots, x_n$ are distinct points on C, and $f_*[C] = \beta \in H_2(X;\mathbb{Z})$. We call β the *degree* of the map. Two maps are equivalent if they differ by an automorphism of the domain $(C, x_1, \ldots, x_n)$. To do intersection theory, we should compactify $\mathcal{M}_{g,n}(X,\beta)$. The standard compactification in Gromov-Witten theory is $\overline{\mathcal{M}}_{g,n}(X,\beta)$, the Kontsevich's moduli space of stable maps $f : (C, x_1, \ldots, x_n) \to X$ of genus g, degree β, where the domain curve C has at most nodal singularities, $x_1, \ldots, x_n$ are distinct smooth points on C, and the map f is *stable* in the sense that the automorphism group of f is finite. When X is projective, the compactified moduli space $\overline{\mathcal{M}}_{g,n}(X,\beta)$ is a proper Deligne-Mumford stack [6].

*Columbia University, Mathematics Department, Room 623, MC 4435, New York, NY 10027. E-mail: ccliu@math.columbia.edu.

1.2. Perfect obstruction theory and the virtual fundamental class. The tangent space T^1 and the obstruction space T^2 at a moduli point $[f : (C, x_1, \ldots, x_n) \to X] \in \overline{\mathcal{M}}_{g,n}(X, \beta)$ fit in the following exact sequence:

$$\begin{aligned} 0 &\to \mathrm{Ext}^0(\Omega_C(x_1 + \cdots + x_n), \mathcal{O}_C) \to H^0(C, f^*T_X) \to T^1 \\ &\to \mathrm{Ext}^1(\Omega_C(x_1 + \cdots + x_n), \mathcal{O}_C) \to H^1(C, f^*T_X) \to T^2 \to 0 \end{aligned}$$

where

- $\mathrm{Ext}^0(\Omega_C(x_1 + \cdots + x_n), \mathcal{O}_C)$ is the space of infinitesimal automorphisms of the domain $(C, x_1, \ldots, x_n)$,
- $\mathrm{Ext}^1(\Omega_C(x_1 + \cdots + x_n), \mathcal{O}_C)$ is the space of infinitesimal deformations of the domain $(C, x_1, \ldots, x_n)$,
- $H^0(C, f^*T_X)$ is the space of infinitesimal deformations of the map f, and
- $H^1(C, f^*T_X)$ is the space of obstructions to deforming the map f.

T^1 and T^2 form sheaves $\mathcal{T}^1$ and $\mathcal{T}^2$ on the moduli space $\overline{\mathcal{M}}_{g,n}(X, \beta)$.

We say X is *convex* if $H^1(C, f^*T_X) = 0$ for all genus 0 stable maps f. Projective spaces $\mathbb{P}^n$, or more generally, generalized flag varieties G/P, are examples of convex varieties. When X is convex and $g = 0$, the moduli space $\overline{\mathcal{M}}_{0,n}(X, \beta)$ is a *smooth* Deligne-Mumford stack (orbifold). In general, $\overline{\mathcal{M}}_{g,n}(X, \beta)$ is a singular Deligne-Mumford stack equipped with a *perfect obstruction theory*: there is a two term complex of locally free sheaves $E \to F$ on $\overline{\mathcal{M}}_{g,n}(X, \beta)$ such that

$$0 \to \mathcal{T}^1 \to F^\vee \to E^\vee \to \mathcal{T}^2 \to 0$$

is an exact sequence of sheaves. The *virtual dimension* d^{vir} of $\overline{\mathcal{M}}_{g,n}(X, \beta)$ is the rank of the virtual tangent bundle $T^{\mathrm{vir}} = F^\vee - E^\vee$. By Riemann-Roch,

$$d^{\mathrm{vir}} = \int_\beta c_1(T_X) + (\dim X - 3)(1 - g) + n \tag{1}$$

There is a *virtual fundamental class*

$$[\overline{\mathcal{M}}_{g,n}(X, \beta)]^{\mathrm{vir}} \in H_{2d^{\mathrm{vir}}}(\overline{\mathcal{M}}_{g,n}(X, \beta); \mathbb{Q}).$$

The virtual fundamental class has been constructed by Li-Tian [40], Behrend-Fantechi [5] in algebraic Gromov-Witten theory, and by Li-Tian [41], Fukaya-Ono [19], Ruan [58], Siebert [61] (more recently, Hofer-Wysocki-Zehnder [26, 27, 28, 29]) in symplectic Gromov-Witten theory. In this paper we will mostly discuss algebraic Gromov-Witten theory.

1.3. GW invariants of compact Calabi-Yau 3-folds. When X is Calabi-Yau, in the sense that the canonical line bundle $\mathcal{O}(K_X)$ of X is trivial, we have $c_1(T_X) = 0$. By (1), the virtual dimension of $\overline{\mathcal{M}}_{g,0}(X,\beta)$ is $(\dim X - 3)(1-g)$, which is independent of the degree β. In particular, when X is a Calabi-Yau 3-fold, the virtual dimension of $\overline{\mathcal{M}}_{g,0}(X,\beta)$ is zero for any genus g and any degree β, and the virtual fundamental class is a degree zero rational homology class:

$$[\overline{\mathcal{M}}_{g,0}(X,\beta)]^{\mathrm{vir}} \in H_0(\overline{\mathcal{M}}_{g,0}(X,\beta);\mathbb{Q}).$$

The genus g, degree β Gromov-Witten invariant of a Calabi-Yau 3-fold X is defined by

$$N^X_{g,\beta} = \int_{[\overline{\mathcal{M}}_{g,0}(X,\beta)]^{\mathrm{vir}}} 1, \tag{2}$$

where $\int$ stands for the pairing between $H_0(\overline{\mathcal{M}}_{g,0}(X,\beta);\mathbb{Q})$ and $H^0(\overline{\mathcal{M}}_{g,0}(X,\beta);\mathbb{Q})$. If $\overline{\mathcal{M}}_{g,0}(X,\beta)$ were a compact complex manifold of dimension zero, it would consist of finitely many points, and the right hand side of (2) would be the number of points in $\overline{\mathcal{M}}_{g,0}(X,\beta)$. In general, $\overline{\mathcal{M}}_{g,0}(X,\beta)$ can be singular and can have positive actual dimension. Then the right hand side of (2) defines the "virtual number" of points in $\overline{\mathcal{M}}_{g,0}(X,\beta)$. In general $N^X_{g,\beta}$ is a rational number instead of an integer because $\overline{\mathcal{M}}_{g,0}(X,\beta)$ is a stack instead of a scheme.

1.4. GW invariants of noncompact Calabi-Yau 3-folds. Let X be a nonsingular projective Calabi-Yau 3-fold. The construction of the virtual fundamental class $[\overline{\mathcal{M}}_{g,0}(X,\beta)]^{\mathrm{vir}}$ requires two properties of the moduli space $\overline{\mathcal{M}}_{g,0}(X,\beta)$: having a prefect obstruction theory, and being proper. When X is *noncompact* nonsingular Calabi-Yau 3-fold, the moduli space $\overline{\mathcal{M}}_{g,0}(X,\beta)$ is usually not proper, but still equipped with a perfect obstruction theory of virtual dimension zero. Therefore, if X is noncompact but the moduli space $\overline{\mathcal{M}}_{g,0}(X,\beta)$ is proper for a particular genus g and degree β, then the Gromov-Witten invariant $N^X_{g,\beta}$ is defined for the particular genus g and degree β.

An important class of examples is the total space of the canonical line bundle of a Fano surface. Let X be the total space of the canonical line bundle over a nonsingular Fano surface S, for example, $S = \mathbb{P}^2$ and X is the total space of $\mathcal{O}_{\mathbb{P}^2}(-3)$. Then X is a noncompact Calabi-Yau 3-fold. Given any $\beta \in H_2(S;\mathbb{Z}) = H_2(X;\mathbb{Z})$ such that $\overline{\mathcal{M}}_{g,0}(S,\beta)$ is nonempty, the inclusion of the zero section $i_0 : S \hookrightarrow X$ induces an inclusion of moduli spaces $\overline{\mathcal{M}}_{g,0}(S,\beta) \hookrightarrow \overline{\mathcal{M}}_{g,0}(X,\beta)$ which is an isomorphism of Deligne-Mumford stacks when $\beta \neq 0$. We conclude that $\overline{\mathcal{M}}_{g,0}(X,\beta)$ is proper when $\beta \neq 0$, so the Gromov-Witten invariants of X are defined for any genus g and any nonzero degree β. Moreover:

1. When $\beta \neq 0$ and $\overline{\mathcal{M}}_{g,0}(S,\beta) = \overline{\mathcal{M}}_{g,0}(X,\beta)$ is nonempty, the perfect obstruction theories on $\overline{\mathcal{M}}_{g,0}(S,\beta)$ and on $\overline{\mathcal{M}}_{g,0}(X,\beta)$ are different: the virtual dimension of $\overline{\mathcal{M}}_{g,0}(S,\beta)$ is $\int_\beta c_1(T_S) + g - 1$, while the virtual

dimension of $\overline{\mathcal{M}}_{g,0}(X,\beta)$ is zero. Indeed, let $\pi : \overline{\mathcal{M}}_{g,1}(S,\beta) \to \overline{\mathcal{M}}_{g,0}(S,\beta)$ be the universal curve, and let $\mathrm{ev} : \overline{\mathcal{M}}_{g,1}(S,\beta) \to S$ be the evaluation map, which sends $[f : (C,x_1) \to S] \in \overline{\mathcal{M}}_{g,1}(S,\beta)$ to $f(x_1) \in S$. Then $\pi_*\mathrm{ev}^*\mathcal{O}_S(K_S) = 0$, so $V_{g,\beta} := R^1\pi_*\mathrm{ev}^*\mathcal{O}_S(K_S)$ is a vector bundle of rank $\int_\beta c_1(T_S) + g - 1$ over $\overline{\mathcal{M}}_{g,0}(X,\beta)$, and

$$[\overline{\mathcal{M}}_{g,0}(X,\beta)]^{\mathrm{vir}} = e(V_{g,\beta}) \cap [\overline{\mathcal{M}}_{g,0}(S,\beta)]^{\mathrm{vir}}, \quad N^X_{g,\beta} = \int_{[\overline{\mathcal{M}}_{g,0}(S,\beta)]^{\mathrm{vir}}} e(V_{g,d}),$$

where $e(V_{g,\beta})$ is the Euler class (i.e. top Chern class) of $V_{g,\beta}$.

2. When $\beta \neq 0$ and $\overline{\mathcal{M}}_{g,0}(S,\beta) = \overline{\mathcal{M}}_{g,0}(X,\beta)$ is empty, we define $N^X_{g,\beta} = 0$.

3. When $\beta = 0$, $\overline{\mathcal{M}}_{g,0}(S,0) = \overline{\mathcal{M}}_{g,0} \times S$ is proper, while $\overline{\mathcal{M}}_{g,0}(X,0) = \overline{\mathcal{M}}_{g,0} \times X$ is not.

When X is a toric Calabi-Yau 3-fold (which must be noncompact), the $(\mathbb{C}^*)^3$-action on X induces a $(\mathbb{C}^*)^3$-action on $\overline{\mathcal{M}}_{g,0}(X,\beta)$. The fixed point set $\overline{\mathcal{M}}_{g,0}(X,\beta)^{(\mathbb{C}^*)^3}$ is a *proper* Deligne-Mumford stack. For $i = 1,2$, let $T^{i,f}$ and $T^{i,m}$ be the fixed and moving parts of the restriction of the sheaf $\mathcal{T}^i$ to the fixed point set $\overline{\mathcal{M}}_{g,0}(X,\beta)^{(\mathbb{C}^*)^3}$. Then $T^{1,f} - T^{2,f}$ is the virtual tangent bundle of $\overline{\mathcal{M}}_{g,0}(X,\beta)^{(\mathbb{C}^*)^3}$, and $N^{\mathrm{vir}} = T^{1,m} - T^{2,m}$ is the virtual normal bundle of $\overline{\mathcal{M}}_{g,0}(X,\beta)^{(\mathbb{C}^*)^3}$ in $\overline{\mathcal{M}}_{g,0}(X,\beta)$. We have

$$H^*_{(\mathbb{C}^*)^3}(\mathrm{pt};\mathbb{Q}) = H^*(B(\mathbb{C}^*)^3;\mathbb{Q}) = H^*((\mathbb{P}^\infty)^3;\mathbb{Q}) = \mathbb{Q}[t_1,t_2,t_3].$$

The genus g, degree β, $(\mathbb{C}^*)^3$-equivariant Gromov-Witten invariant of X is defined to be

$$N^X_{g,\beta}(t_1,t_2,t_3) = \int_{[\overline{\mathcal{M}}_{g,0}(X,\beta)^{(\mathbb{C}^*)^3}]^{\mathrm{vir}}} \frac{1}{e_{(\mathbb{C}^*)^3}(N^{\mathrm{vir}})} \in \mathbb{Q}(t_1,t_2,t_3)$$

where

$$e_{(\mathbb{C}^*)^3}(N^{\mathrm{vir}}) \in H^*_{(\mathbb{C}^*)^3}(\overline{\mathcal{M}}_{g,0}(X,\beta)^{(\mathbb{C}^*)^3};\mathbb{Q}) \cong H^*(\overline{\mathcal{M}}_{g,0}(X,\beta)^{(\mathbb{C}^*)^3};\mathbb{Q}) \otimes_{\mathbb{Q}} \mathbb{Q}[t_1,t_2,t_3]$$

is the $(\mathbb{C}^*)^3$-equivariant Euler class of the virtual normal bundle N^{vir}.

In general, $N^X_{g,\beta}(t_1,t_2,t_3)$ is a rational function in t_1,t_2,t_3 with $\mathbb{Q}$ coefficients, homogeneous of degree 0. In some cases, this equivariant invariant becomes a topological invariant independent of the equivariant parameters t_i.

1. If X is the total space of the canonical line bundle of a toric Fano surface, then $N^X_{g,\beta}(t_1,t_2,t_3)$ is a constant rational number independent of t_1,t_2,t_3.

2. As a consequence of the topological vertex (see Section 3.1 below), for any toric Calabi-Yau 3-fold,

$$N^X_{g,\beta}(t_1,t_2,-t_1-t_2) \tag{3}$$

is a rational number independent of t_1,t_2, for any genus g and degree β. This is a surprising result since a priori (3) is a element in $\mathbb{Q}(t_2/t_1)$.

2. The Quintic 3-fold

The quintic 3-fold Q is a nonsingular degree 5 hypersurface in $\mathbb{P}^4$. It is a nonsingular projective Calabi-Yau 3-fold. By Lefschetz hyperplane theorem, $H_2(Q;\mathbb{Z}) \cong H_2(\mathbb{P}^4;\mathbb{Z}) = \mathbb{Z}\ell$, where ℓ is the class of a projective line. We write $\overline{\mathcal{M}}_{g,0}(Q,d)$ and $\overline{\mathcal{M}}_{g,0}(\mathbb{P}^4,d)$ instead of $\overline{\mathcal{M}}_{g,0}(Q,d\ell)$ and $\overline{\mathcal{M}}_{g,0}(\mathbb{P}^4,d\ell)$, respectively. We define

$$N_{g,d} := N^Q_{g,d\ell} = \int_{[\overline{\mathcal{M}}_{g,0}(Q,d)]^{\mathrm{vir}}} 1.$$

The generating functions F_g of genus g Gromov-Witten invariants of Q are defined by (see e.g. [54, Section 3]):

$$F_0(T) = \frac{5}{6}T^3 + \sum_{d=1}^{\infty} N_{0,d}e^{dT}, \tag{4}$$

$$F_1(T) = -\frac{25}{12}T + \sum_{d=1}^{\infty} N_{1,d}e^{dT}, \tag{5}$$

and for $g \geq 2$,

$$F_g(T) = \frac{-50\cdot(-1)^g\cdot|B_{2g}B_{2g-2}|}{g(2g-2)\cdot(2g-2)!} + \sum_{d=1}^{\infty} N_{g,d}e^{dT},$$

where B_{2g} and B_{2g-2} are Bernoulli numbers.

2.1. Genus $g = 0$. In 1991, P. Candelas, X. de la Ossa, P. Green, and L. Parkes [10] derived the number n_d of rational curves of any degree $d > 0$ from mirror symmetry. Their stunning predictions motivated the development of Gromov-Witten theory. To state their enumerative prediction, we introduce $I_k(t)$ defined by

$$\sum_{k=0}^{4} I_k(t)H^k = \sum_{d=0}^{\infty} e^{(H+d)t}\frac{\prod_{m=1}^{5d}(5H+m)}{\prod_{m=1}^{d}(H+m)^5}$$

where $H \in H^2(\mathbb{P}^4,\mathbb{C})$ is the hyperplane class. For example,

$$I_0(t) = 1 + \sum_{d=1}^{\infty} e^{dt}\frac{(5d)!}{(d!)^5}, \quad I_1(t) = tI_0(t) + \sum_{d=1}^{\infty} e^{dt}\left(\frac{(5d)!}{(d!)^5}\sum_{m=d+1}^{5d}\frac{5}{m}\right).$$

Let $T = I_1(t)/I_0(t)$. In terms of the genus 0 Gromov-Witten invariants of the quintic 3-folds, the prediction in [10] can be stated as

$$F_0(T) = \frac{5}{2}\left(\frac{I_1(t)}{I_0(t)}\frac{I_2(t)}{I_0(t)} - \frac{I_3(t)}{I_0(t)}\right), \tag{6}$$

where $F_0(T)$ is defined by (4) above. The conjectural formula (6) was proved independently by Givental [21], and by Lian-Liu-Yau [38].

We now explain one of the ingredients of the proof: evaluation of $N_{0,d}$ by localization on the moduli space $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$. This approach was proposed by Kontsevich [33], and is fully justified once the foundation of the algebraic Gromov-Witten theory has been established [40, 5, 4].

The inclusion $Q \hookrightarrow \mathbb{P}^4$ induces an inclusion of moduli spaces $\iota : \overline{\mathcal{M}}_{0,0}(Q, d) \to \overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$. The following properties hold when $g = 0$ but fail when $g > 0$.

1. $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$ is a *smooth* proper Deligne-Mumford stack of dimension $5d+1$, and has a fundamental class (instead of a *virtual* fundamental class)
$$[\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)] \in H_{2(5d+1)}(\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d); \mathbb{Q}).$$

2. There is a vector bundle $V_{0,d}$ of rank $5d+1$ over $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$ and a section $\tilde{s}$ of $V_{0,d}$ such that $\overline{\mathcal{M}}_{0,0}(Q, d)$ is the zero locus $\tilde{s}^{-1}(0)$.

3. $\iota_*[\overline{\mathcal{M}}_{0,0}(Q, d)]^{\mathrm{vir}} = e(V_{0,d}) \cap [\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)]$, where $e(V_{0,d}) = c_{5d+1}(V_{0,d})$ is the Euler class. Therefore
$$N_{0,d} = \int_{[\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)]} e(V_{0,d}). \tag{7}$$

We now describe $V_{0,d}$ and the section $\tilde{s}$ explicitly. Let $\pi : \overline{\mathcal{M}}_{0,1}(\mathbb{P}^4, d) \to \overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$ be the universal curve, and let $\mathrm{ev} : \overline{\mathcal{M}}_{0,1}(\mathbb{P}^4, d) \to \mathbb{P}^4$ be the evaluation map. Then $R^1\pi_*\mathrm{ev}^*\mathcal{O}_{\mathbb{P}^4}(5) = 0$, so $V_{0,d} := \pi_*\mathrm{ev}^*\mathcal{O}_{\mathbb{P}^4}(5)$ is a rank $5d+1$ locally free sheaf over $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$ whose fiber over $[f : C \to \mathbb{P}^4]$ is $H^0(C, f^*\mathcal{O}_{\mathbb{P}^4}(5))$. The quintic 3-fold Q is the zero locus of a section $s \in H^0(\mathbb{P}^4, \mathcal{O}_{\mathbb{P}^4}(5))$. The image of a stable map $f : C \to \mathbb{P}^4$ is contained in $Q = s^{-1}(0)$ if and only if $f^*s = 0 \in H^0(C, f^*\mathcal{O}_{\mathbb{P}^4}(5))$. Let $\tilde{s} = \pi_*\mathrm{ev}^*(s) : \overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d) \to V_{0,d}$ be the section whose value at the moduli point $[f : C \to \mathbb{P}^4]$ is f^*s. Then $\tilde{s}^{-1}(0) = \overline{\mathcal{M}}_{0,0}(Q, d)$.

The torus $T = (\mathbb{C}^*)^5$ acts on $\mathbb{P}^4$ by
$$(t_0, t_1, \ldots, t_4) \cdot [z_0, z_1, \ldots, z_4] = [t_0z_0, t_1z_1, \ldots, t_4z_4]$$
for $(t_0, t_1, \ldots, t_4) \in T$ and $[z_0, z_1, \ldots, z_4] \in \mathbb{P}^4$. The T-action on $\mathbb{P}^4$ can be lifted to a T-action on $\mathcal{O}_{\mathbb{P}^4}(5)$. The T-action on $\mathbb{P}^4$ induces a T-action on the moduli space $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$ by moving the image of a stable map, and the T-action on $\mathcal{O}_{\mathbb{P}^4}(5)$ induces a T-action on $V_{0,d}$ such that $V_{0,d} \to \overline{\mathcal{M}}_{0,0}(\mathbb{P}^4)$ is a T-equivariant vector bundle. The fixed points set $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)^T$ of the T-action on $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4, d)$ consists of finitely many isolated points. The genus 0 GW invariants $N_{0,d}$ can

be evaluated using localization:

$$\begin{aligned} N_{0,d} &= \int_{[\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4,d)]} e(V_{0,d}) = \int_{[\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4,d)]} e_T(V_{0,d}) \\ &= \sum_{\xi\in\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4,d)^T} \frac{e_T((V_{0,d})_\xi)}{e_T(T_\xi)}, \end{aligned} \tag{8}$$

where e_T is the T-equivariant Euler class, T_ξ is the tangent space of the moduli space $\overline{\mathcal{M}}_{0,0}(\mathbb{P}^4,d)$ at ξ, and $(V_{0,d})_\xi$ is the fiber of $V_{0,d}$ at ξ. The last equality in (8) follows from the Atiyah-Bott localization formula [3].

2.2. Genus $g = 1$. In 1993, M. Bershadsky, S. Cecotti, H. Ooguri, and C. Vafa [7] made the following prediction on genus one Gromov-Witten invariants of the quintic 3-fold.

$$2F_1(T) = -\frac{25}{6}t + \ln\Big(I_0(t)^{-62/3}(1-5^5e^t)^{-1/6}J_1'(t)^{-1}\Big) \tag{9}$$

where $T = J_1(t) = I_1(t)/I_0(t)$, and $F_1(T)$ is defined by (5).

The conjectural formula (9) was proved by A. Zinger in [68]. To prove (9), Zinger and his collaborators have developed a theory of *reduced* genus one Gromov-Witten invariants for Q. Indeed, the theory is defined for a degree $(r+1)$ hypersurface in $\mathbb{P}^r$, where r is arbitrary.

Unlike $\overline{\mathcal{M}}_{0,n}(\mathbb{P}^r,d)$, $\overline{\mathcal{M}}_{1,n}(\mathbb{P}^r,d)$ is singular and reducible in general. The open substack $\mathcal{M}_{1,n}(\mathbb{P}^r,d)$ of $\overline{\mathcal{M}}_{1,n}(\mathbb{P}^r,d)$ is smooth and irreducible. The closure $\overline{\mathcal{M}}^0_{1,n}(\mathbb{P}^r,d)$ of $\overline{\mathcal{M}}_{1,n}(\mathbb{P}^r,d)$ is called the *main component* $\overline{\mathcal{M}}_{1,n}(\mathbb{P}^r,d)$.

Let Q^{r-1} be a degree $r+1$ smooth hypersurface in $\mathbb{P}^r$. Then $Q^{r-1} = s^{-1}(0)$ for some section $s \in H^0(\mathbb{P}^r, \mathcal{O}_{\mathbb{P}^r}(r+1))$. Note that Q^{r-1} is a Calabi-Yau $(r-1)$-fold, and that the virtual dimension of $\overline{\mathcal{M}}_{1,0}(Q^{r-1},d)$ is zero for any r,d. The (standard) genus one, degree d Gromov-Witten invariant of Q^{r-1} is defined by

$$N_{1,d}^{Q^{r-1}} = \int_{[\overline{\mathcal{M}}_{1,0}(Q^{r-1},d)]^{\mathrm{vir}}} 1.$$

The main component of $\overline{\mathcal{M}}_{1,n}(Q^{r-1},d)$ is given by

$$\overline{\mathcal{M}}^0_{1,n}(Q^{r-1},d) := \overline{\mathcal{M}}_{1,n}(Q^{r-1},d) \times_{\overline{\mathcal{M}}_{1,n}(\mathbb{P}^r,d)} \overline{\mathcal{M}}^0_{1,n}(\mathbb{P}^r,d).$$

The reduced genus one invariants $N_{1,d}^{Q^{r-1},red}$ can be viewed as the contribution to the standard genus one invariant $N_{1,d}^{Q^{r-1}}$ from the main component $\overline{\mathcal{M}}^0_{1,0}(Q^{r-1},d)$. Zinger defined reduced genus one invariants in [66], and related them to standard genus zero and genus one invariants in [67]. In particular,

when $r = 4$, $Q^3 = Q$, Zinger showed that

$$N_{1,d} = N_{1,d}^{red} + \frac{1}{12} N_{0,d},$$

where $N_{0,d} = N_{0,d}^{Q}$, $N_{1,d} = N_{1,d}^{Q}$, and $N_{1,d}^{red} = N_{1,d}^{Q,red}$.

Let $\pi : \overline{\mathcal{M}}_{1,1}(\mathbb{P}^r, d) \to \overline{\mathcal{M}}_{1,0}(\mathbb{P}^r, d)$ be the universal curve, and let $\mathrm{ev} : \overline{\mathcal{M}}_{1,1}(\mathbb{P}^r, d) \to \mathbb{P}^r$ be the evaluation map. Let π_0 and ev_0 denote the restrictions of π and ev to the main component. The sheaf $(V_{1,d})_0 := (\pi_0)_*(\mathrm{ev}_0)^*\mathcal{O}_{\mathbb{P}^r}(r+1)$ is not locally free; it restricts to a locally free sheaf of rank $(r+1)d$ on $\mathcal{M}_{1,0}(\mathbb{P}^r, d)$. Nevertheless, Zinger showed that the Euler class $e((V_{1,d})_0) \in H_{2(r+1)d}(\overline{\mathcal{M}}_{1,0}(\mathbb{P}^r, d); \mathbb{Q})$ is defined. J. Li and Zinger proved in [42] that

$$N_{1,d}^{Q^{r-1},red} = \int_{[\overline{\mathcal{M}}_{1,0}^0(\mathbb{P}^r,d)]^{\mathrm{vir}}} e((V_{1,d})_0).$$

Vakil and Zinger constructed a desingularization $\widetilde{\mathcal{M}}_{1,n}^0(\mathbb{P}^r, d) \to \overline{\mathcal{M}}_{1,n}^0(\mathbb{P}^r, d)$ of the main component [63]. The Vakil-Zinger desingularization has the following nice properties:

1. $\widetilde{\mathcal{M}}_{1,n}^0(\mathbb{P}^r, d)$ is a smooth proper Deligne-Mumford stack of dimension $(r+1)d + n$.
2. The map $\widetilde{\mathcal{M}}_{1,n}^0(\mathbb{P}^r, d) \to \overline{\mathcal{M}}_{1,n}^0(\mathbb{P}^r, d)$ is T-equivariant.
3. Let $\tilde{\pi} : \widetilde{\mathcal{M}}_{1,1}^0(\mathbb{P}^r, d) \to \widetilde{\mathcal{M}}_{1,0}^0(\mathbb{P}^r, d)$ be be the universal family, and let $\tilde{\mathrm{ev}} : \widetilde{\mathcal{M}}_{1,1}^0(\mathbb{P}^r, d) \to \mathbb{P}^r$ be the evaluation map. Then $\tilde{V}_{1,d} = \tilde{\pi}_*\tilde{\mathrm{ev}}^*\mathcal{O}_{\mathbb{P}^r}(r+1)$ is a locally free sheaf of rank $(r+1)d$.
4. $$N_{1,d}^{Q^{r-1},red} = \int_{[\widetilde{\mathcal{M}}_{1,0}(\mathbb{P}^r,d)]} e(\tilde{V}_{1,d}) \tag{10}$$

The right hand side of (10) can be computed by localization.

The above results are proved using the symplectic approach. The algebraic approach to reduced Gromov-Witten invariants is developed in recent work of Y. Hu and J.Li [30] and work in progress of H. Chang and J. Li [11]. In particular, Hu-Li reproved Vakil-Zinger desingularization via an algebraic approach.

2.3. Genus $g \geq 2$. D. Maulik and R. Pandharipande [50] provided a calculation scheme which determines $N_{g,d}$ for any given genus g and degree d in terms of previously determined Gromov-Witten invariants of the following targets: $\mathbb{P}^3$, a K3 surface, three Fano surfaces ($\mathbb{P}^2$, $\mathbb{P}^1 \times \mathbb{P}^1$, the blow-up of $\mathbb{P}^2$ at 6 points), and four curves (including $\mathbb{P}^1$).

In 2006, M. Huang, A. Klemm, and S. Quackenbush [25] determined $F_g(t)$ for $2 \leq g \leq 51$ using string theory. These predictions have not been verified mathematically yet.

3. Toric Calabi-Yau 3-folds

3.1. The topological vertex.

3.1.1. The physical theory. Based on the *large N duality* [22] between the topological string theory on Calabi-Yau 3-folds and the Chern-Simons theory on 3-manifolds, Aganagic-Klemm-Mariño-Vafa proposed the *topological vertex* [1], an algorithm of computing open and closed Gromov-Witten invariants in all genera of any nonsingular toric Calabi-Yau 3-folds. The algorithm of AKMV can be summarized in the following three steps.

O1. *Topological vertex.* There exist certain open Gromov-Witten invariants that count holomorphic maps from bordered Riemann surfaces to $\mathbb{C}^3$ with boundaries mapped to three Lagrangian submanifolds L_1, L_2, L_3. Such invariants depend on the following discrete data:

(i) the topological type of the domain, classified by the genus g and the number h of boundary circles;

(ii) the topological type of the map, described by a triple of partitions $\vec{\mu} = (\mu^1, \mu^2, \mu^3)$ where $\mu^i = (\mu^i_1, \mu^i_2, \ldots)$ are degrees ("winding numbers") of boundary circles in $L_i \cong S^1 \times \mathbb{C}$;

(iii) the "framing" $n_i \in \mathbb{Z}$ of the Lagrangian submanifolds L_i $(i = 1, 2, 3)$.

The *topological vertex* $C_{\vec{\mu}}(\lambda; \mathbf{n})$ is a generating function of such invariants where one fixes the winding numbers $\vec{\mu} = (\mu^1, \mu^2, \mu^3)$ and the framings $\mathbf{n} = (n_1, n_2, n_3)$ and sums over the genus of the domain.

O2. *Gluing algorithm.* Any toric Calabi-Yau 3-fold X can be constructed by gluing $\mathbb{C}^3$ charts. The open and closed Gromov-Witten invariants of X can be expressed in terms of local open Gromov-Witten invariants $C_{\vec{\mu}}(\lambda; \mathbf{n})$ of $\mathbb{C}^3$ by explicit gluing algorithm.

O3. *Closed formula.* By the large N duality, the topological vertex is given by

$$C_{\vec{\mu}}(\lambda; \mathbf{n}) = q^{\frac{1}{2}(\sum_{i=1}^{3} \kappa_{\mu^i} n_i)} \mathcal{W}_{\vec{\mu}}(q), \quad q = e^{\sqrt{-1}\lambda}, \tag{11}$$

where $\kappa_\mu = \sum \mu_i(\mu_i - 2i + 1)$ for a partition $\mu = (\mu_1 \geq \mu_2 \geq \cdots)$, and

$$\mathcal{W}_{\mu^1,\mu^2,\mu^3}(q) = q^{(\kappa_{\mu^2}+\kappa_{\mu^3})/2} \sum c^{\mu^1}_{\eta\rho^1} c^{(\mu^3)^t}_{\eta(\rho^3)^t} \frac{\mathcal{W}_{(\mu^2)^t\rho^1}(q)\mathcal{W}_{\mu^2(\rho^3)^t}(q)}{\mathcal{W}_{\mu^2}(q)}. \tag{12}$$

In (12), $c^{\mu}_{\eta\rho}$ be the Littlewood-Richardson coefficients and $\mathcal{W}_{\mu\nu}$ can be expressed in terms of the skew Schur functions $s_{\mu/\lambda}$ (see [53]):

$$\mathcal{W}_{\mu\nu}(q) = q^{(\kappa_\mu+\kappa_\nu)/2} \sum_{\lambda} s_{\mu^t/\lambda}(q^{-\frac{1}{2}}, q^{-\frac{3}{2}}, \ldots) s_{\nu^t/\lambda}(q^{-\frac{1}{2}}, q^{-\frac{3}{2}}, \ldots).$$

The left hand side of (11) is an infinite series while the right hand side of (11) is a finite sum.

When the toric Calabi-Yau 3-fold is the total space of the canonical line bundle K_S of a toric surfaces S (e.g. $K_{\mathbb{P}^2} = \mathcal{O}_{\mathbb{P}^2}(-3)$), only $\tilde{C}_{\mu^1,\mu^2,\emptyset}$ are required to evaluate its Gromov-Witten invariants. The algorithm in this case was described in [2].

3.1.2. The mathematical theory. In [39], Li-Liu-Liu-Zhou developed a mathematical theory of the topological vertex based on relative Gromov-Witten theory. The relative Gromov-Witten theory has been developed in symplectic geometry by Li-Ruan [34] and Ionel-Parker [31, 32]. In our context, we need to use the algebraic version developed by J. Li [35, 36]. The algorithm of LLLZ can be summarized as follows.

R1. LLLZ defined formal relative and absolute Gromov-Witten invariants for *relative formal toric Calabi-Yau (FTCY) 3-folds*. These invariants are refinements and generalizations of open and closed Gromov-Witten invariants of smooth toric Calabi-Yau 3-folds.

R2. Formal relative and absolute Gromov-Witten invariants satisfy the degeneration formula. In particular, they can be expressed in terms of $\tilde{C}_{\vec{\mu}}(\lambda;\mathbf{n})$, formal relative Gromov-Witten invariants of an indecomposable relative FTCY 3-fold. The degeneration formula agrees with the gluing formula in O2, with $C_{\vec{\mu}}(\lambda;\mathbf{n})$ replaced by $\tilde{C}_{\vec{\mu}}(\lambda;\mathbf{n})$.

R3. $\tilde{C}_{\vec{\mu}}(\lambda;\mathbf{n}) = q^{(\sum_{i=1}^3 \kappa_{\mu^i} n_i)/2}\tilde{\mathcal{W}}_{\vec{\mu}}(q)$, where

$$\begin{aligned}\tilde{\mathcal{W}}_{\rho_1,\rho_2,\rho_3}(q) =& q^{-(\kappa_{\rho_1}-2\kappa_{\rho_2}-\frac{1}{2}\kappa_{\rho_3})/2}\sum c^{\nu^+}_{(\nu^1)^t\rho^2}c^{\rho^1}_{(\eta^1)^t\nu^1}c^{\rho^3}_{\eta^3\nu^3}\\ &\cdot q^{(-2\kappa_{\nu^+}-\frac{\kappa_{\nu^3}}{2})/2}\mathcal{W}_{\nu^+,\nu^3}(q)\frac{1}{z_\mu}\chi_{\eta^1}(\mu)\chi_{\eta^3}(2\mu)\end{aligned} \tag{13}$$

A more detailed survey of [39] can be found in [43] and in [37, Section 3].

3.1.3. Comparison. The equivalence of the physical theory and the mathematical theory of the topological vertex boils down to the following identity of classical symmetric functions:

$$\tilde{\mathcal{W}}_{\mu^1,\mu^2,\mu^3}(q) = \mathcal{W}_{\mu^1,\mu^2,\mu^3}(q). \tag{14}$$

The 1-leg case $\tilde{\mathcal{W}}_{\mu,\emptyset,\emptyset} = \mathcal{W}_{\mu,\emptyset,\emptyset}$ is directly related to a formula of Hodge integrals conjectured by Mariño-Vafa [46] and proved in [44] and in [52] by different methods. The 2-leg case $\tilde{\mathcal{W}}_{\mu^1,\mu^2,\emptyset} = \mathcal{W}_{\mu^1,\mu^2,\emptyset}$ is directly related to a formula of two-partition Hodge integrals (see [39, 45]). The full 3-leg case of (14) follows from the results in [49] (c.f. Section 4.1).

3.2. The BKMP conjecture. Given an affine plane curve

$$C = \{(x, y) \in \mathbb{C}^2 \mid \mathcal{E}(x, y) = 0\},$$

B. Eynard and N. Orantin [16] constructed the k-point correlation functions to order g, $W_k^{(g)}(p_1, \ldots, p_k)$, which are meromorphic multilinear forms on C, for any $g \in \mathbb{Z}_{\geq 0}$ and $k \in \mathbb{Z}_{>0}$. They are determined by the initial values $W_1^{(0)} = 0$, $W_2^{(0)}(p_1, p_2) = B(p_1, p_2)$ (the Bergmann kernel on C), and a recursive equation.

V. Bouchard, A. Klemm, M. Mariño and S. Pasquetti [9] adapted the recursive method of Eynard-Orantin and constructed

$$W^{(g)}_{h_1,\ldots,h_k}(p^1_1, \ldots, p^1_{h_1}; \cdots; p^k_1, \ldots, p^k_{h_k}; n_1, \ldots, n_k)$$

from the mirror curve of a toric Calabi-Yau 3-fold X. They conjectured that after a mirror transformation, the integrated correlation functions

$$\begin{aligned} & A^{(g)}_{h_1,\ldots,h_k}(p^1_1, \ldots, p^1_{h_1}; \cdots; p^k_1, \ldots, p^k_{h_k}; n_1, \ldots, n_k) \\ = & \int W^{(g)}_{h_1,\ldots,h_k}(p^1_1, \ldots, p^1_{h_1}; \cdots; p^k_1, \ldots, p^k_{h_k}; n_1, \ldots, n_k) \end{aligned}$$

are equal to the open Gromov-Witten invariants of X relative to k Lagrangian submanifolds $L_1, \ldots, L_k$ with framings $n_1, \ldots, n_k \in \mathbb{Z}$. By R1 of Section 3.1.2, these open Gromov-Witten invariants can be defined mathematically as formal relative GW invariants of an FTCY 3-fold relative to k divisors. The BKMP conjecture has been proved for the framed 1-legged topological vertex by L. Chen [12] and by J. Zhou [64]. J. Zhou later proved the conjecture for the framed 3-legged topological vertex [65].

Note that the BKMP conjecture is a different algorithm from the topological vertex. For the framed 3-legged vertex, the topological vertex provides a closed formula of a generating function of invariants of all genera for fixed winding numbers μ^1, μ^2, μ^3; the BKMP conjecture provides a closed formula of a generating function of invariants of all winding numbers for a fixed genus g.

4. Generalization to Non-toric Non-Calabi-Yau 3-Folds

Let X be a nonsingular, possibly non-Calabi-Yau, projective 3-fold.

4.1. GW/DT Correspondence. The Donaldson-Thomas invariants of X are defined via integration against a virtual fundamental class over Hilbert schemes of curves in X. Unlike Gromov-Witten invariants, Donaldson-Thomas invariants are *integers*, and are defined only in dimension 3. The foundation of Donaldson-Thomas theory was developed by Donaldson and Thomas [15, 62].

D. Maulik, N. Nekrasov, A. Okounkov, and R. Pandharipande conjectured a correspondence between the GW (Gromov-Witten) and DT (Donaldson-Thomas) theories for any nonsingular projective 3-fold [47, 48]. This correspondence can also be formulated for certain noncompact 3-folds in the presence of a torus action; the correspondence for toric Calabi-Yau 3-folds is equivalent to the algorithm of the topological vertex [47, 53]. For non-Calabi-Yau toric 3-folds the building block is the equivariant vertex (see [47, 48, 55, 56]) which depends on equivariant parameters. D. Maulik, A. Oblomkov, A. Okounkov and R. Pandharipande proved GW/DT correspondence for all toric 3-folds [49].

4.2. GW/PT Correspondence. R. Pandharipande and R.P. Thomas define integral invariants counting pairs (C, D) where $C \subset X$ is an embedded curve and $D \subset C$ is a divisor [55]. Pandharipande-Thomas (PT) invariants are defined via integration against a virtual fundamental class over the moduli space of stable pairs, viewed as objects in the derived category of X. They conjecture that the PT invariants are equal to the reduced DT invariants (obtained, roughly, from DT invariants by removing contributions from zero dimensional subschemes). This leads to a conjectural GW/PT correspondence for nonsingular projective 3-folds, and for certain noncompact 3-folds in the presence of a torus action [55, 56, 57]. D. Maulik, R. Pandharipande, and R.P. Thomas proved the GW/PT correspondence for all toric 3-folds [51].

References

[1] M. Aganagic, A. Klemm, M. Mariño, and C. Vafa, "The topological vertex," Comm. Math. Phys. **254** (2005), no. 2, 425–478.

[2] M. Aganagic, M. Mariño, and C. Vafa, "All loop topological string amplitudes from Chern-Simons theory," Comm. Math. Phys. **247** (2004), no. 2, 467–512.

[3] M.F. Atiyah and R. Bott, "The moment map and equivariant cohomology," Topology **23** (1984), no. 1, 1–28.

[4] K. Behrend, "Gromov-Witten invariants in algebraic geometry," Invent. Math. **127** (1997), no. 3, 601–617.

[5] K. Behrend and B. Fantechi, "Intrinsic normal cone," Invent. Math. **128** (1997), no.1, 45–88.

[6] K. Behrend and Y. Manin, "Stacks of stable maps and Gromov-Witten invariants," Duke Math. J. **85** (1996), no. 1, 1–60.

[7] M. Bershadsky, S. Cecotti, H. Ooguri, and C. Vafa, "Holomorphic anomalies in topological field theories," Nuclear Phys. **B 405** (1993), no. 2-3, 279–304.

[8] M. Bershadsky, S. Cecotti, H. Ooguri, and C. Vafa, "Kodaira-Spencer theory of gravity and exact results for quantum string amplitudes," Comm. Math. Phys. **165** (1994), no. 2, 311–427.

[9] V. Bouchard, A. Klemm, M. Mariño, and S. Pasquetti, "Remodelling the B-model," Comm. Math. Phys. **287** (2009), no. 1, 117–178.

[10] 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 Phys. **B 359** (1991), no. 1, 21–74.

[11] H.-L. Chang and J. Li, in preparation.

[12] L. Chen, "Bouchard-Klemm-Marino-Pasquetti Conjecture for $\mathbb{C}^3$," arXiv:0910.3739.

[13] D.A. Cox and S. Katz, *Mirror symmetry and algebraic geometry*, Mathematical Surveys and Monographs **68**, American Mathematical Society, Providence, RI, 1999.

[14] D.-E. Diaconescu and B. Florea, "Localization and gluing of topological amplitudes," Comm. Math. Phys. **257** (2005), no. 1, 119–149.

[15] S.K. Donaldson and R.P. Thomas, "Gauge theory in higher dimensions," The geometric universe (Oxford, 1996), 31–47, Oxford Univ. Press, Oxford, 1998.

[16] B. Eynard and N. Orantin, "Invariants of algebraic curves and topological expansion," Commun. Number Theory Phys. **1** (2007), no. 2, 347–452.

[17] C. Faber, "Algorithms for computing intersection numbers on moduli spaces of curves, with an application to the class of the locus of Jacobians," New trends in algebraic geometry (Warwick, 1996), 93–109, London Math. Soc. Lecture Note Ser., 264, Cambridge Univ. Press, Cambridge, 1999.

[18] C. Faber and R. Pandharipande, "Hodge integrals and Gromov-Witten theory," Invent. Math **139** (2000), no.1, 173–199.

[19] K. Fukaya and K. Ono, "Arnold conjecture and Gromov-Witten invariant," Topology **38** (1999), no. 5, 933–1048.

[20] W. Fulton and R. Pandharipande, "Notes on stable maps and quantum cohomology," Algebraic geometry—Santa Cruz 1995, 45–96, Proc. Sympos. Pure Math., **62**, Part 2, Amer. Math. Soc., Providence, RI, 1997.

[21] A.B. Givental, "Equivariant Gromov-Witten invariants," Internat. Math. Res. Notices **1996**, no. 13, 613–663.

[22] R. Gopakumar and C. Vafa, "On the gauge theory/geometry correspondence," Adv. Theor. Math. Phys. **3** (1999), no. 5, 1415–1443.

[23] R. Gopakumar and C. Vafa, "M-Theory and Topological Strings–II," arXiv:hep-th/9812127.

[24] T. Graber and R. Pandharipande, "Localization of virtual classes," Invent. Math. **135** (1999), no. 2, 487–518.

[25] M.-X. Huang, A. Klemm, and S. Quackenbush, "Topological String Theory on Compact Calabi-Yau: Modularity and Boundary Conditions," arXiv:hep-th/0612125.

[26] H. Hofer, K. Wysocki, and E. Zehnder, "A General Fredholm Theory I: A Splicing-Based Differential Geometry," J. Eur. Math. Soc. (JEMS) **9** (2007), no. 4, 841–876.

[27] H. Hofer, K. Wysocki, and E. Zehnder, "A General Fredholm Theory II: Implicit Function Theorems," Geom. Funct. Anal. **19** (2009), no. 1, 206–293.

[28] H. Hofer, K. Wysocki, and E. Zehnder, "A general Fredholm theory. III. Fredholm functors and polyfolds," Geom. Topol. **13** (2009), no. 4, 2279–2387.

[29] H. Hofer, K. Wysocki, and E. Zehnder, "Integration Theory for Zero Sets of Polyfold Fredholm Sections," Math. Ann. **346** (2010), no. 1, 139–198.

[30] Y. Hu and J. Li, "Genus-One Stable Maps, Local Equations, and Vakil-Zinger's desingularization," arXiv:0812.4286.

[31] E.-N. Ionel and T. Parker, "Relative Gromov-Witten invariants," Ann. of Math. (2) **157** (2003), no. 1, 45–96.

[32] E.-N. Ionel and T. Parker, "The symplectic sum formula for Gromov-Witten invariants," Ann. of Math. (2) **159** (2004), no. 3, 935–1025.

[33] M. Kontsevich, "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.

[34] A. Li and Y. Ruan, "Symplectic surgery and Gromov-Witten invariants of Calabi-Yau 3-folds," Invent. Math. **145** (2001), no. 1, 151–218.

[35] J. Li, "Stable Morphisms to singular schemes and relative stable morphisms," J. Diff. Geom. **57** (2001), 509–578.

[36] J. Li, "A degeneration formula of Gromov-Witten invariants," J. Diff. Geom. **60** (2002), 199–293.

[37] J. Li, "Recent progress in GW-invariants of Calabi-Yau threefolds," Current developments in mathematics, 2007, 77–99, Int. Press, Somerville, MA, 2009.

[38] B.H. Lian, K. Liu, and S.-T. Yau, "Mirror principle I," Asian J. Math. **1** (1997), no. 4, 729–763.

[39] J. Li, C.-C.M. Liu, K. Liu, and J. Zhou, "A mathematical theory of the topological vertex," Geom. Topol. **13** (2009), no. 1, 527–621.

[40] J. Li and G. Tian, "Virtual moduli cycles and Gromov-Witten invariants of algebraic varieties," J. Amer. Math. Soc. **11** (1998), no. 1, 119–174.

[41] J. Li and G. Tian, "Virtual moduli cycles and Gromov-Witten invariants of general symplectic manifolds," Topics in symplectic 4-manifolds (Irvine, CA, 1996), 47–83, First Int. Press Lect. Ser., I, Int. Press, Cambridge, MA, 1998.

[42] J. Li and A. Zinger, "On the genus-one Gromov-Witten invariants of complete intersections," J. Differential Geom. **82** (2009), no. 3, 641–690.

[43] C.-C.M. Liu, "Gromov-Witten invariants of toric Calabi-Yau 3-folds," arXiv:0811.4713, to appear in Handbook of Geometry Analysis (Vol II) ALM **13**.

[44] C.-C.M. Liu, K. Liu, and J. Zhou, "A proof of a conjecture of Mariño-Vafa on Hodge Integrals," J. Differential Geom. **65** (2003), no. 2, 289–340.

[45] C.-C.M. Liu, K. Liu, and J. Zhou, "A formula of two-partition Hodge integrals," J. Amer. Math. Soc. **20** (2007), no. 1, 149–184.

[46] M. Mariño and C. Vafa, "Framed knots at large N," Orbifolds in mathematics and physics (Madison, WI, 2001), 185–204, Contemp. Math., **310**, Amer. Math. Soc., Providence, RI, 2002.

[47] D. Maulik, N. Nekrasov, A. Okounkov, and R. Pandharipande, "Gromov-Witten theory and Donaldson-Thomas theory I," Compos. Math. **142** (2006), no. 5, 1263–1285.

[48] D. Maulik, N. Nekrasov, A. Okounkov, R. and Pandharipande, "Gromov-Witten theory and Donaldson-Thomas theory II," Compos. Math. **142** (2006), no. 5, 1286–1304.

[49] D. Maulik, A. Oblomkov, A. Okounkov, and R. Pandharipande, "Gromov-Witten/Donaldson-Thomas correspondence for toric 3-folds," arXiv:0809.3976.

[50] D. Maulik and R. Pandharipande, "A topological view of Gromov-Witten theory," Topology **45** (2006), no. 5, 887–918.

[51] D. Maulik, R. Pandharipande, and R.P. Thomas, "Curves on K3 surfaces and modular forms," arXiv:1001.2719.

[52] A. Okounkov and R. Pandharipande, "Hodge integrals and invariants of the unknot," Geom. Topol. **8** (2004), 675–699.

[53] A. Okounkov, N. Reshetikhin, and C. Vafa, "Quantum Calabi-Yau and classical crystals," The unity of mathematics, 597–618, Progr. Math., **244**, Birkhäuser Boston, Boston MA, 2006.

[54] R. Pandharipande, "Three questions in Gromov-Witten theory," Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002), 503–512, Higher Ed. Press, Beijing, 2002.

[55] R. Pandharipande and R.P. Thomas, "Curve counting via stable pairs in the derived category," Invent. Math. **178** (2009), no. 2, 407–447.

[56] R. Pandharipande and R.P. Thomas, "The 3-fold vertex via stable pairs," Geom. Topol. **13** (2009), no. 4, 1835–1876.

[57] R. Pandharipande and R.P. Thomas, "Stable pairs and BPS invariants," J. Amer. Math. Soc. **23** (2010), no. 1, 267–297.

[58] Y. Ruan, "Virtual neighborhoods and pseudo-holomorphic curves," Proceedings of 6th Gökova Geometry-Topology Conference, Turkish J. Math. **23** (1999), no. 1, 161–231.

[59] Y. Ruan, G. Tian, "A mathematical theory of quantum cohomology," J. Differential Geom. **42** (1995), no. 2, 259–367.

[60] Y. Ruan, G. Tian, "Higher genus symplectic invariants and sigma models coupled with gravity," Invent. Math. **130** (1997), no. 3, 455–516.

[61] B. Siebert, "Gromov-Witten invariants of general symplectic manifolds," arXiv:dg-ga/9608005.

[62] R.P. Thomas, "A holomorphic Casson invariant for Calabi-Yau 3-folds, and bundles on $K3$ fibrations," J. Differential Geom. **54** (2000), no. 2, 367–438.

[63] R. Vakil and A. Zinger, "A desingularization of the main component of the moduli space of genus-one stable maps into $\mathbb{P}^n$," Geom. Topol. **12** (2008), no. 1, 1–95.

[64] J.Zhou, "Local Mirror Symmetry for One-Legged Topological Vertex," arXiv:0910.4320.

[65] J. Zhou, "Local Mirror Symmetry for the Topological Vertex," arXiv:0911.2343.

[66] A. Zinger, "Reduced genus-one Gromov-Witten invariants," J. Differential Geom. **83** (2009), no. 2, 407–460.

[67] A. Zinger, "Standard vs. reduced genus-one Gromov-Witten invariants," Geom. Topol. **12** (2008), no. 2, 1203–1241.

[68] A. Zinger, "The reduced genus 1 Gromov-Witten invariants of Calabi-Yau hypersurfaces," J. Amer. Math. Soc. **22** (2009), no. 3, 691–737.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Flips and Flops

Christopher D. Hacon[†] and James M^c^Kernan[*]

Abstract

Flips and flops are elementary birational maps which first appear in dimension three. We give examples of how flips and flops appear in many different contexts. We describe the minimal model program and some recent progress centred around the question of termination of flips.

Mathematics Subject Classification (2010). Primary 14E30

Keywords. Flips, Flops, Minimal model program, Mori theory.

1. Birational Geometry

1.1. Curves and Surfaces. Before we start talking about flips perhaps it would help to understand the birational geometry of curves and surfaces. For the purposes of exposition we work over the complex numbers, and we will switch freely between the algebraic and holomorphic perspective.

Example 1.1. *Consider the function*

$$\phi\colon \mathbb{C}^2 \dashrightarrow \mathbb{C} \qquad \textit{defined by the rule} \qquad (x,y) \longrightarrow y/x.$$

Geometrically this is the function which assigns to every point (x,y) *the slope of the line connecting* $(0,0)$ *to* (x,y)*. This function is not defined where* $x=0$ *(the slope is infinite here). One can partially remedy this situation by replacing the complex plane* $\mathbb{C}$ *by the Riemann sphere* $\mathbb{P}^1 = \mathbb{C} \cup \{\infty\}$*. We get a function*

$$\phi\colon \mathbb{C}^2 \dashrightarrow \mathbb{P}^1 \qquad \textit{defined by the rule} \qquad (x,y) \longrightarrow [X:Y].$$

[†]Department of Mathematics, University of Utah, 155 South 1400 East, JWB 233, Salt Lake City, UT 84112, USA. E-mail: hacon@math.utah.edu.

[*]Department of Mathematics, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. E-mail: mckernan@math.mit.edu.

Christopher Hacon was supported by NSF grant 0757897. James M^c^Kernan was supported by NSF Grant 0701101 and by the Clay Mathematics Institute. We would like to thank Chenyang Xu for his help with much of the contents of this paper. We would also like to thank the referee for some helpful comments.

However ϕ is still not defined at the origin of $\mathbb{C}^2$. Geometrically this is clear, since it does not make sense to ask for the slope of the line connecting the origin to the origin. In fact, if one imagines approaching the origin along a line through the origin then ϕ is constant along any such line and picks out the slope of this line. So it is clear that we cannot extend ϕ to the whole of $\mathbb{C}^2$ continuously.

It is convenient to have some notation to handle functions which are not defined everywhere:

Definition 1.2. *Let X be an irreducible quasi-projective variety and let Y be any quasi-projective variety. Consider pairs (f, U), where $U \subset X$ is an open subset and $f\colon U \longrightarrow Y$ is a morphism of quasi-projective varieties. We say two pairs (f, U) and (g, V) are* ***equivalent*** *if there is an open subset $W \subset U \cap V$ such that $f|_W = g|_W$.*

A ***rational map*** *$\phi\colon X \dashrightarrow Y$ is an equivalence class of pairs (f, U).*

In fact if ϕ is represented by (f, U) and (g, V) then ϕ is also represented by $(h, U \cup V)$ where

$$h(x) = \begin{cases} f(x) & x \in U \\ g(x) & x \in V. \end{cases}$$

So there is always a largest open subset where ϕ is defined, called the **domain** of ϕ, denoted $\operatorname{dom}\phi$. The locus of points not in the domain of ϕ is called the **indeterminacy** locus.

Example 1.3. *Let C be the conic in $\mathbb{P}^2$ defined by the equation*

$$X^2 + Y^2 = Z^2.$$

Consider the rational map

$$\phi\colon C \dashrightarrow \mathbb{P}^1 \qquad \text{defined by the rule} \qquad [X : Y : Z] \longrightarrow [X : Z - Y].$$

It would seem that ϕ is not defined where both $X = 0$ and $Y = Z$ and of course $X^2 + Y^2 = Z^2$, that is, at the point $[0 : 1 : 1]$.

If one passes to the open subset $U = \mathbb{C}^2$, where $Z \neq 0$, and introduces coordinates $x = X/Z$ and $y = Y/Z$ then $C_0 = C \cap U$ is defined by the equation $x^2 + y^2 = 1$ and the map above reduces to the function

$$C_0 \dashrightarrow \mathbb{C} \qquad \text{defined by the rule} \qquad (x, y) \longrightarrow x/(1 - y),$$

which would again not seem to be defined at the point $(0, 1)$ of the curve C_0.

However note that $(Z - Y)(Y + Z) = Z^2 - Y^2 = X^2$ on the curve C. Therefore, on the open set $C - \{[0 : 1 : 1], [0 : -1 : 1]\}$,

$$[X(Y + Z) : (Z - Y)(Y + Z)] = [X(Y + Z) : X^2] = [Y + Z : X].$$

Thus ϕ is also the function

$$\phi\colon C \dashrightarrow \mathbb{P}^1 \qquad \textit{defined by the rule} \qquad [X : Y : Z] \longrightarrow [Y + Z : X].$$

It is then clear that ϕ is in fact a morphism, defined on the whole of the smooth curve C.

In fact the most basic result in birational geometry is that every map from a smooth curve to a projective variety always extends to a morphism:

Lemma 1.4. *Let $f\colon C \dashrightarrow X$ be a rational map from a smooth curve to a projective variety. Then f is a morphism, that is, the domain of f is the whole of C.*

Proof. As X is a closed subset of $\mathbb{P}^n$, it suffices to show that the composition $C \dashrightarrow \mathbb{P}^n$ is a morphism. So we might as well assume that $X = \mathbb{P}^n$. C is abstractly a Riemann surface. Working locally we might as well assume that $C = \Delta$, the unit disk in the complex plane $\mathbb{C}$. We may suppose that f is defined outside of the origin and we want to extend f to a function on the whole unit disk. Let z be a coordinate on the unit disk. Then f is locally represented by a function

$$z \longrightarrow [f_0 : f_1 : \cdots : f_n],$$

where each f_i is a meromorphic function of z with a possible pole at zero. It is well known that $f_i(z) = z^{m_i} g_i(z)$, where $g_i(z)$ is holomorphic and does not vanish at zero and $m_0, m_1, \ldots, m_n$ are integers. Let $m = \min m_i$. Then f is locally represented by the function

$$z \longrightarrow [h_0 : h_1 : \cdots : h_n],$$

where $h_i(z) = z^{-m} f_i(z)$. As $h_0, h_1, \ldots, h_n$ are holomorphic functions and at least one of them does not vanish at zero, it follows that f is a morphism. $\square$

Note that the birational classification of curves is easy. If two curves are smooth and birational then they are isomorphic. In particular, two curves are birational if and only if their normalisations are isomorphic.

Definition 1.5. *Let $\phi\colon X \dashrightarrow Y$ be a rational map between two irreducible quasi-projective varieties. The* ***graph of*** *ϕ, denoted Γ_ϕ, is the closure in $X \times Y$ of the graph of the function $f\colon U \longrightarrow Y$, where ϕ is represented by the pair (f, U). We say that ϕ is* ***proper*** *if the projection of Γ_ϕ down to Y is a proper morphism.*

Note that the inclusion of $\mathbb{C}$ into $\mathbb{P}^1$ is not proper. In this paper, we will only be concerned with proper rational maps.

Definition 1.6. *Consider the rational function*

$$\phi\colon \mathbb{C}^2 \dashrightarrow \mathbb{C} \qquad \textit{defined by the rule} \qquad (x, y) \longrightarrow y/x,$$

which appears in (1.1). *Then the graph* $\Gamma_\phi \subset \mathbb{C}^2 \times \mathbb{P}^1$ *is the zero locus of the polynomial* $xT = yS$, *where* (x, y) *are coordinates on* $\mathbb{C}^2$ *and* $[S : T]$ *are homogeneous coordinates on* $\mathbb{P}^1$. *Consider projection onto the first factor* $\pi\colon \Gamma_\phi \longrightarrow \mathbb{C}^2$. *Away from the origin this morphism is an isomorphism but the inverse image* E *of the origin is a copy of* $\mathbb{P}^1$. π *is called the* ***blow up*** *of the origin and* E *is called the* ***exceptional divisor***.

We note that π *has a simple description in terms of toric geometry.* $\mathbb{C}^2$ *corresponds to the cone spanned by* $(0, 1)$ *and* $(1, 0)$. Γ_ϕ *is the union of the two cones spanned by* $(1, 0)$ *and* $(1, 1)$ *and* $(1, 1)$ *and* $(0, 1)$*; it is obtained in an obvious way by inserting the vector* $(1, 1)$.

Given any smooth surface S, *we can define the blow up of a point* $p \in S$ *by using local coordinates. More generally given any smooth quasi-projective variety* X *and a smooth subvariety* V, *we may define the blow up* $\pi\colon Y \longrightarrow X$ *of* V *inside* X. π *is a birational morphism, which is an isomorphism outside* V. *The inverse image of* V *is a divisor* E*; the fibres of* E *over* V *are projective spaces of dimension one less than the codimension of* V *in* X *and in fact* E *is a projective bundle over* V. V *is called the* ***centre of*** E.

For example, to blow up one of the axes in $\mathbb{C}^3$, *the toric picture is again quite simple. Start with the cone spanned by* $(1, 0, 0)$, $(0, 1, 0)$ *and* $(0, 0, 1)$, *corresponding to* $\mathbb{C}^3$ *and insert the vector* $(1, 1, 0) = (1, 0, 0) + (0, 1, 0)$. *We get two cones one spanned by* $(1, 0, 0)$, $(1, 1, 0)$ *and* $(0, 0, 1)$ *and the other spanned by* $(0, 1, 0)$, $(1, 1, 0)$ *and* $(0, 0, 1)$. *To blow up the origin, insert the vector* $(1, 1, 1)$. *There are then three cones. One way to encode this data a little more efficiently is to consider the triangle (two dimensional simplex) spanned by* $(1, 0, 0)$, $(0, 1, 0)$ *and* $(0, 0, 1)$ *and consider the intersection of the corresponding cones with this triangle.*

1.2. Strong and weak factorisation.

We have the following consequence of resolution of singularities, see [16]:

Theorem 1.7 (Resolution of indeterminancy; Hironaka). *Let* $\phi\colon X \dashrightarrow Y$ *be a rational map between two quasi-projective varieties.*

If X *is smooth, then there is a sequence of blow ups* $\pi\colon W \longrightarrow X$ *along smooth centres such that the induced rational map* $\psi\colon W \longrightarrow Y$ *is a morphism.*

For surfaces we can do much better in the case of a birational map:

Theorem 1.8. *Let* $\phi\colon S \dashrightarrow T$ *be a birational map between two smooth quasi-projective surfaces.*

Then there is a smooth surface W *and two birational morphisms* $\pi\colon W \longrightarrow S$ *and* $\pi'\colon W \longrightarrow T$, *both of which are compositions of blow ups along smooth centres.*

Example 1.9. *Consider the function*

$$\phi\colon \mathbb{P}^2 \dashrightarrow \mathbb{P}^2 \quad \textit{defined by the rule} \quad [X : Y : Z] \longrightarrow [X^{-1} : Y^{-1} : Z^{-1}].$$

Then ϕ is a birational map, an involution of $\mathbb{P}^2$. As

$$[X^{-1} : Y^{-1} : Z^{-1}] = [YZ : XZ : YZ],$$

it is not hard to see that ϕ sends the three coordinate axes to the coordinate points. But then it follows that the coordinate points $[1 : 0 : 0]$, $[0 : 1 : 0]$ and $[0 : 0 : 1]$ are part of the indeterminacy locus of ϕ. In fact, if we blow up $\pi\colon W \longrightarrow \mathbb{P}^2$ the three coordinate points, then ϕ blows down the strict transform of the three coordinate axes $\pi'\colon W \longrightarrow \mathbb{P}^2$.

Consider the standard fan for $\mathbb{P}^2$, given by the union of the three cones spanned by $(1,0)$, $(0,1)$ and $(-1,-1)$. Blowing up the coordinate points, corresponds to inserting the three vectors $(1,1) = (1,0) + (0,1)$, $(0,-1) = (1,0)+(-1,-1)$ and $(-1,0) = (-1,-1)+(0,1)$. The resulting fan is the fan for the toric variety W. Note that the strict transforms of the three coordinate axes are now contractible as $(1,0) = (1,1) + (0,-1)$, $(-1,-1) = (-1,0) + (0,-1)$ and $(0,1) = (1,1) + (-1,0)$.

1.3. Flips and Flops. It is conjectured that a result similar to (1.8) holds in all dimensions:

Conjecture 1.10 (Strong factorisation). *Let $\phi\colon X \dashrightarrow Y$ be a birational map between two quasi-projective varieties.*

Then there is a quasi-projective variety W and two birational morphisms $\pi\colon W \longrightarrow X$ and $\pi'\colon W \longrightarrow Y$ which are both the composition of a sequence of blow ups of smooth centres.

Unfortunately we only know a weaker statement, see [1] and [44]:

Theorem 1.11 (Weak factorisation: Abramovich, Karu, Matsuki, Włodarczyk; Włodarczyk). *Let $\phi\colon X \dashrightarrow Y$ be a birational map between two quasi-projective varieties.*

Then we may factor ϕ into a sequence of birational maps $\phi_1, \phi_2, \ldots, \phi_m$, $\phi_i\colon X_i \dashrightarrow X_{i+1}$ and there are quasi-projective varieties $W_1, W_2, \ldots, W_m$ and two birational morphisms $\pi_i\colon W_i \longrightarrow X_i$ and $\pi'_i\colon W_i \longrightarrow X_{i+1}$ which are both the composition of a sequence of blow ups of smooth centres.

The problem is that beginning with threefolds there are birational maps which are isomorphisms in codimension two:

Example 1.12. *Suppose we start with $\mathbb{C}^3$ and blow up both the x-axis and the y-axis. Suppose we first blow up the x-axis and then the y-axis to get $X \longrightarrow \mathbb{C}^3$. Let E_x be the exceptional divisor over the x-axis, with strict transform E'_x and let E_y be the exceptional divisor over the y-axis. The strict transform of the y-axis intersects E_x in a point. When we blow this up, we also blow up this point of E_x. So E'_x has one reducible fibre with two components and E_y is a $\mathbb{P}^1$-bundle over the y-axis. If we blow up $Y \longrightarrow \mathbb{C}^3$ in the opposite order then E_x is now a $\mathbb{P}^1$-bundle and the strict transform E'_y contains one reducible fibre. The*

resulting birational map $X \dashrightarrow Y$ is an isomorphism outside the extra copies of $\mathbb{P}^1$ belonging to E'_x and E'_y. On the other hand it is not an isomorphism along these curves. This is the simplest example of a flop.

The language of fans and toric geometry is very convenient. We start with the cone spanned by $e_1 = (1,0,0)$, $e_2 = (0,1,0)$ and $e_3 = (0,0,1)$. Blowing up the x-axis corresponds to inserting the vector $e_2 + e_3$ and we get two cones, σ_1, spanned by e_1, $e_2 + e_3$ and e_2 and σ_2 spanned by e_1, $e_2 + e_3$ and e_3. Blowing up the y-axis we insert the vector $e_1 + e_3$, so that we subdivide σ_2 into two more cones, one spanned by e_1, $e_2 + e_3$ and $e_1 + e_3$ and the other spanned by e_3, $e_2 + e_3$, $e_1 + e_3$.

Now suppose that we reverse the order. At the first step we insert the vector $e_1 + e_3$ and we get two cones, τ_1 spanned by e_1, $e_1 + e_3$ and e_2 and τ_2 spanned by e_2, $e_1 + e_3$ and e_3. At the next step we insert the vector $e_2 + e_3$, and subdivide τ_2 into two cones, one spanned by e_2, $e_1 + e_3$ and $e_2 + e_3$ and the other spanned by e_3, $e_1 + e_3$ and $e_2 + e_3$.

In fact to prove (1.10) it suffices to prove it in the special case of toric varieties. For an interesting explanation of the difficulties in proving strong factorisation, see [17].

There is another way to construct this flop:

Example 1.13. *Let Q be the quadric cone $xz - yt = 0$ inside $\mathbb{C}^4$. If we blow up the origin we get a birational morphism $W \longrightarrow Q$ with exceptional E divisor isomorphic $\mathbb{P}^1 \times \mathbb{P}^1$. We can partially contract E, by picking one of the projection maps, $W \longrightarrow X$ and $W \longrightarrow Y$. The resulting birational map $X \dashrightarrow Y$ is the same as the flop introduced above.*

Perhaps the easiest way to see this is to use toric geometry. Q corresponds to the cone spanned by four vectors v_1, v_2, v_3 and v_4 in $\mathbb{R}^3$, belonging to the standard lattice $\mathbb{Z}^3$, any three of which span the standard lattice, such that $v_1 + v_3 = v_2 + v_4$. W corresponds to inserting the vector $v_1 + v_3$ and subdividing the cone into four subcones. X and Y correspond to the two different ways to pair off the four maximal cones into two cones.

One particularly nice feature of the toric description is that we can modify the picture above to get lots of examples of flips and flops. Suppose we pick any four vectors v_1, v_2, v_3 and v_4 belonging to the standard lattice which span a strongly convex cone. Then $a_1v_1 + a_3v_3 = a_2v_2 + a_4v_4$, for some positive integers a_1, a_2, a_3 and a_4. Once again we can insert the vector $a_1v_1 + a_3v_3$ and pair off the resulting cones to get two different toric threefolds X and Y which are isomorphic in codimension one.

The simplest example of a flip is when $2v_1 + v_3 = v_2 + v_4$. If we start with the wall connecting v_2 and v_4 then the flip corresponds to replacing this by the wall connecting v_1 and v_3. X has one singular point, which is a $\mathbb{Z}^2$-quotient singularity, corresponding to the cone spanned by v_2, v_3 and v_4. Indeed, $2v_1$ is an integral linear combination of these vectors but not v_1. On the other hand, Y is smooth.

Another place that flops appear naturally is in the example of a Cremona transformation of $\mathbb{P}^3$.

Example 1.14. *Consider the function*

$$\phi\colon \mathbb{P}^3 \dashrightarrow \mathbb{P}^3 \quad \textit{defined by the rule} \quad [X:Y:Z:T] \longrightarrow [X^{-1}:Y^{-1}:Z^{-1}:T^{-1}].$$

Then ϕ is a birational automorphism of $\mathbb{P}^3$. The graph of this function first blows up the four coordinate points, to get four copies of $\mathbb{P}^2$, then the six coordinate axes, to get six copies of $\mathbb{P}^1 \times \mathbb{P}^1$. The reverse map then blows down those six copies of $\mathbb{P}^1 \times \mathbb{P}^1$, but this time using the other projection and then we finally blow down the strict transforms of the four coordinate planes.

Note that if we just blow up the four coordinate points on both sides then the resulting threefolds are connected by six flops. All of this is easy to describe using toric geometry; the picture is similar to the picture above of the Cremona transformation of $\mathbb{P}^2$.

One can use flops to construct some interesting examples.

Example 1.15 (Hironaka). *Suppose we start with $X = \mathbb{P}^3$ and two conics C_1 and C_2 which intersect in two points p and q. Imagine blowing up both C_1 and C_2 but in a different order at p and q. Suppose we blow up first C_1 and then C_2 over p but first C_2 and then C_1 over q. Let $\pi\colon M \longrightarrow \mathbb{P}^3$ be the resulting birational map.*

We claim that even the exceptional locus $E_1 \cup E_2$ is not a projective variety. Let l be general fibre of the exceptional divisor E_1 over C_1, let $l_1 + l_2$ be the reducible fibre over p, let m be the general fibre of E_2 over C_2 and let m_1+m_2 be the reducible fibre over q. Suppose the irreducible fibre of E_2 over p is attached to l_1 and the irreducible fibre of E_1 over q is attached to m_1. Note that

$$m_1 \equiv l \equiv l_1 + l_2 \equiv m + l_2 \equiv m_1 + m_2 + l_2,$$

where $\equiv$ denotes numerical equivalence. This implies that $l_2 + m_2 \equiv 0$. If M is projective then a hyperplane class H would intersect $l_2 + m_2$ positively, a contradiction.

Note that M is related to a projective variety Y over X by an (analytic) flop. Just flop either l_2 or m_2.

Example 1.16 (Atiyah). *Suppose that we start with a family of quartic surfaces in $\mathbb{P}^3$ degenerating to a quartic surface with a simple node (a singularity which in local analytic coordinates resembles $x^2 + y^2 + z^2 = 0$ in $\mathbb{C}^3$). It is a simple matter to find a degeneration whose total space has a singularity locally of the form $xz - yt$. In this case we can blow up this singularity in two different ways, see* (1.13)*, to get two different families of smooth K3 surfaces, which are connected by a flop.*

But now we have two distinct families of K3 surfaces, which agree outside one point. In fact even though the families are different they have isomorphic fibres. It follows that the moduli space of K3 surfaces is not Hausdorff.

Example 1.17 (Reid). *Let $X_0 \subset \mathbb{C}^4$ be the smooth threefold given by the equation*

$$y^2 = ((x-a)^2 - t_1)((x-b)^2 - t_2),$$

where x, y, t_1, t_2 are coordinates on $\mathbb{C}^4$ and $a \neq b$ are constants. Let X be the closure of X_0 in $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{C}^2$. Projection $\pi\colon X \longrightarrow \mathbb{C}^2$ down to $\mathbb{C}^2$ with coordinates t_1 and t_2 realises X as a family of projective curves of genus one over $\mathbb{C}^2$. If $t_1 t_2 \neq 0$ then we have a smooth curve of genus one, that is an elliptic curve. If $t_1 = 0$ and $t_2 \neq 0$ or $t_2 = 0$ and $t_1 \neq 0$ then we get a nodal rational curve (a copy of $\mathbb{P}^1$ with two points identified). If $t_1 = t_2 = 0$ then we get a pair $C_1 \cup C_2$ of copies of $\mathbb{P}^1$ joined at two points.

One can check that both C_1 and C_2 can be contracted individually to a simple node. Therefore we can flop either C_1 or C_2. Suppose that we flop C_1. Since C_1 is contracted by π this flop is over S so that the resulting threefold Y admits a morphism to $\psi\colon Y \longrightarrow \mathbb{C}^2$. We haven't changed the morphism π outside s and one can check that the fibre over $(0,0)$ of ψ is a union $D_1 \cup D_2$ of two copies of $\mathbb{P}^1$ which intersect in two different points. Once again we can flop either of these curves. Suppose that D_2 is the strict transform of C_2 so that D_1 is the flopped curve. If we flop D_1 then we get back to X but if we flop D_2 then we get another threefold which fibres over S. Continuing in this way we get infinitely many threefolds all of which admit a morphism to S and all of which are isomorphic over the open set $S - \{s\}$. Let G be the graph whose vertices are these threefolds, where we connect two vertices by an edge if there is a flop between the two threefolds over S. Let G' be the graph whose vertices are the integers where we connect two vertices i and j if and only if $|i-j| = 1$. Then G and G' are isomorphic.

2. Minimal Model Program

The idea behind the minimal model program (which we will abbreviate to MMP) is to find a particularly simple birational representative of every projective variety. For curves we have already seen that two smooth curves are birational if and only if they are isomorphic. For surfaces there are non-trivial birational maps, but by (1.8) only if there are rational curves (non-constant images of $\mathbb{P}^1$). Roughly speaking, simple means that we cannot contract any more rational curves. In practice it turns out that we don't want to contract every curve, just those curves on which the canonical divisor is negative.

Definition 2.1. *Let X be a normal projective variety. A **divisor** $D = \sum n_i D_i$ is a formal linear combination of codimension one subvarieties.*

*The **canonical divisor** K_X is the divisor associated to the zeroes and poles of any meromorphic differential form ω.*

Note that the canonical divisor is really an equivalence class of divisors.

Example 2.2. *If $X = \mathbb{P}^1$ and z is the standard coordinate on $\mathbb{C}$ then dz/z is a meromorphic differential form. It has a pole at zero and a pole at infinity, since*

$$\frac{d(1/z)}{1/z} = -\frac{dz}{z}.$$

If p represents zero and q infinity then $K_{\mathbb{P}^1} = -p - q$. If we started with dz/z^2 then $K_{\mathbb{P}^1} = -2p$ (a double pole at zero) but if we start with dz then $K_{\mathbb{P}^1} = -2q$ (a double pole at infinity). And so on. If X is an elliptic curve E then it is a one dimensional complex torus, the quotient of $\mathbb{C}$ by a lattice isomorphic to $\mathbb{Z}^2$. In this case the differential form dz descends to the torus (as it is translation invariant) and $K_E = 0$ (no zeroes or poles). If C has genus $g \geq 2$ then $\deg K_C = 2g - 2 > 0$.

For $\mathbb{P}^n$ we have $K_{\mathbb{P}^n} = -(n+1)H$, where H is the class of a hyperplane. More generally still, suppose X is a projective toric variety. Then a dense open subset of X is isomorphic to a torus $(\mathbb{C}^*)^n$. A natural holomorphic differential n-form which is invariant under the action of the torus is

$$\frac{dz_1}{z_1} \wedge \frac{dz_2}{z_2} \wedge \cdots \wedge \frac{dz_n}{z_n}.$$

This form extends naturally to a meromorphic differential on the whole toric variety with simple poles along the invariant divisors. In other words,

$$K_X + \Delta \sim_{\mathbb{Q}} 0,$$

where $\Delta = \sum D_i$ is a sum of the invariant divisors. In the case of $\mathbb{P}^n$ there are $n+1$ invariant divisors corresponding to the $n+1$ coordinate hyperplanes.

One of the most useful ways to compute the canonical divisor is the adjunction formula. If M is a smooth variety and X is a smooth divisor then

$$(K_M + X)|_X = K_X.$$

For example, if X is a quartic surface in $\mathbb{P}^3$ then

$$K_X = (K_{\mathbb{P}^3} + X)|_X = (-4H + 4H)|_X = 0.$$

Together with the fact that smooth hypersurfaces of dimension at least two are simply connected this implies that X is a K3 surface.

Suppose that $T \longrightarrow S$ is the blow up of a point with exceptional divisor $E \simeq \mathbb{P}^1$. It is straightforward to check that the self-intersection $E^2 = E \cdot E = -1$. By adjunction we have

$$-2 = K_{\mathbb{P}^1} = K_E = (K_T + E)|_E = K_T \cdot E + E^2.$$

It follows that $K_T \cdot E = -1$. For obvious reasons we call any such curve a -1-curve. The idea of the MMP is to only contract curves on which the canonical divisor is negative.

Definition 2.3. *Let X be a normal projective variety and let D be a Cartier divisor (something locally defined by a single equation). We say that D is* ***nef*** *if $D \cdot C \geq 0$ for every curve $C \subset X$.*

Let us first see how the minimal model program works for surfaces.

Step 0: Start with a smooth surface S.

Step 1: Is K_S nef? If yes, then stop. S is a minimal model.

Step 2: If no, then there must be a curve C such that $K_S \cdot C < 0$. We can always choose C so that there is a contraction morphism $\pi\colon S \longrightarrow T$ which contracts C and we are in of the following three cases:

(i) $S = \mathbb{P}^2$, T is a point and C is a line.

(ii) T is a curve, S is a $\mathbb{P}^1$-bundle over T and C is a fibre.

(iii) T is a smooth surface, π is a blow up of a point on T and C is the exceptional divisor.

Step 3: If we are in case (i) or (ii), then stop. Otherwise replace S by T and go back to Step 1.

The fact that we can always find a curve C to contract is a non-trivial result, due to the Italian school of algebraic geometry. It is possible that at Step 1 there is more than one choice of π.

Example 2.4. *Suppose that we start with the blow up S of $\mathbb{P}^2$ at two different points p and q. There are three relevant curves, E and F the exceptional divisors over p and q and L, the strict transform of the line connecting p and q.*

At the first step of the K_S-MMP we are presented with three choices. We can choose to contract E, F or L, since all three of these curves are -1-curves. If we contract E, $\pi\colon S \longrightarrow T$, then at the next step we are presented with two choices of curves to contract on T. We can either contract the image of F, in which case the end product of the MMP is the original $\mathbb{P}^2$. On the other hand, there is a morphism $T \longrightarrow \mathbb{P}^1$. Every fibre is isomorphic to $\mathbb{P}^1$, L is a fibre and F is a section of this morphism. This is a possible end product of the MMP. If instead we decide to contract F, then we get almost exactly the same picture; note however that even though the two $\mathbb{P}^1$-bundles we get are isomorphic, the induced birational map between them is not an isomorphism. However if we choose to contract L then the resulting surface is isomorphic to $\mathbb{P}^1 \times \mathbb{P}^1$. Projection to either factor $\mathbb{P}^1 \times \mathbb{P}^1 \longrightarrow \mathbb{P}^1$ are two possible other end products of the MMP.

Once again the language of toric geometry gives a convenient way to encode this picture. S corresponds to the fan with one dimensional rays spanned by $(1,0)$, $(0,1)$, $(-1,0)$, $(-1,-1)$ and $(0,-1)$. Blowing down E and F corresponds to removing the two rays $(-1,0)$ and $(0,-1)$. Blowing down E corresponds to removing $(-1,0)$ and the morphism to $\mathbb{P}^1$ corresponds to the projection of $\mathbb{R}^2$

onto the x-axis. Contracting L corresponds to removing $(-1,-1)$; the resulting fan is clearly the fan for $\mathbb{P}^1 \times \mathbb{P}^1$.

The most important feature of any algorithm is termination. Termination for surfaces is clear. Every time we contract a copy of $\mathbb{P}^1$, topologically we are replacing a copy of the sphere S^2 by a point. Consequently the second Betti number $b_2(S)$ drops by one and so the MMP terminates after at most $b_2(S)$-steps. Equivalently the Picard number drops by one at every step.

One interesting application of the MMP for surfaces is in the construction of a compactification $\overline{M}_g$ of the moduli space of curves M_g of genus $g \geq 2$. In particular suppose we are given a family $\pi\colon S_0 \longrightarrow C_0$ of smooth projective curves over a smooth affine curve C_0. Then there is a unique projective curve C which contains C_0 as an open subset. We would like to complete S_0 to a family of curves $\pi\colon S \longrightarrow C$, which makes the following diagram commute:

$$\begin{array}{ccc} S_0 & \longrightarrow & S \\ {\scriptstyle \pi_0}\downarrow & & \downarrow{\scriptstyle \pi} \\ C_0 & \longrightarrow & C. \end{array}$$

Here the horizontal arrows are inclusions. We would like the fibres of π to be nodal projective curves, whose canonical divisor is ample. The first observation is that this is not in fact possible. In general we can only fill in this family after a finite cover of C_0.

Here is the general algorithm. The first step is to pick any compactification of S_0 and of the morphism π_0. The next step is to blow up S, so that the reduced fibres are curves with nodes. After this we take a cover of C and replace S by the normalisation of the fibre product. If the cover of C is sufficiently ramified along the singular fibres of π this step will eliminate the multiple fibres. The penultimate step is to run the MMP over C. This has the effect of contracting all -1-curves contained in the fibres of π. The final step is to contract all -2-curves, that is, all copies of $\mathbb{P}^1$ with self-intersection -2, which are fibres of π.

We now consider the MMP in higher dimension. There is a similar picture, except that we also encounter flips:

Definition 2.5. *Let $\pi\colon X \longrightarrow Z$ be a birational morphism. We say that π is* ***small*** *if π does not contract a divisor. We say that π is a* ***flipping contraction*** *if $-K_X$ is ample over Z and the relative Picard number is one. The* ***flip*** *of π is another small birational morphism $\psi\colon Y \longrightarrow Z$ of relative Picard number one such that K_Y is ample over Z.*

The relative Picard number is the difference in the Picard numbers. The relative Picard number is one if and only if every two curves contracted by π are numerically multiples of each other. In this case a $\mathbb{Q}$-Cartier divisor D is ample over Z if and only if $D \cdot C > 0$ for one curve C contracted by π.

Flops are defined similarly, except that now K_X and K_Y are trivial over Z and yet the induced birational map $X \dashrightarrow Y$ is not an isomorphism. The MMP in higher dimensions proceeds as follows:

Step 0: Start with a smooth projective variety X.

Step 1: Is K_X nef? If yes, then stop. X is a minimal model.

Step 2: If no, then there must be a curve C such that $K_X \cdot C < 0$. We can always choose C so that there is a contraction morphism $\pi\colon X \longrightarrow Z$ which contracts C and there are two cases:

(i) $\dim Z < \dim X$. C is contained in a fibre. The fibres F of π are Fano varieties, so that $-K_F$ is ample. π is a Mori fibre space.

(ii) $\dim Z = \dim X$. In this case π is birational and there are two sub cases:

  (a) π contracts a divisor E.

  (b) π is small.

Step 3: If we are in case (i), then stop. If we are in case (a) then replace X by Z and go back to (1). If we are in case (b) then replace X by the flip $X \dashrightarrow Y$ and go back to (1).

The fact that we may find C and π at step 2 is quite subtle, and is due to the work of many people, including Kawamata, Kollár, Miyaoka, Mori, Reid, Shokurov and many others. For more details see, for example, the book by Kollár and Mori, [25]. For an excellent survey of flips and flops, especially for threefolds, see [22]. We should also point out that if we are in step 3 it is possible (and in fact common) for Z to be singular, even if we just contract a divisor. However the singularities are mild ($\mathbb{Q}$-factorial terminal singularities) and this algorithm works with these singularities.

Existence of terminal 3-fold flips was first proved by Mori, [31]. Kollár and Mori give a complete classification of all terminal 3-fold flips in [24], at least when the flipping curve is irreducible. Shokurov proved the existence of 4-fold flips, [40]. Existence in all dimensions was proved in [14] and [15]:

Theorem 2.6 (Existence: Hacon, M^c^Kernan). *Flips exist in all dimensions.*

Actually stating things this way is a considerable simplification; we also need the main result of [6] to finish a somewhat involved induction. The proof of (2.6) draws considerable inspiration and ideas from two sources. First, Siu's theory of multiplier ideals and his proof of deformation invariance of plurigenera, see [43], especially the recasting of these ideas in the algebraic setting [19], due to Kawamata. Second, Shokurov's theory of saturation of the restricted algebras and his proof of the existence of flips for fourfolds, [40], all of which is succinctly explained in Corti's book, [9].

We have already seen (2.4) that the end product of the MMP is not unique. For surfaces the minimal model is unique. If X is a threefold and $X \dashrightarrow Y$ is a

flop then X is minimal if and only if Y is minimal, so there is often more than minimal model. In fact, Kawamata [20] proved that any two minimal models are connected by a sequence of flops.

Example 2.7. *Suppose we start with the elliptic fibration $\pi\colon X \longrightarrow S$ given in* (1.17)*. Possibly replacing S by a finite cover, we may assume that S contains no rational curves. Suppose that we run the K_X-MMP. At every step of the MMP the locus we contract is covered by rational curves. It follows that every step of the MMP is over S and the end product of the MMP is a minimal model. The MMP preserves the property that one isolated fibre is the union of two copies of $\mathbb{P}^1$ meeting in two points. It follows that X has infinitely many minimal models.*

Kawamata has similar examples of Calabi-Yau threefolds with infinitely many minimal models, [18].

If we get down to a Mori fibre space the situation is considerably more complicated, as (1.9) and (1.14) demonstrate. However Sarkisov proposed a way to use the MMP to connect any two birational Mori fibre spaces by a sequence of four types of elementary links, see [8]. The Sarkisov program was recently shown to work in all dimensions in [13].

Note that termination of the MMP is far more subtle in dimension at least three. It is clear that we cannot keep contracting divisors. As in the case of surfaces the relative Picard number drops every time we contract a divisor and is unchanged under flips and so we can only contract a divisor finitely many times. However it is far less clear which discrete invariants improve after each flip.

Conjecture 2.8. *There is no infinite sequence of flips.*

The rest of this paper will be devoted to exploring (2.8).

We know that the MMP always works for toric varieties, due to the work of Reid, [35] and Kawamata, Matsuda and Matsuki, [21]. The proof is almost entirely combinatorial.

3. Local Approach to Termination

We review the first approach to the termination of flips. The idea is to find an invariant of X which has three properties:

1. The invariant takes values in an ordered set I.
2. The invariant always increases after a flip.
3. The set I satisfies the ascending chain condition (abbreviated to ACC).

Typically the invariant is some measure of the complexity of the singularities of X. Usually it is not hard to ensure that properties (1) and (2) hold. There are many sensible ways to measure the complexity of a singularity and flips

tend to improve singularities. The most subtle part seems to be checking that (3) holds as well.

The most naive invariant of any singularity is the multiplicity. If $X \subset \mathbb{C}^{n+1}$ and X is defined by the analytic function $f(z_1, z_2, \dots, z_n)$ the **multiplicity** m of X at the origin is the smallest positive integer such that $f \in \mathfrak{m}^m = \langle z_1, z_2, \dots, z_n \rangle^m$. If we take the reciprocal of the multiplicity then the set

$$I = \{ \frac{1}{m} \mid m \in \mathbb{N} \},$$

is naturally ordered and clearly satisfies the ACC. Unfortunately it is hard to keep track of the behaviour of the multiplicity under flips.

The idea is to pick an invariant which is more finely-tuned to the canonical divisor:

Definition 3.1. *Let X be a normal quasi-projective variety. A **log resolution** is a projective morphism $\pi\colon Y \longrightarrow X$ such that Y and the exceptional locus is **log smooth**, that is, Y is smooth and the exceptional locus is a divisor with simple normal crossings.*

If K_X is $\mathbb{Q}$-Cartier then we may write

$$K_Y + E = \pi^* K_X + \sum a_i E_i,$$

*where $E = \sum E_i$ and a_i are rational numbers. The **log discrepancy of** E_i **with respect to** K_X is a_i. The **log discrepancy** of X is the infimum of the a_i, over all exceptional divisors on all log resolutions.*

*We say that X is **terminal**, **canonical**, **log terminal**, **log canonical** if $a > 1$, $a \geq 1$, $a > 0$ and $a \geq 0$.*

*If $V \subset X$ is a closed subset, then the **log discrepancy** of X at V is the infimum of the a_i, over all exceptional divisors whose image is V, and all log resolutions.*

*The **log discrepancy along** V is the infimum of the a_i, over all exceptional divisors whose image is contained in V, and all log resolutions.*

Let us start with some simple examples.

Example 3.2. *Let S be a smooth surface and let $p \in S$. Let $\pi\colon T \longrightarrow S$ blow up p, with exceptional divisor E. Suppose we write*

$$K_T + E = \pi^* K_S + aE.$$

If we intersect both sides with E then we get

$$-2 = K_{\mathbb{P}^1} = K_E = (K_T + E) \cdot E = (\pi^* K_S + aE) \cdot E = aE^2 = -a.$$

So $a = 2$. It is a simple matter to check that if we blow up more over the point p then every exceptional divisor has log discrepancy greater than two. So the log

discrepancy of a smooth surface is 2. *It is also not hard to check that if* X *is not log canonical then the log discrepancy is* $-\infty$ *and that if* X *is log canonical the log discrepancy is the minimum of the log discrepancy of the exceptional divisors of any log resolution.*

If X *is an affine toric variety, corresponding to the cone* σ, *then* K_X *is* $\mathbb{Q}$-*Cartier if the primitive generators of the one dimensional faces of* $\sigma \subset \mathbb{R}^n$ *lie in an affine hyperplane (this is always the case if* σ *is simplicial). In this case there is a linear functional* $\phi\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *which takes the value* 1 *on this hyperplane. The log discrepancy of any toric divisor is the value of* ϕ *on the primitive generator of the extremal ray corresponding to this divisor. In particular* X *is log terminal.*

For example if X *is smooth of dimension* n, *then* X *corresponds to the cone spanned by the standard generators* $e_1, e_2, \ldots, e_n$ *of the standard lattice* $\mathbb{Z}^n \subset \mathbb{C}^n$. *If we insert the vector* $e_1 + e_2$ *then the log discrepancy of the exceptional divisor of the blow up of the corresponding codimension two coordinate subspace is* 2 *and this is the log discrepancy of* X. *If we insert the sum* $e_1 + e_2 + \cdots + e_n$ *this corresponds to blowing up the origin. The log discrepancy of the exceptional divisor is* n *and this is the log discrepancy of* X *at the origin.*

Lemma 3.3. *Let* X *be a normal variety.*

Then X *is the disjoint union of finitely many locally closed subsets* $Z_1, Z_2, \ldots, Z_m$ *and there is a function* f *such that the log discrepancy of* X *at a subvariety* V *is equal to* $f(i, d)$, *where* V *has dimension* d *and* $1 \leq i \leq m$ *is the unique index such that* $V \cap Z_i$ *is dense in* V.

Proof. Pick a log resolution $\pi\colon Y \longrightarrow X$. If V is a subvariety of X then the log discrepancy at V is either computed by an exceptional divisor of π, or it is computed by some divisor which is exceptional over Y. There are only finitely many divisors extracted by π and the log discrepancy of a subvariety over Y is just determined by its dimension and the list of exceptional divisors which contain it. □

Proposition 3.4. *If* $\pi\colon X \dashrightarrow Y$ *is a flip, then the log discrepancy of any divisor* E *never goes down and always goes up if the centre of* E *is contained in the indeterminacy locus of* π.

Proof. See (5.11) of [21]. □

Definition 3.5 (Shokurov). *Let* X *be a threefold with canonical singularities. The* ***difficulty*** *of* X *is the number of divisors of log discrepancy less than two.*

Lemma 3.6. *Let* X *be a threefold with canonical singularities.*

Then

1. *the difficulty is finite, and*
2. *the difficulty always goes down under flips.*

Proof. It is easy to check that (1) holds by direct computation on a log resolution.

Let $\phi\colon X \dashrightarrow Y$ be a flip. Let C be a flipped curve, that is, a curve contained in the indeterminacy locus of ϕ^{-1}. As the log discrepancy goes up under flips, Y is terminal about a general point of C. It follows that Y is smooth along the generic point of C so that there is an exceptional divisor E with centre C of log discrepancy two. The log discrepancy of E with respect to X must be less than two, by (3.4). It follows that the difficulty decreases by at least one, which is (2). □

Note that (3.6) easily implies that there is no infinite sequence of flips, starting with a threefold with canonical singularities. There have been many papers which extend Shokurov's work to higher dimensions, most especially to dimension four, for example [28], [11] and [3]. Unfortunately in higher dimensions there are infinitely many divisors of log discrepancy at most two and singular varieties of log discrepancy greater than two. It seems hard to control the situation using only the difficulty.

To remedy this situation, Shokurov has proposed some amazing conjectural properties of the log discrepancy:

Conjecture 3.7 (Shokurov). *Fix a positive integer n. The set*

$$L_n = \{\, a \in \mathbb{Q} \mid a \text{ is the log discrepancy at a subvariety of a normal variety of dimension } n \,\},$$

satisfies the ACC.

Conjecture 3.8 (Ambro, Shokurov). *Let X be a quasi-projective variety. The function*

$$a\colon X \longrightarrow \mathbb{Q},$$

which sends a point x to the log discrepancy of X at x is lower semi-continuous.

Theorem 3.9 (Shokurov). *Assume* $(3.7)_n$ *and* $(3.8)_n$.

Then every sequence of flips in dimension n terminates, that is, $(2.8)_n$ *holds.*

Proof. We sketch Shokurov's beautiful argument.

Suppose not, that is, suppose we are given an infinite sequence of flips $\phi_i\colon X_i \dashrightarrow X_{i+1}$. Let E_i be the locus of indeterminacy of ϕ_i. Then E_i is a closed subset of X_i.

Let a_i be the log discrepancy of X_i along E_i. Let

$$\alpha_i = \inf\{\, a_j \mid j \geq i \,\}.$$

Then $\alpha_i \leq \alpha_{i+1}$, with equality, unless $\alpha_i = a_i$. As we are assuming $(3.7)_n$ it follows that α_i is eventually constant. Suppose that $\alpha_i = a$, for i sufficiently large. (3.3) implies that the sets

$$I_i = \{\, l \mid l \text{ is the log discrepancy at a subvariety of } X_i \,\},$$

are finite. If $j > i$ and $a_j \leq a_l$ for all $i \leq l \leq j$, then (3.4) implies that $a_j \in I_i$. Since I_i is finite, it follows that $a_i \geq a$ for all i, with equality for infinitely many i. Let

$$J = \{\, i \in \mathbb{N} \mid a_i = a \,\}.$$

By assumption for each $i \in J$ there is a log resolution and at least one exceptional divisor F_i whose centre is contained in E_i such that the log discrepancy of F_i is a. Let d_i be the maximal dimension of the centre on X_i of any such exceptional divisor F_i. Pick d such that $d_i \leq d$ for all but finitely many $i \in J$, with equality for infinitely many $i \in J$.

Let

$$W_i' = \{\, x \in X_i \mid x \in V,\ \dim V = d,\ \text{log discrepancy of } X \text{ at } V \text{ is at most } a \,\}.$$

As we are assuming $(3.8)_n$, $W_i' \subset X_i$ is a closed subset. Let W_i be the union of those components of W_i' for which there is a subvariety V of dimension d such that the log discrepancy of X_i at V is a. Then (3.3) implies that if $V \subset W_i$ is a closed subset of dimension d, then the log discrepancy of X_i at V is at most a with equality if V passes through the general point of W_i.

(3.4) implies that every component of W_{i+1} is birational to a unique component of W_i. It follows that eventually W_i and W_{i+1} have the same number of components. Let $\phi_i \colon W_i \dashrightarrow W_{i+1}$ be the induced birational map. ϕ_i is eventually an isomorphism along any centre in W_i of dimension d. If $V \subset W_{i+1}$ is of dimension d then the log discrepancy of X_{i+1} at V is at most a. It follows that ϕ_i^{-1} must be an isomorphism along V, since the log discrepancy of X_i along E_i is a and log discrepancies only go up under flips, (3.4).

If ϕ_i is not an isomorphism in dimension d, then it must contract a subvariety of dimension d. But this cannot happen infinitely often, a contradiction. $\square$

Note that there are more general versions of (3.7) and (3.8), which involve log pairs (X, Δ) and that Shokurov proves that if one assumes these more general conjectures then any sequence of log flips terminates. For more details see [41].

Unfortunately both (3.7) and (3.8) seem to be hard conjectures. We know $(3.7)_2$ and $(3.8)_2$, by virtue of Alexeev's classification of log canonical surface singularities. We know that

$$L_3 \cap [1, \infty) = \left\{ 1 + \frac{1}{r} \mid r \in \mathbb{N} \cup \{\infty\} \right\} \cup \{3\},$$

by virtue of the classification of terminal singularities due to Mori, [30] and Reid [36] and a result of Kawamata, see the appendix to [37]. Borisov [7] proved that (3.7) holds for toric varieties. Ambro proved [4] that $(3.8)_3$ holds and that (3.8) holds for toric varieties.

One interesting consequence of (3.8) is the following:

Conjecture 3.10 (Shokurov). *Let X be a normal quasi-projective variety of dimension n.*

Then the log discrepancy of any point is at most n.

Indeed if $x \in X$, then pick a curve C which contains x and intersects the smooth locus X_0 of X. Then x is the limit of points $y \in C \cap X_0$. We have already seen that the log discrepancy of X at y is n. So if we assume $(3.8)_n$ then the log discrepancy of X at x is at most n.

Note that to prove (3.10) we may assume that the log discrepancy is greater than one, that is, we may assume that X is terminal. Even though (3.10) would appear to be much weaker than (3.8), we only know that $(3.10)_3$ holds by virtue of Mori's classification of threefold terminal singularities and a result of Markushevich, [27].

4. Global Approach to Termination

Instead of focusing on showing that some invariant satisfies the ACC, the global approach to termination tries to use the global geometry of X. At this point it is convenient to work with:

Definition 4.1. *A **log pair** (X, Δ) is a normal variety together with a divisor $\Delta \geq 0$ such that $K_X + \Delta$ is $\mathbb{R}$-Cartier.*

One can define the log discrepancy and the various flavours of log terminal, just as for the canonical divisor.

Example 4.2. *Let X be a toric variety and let $\Delta = \sum D_i$ be the sum of the invariant divisors. Then $K_X + \Delta \sim_{\mathbb{Q}} 0$ so that (X, Δ) is a log pair. We may find $\pi\colon Y \longrightarrow X$ a toric log resolution. Note that*

$$K_Y + \Gamma = \pi^*(K_X + \Delta),$$

where $\Gamma = \sum G_i$ is the sum of the invariant divisors on Y, since both sides are zero. As π is toric, Γ contains all of the exceptional divisors with coefficient one. It follows that (X, Δ) is log canonical.

We use the following finiteness result:

Theorem 4.3 (Birkar, Cascini, Hacon, McKernan). *Let X be a smooth projective variety. Fix an ample divisor A and finitely many divisors $B_1, B_2, \ldots, B_k$ such that $(X, \sum B_i)$ is log smooth.*

Then there are finitely many $1 \leq i \leq m$ rational maps $\phi_i\colon X \dashrightarrow Y_i$ such that if $(b_1, b_2, \ldots, b_k) \in [0,1]^k$ and $\phi\colon X \dashrightarrow Y$ is a weak log canonical model of $K_X + A + \sum b_i B_i$ then $\phi = \phi_i$ for some index $1 \leq i \leq m$.

We have already remarked (2.7) that there are examples due to Reid of threefolds with infinitely many minimal models. The presence of the divisor A is therefore important in the statement of (4.3). However Shokurov [38] proves a similar result for threefolds, but now without the ample divisor A and shows that the same result holds in all dimensions if one knows the abundance conjecture, (5.7). In a related direction, Kawamata, [18] and Morrison [32] have conjectured that the number of minimal models is finite up to birational automorphisms of X, when X is Calabi-Yau and Δ is empty.

We use (4.3) to run a special MMP, known as the MMP with scaling.

Step 0: Start with a projective variety X, an ample divisor A, a divisor $B = \sum b_i B_i$, where $(X, \sum B_i)$ is log smooth and $(b_1, b_2, \ldots, b_k) \in [0,1]^k$ and an ample divisor H such that $K_X + A + B + H$ is nef.

Step 1: Let

$$\lambda = \inf\{\, t \in [0,1] \mid K_X + A + B + tH \text{ is nef}\},$$

be the nef threshold.

Step 2: Is $\lambda = 0$? If yes, then stop.

Step 3: If no, then there must be a curve C such that $(K_X + A + B) \cdot C < 0$ and $(K_X + A + B + \lambda H) \cdot C = 0$. We can always choose C so that there is a contraction morphism $\pi\colon X \longrightarrow Z$ which contracts C and there are two cases:

(i) $\dim Z < \dim X$. C is contained in a fibre. $-(K_X + A + B)$ is ample on a fibre.

(ii) $\dim Z = \dim X$. In this case π is birational and there are two subcases:

  (a) π contracts a divisor E.

  (b) π is an isomorphism in codimension at least two.

Step 4: If we are in case (i), then stop. If we are in case (a) then replace X by Z and go back to (2). If we are in case (b) then replace X by the flip $X \dashrightarrow Y$ and go back to (2).

Note that if H is any ample divisor then $K_X + A + B + tH$ is ample for any t sufficiently large. So finding an ample divisor H such that $K_X + A + B + H$ is nef is never an issue. Note also that if $\lambda = 0$ then $K_X + A + B$ is nef and we have arrived at a log terminal model. The only significant difference between the MMP with scaling and the usual MMP is that we only choose to contract those curves on which $K_X + A + B + \lambda H$ is zero. With this choice, it is easy to see that we keep the condition that $K_X + A + B + \lambda H$ is nef. More to the point, every step of the MMP is a weak log canonical model of $K_X + A + (B + \lambda H)$, for some choice of $\lambda \in [0,1]$. Finiteness of models, (4.3) and the fact that we never return to the same model, (3.4), implies that the MMP with scaling always terminates.

To run the MMP with scaling, we need the ample divisor A. If we start with $K_X + \Delta$ kawamata log terminal, we can find A ample and $B \geq 0$ such that $K_X + \Delta \sim_{\mathbb{R}} K_X + A + B$, where $K_X + B$ is kawamata log terminal if and only if Δ is big. If we start with a birational map $\pi\colon X \longrightarrow Y$ then every divisor is big over Y and so the MMP with scaling always applies if we work over Y.

For example, we may use the MMP with scaling to show that every complex manifold which is birational to a projective variety but which is not a projective variety must contain a rational curve. For example, one might modify Hironaka's example, (1.15), by starting with any smooth projective threefold X with two curves intersecting transversely at two points. It is easy to find many examples which don't contain any rational curves. But the next step involves blowing up both curves and so M contains lots of rational curves.

Shokurov [39] proved the following result assuming the full MMP and our proof is based heavily on his ideas:

Theorem 4.4 (Birkar, Cascini, Hacon, McKernan). *Let M be a complex manifold. Suppose there is a proper birational map $\pi\colon X \longrightarrow M$ such that X is smooth and projective.*

If M does not contain a rational curve then M is projective.

Proof. Pick an ample divisor H such that $K_X + H$ is ample. We run the K_X-MMP with scaling of H. Suppose that $\pi\colon X \longrightarrow Y$ is a K_X-negative contraction. By a result of Miyaoka and Mori, [29], the locus contracted by π is covered by rational cuves. As M does not contain a rational curve, it follows that π is a morphism over M. In particular the $(K_X + H)$-MMP is automatically a MMP over Y. As π is birational, it follows that the MMP with scaling terminates, as observed above. At the end we have a projective variety Y such that K_Y is nef and a birational morphism $f\colon Y \longrightarrow M$. As M is smooth it follows that f is an isomorphism so that M is a projective variety. □

5. Local-global Approach to Termination

Even though the MMP with scaling is useful, it is becoming increasingly clear that we would still like to have the full MMP, even in the special case when Δ is big. This would be useful in the construction of the moduli space of varieties of general type. One possible approach is to try to blend both the local and the global approach to termination of flips.

We have already seen that the log discrepancy always improves under flips. However the most fundamental invariant of any singularity would seem to be the multiplicity. The log canonical threshold is a more sophisticated version of the multiplicity which takes into account higher terms and is at the same time more adapted to the canonical divisor. If $X \subset \mathbb{C}^n$ is a hypersurface, then the log canonical threshold λ of X at the origin, is the largest t such that $(\mathbb{C}^n, tX)$ is log canonical in a neighbourhood of the origin. If X has multiplicity m at

the origin, then we have

$$\frac{1}{m} \leq \lambda \leq \frac{n}{m}.$$

Shokurov has conjectured that the set of log canonical thresholds should satisfy the ACC:

Conjecture 5.1 (Shokurov). *Fix a positive integer n and a subset $I \subset [0,1]$ which satisfies the descending chain condition (abbreviated to DCC).*

Then there is a finite set $I_0 \subset I$ such that if

1. *X is a variety of dimension n,*
2. *(X, Δ) is log canonical,*
3. *every component of Δ contains a non kawamata log terminal centre of (X, Δ), and*
4. *the coefficients of Δ belong to I,*

then the coefficients of Δ belong to I_0.

Example 5.2. *Let $X \subset \mathbb{C}^n$ be the hypersurface given by the equation*

$$x_1^{a_1} + x_2^{a_2} + \cdots + x_n^{a_n} = 0.$$

Then the log canonical threshold is

$$\min\left(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}, 1\right).$$

It is elementary to check that these numbers satisfy the ACC.

Theorem 5.3 (Special termination; Shokurov). *Assume $(2.8)_{n-1}$.*

Let (X, Δ) be a projective log canonical pair of dimension n. Let $\phi_i \colon X_i \dashrightarrow X_{i+1}$ be a sequence of flips. Let $Z_{ij} \subset X_i$ be the locus where the induced birational map $X_i \dashrightarrow X_j$ is not an isomorphism. Let

$$Z_i = \bigcup_{j>i} Z_{ij}.$$

Let V_i be the locus where $K_{X_i} + \Delta_i$ is not kawamata log terminal.

Then V_i and Z_i eventually don't intersect.

Note that V_i is a closed subset of X_i, whilst Z_i is a countable union of closed subsets of X_i.

Theorem 5.4 (Birkar). *Assume $(2.8)_{n-1}$ and $(5.1)_n$. Let (X, Δ) be a projective kawamata log terminal pair of dimension n.*

If there is a divisor $M \geq 0$ which is numerically equivalent to $K_X + \Delta$, then every sequence of $(K_X + \Delta)$-flips terminates.

Proof. We sketch Birkar's ingenious argument.

Let $\phi_i \colon X_i \dashrightarrow X_{i+1}$ be a sequence of $(K_X + \Delta)$-flips. Let $Z_{ij} \subset X_i$ be the locus where the induced birational map $X_i \dashrightarrow X_j$ is not an isomorphism. Let

$$Z_i = \bigcup_{j>i} Z_{ij}.$$

Let Δ_i and M_i be the strict transforms of Δ and M. Note that $K_{X_i} + \Delta_i$ is numerically equivalent to M_i. In particular $\phi_1, \phi_2, \dots$ is also a sequence of $(K_X + \Delta + tM)$-flips for any $t \geq 0$. Let

$$\lambda_i = \sup\{\, t \in \mathbb{R} \mid K_{X_i} + \Delta_i + tM_i \text{ is log canonical along } Z_i \,\},$$

be the log canonical threshold along Z_i. Note that $\lambda_i \leq \lambda_{i+1}$, as log discrepancies only go up under flips. In particular if I is the set of all coefficients of $\Delta_i + \lambda_i M_i$, then I satisfies the DCC. As we are assuming $(5.1)_n$, it follows that $\lambda_1, \lambda_2, \dots$ is eventually constant. Suppose that $\lambda_i = \lambda$, for all $i \geq i_0$. As we are assuming $(2.8)_{n-1}$, (5.3) implies that $V_i \cap Z_i$ is eventually empty, that is, the sequence of flips is finite. □

Note that we cheated a little in the proof of (5.4). Eventually $K_{X_i} + \Delta_i + \lambda_i M_i$ is not log canonical, so that strictly speaking (5.3) does not apply. In practice one can get around this by passing to a log terminal model. For more details, see [5].

To give a complete proof of termination of flips using (5.4), note that we need to do two things. Obviously we need to prove (5.1). However to complete the induction we need to deal with the case when $K_X + \Delta$ is not numerically equivalent to a divisor $M \geq 0$. This part breaks up into two separate pieces.

Definition 5.5. *Let X be a normal projective variety. We say that D is* ***pseudo-effective*** *if D is a limit of big divisors.*

Conjecture 5.6. *Suppose that $K_X + \Delta$ is kawamata log terminal.*

If $K_X + \Delta$ is pseudo-effective then there is a divisor $M \geq 0$ such that $K_X + \Delta \sim_{\mathbb{R}} M \geq 0$.

One should understand this conjecture as being part of the abundance conjecture:

Conjecture 5.7 (Abundance). *Let (X, Δ) be a projective log canonical pair.*

If $K_X + \Delta$ is nef then it is semiample.

In particular (5.6) seems very hard. One way to get around this gap in our knowledge is to assume that Δ is big. In this case we have, [6] and [42]:

Theorem 5.8 (Birkar, Cascini, Hacon, M^c^Kernan; Siu). *Suppose that $K_X + \Delta$ is kawamata log terminal.*

If $K_X + \Delta$ is pseudo-effective and Δ is big then there is a divisor $M \geq 0$ such that $K_X + \Delta \sim_{\mathbb{R}} M \geq 0$.

Lazić [26] and Păun [34] have since given simpler proofs of (5.8). Note that the steps of the MMP preserve the property that Δ is big. The final piece of the puzzle is to deal with the case that $K_X + \Delta$ is not pseudo-effective. It seems that ideas from bend and break, [29], might prove useful in this case.

Part of the appeal of this approach to termination is that (5.1) seems far more tractable than (3.7). We know (5.1) in some highly non-trivial examples. For example, Alexeev proved $(5.1)_3$ [2], using boundedness of log del Pezzo surfaces whose log discrepancy is bounded away from zero. Further, de Fernex, Ein and Mustaţă, have proved the case when X is smooth, see [10] and the references therein.

We end with some speculation about a way to attack (5.1). We first note a reduction step due originally to Shokurov, see [33]. To prove $(5.1)_n$ we just need to prove:

Corollary 5.9. *Fix a positive integer n and a subset $I \subset [0,1]$ which satisfies DCC.*

Then there is a finite set $I_0 \subset I$ such that if

1. *X is a projective variety of dimension n,*
2. *(X, Δ) is kawamata log terminal,*
3. *Δ is big,*
4. *the coefficients of Δ belong to I, and*
5. *$K_X + \Delta$ is numerically trivial,*

then the coefficients of Δ belong to I_0.

in dimension $n-1$. To this end, consider:

Conjecture 5.10. *Fix a positive integer n.*

Then there is a constant m such that if

- *X is a projective variety of dimension n,*
- *(X, Δ) is log canonical and log smooth,*
- *$K_X + \Delta$ is big and*
- *r is a positive integer such that $r(K_X + \Delta)$ is Cartier,*

then the rational map determined by the linear system $|mr(K_X + \Delta)|$ is birational.

Note that (5.10) closely resembles some results and conjectures stated in [12]. We note that this is slightly deceptive, since (5.10) seems quite a bit harder than these conjectures. Hopefully (5.10) has a better formulation, which is more straightforward to prove and has the same consequences. Note that if

we add the condition that $K_X + \Delta$ is nef then the existence of m is a result due to Kollár, [23], an effective version of the base point free theorem.

The following is standard:

Lemma 5.11. *Let X be a smooth projective variety of dimension n and let D be a Cartier divisor on X such that ϕ_D is birational.*

Then $\phi_{K_X+(2n+1)D}$ is birational.

The hope is to prove $(5.9)_n$ using:

Lemma 5.12. *Assume $(5.10)_n$. Let $I \subset [0,1]$ be a finite set and let n be a positive integer. Suppose that $I \cup \{1\}$ are linearly independent real numbers over the rationals.*

Then there is a positive real number $\epsilon > 0$ such that if

- *X is a projective variety of dimension n,*
- *(X, Δ) is log canonical and log smooth,*
- *the coefficients of Δ belong to I, and*
- *$K_X + \Delta$ is big*

then $K_X + (1-\epsilon)\Delta$ is big.

Proof. Let m be the constant given by (5.10). By simultaneous Diophantine approximation applied to the finite set I, we may pick a positive integer r with the following properties: if $a \in I$ then there is a rational number $b \geq a$ such that rb is an integer and

$$b - a < \frac{1}{2m(2n+1)r}.$$

If we set

$$t = m(2n+1)r,$$

then we may pick $\Theta \geq \Delta$ such that

$$\|\Delta - \Theta\| < \frac{1}{2t},$$

where $r\Theta$ is Cartier. By (5.11),

$$K_X + m(2n+1)r(K_X + \Theta) = (t+1)\left(K_X + \frac{t}{t+1}\Theta\right),$$

defines a birational map. In particular

$$K_X + \left(1 - \frac{1}{2t}\right)\Theta,$$

is big. So we may take

$$\epsilon = \frac{1}{2m(2n+1)r}.$$

□

References

[1] D. Abramovich, K. Karu, K. Matsuki, and J. Włodarczyk, *Torification and factorization of birational maps*, J. Amer. Math. Soc. **15** (2002), no. 3, 531–572 (electronic).

[2] V. Alexeev, *Boundedness and K^2 for log surfaces*, International J. Math. **5** (1994), 779–810.

[3] V. Alexeev, C. Hacon, and Y. Kawamata, *Termination of (many) 4-dimensional log flips*, Invent. Math. **168** (2007), no. 2, 433–448.

[4] F. Ambro, *On minimal log discrepancies*, Math. Res. Lett. **6** (1999), no. 5–6, 573–580.

[5] C. Birkar, *Ascending chain condition for log canonical thresholds and termination of log flips*, Duke Math. J. **136** (2007), no. 1.

[6] C. Birkar, P. Cascini, C. Hacon, and J. McKernan, *Existence of minimal models for varieties of log general type*, J. Amer. Math. Soc. **23** (2010), no. 2, 405–468, arXiv:math.AG/0610203.

[7] A. Borisov, *Minimal discrepancies of toric singularities*, Manuscripta Math. **92** (1987), no. 1, 33–45.

[8] A. Corti, *Factoring birational maps of threefolds after Sarkisov*, J. Algebraic Geom. **4** (1995), no. 2, 223–254.

[9] ———, *3-fold flips after Shokurov*, Flips for 3-folds and 4-folds (Alessio Corti, ed.), Oxford University Press, 2005, pp. 13–40.

[10] T. de Fernex, L. Ein, and M. Mustaţă, *Shokurov's ACC Conjecture for log canonical thresholds on smooth varieties*, Duke Math. J. **152** (2010), no. 1, 93–114, arXiv:0905.3775v3.

[11] O. Fujino, *Termination of 4-fold canonical flips*, Publ. Res. Inst. Math. Sci. **40** (2004), no. 1, 231–237.

[12] C. Hacon and J. McKernan, *Boundedness results in birational geometry.*

[13] ———, *The Sarkisov program*, arXiv:0905.0946v1.

[14] ———, *Extension theorems and the existence of flips*, Flips for 3-folds and 4-folds (Alessio Corti, ed.), Oxford University Press, 2007, pp. 79–100.

[15] ———, *Existence of minimal models for varieties of log general type II*, J. Amer. Math. Soc. **23** (2010), no. 2, 469–490, arXiv:0808.1929.

[16] H. 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.

[17] K. Karu and S. Silva, *On Oda's strong factorization conjecture*, arXiv:0911.4693v1.

[18] Y. Kawamata, *On the cone of divisors of Calabi-Yau fiber spaces*, Internat. J. Math. **8** (1997), no. 5, 665–687.

[19] ———, *On the extension problem of pluricanonical forms*, Algebraic geometry: Hirzebruch 70 (Warsaw, 1998), Contemp. Math., vol. 241, Amer. Math. Soc., Providence, RI, 1999, pp. 193–207.

[20] ———, *Flops connect minimal models*, Publ. Res. Inst. Math. Sci. **44** (2008), no. 2, 419–423.

[21] Y. Kawamata, K. Matsuda, and K. Matsuki, *Introduction to the minimal model program*, Algebraic Geometry, Sendai (T. Oda, ed.), Kinokuniya-North-Holland, 1987, Adv. Stud. Pure Math, vol 10, pp. 283–360.

[22] J. Kollár, *Flips, flops, minimal models, etc*, Surveys in differential geometry (Cambridge, MA, 1990), Lehigh Univ., Bethlehem, PA, 1991, pp. 113–199.

[23] ———, *Effective base point freeness*, Math. Ann. **296** (1993), no. 4, 595–605.

[24] J. Kollár and S. Mori, *Classification of three-dimensional flips*, J. Amer. Math. Soc. **5** (1992), 533–703.

[25] ———, *Birational geometry of algebraic varieties*, Cambridge tracts in mathematics, vol. 134, Cambridge University Press, 1998.

[26] V. Lazić, *Adjoint rings are finitely generated*, arXiv:0905.2707v2.

[27] D. Markushevich, *Minimal discrepancy for a terminal cDV singularity is* 1, J. Math. Sci. Univ. Tokyo **3** (1996), no. 2, 445–456.

[28] K. Matsuki, *Termination of flops for* 4*-folds*, Amer. J. Math. **113** (1991), no. 5, 835–859.

[29] Y. Miyaoka and S. Mori, *A numerical criterion for uniruledness*, Ann. of Math. **124** (1986), 65–69.

[30] S. Mori, *On* 3*-dimensional terminal singularities*, Nagoya Math. J. **98** (1985), 43–66.

[31] ———, *Flip theorem and the existence of minimal models for* 3*-folds*, J. Amer. Math. Soc. **1** (1988), no. 1, 117–253.

[32] D. Morrison, *Compactifications of moduli spaces inspired by mirror symmetry*, Astérisque (1993), no. 218, 243–271, Journées de Géométrie Algébrique d'Orsay (Orsay, 1992).

[33] J. M^c^Kernan and Y. Prokhorov, *Threefold Thresholds*, Manuscripta Math. **114** (2004), no. 3, 281–304.

[34] M. Păun, *Relative critical exponents, non-vanishing and metrics with minimal singularities*, arXiv:0807.3109v1.

[35] M. Reid, *Decomposition of toric morphisms*, Arithmetic and geometry, Vol. II, Progr. Math., vol. 36, Birkhäuser Boston, pp. 395–418.

[36] ———, *Canonical* 3*-folds*, Journées de Géométrie Algébrique d'Angers (Alphen aan den Rijn) (A. Beauville, ed.), Sithoff and Noordhoff, 1980, pp. 273–310.

[37] V. V. Shokurov, *Three-dimensional log perestroikas*, Izv. Ross. Akad. Nauk Ser. Mat. **56** (1992), no. 1, 105–203.

[38] ———, 3*-fold log models*, J. Math. Sci. **81** (1996), no. 3, 2667–2699, Algebraic geometry, 4.

[39] ———, *Letters of a bi-rationalist. I. A projectivity criterion*, Birational algebraic geometry (Baltimore, MD, 1996), Contemp. Math., vol. 207, Amer. Math. Soc., Providence, RI, 1997, pp. 143–152.

[40] ———, *Prelimiting flips*, Proc. Steklov Inst. of Math. **240** (2003), 82–219.

[41] ———, *Letters of a bi-rationalist. V. Minimal log discrepancies and termination of log flips*, Tr. Mat. Inst. Steklova **246** (2004), no. Algebr. Geom. Metody, Svyazi i Prilozh., 328–351.

[42] Y-T. Siu, *A General Non-Vanishing Theorem and an Analytic Proof of the Finite Generation of the Canonical Ring*, arXiv:math.AG/0610740.

[43] ———, *Invariance of plurigenera*, Invent. Math. **134** (1998), no. 3, 661–673.

[44] J. Włodarczyk, *Toroidal varieties and the weak factorization theorem*, Invent. Math. **154** (2003), no. 2, 223–331.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Quantitative Extensions of Twisted Pluricanonical Forms and Non-vanishing

Mihai Păun*

Abstract

We will discuss here a few recent applications of the analytic techniques in algebraic geometry.

Mathematics Subject Classification (2010). 14C30, 32J25, 32QXX.

Keywords. L^2 estimates, extension theorems, non-vanishing, closed positive currents, metrics with minimal singularities.

1. L^2 Extension Results

One of the most important achievements of the L^2 theory is the extension result established by T. Ohsawa and K. Takegoshi in 1987 (cf. [40]). This theorem was subsequently refined in [2], [17], [32] [35], [36], [41], [42], [48] and its implications to both algebraic and analytic geometry turned out to be fundamental: various forms of approximation of closed positive currents, study of the adjoint linear systems, deformational invariance of plurigenera and positivity of direct images, to quote only a few.

We first recall here the original result [40], and then we state a few generalizations which will be needed in the following paragraphs.

Let $\Omega \subset \mathbb{C}^n$ be a ball of radius r and let $h : \Omega \to \mathbb{C}$ be a holomorphic function, such that $\sup_\Omega |h| \leq 1$; moreover, we assume that the gradient ∂h of h is nowhere zero on the set $V := (h = 0)$. We denote by φ a plurisubharmonic function, such that its restriction to V is well-defined (i.e., $\varphi_{|V} \not\equiv -\infty$). Then the extension theorem in [40] is as follows.

*Institut Elie Cartan, Université Henri Poincaré, Nancy and Korea Institute for Advanced Studies, Seoul. E-mail: paun@iecn.u-nancy.fr.

Theorem 1.1 ([40]). *For any holomorphic function f on V, there exists a function F, holomorphic in all of Ω, such that $F = f$ on V, and moreover*

$$\int_\Omega |F|^2 \exp(-\varphi) d\lambda \leq C_0 \int_V |f|^2 \exp(-\varphi) \frac{d\lambda_V}{|\partial h|^2}.$$

Here, C_0 is an *absolute constant*: this is the main point in all applications of the result above we are aware of.

The first "geometric form" of 1.1 was obtained by Manivel in [35], and it was further refined in [17], and in [36]. In algebraic geometry, the *quantitative part* of this result (i.e., the estimates for the extension) is not available. However, the Kawamata-Viehweg vanishing theorem is used as a substitute in many applications (see e.g. [33]).

Theorem 1.2 ([35], [17], [36]). *Let X be a projective manifold, and let $S \subset X$ be the zero set of a holomorphic section $s \in H^0(X, E)$ of a line bundle $E \to X$; the hypersurface S is assumed to be non-singular. Let (L, h_L) be a line bundle, endowed with a (possibly singular) metric h, such that:*

$$\Theta_h(L) \geq 0, \quad \Theta_h(L) \geq \frac{1}{\alpha}\Theta(E) \tag{1}$$

and such that $|s|^2 \leq \exp(-\alpha)$ on X. We assume that the restriction of the metric h_L to S is well defined.
Then every section $u \in H^0\big(S, (K_X + S + L_{|S}) \otimes \mathcal{I}(h_{L|S})\big)$ admits an extension U to X such that

$$\int_X \frac{|U|^2 e^{-\varphi_L - \varphi_S}}{|s|^2(-\log|s|)^2} \leq C_0 \int_S \frac{|u|^2 e^{-\varphi_L}}{|ds|^2},$$

where $h = e^{-\varphi_L}$, provided the right hand side is finite.

Let $p : \mathcal{X} \to \mathbb{D}$ be a projective family over the unit disk, for which $0 \in \mathbb{D}$ is a regular value. We state next the version of 1.1 established by Y.-T. Siu.

Theorem 1.3 ([48]). *Let (L, h_L) be a line bundle on $\mathcal{X}$, endowed with a positively curved metric, whose restriction to the central fiber is not identically $+\infty$. Let u be a holomorphic section of the bundle $K_{\mathcal{X}} + L_{|\mathcal{X}_0}$ which is L^2 with respect to h_L, i.e.*

$$\int_{\mathcal{X}_0} |u|^2 e^{-\varphi_L} < \infty. \tag{2}$$

Then there exists a section U of the bundle $K_{\mathcal{X}} + L$ whose restriction to $\mathcal{X}_0$ is equal to u, and such that

$$\int_{\mathcal{X}} |U|^2 e^{-\varphi_L} \leq C_0 \int_{\mathcal{X}_0} |u|^2 e^{-\varphi_L}.$$

Let $m \geq 1$ be a real number. We conclude this paragraph by an Ohsawa-Takegoshi-type theorem with $L^{\frac{2}{m}}$ estimates, which is derived in [5] from the original result by a fixed point method.

Proposition 1.4 ([5]). *For any holomorphic function $f : V \to \mathbb{C}$ with the property that*

$$\int_V |f|^{2/m} \exp(-\varphi) \frac{d\lambda_V}{|\partial h|^2} < \infty,$$

there exists a function $F \in \mathcal{O}(\Omega)$ such that:

(i) $F_{|V} = f$ *i.e. the function F is an extension of f;*

(ii) *The next $L^{2/m}$ bound holds*

$$\int_\Omega |F|^{2/m} \exp(-\varphi) d\lambda \leq C_0 \int_V |f|^{2/m} \exp(-\varphi) \frac{d\lambda_V}{|\partial h|^2},$$

where C_0 is the same constant as in theorem 1.1.

Proof. We include here a sketch of the proof of the proposition above, as it is simple and flexible enough to be used in other contexts.

In the first place we can assume that the function φ is smooth, and that the functions h (respectively f) can be extended in a neighbourhood of Ω (of V inside $V \cap \overline{\Omega}$, respectively). Once the result is established under these additional assumptions, the general case follows by approximations and standard normal families arguments.

Since Ω is a bounded Stein subset of $\mathbb{C}^n$ we can certainly construct holomorphic extensions of f; among all of these extensions, we consider one which *minimizes* the following semi-norm

$$\|g\|_m^2 := \left(\int_\Omega |g|^{\frac{2}{m}} \exp(-\varphi) d\lambda \right)^m$$

and we call it F. The minimal extension F automatically satisfies the estimate (ii). Indeed, we consider the psh function $\varphi_1 := \varphi + (1 - 1/m) \log |F|^2$ on Ω, and then we have

$$\int_V |f|^2 \exp(-\varphi_1) \frac{d\lambda_V}{|\partial h|^2} = \int_V |f|^{2/m} \exp(-\varphi) \frac{d\lambda_V}{|\partial h|^2} < \infty.$$

By theorem 1.1, there exists an extension F_1 of f, such that

$$\int_\Omega \frac{|F_1|^2}{|F|^{2\frac{m-1}{m}}} \exp(-\varphi) d\lambda \leq C_0 \int_V |f|^{2/m} \exp(-\varphi) \frac{d\lambda_V}{|\partial h|^2}, \tag{3}$$

and we claim that the left hand side of (3) *is greater than* $\|F\|_m^{\frac{2}{m}}$. If this is not the case, then we have

$$\|F_1\|_m^2 < \|F\|_m^2$$

by Hölder inequality, which in turn contradicts the choice of F; the result is therefore proved. $\square$

During the next two sections, we will highlight some algebro-geometric settings for which the results above are very useful.

2. Effective Pseudoeffectivity of Relative Pluricanonical Bundles

Let $p : X \to Y$ be a surjective projective map between the non-singular manifolds X and Y. We denote by $K_{X/Y} := K_X - p^\star K_Y$ the relative canonical bundle of p. Let (L, h_L) be a line bundle on X, endowed with a (possibly singular) metric with positive curvature current, and let m be a positive integer.

In this paragraph we indicate the construction of a *natural, positively curved metric* on the bundle

$$mK_{X/Y} + L \tag{4}$$

by using *sections* of its restriction to the fibers X_y which are $L^{2/m}$-integrable with respect to h_L. Here $y \in Y$ is a very general point, and X_y is the fiber of p over y. Part of the motivation for considering this metric is to get a better understanding of the positivity properties of twisted relative canonical bundles in algebraic geometry.

To start with, we will assume that h_L *is a genuine metric (i.e. non-singular) and that the map p is a smooth fibration.*

The above data induces a metric $h^{(m)}_{X/Y}$ on the bundle $mK_{X/Y} + L$. The corresponding dual metric can be described intrinsically as follows. Let ξ be a vector in the fiber $-(mK_{X/Y} + L)_x$; then we define its norm

$$\|\xi\|^2 := \sup |U(x) \cdot \xi|^2$$

the "sup" being taken over all sections u of $mK_{X_y} + L$ such that

$$\|u\|^{2/m}_{m,y} := \int_{X_y} |u|^{\frac{2}{m}} e^{-\frac{1}{m}\varphi_L} \leq 1; \tag{5}$$

in the above equality we assume that $p(x) = y$ and we denote by U the section of the bundle $mK_{X/Y} + L_{|X_y}$ corresponding to u via the standard identification (see e.g. [4], section 1). To our knowledge, this kind of metrics first appeared in the article [39] by Narasimhan-Simha.

We now describe the local weights of the metric $h^{(m)}_{X/Y}$; we denote by (z^j) and (t^i) respectively some local coordinates centered at x, respectively y. These coordinates provides us with a trivialization of the relative canonical bundle, with respect to which the weight of $h^{(m)}_{X/Y}$ reads as

$$\exp\left(\varphi^{(m)}_{X/Y}(x)\right) = \sup_{\|u\|_{m,y} \leq 1} |f_U(x)|^2 \tag{6}$$

where $u \in H^0\big(X_y, mK_{X_y} + L\big)$ is a global section; the notations are as follows: $U := u \wedge (p^*dt)^{\otimes m}/(dt)^{\otimes m}$ is the section of bundle $mK_{X/Y} + L_{|X_y}$ corresponding to u, and the above (local) holomorphic function f_U appearing in (5) is defined by the equality $U = f_U(dz)^{\otimes m}$.

The above construction has a meaning even if the metric h_L we start with is allowed to be singular (but we still assume that the map p is a non-singular fibration). In this case some fibers X_y may be contained in the unbounded locus of h_L, i.e. $h_{L|X_y} \equiv \infty$, but for such $y \in Y$ we adopt the convention that the metric $h^{(m)}_{X/Y}$ is identically $+\infty$ as well. As for the fibers in the complement of this set, the family of sections we consider in order to define the metric consists in twisted pluricanonical forms whose m^{th} root is in L^2, normalized as in (5).

In this context, the result proved in [4] (see also [53] and the references therein) is the next one.

Theorem 2.1 ([4]). *Let $p : X \to Y$ be a proper projective non-singular fibration, and let $L \to X$ be a line bundle endowed with a metric h_L such that:*

(1) *The curvature current of the bundle (L, h_L) is positive, i.e. $\Theta_{h_L}(L) \geq 0$;*

(2) *For each $y \in Y$, all the sections of the bundle $mK_{X_y} + L$ extend near y;*

(3) *There exist $z \in Y$ and a section $u \in H^0\big(X_z, mK_{X_z} + L\big)$ such that*

$$\int_{X_y} |u|^{\frac{2}{m}} e^{-\frac{1}{m}\varphi_L} < \infty.$$

Then the above metric $h^{(m)}_{X/Y}$ has semi-positive curvature current.

The main technical tool in the proof of the theorem above is the semi-positivity of the direct image of bundles of type $K_{\mathcal{X}/\mathcal{Y}} + L$, established for smooth and proper Kähler families $p : \mathcal{X} \to \mathcal{Y}$ by B. Berndtsson in [3]. Actually, this field has a very rich and interesting history: on the analytic side, we mention the articles by F. Maitani, H. Yamaguchi, cf. [34], [59]; their results were understood geometrically and widely generalized by B. Berndtsson in [3] (see also the "foliated" version due to M. Brunella in [12]). On the algebraic geometry side, we refer to the important contributions of F. Campana, T. Fujita, P. Griffiths, A. Höring, Y. Kawamata, J. Kollár, C. Mourougane, M. S. Narasimhan-R. R. Simha, G. Schumacher, S. Takayama and E. Vieweg among many others, cf. [14], [22], [23], [24], [26], [31], [37], [39], [57], [58].

In case of a smooth projective family p, theorem 2.1 is proved for $m = 1$ by using the ideas in [3], together with a regularization argument (actually, this is the reason why we cannot get the same result for Kähler families).

If $m \geq 2$, then the "classical" argument in algebraic geometry consists in a reduction to the case $m = 1$ via a ramified cover trick. This is not how we proceed in [4], for several reasons. In order indicate our proof, we assume that

Y is equal to the unit disk $\mathbb{D}$, so that p is a smooth projective family over $\mathbb{D}$. We then consider the space

$$\mathcal{Y} := \mathbb{D} \times H^0(X, mK_{X/\mathbb{D}} + L)$$

and we regard the direct image $\mathcal{E}$ of $mK_{X/\mathbb{D}} + L$ as a (trivial) vector bundle over $\mathcal{Y}$, endowed with the non-trivial metric

$$\begin{aligned} \|u\|^2_{y,v} &:= \int_{X_y} \frac{|u|^2}{|v|^{2\frac{m-1}{m}}} e^{-\frac{\varphi_L}{m}} \\ &= \int_{X_y} |u|^2 e^{-\frac{\varphi_L}{m} - \frac{m-1}{m}\log|v|^2}. \end{aligned} \tag{7}$$

We remark that the semi-norm in (5) is the "restriction to the diagonal" (i.e. $u = v$) of the expression (7). In [4], we show that $\mathcal{E}$ endowed with the metric (7) is semi-positively curved, and moreover that this implies theorem 2.1. □

We come back here to the case of an arbitrary map $p : X \to Y$; we argue as follows. By standard semi-continuity results, there exists a (non-empty) Zariski open set $Y_0 \subset Y$ such that the condition (2) of 2.1 above is fulfilled; by restricting further the set Y_0 we can assume that if we denote by X_0 the inverse image of the set Y_0 via the map p, then $p : X_0 \to Y_0$ is a non-singular fibration.

Then we use theorem 2.1 (where we replace Y by Y_0 and X by X_0) and obtain a metric $h^{(m)}_{X_0/Y_0}$ is explicitly given over the fibers X_y, as soon as $y \in Y_0$ and the restriction of h_L to X_y is well defined. This metric *admits an extension to the whole manifold* X: the justification of this fact requires the estimate in (ii) of proposition 1.4, which provides us with an uniform $L^{\frac{2}{m}}$ bound of the functions f_U (since the sections we are using in order to construct the metric are normalized by the relation (5)). In conclusion, we have.

Theorem 2.2 ([5]). *Let $p : X \to Y$ be a proper projective fibration, and let $L \to X$ be a line bundle endowed with a metric h_L such that the curvature current of the bundle (L, h_L) is positive. Then the bundle $mK_{X/Y} + L$ admits a metric $h^{(m)}_{X/Y}$ with semi-positive curvature current, provided that there exists a general point $z \in Y$ and a section $u \in H^0\big(X_z, mK_{X_z} + L\big)$ such that*

$$\int_{X_y} |u|^{\frac{2}{m}} e^{-\frac{1}{m}\varphi_L} < \infty.$$

Moreover, the weights of the metric $h^{(m)}_{X/Y}$ are explicitly described by (6) *if $y \in Y_0$.*

We refer to [5] for the details of the proof. In the paragraph 5 of the recent preprint [50], the relevance of these topics in the context of the abundance conjecture is discussed (see also [7], [45]). □

We add here a few observations concerning the metric $h^{(m)}_{X/Y}$ (2.4 and 2.5 below turned out to be useful in [6]).

Remark 2.3. The fact that the metric $h^{(m)}_{X/Y}$ is "the right one" is also illustrated by the example provided in [4]: if $p : \widehat{X} \to X$ is the blow-up of e.g. a point, then the corresponding metric on the relative canonical bundle of p is precisely the one induced by the exceptional divisor. □

Remark 2.4. Let (L_m, h_m) be a sequence of positively curved bundles, such that the corresponding metrics have the next uniformity property: *there exists a constant $C > 0$ such that $\varphi_{L_m} \leq Cm$.* The above arguments apply to the sequence $mK_{X/Y} + L_m$; the upshot is that the weights $\varphi^{(m)}_{X/Y}$ of the corresponding metrics are *bounded by a function which is linear with respect to m.* If we denote by L_∞ the limit of $\frac{1}{m}L_m$, then weak limits of the sequence of metrics $\left(\frac{1}{m}\varphi^{(m)}_{X/Y}\right)$ exist, and provides the Chern class of the bundle $K_{X/Y} + L_\infty$ with a closed positive current, see [5]. □

Remark 2.5. One drawback of the construction of the metric $h^{(m)}_{X/Y}$ is the fact that over a (proper) Zariski closed subset $Y_1 \subset Y$ we ignore everything about its singularities. We show here that in some special cases this can be improved (see [5] for a more ample discussion).

Let $y \in Y$ be a regular value of p, which does not necessarily belong to the set Y_0, but such that the multiplier ideal sheaf associated to $h_{L|X_y}$ is the structural sheaf. Let U be a holomorphic section of the bundle $mK_{X/Y} + L$ over the whole family X. Then (modulo an abuse of notation) we have

$$\frac{|U(x)|^2 e^{-\varphi^{(m)}_{X/Y}(x)}}{\left(\int_{X_y} |U|^{2/m} e^{-\frac{\varphi_L}{m}} d\lambda\right)^m} \leq 1 \tag{8}$$

where $x \in X_y$ is an arbitrary point.

Indeed, if $y \in Y_0$, then the above claim is a consequence of the definition. If not, then we use in [5] a limit argument–since the weights $\varphi^{(m)}_{X/Y}$ are *upper semi-continuous*. Hence the extendable sections of the bundle $mK_{X/Y} + L_{|X_y}$ provides us with a *bound of the singularities* of $\varphi^{(m)}_{X/Y}$ over X_y. □

3. Minimal Singularities Metrics and their Restriction Properties

In a series of articles (cf. [47] and [48]), Y.-T. Siu established the *deformational invariance of plurigenera of smooth, projective families*, a long-standing conjecture concerning the classification of algebraic manifolds. This result plays

a key rôle in many of the recent developments in algebraic geometry (effective birationality of pluricanonical maps for general type manifolds, finiteness of the canonical algebra...). The original techniques invented to prove it can be adapted to a wide range of geometric situations, as it is illustrated by the articles [13], [18], [21], [25], [27], [30], [43], [44], [52], [53], [55].

We first recall here the result in [48]; let $p : \mathcal{X} \to \mathbb{D}$ be a smooth, projective family, and let $u \in H^0(\mathcal{X}_0, mK_{\mathcal{X}_0})$ be a pluricanonical section over the fiber $\mathcal{X}_0$. Then we have.

Theorem 3.1 ([48]). *There exists a section U of the bundle $mK_{\mathcal{X}}$, whose restriction to $\mathcal{X}_0$ is equal to $u \wedge dp^{\otimes m}$.*

For a simplified proof of this result we refer to [43]. The main technical tool in the proof of 3.1 is theorem 1.3: if $m \geq 2$, we write $mK_{\mathcal{X}} = K_{\mathcal{X}} + (m-1)K_{\mathcal{X}}$, so the heart of the matter is to construct a positively curved metric h_L on the bundle $L := (m-1)K_{\mathcal{X}}$, such that u satisfies the L^2 condition (2). The construction of the metric is done via an algorithm based on 1.3, which will not be detailed here. We rather remark:

(a) the absence of any strict positivity hypothesis in 3.1;

(b) that in [48], the metric h_L *depends on the section u we want to extend.*

Motivated by applications in algebraic geometry, one has to generalize this kind of results for *twisted pluricanonical forms*. In other words, we are asking for the analogue of 3.1 if u is a section of the bundle

$$m(K_{\mathcal{X}} + \Delta + E)_{|\mathcal{X}_0}, \tag{9}$$

where $\Delta := \sum_{j \in J} \nu^j Y_j$ is an effective $\mathbb{Q}$-divisor, such that $\nu^j < 1$ and such that the hypersurfaces (Y_j) are non-singular and mutually disjoint; the $\mathbb{Q}$-divisor E is (usually) assumed to be ample, and m is assumed to be divisible enough.

As shown by examples in [21], in general it is not true that all the sections of the bundle (9) extend to $\mathcal{X}$ (strangely enough, this has something to do with the diophantine properties of the cohomology class associated to $\Delta + E$). However, it is established in the articles [21], [25] the existence of an effective $\mathbb{Q}$-divisor

$$\Xi := \sum_{j \in J} \rho^j Y_{j|\mathcal{X}_0}, \tag{10}$$

with $\rho^j \leq \nu^j$, such that *a section u of the bundle* (9) *extends to $\mathcal{X}$ if and only if its zero divisor is greater than $m\Xi$.*

In order to formulate and discuss our results, we need first to recast 3.1 in metric terms. To start with, we recall the next important notion.

Definition 1. Let F be a line bundle, endowed with a non-singular, reference metric h_∞. We assume that F is pseudo-effective, and then we define the *metric with minimal singularities in the sense of Demailly* as follows

$$h^{\mathcal{D}}_{\min} := e^{-\psi^{\mathcal{D}}_{\min}} h_\infty, \quad \psi^{\mathcal{D}}_{\min} := \sup \psi \tag{11}$$

where the real-valued functions ψ above belong to the space $L^1(X)$, they are normalized by the condition $\sup_X \psi = 0$, and the curvature current of the associated metric $e^{-\psi} h_\infty$ is positive.

Definition 2. If the Kodaira dimension of F is positive, then the *metric with minimal singularities in the sense of Siu* will be denoted by $h^{\mathcal{S}}_{\min}$; it is defined in a similar fashion, modulo the fact that in (11) we only use functions $\psi = \dfrac{1}{m} \log |u|^2_{h_\infty}$, where u is a section of mL, for all m.

Thus, from metric point of view, the results in [48], [21] suggest that if $F := K_{\mathcal{X}} + \Delta + E$, then the restriction $\varphi^{\mathcal{S}}_{\min|\mathcal{X}_0}$ is well-defined, and it is equivalent to the minimal metric corresponding to $K_{\mathcal{X}} + \Delta' + E_{|\mathcal{X}_0}$ plus the tautological function associated to Ξ, where $\Delta' := \Delta - \Xi$.

We show in [6] that this is indeed the case, in a more general framework. We will analyze here the extension of sections of (9) under the hypothesis that the curvature form of E is only assumed to be semi-positive. Hence, unlike the usual setting, the bundle E (or its restriction to the central fiber) is not necessarily big, but a *natural vanishing assumption* for the section to be extended is needed. Our next result can be seen as an effective version of the L. Ein-M. Popa theorem in [21]; also, it is a generalization of results due to J.-P. Demailly and H. Tsuji in [19], respectively [53], [54].

Let $L \to \mathcal{X}$ be a hermitian line bundle such that $c_1(L)$ contains the current

$$m([\Delta] + \alpha) \in c_1(L) \tag{12}$$

where the notations are as follows.

(**a**) $\Delta := \sum_{j \in J} \nu^j Y_j$ is an *effective* $\mathbb{Q}$-divisor, such that $m > m\nu^j \in \mathbb{Z}$ for any $j \in J$; the hypersurfaces $Y_j \subset \mathcal{X}$ are mutually disjoint, and they intersect $\mathcal{X}_0$ transversally.

(**b**) α is a closed, non-singular, semi-positive form of (1,1)-type, with the property that $\{m\alpha\} \in H^2(\mathcal{X}, \mathbb{Z})$.

Furthermore, we assume that the bundle $K_{\mathcal{X}} + 1/mL$ is pseudoeffective, and let $h_{\min} := h^{\mathcal{D}}_{\min}$ be a metric with minimal singularities corresponding to it; we denote by $\Theta_{\min}$ its curvature current. We assume that

$$\nu_{\min}\big(\{K_{\mathcal{X}} + 1/mL\}, \mathcal{X}_0\big) = 0 \tag{13}$$

that is to say, the minimal multiplicity of the class $\{K_{\mathcal{X}} + 1/mL\}$ along the central fiber $\mathcal{X}_0$ is equal to zero (see e.g. [11], [38]). Let $A \to \mathcal{X}$ be an ample line

bundle. The assumption (13) implies that the metric with minimal singularities $h_{\min,\varepsilon}$ corresponding to the class $K_{\mathcal{X}} + 1/mL + \varepsilon A$ is not identically $+\infty$ when restricted to $\mathcal{X}_0$, so that we can write

$$\Theta_{\min,\varepsilon|\mathcal{X}_0} = \sum_{j\in J} \rho^j_{\min,\varepsilon}[Y_{j0}] + \Lambda_{0\varepsilon} \tag{14}$$

where $Y_{j0} := Y_j \cap \mathcal{X}_0$ and where $(\rho^j_{\min,\varepsilon})$ are positive real numbers. For each j, the sequence $(\rho^j_{\min,\varepsilon})$ is decreasing, and we define

$$\rho^j_{\min,\infty} := \lim_{\varepsilon\to 0} \rho^j_{\min,\varepsilon}. \tag{15}$$

We introduce the notation $J' := \{j \in J : \rho^j_{\min,\infty} < \nu^j\}$.

Let $h_0 = e^{-\varphi_0}$ be a metric on the $\mathbb{Q}$-bundle $K_{\mathcal{X}_0} + 1/mL$ with the property that

$$\Theta_{h_0}(K_{\mathcal{X}_0} + 1/mL) \geq 0$$

and such that the following inequality is satisfied

$$\varphi_0 \leq \sum_{j\in J'} \rho^j_{\min,\infty} \log|f_{Y_j}|^2 + \sum_{j\in J\setminus J'} \nu^j \log|f_{Y_j}|^2. \tag{16}$$

We remark that φ_0 plays the role of the section u of the bundle (9). The inequality (16) is the analogue of the vanishing of u along the *obstruction to extension* divisor Ξ mentioned above.

We denote by φ_L the singular metric on L induced by the decomposition (12), and for each j we denote by f_{Y_j} in (16) the local equations of the hypersurface Y_j; we have.

Theorem 3.2 ([6]). *Under the hypothesis* (**a**), (**b**) *and* (12) − (16) *above, the restriction* $\varphi_{\min|\mathcal{X}_0}$ *is well-defined, and there exists a constant* $C < 0$ *such that the following inequality holds at each point of* $\mathcal{X}_0$

$$\varphi_{\min|\mathcal{X}_0} \geq C + \varphi_0. \tag{17}$$

In particular, given any section u *of the bundle* $mK_{\mathcal{X}_0} + L$ *whose zero divisor is greater than*

$$m\sum_{j\in J'} \rho^j_{\min,\infty}[Y_{j0}] + m\sum_{j\in J\setminus J'} \nu^j[Y_{j0}] \tag{$\star$}$$

there exists a section U *of* $mK_{\mathcal{X}} + L$ *extending* u*, and such that*

$$\int_{\mathcal{X}} |U|^{\frac{2}{m}} e^{-\frac{1}{m}\varphi_L} \leq C_0 \int_{\mathcal{X}_0} |u|^{\frac{2}{m}} e^{-\frac{1}{m}\varphi_L}. \qquad \square$$

We remark that as a consequence of (17) we obtain Ohsawa-Takegoshi type estimates for the extension U, provided that the section u vanishes along the divisor $(\star)$.

If the form α in (b) is strictly positive, then the second part of the preceding result was established in [21], [25]. Also, we refer to the section 17 of the article [19] (and the references therein) for an enlightening introduction and related results around this circle of ideas. $\square$

In order to give an interpretation of the result 3.2, we define

$$L' = L_{|\mathcal{X}_0} - m \sum_{j \in J'} \rho^j_{\min,\infty}[Y_{j0}] - m \sum_{j \in J \setminus J'} \nu^j [Y_{j0}];$$

it is not too difficult to show that the bundle $K_{\mathcal{X}_0} + 1/mL'$ is pseudoeffective. We denote by $\varphi'_{\min}$ the metric with minimal singularities in the sense of Demailly corresponding to the bundle $K_{\mathcal{X}_0} + 1/mL'$; then we have

$$\left| \varphi_{\min|\mathcal{X}_0} - \sum_{j \in J'} \rho^j_{\min,\infty} \log |f_j|^2 - \sum_{j \in J \setminus J'} \nu^j \log |f_j|^2 - \varphi'_{\min} \right| \leq C. \tag{18}$$

so the singularities of the restriction $\varphi_{\min|\mathcal{X}_0}$ are completely understood in terms of the extremal metric $\varphi'_{\min}$. Except for the *rationality* of the coefficients $\rho^j_{\min,\infty}$, the relation (18) is the metric version of the description of the *restricted algebra* in [1], [21], [25]. $\square$

Furthermore, we show in [6] that the inequality (17) of 3.2 has a compact counterpart, i.e. when the couple $(\mathcal{X}, \mathcal{X}_0)$ is replaced by (X, S), where we denote $S \subset X$ a non-singular hypersurface of the projective manifold X. The bundle $L \to X$ is assumed to have the properties $(\mathbf{a}), (\mathbf{b})$ above; *in addition*, we assume that we have

$$\alpha \geq \gamma \Theta_h\big(\mathcal{O}(\mathcal{S})\big), \tag{$\dagger$}$$

where γ is a positive real, and h is a non-singular metric on the bundle $\mathcal{O}(\mathcal{S})$ associated to S.

The hypothesis concerning $\{K_X + S + \frac{1}{m}L\}$, its corresponding minimal metric $\varphi_{\min}$ and the metric φ_0 on $K_X + S + \frac{1}{m}L_{|S}$ encoded in relations (12)-(16) are assumed to hold transposed in the actual setting. In this case, the perfect analogue of (17) is true, as follows: we have

$$\varphi_{\min|\mathrm{S}} \geq C + \varphi_0 \tag{19}$$

as it is shown by theorem B.9 in [6]. $\square$

We will not reproduce here the arguments for theorem 3.2 or (19); we just mention that the proof is a *sophisticated version* of the usual arguments, together with an additional input needed in order to gain a good enough control

of some constants, which in turn justifies a limit process. In case of 3.2, the additional argument needed relies on theorem 2.1, together with remarks 2.4 and 2.5. As for the inequality (19), the theorem 2.1 is not available in this context, but instead we use the a version of proposition 1.4, which we explain next.

We assume that we find ourselves in the following context: X is a projective manifold, $(s = 0) := S \subset X$ is a non-singular hypersurface, $\Delta := \sum_j \nu^j Y_j$ is an effective $\mathbb{Q}$-divisor on X, such that the pairs (X, Δ) and $(S, \Delta_{|S})$ are klt (so in particular, S does not belong to the support of Δ). Let E be a $\mathbb{Q}$-bundle on X, whose Chern class contains a non-singular form θ_E, with the following positivity properties

$$\theta_E \geq 0, \quad \theta_E \geq \delta\Theta\big(\mathcal{O}(\mathcal{S})\big) \tag{20}$$

for a real $0 < \delta < 1$. We assume moreover that the section

$$u \in H^0\big(S, m(K_S + \Delta + E_{|S})\big) \tag{21}$$

admits *some extension* to X; then there exists a section

$$U \in H^0\big(X, m(K_X + S + \Delta + E)\big) \tag{22}$$

such that $U_{|S} = u \wedge (ds)^{\otimes m}$ and such that

$$\int_X |U|^{\frac{2}{m}} e^{-\varphi_\Delta - \varphi_E - \varphi_S} \leq C \int_S |u|^{\frac{2}{m}} e^{-\varphi_\Delta - \varphi_E}. \tag{23}$$

In (23), we denote by φ_S a non-singular metric on $\mathcal{O}(\mathcal{S})$, and the constant C depends only on δ in (20) and the norm of the section s with respect to φ_S (hence it is independent on the particular section u, and independent on m as well). The existence of the extension U is obtained precisely as in the proof of 1.4: for example, one can consider the extension which minimizes the semi-norm on the left hand side of (23). Therefore, by this simple procedure we convert a non-effective extension of u into an effective one; we refer to [6] for the relevance of this observation.

A last remark here is that once the inequality (17) is established, the second part of the theorem 3.2 follows by an immediate application of 1.3, together with (an appropriate version of) 1.4.

4. Non-vanishing

In this last paragraph we will present our version of the so-called *non-vanishing theorem* in [44]. The statement is the following.

Theorem 4.1. *Let X be a projective manifold, and let $\theta_L \in \mathrm{NS}_{\mathbb{R}}(X)$ be a cohomology class in the real Neron-Severi space of X, such that:*

(a) *The adjoint class $c_1(K_X) + \theta_L$ is pseudoeffective, i.e. there exist a closed positive current*
$$\Theta_{K_X+L} \in c_1(K_X) + \theta_L;$$

(b) *The class θ_L contains a Kähler current Θ_L such that we have*
$$e^{\varphi_{K_X+L}-\varphi_L} \in L^1(X, x)$$
where φ_{K_X+L} (resp. φ_L) is a local potential of the current Θ_{K_X+L} (resp. Θ_L) locally near $x \in X$.

Then the adjoint class $c_1(K_X) + \theta_L$ contains an effective $\mathbb{R}$–divisor, i.e. there exists a finite family of positive reals μ^j and hypersurfaces $W_j \subset X$ such that
$$\sum_{j=1}^{N} \mu^j [W_j] \in c_1(K_X) + \theta_L.$$

In connection with this result, we mention here the following theorem, established in [9] by C. Birkar, P. Cascini, C. Hacon and J. McKernan (which in some sense was part of the motivation for our work [44]; see also [8], [10], [20], [28]): *let (X, Δ) be a klt pair, where Δ is a big $\mathbb{R}$–divisor such that $K_X + \Delta$ is pseudo-effective. Then $K_X + \Delta$ is $\mathbb{R}$–linearly equivalent to an effective divisor.*

The integral hypothesis (b) in 4.1 is much more general than the klt assumption of the pair (X, Δ). However, it is enough to consider this latter case: this is a consequence of a theorem due to H. Skoda. Nevertheless, we prefer to state our result in this form, since the hypothesis (b) is *canonical*, in the sense that it concerns a global measure on X. Also, the important aspect of our proof is that it is direct and Char p-free, we avoid the *explicit* use of the minimal model program algorithm. Finally, many arguments in [44] are borrowed from Y.-T. Siu's analogue statement in [49], [50]. □

In order to put 4.1 in a proper perspective, we will highlight next its relationship with the classical "non-vanishing" theorems of V. Shokurov and Y. Kawamata, cf. [46], [26]; we have.

Theorem 4.2 ([46]). *Let D and G be a nef line bundle, respectively a $\mathbb{Q}$-divisor. We assume that the following relations hold :*

(i) *The $\mathbb{Q}$–divisor $D + G - K_X$ is nef and big ;*

(ii) *The pair $(X, -G)$ is klt.*

Then for all large enough integers $m \in \mathbb{Z}_+$, the bundle $mD + \lceil G \rceil$ is effective.

We can assume that the support of the divisor G has normal crossings, and we write $G = G_+ - G_-$ as a difference of two effective divisors. Let $L := D + G_+ - K_X$; according to (i), (ii), the pair (X, L) is big and klt. Then the equality

$$K_X + L = D + G_+$$

holds, hence the non-vanishing statements in [9], [49], [44] appear to be a natural generalization of Shokurov's result, in the sense that the divisor D is only assumed to be *pseudoeffective* rather than *numerically effective*.

To push a bit further the analogy, a quick glance at the proof of 4.2 in [46] shows that the main use of the nefness of D is for Kawamata-Viehweg theorem to hold: the vanishing of an h^1 cohomology group makes possible extension of twisted pluricanonical sections from subvarieties. In [44], the heart of our proof is to show that under some precise circumstances, the extension of pluricanonical forms is still possible, even if the vanishing of h^1 is not known to hold (cf. 4.3, 4.4 below). From our point of view, this is why the generalization of the classical non-vanishing to the pseudoeffective case can be achieved. □

We will not discuss here the structure of the proof of 4.1, since it is quite detailed in [44], as well as in [10]. Instead, we will extract from it *two pseudoeffectivity criteria*, which are implicit in [44], and besides from the proof of 4.1 may have some independent interest. The framework is as follows.

Let X be a projective manifold, and let S, Y_j be a set of strictly normal crossing hypersurfaces, such that $Y_j \cap Y_k = \emptyset$ if $j \neq k$. We fix a $\mathbb{Q}$-bundle A on X, such that for every $\delta > 0$, there exists a set $0 < \delta^j < \delta$ of positive real numbers, such that $A - \sum_j \delta^j Y_j$ is ample. Then we have the next statements.

Theorem 4.3 ([44]). *Let $0 \leq \nu^j < 1$; we assume that for all $\varepsilon \ll 1$, there exists a current*

$$T_\varepsilon \in \Big\{ K_X + S + \sum_j \nu^j Y_j + A \Big\}$$

whose Lelong number along S is equal to zero, and such that $T_\varepsilon \geq -\varepsilon\omega$, where ω is a Kähler form on X. Then the class $\{K_X + S + \sum_j \nu^j Y_j + A\}$ contains an effective $\mathbb{R}$-divisor whose support does not include S.

One of the important tools in the proof of the above theorem is the following result (see [44], paragraph 1.H).

Theorem 4.4 ([44]). *We assume that the numbers ν^j above are rational; there exists a positive real $\varepsilon \ll 1$ such that the following property holds true.*

Any section $u \in H^0\big(S, q(K_X + S + \sum_j \nu^j Y_j + A)_{|S}\big)$ extends to X, provided that there exists $T \in \{K_X + S + \sum_j \nu^j Y_j + A\}$ a closed current whose restriction to S is well-defined, such that $T \geq -\varepsilon/q\omega$ and such that

$$\mathrm{ord}_{Y_{j|S}}(u) \geq q \min\Big(\nu\big(T_{|S}, Y_{j|S}\big), \nu^j\Big) - \varepsilon$$

for all j. In particular, the bundle $K_X + S + \sum_j \nu^j Y_j + A$ is (pseudo)effective, if a couple (u, T) as above exists.

We remark here the differences between the statement 4.4 and the results quoted in the preceding paragraph: the current T is allowed to have a slightly negative part, and the vanishing of the section is assumed to be smaller than what it should. The reason why the *extension of pluricanonical forms algorithm* can be applied is that the negativity of T, respectively the lack of vanishing of u are "errors" which can be absorbed by the ample part A of the boundary, provided that their relationship with the degree q of u are as indicated in 4.4.

We stress on the fact that the hypothesis of the statement above may look artificial, but in the context of the proof in [44] they appear to be very natural. Indeed, our arguments involve a diophantine approximation process, which induce a *loss of positivity* quantified as in 4.4. □

Another important consequence of the techniques developed in [44] is the equivalence of the metrics $h^{\mathcal{D}}_{\min}$ and $h^{\mathcal{S}}_{\min}$ for effective bundles of type $K_X + L$, where L is a big and klt $\mathbb{Q}$-divisor. The proof relies on the fact that the algebra associated to $K_X + L$ is of finite type (cf. [9]); translated in metric language, this means that $h^{\mathcal{S}}_{\min}$ has analytic singularities. This fact is used as follows: if the two metrics above are not equivalent, then $h^{\mathcal{S}}_{\min}$ can be seen as an *incomplete linear system* (i.e., when both *sections and metrics* are taken into account), and the procedure in [44] allows the construction of a $\mathbb{Q}$-section of $K_X + L$, whose vanishing at some point is strictly smaller than the one of the metric $h^{\mathcal{S}}_{\min}$; this is impossible.

As pointed out in [53], it would be extremely interesting to establish a similar relationship between $h^{\mathcal{D}}_{\min}$ and $h^{\mathcal{S}}_{\min}$ without the strict positivity of L. □

References

[1] F. Ambro, *Restrictions of log canonical algebras of general type*, arXiv:math/0510212.

[2] B. Berndtsson, *On the Ohsawa-Takegoshi extension theorem* Ann. Inst. Fourier (1996).

[3] B. Berndtsson, *Curvature of Vector bundles associated to holomorphic fibrations,* to appear in Ann. of Maths. (2007).

[4] B. Berndtsson, M. Păun, *Bergman kernels and the pseudo-effectivity of the relative canonical bundles,* arXiv:math/0703344, Duke Math. Journal 2008.

[5] B. Berndtsson, M. Păun, *Bergman kernels and subadjunction*, arXiv 2009.

[6] B. Berndtsson, M. Păun, *Qualitative extensions of twisted pluricanonical forms and closed positive currents*, arXiv 2009.

[7] B. Berndtsson, *Strict and non strict positivity of direct image bundles,* arXiv:1002.4797, 2009.

[8] C. Birkar, *On existence of log minimal models,* arXiv:math/0610203v2.

[9] C. Birkar, P. Cascini, C. Hacon, J. McKernan, *Existence of minimal models for varieties of log general type,* arXiv:math/0610203v2, J. Amer. Math. Soc. 23, 2010.

[10] C. Birkar, M. Păun, *Minimal models, flips and finite generation: a tribute to V. V. SHOKUROV and Y.-T. SIU,* 2009.

[11] S. Boucksom, *Divisorial Zariski decompositions on compact complex manifolds,* Ann. Sci. Ecole Norm. Sup. (4) 2004.

[12] M. Brunella, *Uniformisation of foliations by curves,* arXiv:0802.4432.

[13] B. Claudon, *Invariance for multiples of the twisted canonical bundle,* Ann. Inst. Fourier (Grenoble) 57, 2007.

[14] F. Campana, *Special Varieties and Classification Theory* Annales de l'Institut Fourier 54, 2004.

[15] F. Campana, T. Peternell, *Geometric stability of the cotangent bundle and the universal cover of a projective manifold* arXiv:math/0405093.

[16] J.-P. Demailly, *Singular hermitian metrics on positive line bundles,* Proc. Conf. Complex algebraic varieties (Bayreuth, April 26, 1990), edited by K. Hulek, T. Peternell, M. Schneider, F. Schreyer, Lecture Notes in Math., Vol. 1507, Springer-Verlag, Berlin, 1992.

[17] J.-P. Demailly, *On the Ohsawa-Takegoshi-Manivel extension theorem,* Proceedings of the Conference in honour of the 85th birthday of Pierre Lelong, Paris, September 1997.

[18] J.-P. Demailly, *Kähler manifolds and transcendental techniques in algebraic geometry,* Plenary talk and Proceedings of the Internat. Congress of Math., Madrid 2006, volume I.

[19] J.-P. Demailly, *Analytic methods in algebraic geometry,* on the web page of the author, December 2009.

[20] S. Druel, *Existence de modèles minimaux pour les variétés de type général,* Exposé 982, Séminaire Bourbaki, 2007/08.

[21] L. Ein, M. Popa, *Adjoint ideals and extension theorems,* personal communication june 2007, arXiv: 0811.4290.

[22] T. Fujita, *On Kahler fiber spaces over curves,* J. Math. Soc. Japan 30 (1978), no. 4, 779794.

[23] P. Griffiths, *Periods of integrals on algebraic manifolds. III. Some global differential-geometric properties of the period mapping* Inst. Hautes Etudes Sci. Publ. Math. No. 38 1970.

[24] A. Höring, *Positivity of direct image sheaves - a geometric point of view,* in preparation, available on the author's home page.

[25] C.D. Hacon, J. McKernan, *Existence of minimal models for varieties of general type, II, (Existence of pl-flips)* arXiv:math.AG/0808, 2009.

[26] Y. Kawamata; *Kodaira dimension of algebraic fiber spaces over curves,* Invent. Math. 66, 1982.

[27] Y. Kawamata, *On the extension problem of pluricanonical forms.* Algebraic geometry: Hirzebruch 70 (Warsaw, 1998), Contemp. Math., vol. 241, Amer. Math. Soc., Providence, RI, 1999.

[28] Y. Kawamata, *Finite generation of a canonical ring.* arXiv:0804.3151.

[29] Y. Kawamata, *Deformation of canonical singularities,* J. Amer. Math. Soc., 12,1999.

[30] D. Kim, *L^2 extension of adjoint line bundle sections,* arXiv:0802.3189, to appear in Ann. Inst. Fourier.

[31] J. Kollár, *Higher direct images of dualizing sheaves, I and II,* Ann. of Math. (2) 124, 1986.

[32] V. Koziarz, *Extensions with estimates of cohomology classes,* preprint available on the web page of the author, 2009.

[33] R. Lazarsfeld, *Positivity in Algebraic Geometry,* Springer, Ergebnisse der Mathematik und ihrer Grenzgebiete.

[34] F. Maitani, H. Yamaguchi *Variation of Bergman metrics on Riemann surfaces,* Math. Ann. 330, 2004.

[35] L. Manivel, *Un théorème de prolongement L2 de sections holomorphes d'un fibré hermitien* Math. Zeitschrift,1993.

[36] J. McNeal, D. Varolin; *Analytic inversion of adjunction: L^2 extension theorems with gain* Ann. Inst. Fourier (Grenoble) **57** (2007), no. 3, 703–718.

[37] C. Mourougane, S. Takayama, *Extension of twisted Hodge metrics for Kähler morphisms* arXiv: 0809.3221.

[38] N. Nakayama *Zariski decomposition and abundance,* MSJ Memoirs **14**, Tokyo (2004).

[39] M. S. Narasimhan, R. R. Simha, *Manifolds with ample canonical class,* Inventiones Math. 1968.

[40] T. Ohsawa, K. Takegoshi, *On the extension of L^2 holomorphic functions* Math. Z., 1987.

[41] T. Ohsawa, *On the extension of L^2 holomorphic functions. VI. A limiting case,* Contemp. Math., Amer. Math. Soc., Providence, 2003.

[42] T. Ohsawa, *Generalization of a precise L2 division theorem,* Complex analysis in several variables Memorial Conference of Kiyoshi Okas Centennial Birthday, 249261, Adv. Stud. Pure Math., 42, Math. Soc. Japan, Tokyo, 2004.

[43] M. Păun, *Siu's Invariance of Plurigenera: a One-Tower Proof* preprint IECN, 2005, J. Differential Geom., 2007.

[44] M. Păun, *Relative critical exponents, non-vanishing and metrics with minimal singularities* arXiv 2008.

[45] G. Schumacher, *Curvature of higher direct images and applications,* arXiv:1002.4858.

[46] V. Shokurov, *A non-vanishing theorem.* Izv. Akad. Nauk SSSR (49) 1985.

[47] Y-T. Siu, *Invariance of plurigenera.* Invent. Math. 134, 1998.

[48] Y.-T. Siu, *Extension of twisted pluricanonical sections with plurisubharmonic weight and invariance of semipositively twisted plurigenera for manifolds not necessarily of general type* Complex geometry, 223–277, Springer, Berlin, 2002.

[49] Y.-T. Siu, *A General Non-Vanishing Theorem and an Analytic Proof of the Finite Generation of the Canonical Ring* arXiv:math/0610740.

[50] Y.-T. Siu, *Finite Generation of Canonical Ring by Analytic Method,* arXiv:0803.2454.

[51] Y.-T. Siu, *Abundance Conjecture,* arXiv 2010.

[52] S. Takayama, *Pluricanonical systems on algebraic varieties of general type,* Invent. Math. 2005.

[53] H. Tsuji, *Extension of log pluricanonical forms from subvarieties* math.CV/0511342.

[54] H. Tsuji, *Canonical singular hermitian metrics on relative canonical bundles* prepublication Sophia University, Japan, arXiv:math.AG/0704.0566.

[55] D. Varolin, *A Takayama-type extension theorem.* math.CV/0607323, to appear in Comp. Math.

[56] E. Viehweg, *Vanishing theorems* J. Reine Angew. Math. 335, 1982.

[57] E. Viehweg, *Weak positivity and the additivity of the Kodaira dimension for certain fibre spaces,* Proc. Algebraic Varieties and Analytic Varieties, Adv. Studies in Math. 1, Kinokunya–North-Holland Publ., 1983.

[58] E. Viehweg, *Quasi-Projective Moduli for Polarized Manifolds,* Springer-Verlag, Berlin, Heidelberg, New York, 1995.

[59] H. Yamaguchi, *Variations of pseudoconvex domains over* $\mathbb{C}^n$, Michigan Math J., 1989.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Cohomological Hasse Principle and Motivic Cohomology of Arithmetic Schemes

Shuji Saito*

Abstract

In 1985 Kazuya Kato formulated a fascinating framework of conjectures which generalize the Hasse principle for the Brauer group of a global field to the so-called cohomological Hasse principle for an arithmetic scheme X. He defined an invariant $KH_a(X)$ $(a \geq 0)$, called the Kato homology of X, that reflects the arithmetic nature of X. As a generalization of the classical Hasse principle, Kato conjectured the vanishing of $KH_a(X) = 0$ for $a > 0$, when X is a proper smooth variety over a finite field, or a regular scheme proper and flat over the ring of integers in a number field or in a local field. The conjecture turns out to play a significant rôle in arithmetic geometry. We will explain recent progress on the conjecture and its implications on finiteness of motivic cohomology, special values of zeta functions, a generalization of higher dimensional class field theory, and a geometric application to quotient singularities.

Mathematics Subject Classification (2010). 19F27, 19E15, 14C25, 14F42

Keywords. Hasse principle, motivic cohomology, zeta function, higher class field theory

Introduction

A fundamental fact in number theory is the Hasse principle for the Brauer group of a global field K, which is a global-local principle for a central simple algebra A over K:

$$A \simeq M_n(K) \text{ if and only if } A \otimes_K K_x \simeq M_n(K_x) \text{ for all places } x \text{ of } K,$$

*Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Tokyo, 153-8914 Japan. E-mail: sshuji@msb.biglobe.ne.jp.

where $M_n(*)$ is the matrix algebra and K_x is the completion of K at x. In 1985 Kazuya Kato [K] formulated a fascinating framework of conjectures which generalizes this fact to higher dimensional *arithmetic schemes,* namely schemes of finite type over a finite field or the ring of integers in a number field or a local field. For an integer $n > 0$ and for an arithmetic scheme X, he defined a collection of $\mathbb{Z}/n\mathbb{Z}$-modules

$$KH_a(X, \mathbb{Z}/n\mathbb{Z}) \quad (a \geq 0)$$

which we call the Kato homology of X. The Hasse principle for the Brauer group of a global field K is equivalent to the vanishing $KH_1(X, \mathbb{Z}/n\mathbb{Z}) = 0$ for all $n > 0$, where $X = \mathrm{Spec}(\mathcal{O}_K)$ with the ring $\mathcal{O}_K$ of integers in K. As a generalization of this fact, he proposed the following conjecture called the cohomological Hasse principle.

Conjecture 0.1. *Let X be either a proper smooth variety over a finite field, or a regular scheme proper flat over the ring of integers in a number field or in a local field. Then*

$$KH_a(X, \mathbb{Z}/n\mathbb{Z}) = 0 \quad \textit{for } a > 0.$$

There is work on the conjecture by Kato [K], Colliot-Thélène [CT] and Jannsen-Saito [JS1], where the vanishing $KH_a(X, \mathbb{Z}/n\mathbb{Z}) = 0$ for small degree a is shown. The first aim of this article is to report on the recent progress on the conjecture, the work of U. Jannsen, M. Kerz and the author, which proves the vanishing in all degrees under suitable conditions. The second aim is to give applications of these results. It turns out that the cohomological Hasse principle plays a significant rôle in arithmetic geometry, in particular in the study of motivic cohomology of arithmetic schemes.

Motivic cohomology is an important object to study in arithmetic geometry. It includes the ideal class group and the unit group of a number field, and the Chow groups of algebraic varieties. It is closely related to zeta-functions of algebraic varieties over a finite field or an algebraic number field. One of the important open problems is the conjecture that motivic cohomology of regular arithmetic schemes is finitely generated, a generalization of the known finiteness results on the ideal class group and the unit group of a number field (Minkowski and Dirichlet), and the group of the rational points on an abelian variety over a number field (Mordell-Weil). There have been only few results on the conjecture except the cases stated above and the one-dimensional case (Quillen). In [JS2] it was found that the Kato homology $KH_a(X, \mathbb{Z}/n\mathbb{Z})$ fills a gap between motivic cohomology with finite coefficient and étale cohomology of X. Thus, thanks to known finiteness results on étale cohomology, the cohomological Hasse principle implies new finiteness results on motivic cohomology.

We will also give other implications. One is a result on special values of the zeta function $\zeta(X, s)$ of a smooth projective variety over a finite field, which expresses

$$\zeta(X, 0)^* := \lim_{s \to 0} \zeta(X, s) \cdot (1 - q^{-s})$$

by the cardinalities of the torsion subgroups of motivic cohomology groups of X. It may be viewed as a geometric analogue of the analytic class number formula for the Dedekind zeta function of a number field.

Another application is a generalization of the higher dimensional class field theory by Schmidt-Spiess [ScSp] which describes the abelian fundamental group of a smooth variety over a finite field by using its Suslin homology of degree 0. Suslin homology is an algebraic analogue of singular homology for topological spaces and is compared to motivic homology defined by Voevodsky. We generalize the work of Schmidt-Spiess to its higher-degree variant and establish an isomorphism between Suslin homology of higher degree and the dual of étale cohomology.

Finally we give an application to a geometric problem on singularities. A consequence is the vanishing of weight homology groups of the exceptional divisors of desingularizations of quotient singularities.

The paper is organized as follows.

In §1 we give a brief review on motivic cohomology. There are mainly two ways of definition. The first one is due to Voevodsky [V1] who constructed the triangulated category of motives and defined motivic (co)homology as the space of maps in this category. We will not go into details of this construction but we explain another (more concrete) definition of motivic (co)homology given by Bloch's higher Chow group and Suslin's homology.

In §2 we state the finiteness conjecture of motivic cohomology and recall some known results on the conjecture. As a tool to approach the conjecture, we introduce the cycle class map from motivic cohomology to étale cohomology constructed by Bloch [B1] and Geisser-Levine [GL] and K.Sato [Sat2].

In §3 we state the Kato conjectures on the cohomological Hasse principle together with a lemma which affirms that the Kato homology controls the kernel and cokernel of the cycle class map introduced in §2.

In §4 we recall all known results on the Kato conjectures and give a very rough sketch of the proof of the most recent result due to Kerz-Saito [KeS], [Sa3].

In §5 we state some new results on the finiteness conjecture of motivic cohomology as an application of the result of Kerz-Saito.

In §6 we give its application to special values of the zeta function of a smooth projective variety over a finite field.

In §7 we give as another application a higher-degree variant of the higher dimensional class field theory of Schmidt-Spiess [ScSp].

In §8 we explain a geometric application to quotient singularities.

The author is grateful to Prof. J.-L. Colliot-Thélène and Prof. T. Geisser for their helpful comments on the first version of this paper.

1. Motivic Cohomology

The purpose of this section is to give a quick review on motivic cohomology. We start with the class number formula for an algebraic number field K:

$$\lim_{s\to 0} \zeta_K(s)\cdot s^{-\rho_0} = -\frac{|Cl(K)|\cdot R_K}{|(\mathcal{O}_K^\times)_{\mathrm{tors}}|} \tag{1.1}$$

where $\zeta_K(s)$ is the Dedekind zeta function of K, ρ_0 is the rank of the unit group $\mathcal{O}_K^\times$ of the ring $\mathcal{O}_K$ of integers, $(\mathcal{O}_K^\times)_{\mathrm{tors}}$ is the torsion part of $\mathcal{O}_K^\times$ (namely the group of the roots of unity in K), $Cl(K)$ is the ideal class group of $\mathcal{O}_K$, and R_K is Dirichlet's regulator

The philosophical question arises whether one could view the above formula as an arithmetic index theorem:

$$\boxed{\text{index (analytic invariant)}} = \boxed{\text{characteristic class (e.g. Euler characteristic)}}$$

An answer to the question is given by motivic cohomology

$$H_M^i(X,\mathbb{Z}(r))$$

which is defined for a scheme X (satisfying a reasonable condition) and for integers i and r. Indeed, in case $X = \mathrm{Spec}(\mathcal{O}_K)$ with $\mathcal{O}_K$ as above, we have

$$Cl(K) = H_M^2(X,\mathbb{Z}(1)), \quad \mathcal{O}_K^\times = H_M^1(X,\mathbb{Z}(1)).$$

Motivic cohomology theory may be considered universal cohomology theory in view of the existence of regulator maps to other cohomology theories, defined according to the context where X lives:

$$H_M^i(X,\mathbb{Z}(r)) \;\to\; \begin{array}{l} H_B^i(X,\mathbb{Z}(r)) \text{ (Betti cohomology)} \\ H_D^i(X,\mathbb{Z}(r)) \text{ (Deligne cohomology)} \\ H_{\text{ét}}^i(X,\mathbb{Z}_\ell(r)) \text{ (étale cohomology)} \\ H_{crys}^i(X/W(k)) \text{ (crystalline cohomology)} \\ \cdots\cdots\cdots \end{array}$$

Dirichlet's regulator map that defines R_K in (1.1) can be viewed as a special case of the regulator map to Deligne cohomology.

Another important property of motivic cohomology is its relation to algebraic K-theory via the spectral sequence for smooth X

$$E_2^{p,q} = H_M^p\left(X,\mathbb{Z}\left(-\frac{q}{2}\right)\right) \Rightarrow K_{-p-q}(X) \tag{1.2}$$

which is an algebraic analogue of the Atiyah-Hirzebruch spectral sequence for toplogical K-theory (see [Gra2] and [Le]).

Here we introduce two kinds of constructions of motivic cohomology. The first one is due to Voevodsky [V1] who constructed $DM(k)$, the triangulated category of motives over a field k. It is a tensor category equipped with a functor

$$M : Sm/k \to DM(k)\ ;\ X \to M(X)$$

where Sm/k is the category of smooth schemes over the field k. Motivic cohomology and homology of $X \in Sm/k$ are then defined as the space of maps in $DM(k)$:

$$H^i_M(X, \mathbb{Z}(r)) = \mathrm{Hom}_{DM(k)}(M(X), \mathbb{Z}(r)[i]),$$
$$H^M_i(X, \mathbb{Z}(r)) = \mathrm{Hom}_{DM(k)}(\mathbb{Z}(r)[i], M(X))$$

respectively, where $\mathbb{Z}(1)$ is a distinguished object in $DM(k)$ called the Tate object. It is invertible for the tensor structure and $\mathbb{Z}(r)$ for $r \in \mathbb{Z}$ is the r-th tensor power of $\mathbb{Z}(1)$. We do not go into details on $DM(k)$.

Another (more concrete) definition of motivic (co)homology is given by

$$\mathrm{CH}^r(X, q), \quad \text{Bloch's higher Chow group ([B2], [Le])}$$
$$H^S_i(X, \mathbb{Z}), \quad \text{Suslin homology ([SV1], [Sc])}$$

defined for a scheme X of finite type over a field or a Dedekind domain. We note that $\mathrm{CH}^r(X, q)$ for $q = 0$ is the Chow group of algebraic cycles on X of codimension r modulo rational equivalence. We have the following comparison result ([V4], [MVW], Lecture 19):

Theorem 1.1. *For a smooth scheme X over a field, we have natural isomorphisms*

$$H^i_M(X, \mathbb{Z}(r)) \simeq \mathrm{CH}^r(X, 2r - i), \quad H^M_i(X, \mathbb{Z}(0)) \simeq H^S_i(X, \mathbb{Z}).$$

Before going to a brief review of the definition of Bloch's higher Chow group and Suslin homology, we first recall the singular homology of a topological space X:

$$H_q(X, \mathbb{Z}) := H_q(s(X, \bullet))$$

where $s(X, \bullet)$ is the singular chain complex:

$$\cdots \to s(X, q) \xrightarrow{\partial} s(X, q-1) \xrightarrow{\partial} \cdots \to s(X, 0),$$

$$s(X, q) = \bigoplus_{\Gamma} \mathbb{Z}[\Gamma], \quad \Gamma \text{ ranges over all continuous maps } \Delta^q_{top} \to X.$$

Here

$$\Delta^q_{top} = \left\{ (x_0, x_1, \ldots, x_q) \in \mathbb{R}^{q+1} \;\middle|\; \sum_{0 \le i \le q} x_i = 1,\ x_i \ge 0 \right\}$$

is the standard simplex and the boundary map ∂ is the alternating sum of the restrictions to the faces of codimension 1 in Δ^q_{top}.

The definition of Bloch's higher Chow group and Suslin homology is an algebraic analogue of the above construction. Here we assume that X is of finite type over a field k while it is possilbe to treat more general cases (cf. [Le] and [Sc]). The standard simplex is replaced by its algebraic analogue

$$\Delta^q = \mathrm{Spec}\left(k[t_0,\dots,t_q]/\left(\sum_{i=0}^{q} t_i - 1\right)\right),$$

whose faces are $\Delta^s = \{t_{i_1} = \cdots = t_{i_{q-s}} = 0\} \subset \Delta^q$. We have two kinds of analogues of $s(X,q)$ given by the spaces of algebraic cycles on $X \times \Delta^q$:

$$z^r(X,q) = \bigoplus_{\Gamma \subset X \times \Delta^q} \mathbb{Z}[\Gamma], \quad c_0(X,q) = \bigoplus_{\Xi \subset X \times \Delta^q} \mathbb{Z}[\Xi]$$

where Γ (resp. Ξ) ranges over all integral closed subschemes of $X \times \Delta^q$, which have codimension r and intersect properly all faces $\Delta^s \subset \Delta^q$ (resp. which are finite surjective over Δ^q). One may be tempted to take Γ and Ξ as maps of schemes $f : \Delta^q \to X$ but this does not give a correct answer (such f give rise to algebraic cycles on $X \times \Delta^q$ by taking its graphs but there are not sufficiently many maps of schemes).

These groups fit into the so-called *cycle complexes* (graded homologically)

$$z^r(X,\bullet) \ : \ \cdots \to z^r(X,q) \xrightarrow{\partial} z^r(X,q-1) \xrightarrow{\partial} \cdots \xrightarrow{\partial} z^r(X,0),$$

$$c_0(X,\bullet) \ : \ \cdots \to c_0(X,q) \xrightarrow{\partial} c_0(X,q-1) \xrightarrow{\partial} \cdots \xrightarrow{\partial} c_0(X,0).$$

Bloch's higher Chow group and Suslin homology are defined as the homology groups of these complexes:

$$\mathrm{CH}^r(X,q) := H_q\big(z^r(X,\bullet)\big),$$

$$H_q^S(X,\mathbb{Z}) := H_q\big(c_0(X,\bullet)\big).$$

One may also consider the versions with finite coefficients:

$$\mathrm{CH}^r(X,q;\mathbb{Z}/n\mathbb{Z}) := H_q\big(z^r(X,\bullet) \otimes \mathbb{Z}/n\mathbb{Z}\big),$$

$$H_q^S(X,\mathbb{Z}/n\mathbb{Z}) := H_q\big(c_0(X,\bullet) \otimes \mathbb{Z}/n\mathbb{Z}\big).$$

In what follows, for a regular scheme of finite type over a perfect field or a Dedekind domain, we denote (cf. Theorem 1.1)

$$\begin{aligned} &H_M^i(X,\mathbb{Z}(r)) = \mathrm{CH}^r(X,2r-i), \\ &H_M^i(X,\mathbb{Z}/n\mathbb{Z}(r)) = \mathrm{CH}^r(X,2r-i;\mathbb{Z}/n\mathbb{Z}). \end{aligned} \tag{1.3}$$

We have an exact sequence

$$0 \to H^i_M(X, \mathbb{Z}(r))/n \to H^i_M(X, \mathbb{Z}/n\mathbb{Z}(r)) \to H^{i+1}_M(X, \mathbb{Z}(r))[n] \to 0 \qquad (1.4)$$

where $M[n] = \mathrm{Ker}(M \xrightarrow{n} M)$ for an abelian group M.

2. Finiteness Conjecture on Motivic Cohomology

A fundamental question in arithmetic geometry is the following.

Conjecture 2.1. *For a regular scheme X of finite type over $\mathbb{F}_p$ or $\mathbb{Z}$, $H^q_M(X, \mathbb{Z}(r))$ is finitely generated.*

In view of the spectral sequence (1.2), the conjecture would imply that the algebraic K-groups $K_i(X)$ of X are finitely generated, which is the so-called Bass conjecture. The above conjecture is a basis of the conjectures on special values of zeta functions of arithmetic varieties due to Beilinson and Bloch-Kato.

Remark 2.2. For a (not necessarily regular) scheme X of finite type over $\mathbb{F}_p$ or $\mathbb{Z}$, $\mathrm{CH}^r(X, q)$ is conjectured to be finitely generated. Indeed this follows from Conjecture 2.1 by the localization sequence for higher Chow groups.

In §5 we will present new finiteness results on motivic cohomology. Very little had been known about the conjecture except the following results. Let X be a regular scheme of finite type over $\mathbb{F}_p$ or $\mathbb{Z}$.

Theorem 2.3. *$H^q_M(X, \mathbb{Z}(1))$ is finitely generated for all integers q.*

In factCwe have

$$H^q_M(X, \mathbb{Z}(1)) = \mathrm{CH}^1(X, 2-q) = \begin{cases} \mathrm{Pic}(X) & q = 2 \\ \Gamma(X, \mathcal{O}_X^\times) & q = 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore the above theorem is a consequence of the finiteness results on the ideal class group and the unit group for the ring of integers in a number field (Minkowski and Dirichlet), and the Mordell-Weil theorem on the rational points of an abelian variety over a number field.

Theorem 2.4. *If $\dim(X) = 1$, $H^q_M(X, \mathbb{Z}(r))$ is finitely generated up to torsion.*

This follows from the fact that $K_i(X)$ is finitely generated, due to Quillen [Q] (see also [Gra1]), together with a result on the degeneracy of the spectral sequence (1.2) up to torsion ([Le], Theorem 11.7). As for the torsion part of $H^q_M(X, \mathbb{Z}(r))$, one can show that it is finite assuming the Bloch-Kato conjecture stated later in this section (see Theorem 2.7 and [Le], Theorems 14.3 and 14.5).

Theorem 2.5. *$H^{2d}_M(X, \mathbb{Z}(d))$ is finitely generated where $d = \dim(X)$.*

Note that $H^{2d}_M(X, \mathbb{Z}(d))$ coincides with the Chow group $\mathrm{CH}_0(X)$ of zero cycles on X modulo rational equivalence. Theorem 2.5 is a consequence of higher unramified class field theory due to Bloch[B1] and Kato-Saito[KS1]:

Theorem 2.6. *Let X be a regular scheme proper over $\mathbb{F}_p$ or $\mathbb{Z}$. Assume $X(\mathbb{R}) = \emptyset$ for simplicity. Then the higher reciprocity map*

$$\rho_X \ : \ \mathrm{CH}_0(X) \to \pi_1^{ab}(X)$$

is an isomorphism if X is flat over $\mathbb{Z}$, and injective with dense image otherwise.

Here $\pi_1^{ab}(X)$ is the abelian fundamental group of X and the definition of ρ_X will be given in §8. Theorem 2.5 follows from Theorem 2.6 and the finiteness result of $\pi_1^{ab}(X)$ due to Katz-Lang.

A way to approach Conjecture 2.1 is to use the cycle class map. Let X be a regular scheme of finite type over a perfect field or the ring $\mathcal{O}_k$ of integers in a number field or in a local field. Under a technical condition (which is necessary only in the case X is flat over $\mathcal{O}_k$ and n is not invertible in $\mathcal{O}_k$), there is a cycle class map

$$\rho_X : H^i_M(X, \mathbb{Z}/n\mathbb{Z}(r)) \to H^i_{\text{ét}}(X, \mathbb{Z}/n\mathbb{Z}(r)) \tag{2.1}$$

from the motivic cohomology with finite coefficient to the étale cohomology with suitable coefficient (explained below). The constructions of the cycle class map are due to Bloch [B1] and Geisser-Levine [GL] and K.Sato [Sat2]. The target group of the cycle class map varies according to the context: In case n is invertible on X,

$$H^i_{\text{ét}}(X, \mathbb{Z}/n\mathbb{Z}(r)) = H^i_{\text{ét}}(X, \mu_n^{\otimes r}),$$

where μ_n is the étale sheaf of the n-th roots of unity. In case X is smooth over a perfect field k and $n = mp^\nu$ with $p = \mathrm{ch}(k)$ and $(p, m) = 1$,

$$H^i_{\text{ét}}(X, \mathbb{Z}/n\mathbb{Z}(r)) = H^i_{\text{ét}}(X, \mathbb{Z}/m\mathbb{Z}(r)) \oplus H^{i-r}(X, W_\nu\Omega^r_{X,log}), \tag{2.2}$$

where $W_\nu\Omega^r_{X,log}$ is the logarithmic part of the de Rham-Witt sheaf $W_\nu\Omega^r_X$ ([Il1], I 5.7). Finally, in case X is flat over $\mathcal{O}_k$ and and n is not invertible on $\mathcal{O}_k$, $H^i_{\text{ét}}(X, \mathbb{Z}/n\mathbb{Z}(r))$ is the hyper cohomology of a certain object of the derived category of complexes of étale sheaves, which is defined by K. Sato [Sat1] as an étale incarnation of the motivic complex on X with finite coefficient.

We note that the target group of the cycle class map is known to be finite. Thus the injectivity of the map would imply a finiteness result for motivic cohomology of X. Indeed we have the following result due to Suslin-Voevodsky [SV2] and Geisser-Levine [GL] (see also K.Sato [Sat2]).

Theorem 2.7. *Let X be as above. Assume $\mathbf{(BK)}^r_{X,\ell}$ (see below) for every prime ℓ dividing n. Then the cycle class map*

$$\rho_X : H^i_M(X, \mathbb{Z}/n\mathbb{Z}(r)) \to H^i_{\acute{e}t}(X, \mathbb{Z}/n\mathbb{Z}(r))$$

is an isomorphism for $i \leq r$ and injective for $i = r + 1$.

In case X is smooth over a perfect field of characteristic $p > 0$ and $n = p^r$, this is a theorem of Geisser-Levine, which is used in Sato's work for the mixed characteristic case.

Corollary 2.8. *Let X be as above and assume X is of finite type over $\mathbb{F}_p$ or $\mathbb{Z}$. Then $H^i_M(X, \mathbb{Z}/n\mathbb{Z}(r))$ is finite for $i \leq r+1$.*

We now explain the condition $\mathbf{(BK)}^t_{X,\ell}$. For a field L and a prime ℓ and an integer $t > 0$, we have the Galois symbol map

$$h^t_{L,\ell} : K^M_t(L)/\ell \to H^t(L, \mathbb{Z}/\ell\mathbb{Z}(t))$$

where $H^*(L, \mathbb{Z}/\ell\mathbb{Z}(t)) = H^*_{\text{ét}}(\mathrm{Spec}(L), \mathbb{Z}/\ell\mathbb{Z}(t))$ is the Galois cohomology of L and $K^M_t(L)$ denotes the Milnor K-group of L. It is conjectured that $h^t_{L,\ell}$ is surjective. The conjecture is called the Bloch-Kato conjecture in case $l \neq \mathrm{ch}(L)$ (the case $l = \mathrm{ch}(L)$ is known to hold due to Bloch-Gabber-Kato [BK]). The surjectivity of $h^t_{L,\ell}$ is known if $t = 1$ (the Kummer theory) or $t = 2$ (Merkurjev-Suslin [MS]) or $\ell = 2$ (Voevodsky [V1]). Recently a proof of the conjecture has been announced by Rost and Voevodsky (see [SJ] and [V2], and [HW], [V3], [We1] and [We2] for details).

For a scheme X, we introduce the condition:

$\mathbf{(BK)}^t_{X,\ell}$: $h^t_{L,\ell}$ is surjective for any field L finitely generated over a residue field of X.

3. Cohomological Hasse Principle

In this section we discuss the cohomological Hasse principle which generalizes the following theorem of Hasse-Minkowski to higher dimensional arithmetic schemes. It plays an important role in the study of motivic cohomology of arithmetic schemes (see Lemma 3.6 below).

Theorem 3.1. *A quadratic form with rational coefficients*

$$a_1X_1^2 + \cdots a_nX_n^2 \quad (a_1, \ldots, a_n \in \mathbb{Q})$$

has a non-trivial zero in $\mathbb{Q}$ if and only if it has in $\mathbb{R}$ and $\mathbb{Q}_p$ for every prime p.

In general, a quadratic form over a field k with $\mathrm{ch}(k) \neq 2$:

$$X^2 - aY^2 - bZ^2 \quad (a, b \in k^\times)$$

has a non-trivial zero in k if and only if

$$h(a) \cup h(b) = 0 \in H^2(k, \mathbb{Z}/2\mathbb{Z})$$

where $h : k^\times/2 \simeq H^1(k, \mathbb{Z}/2\mathbb{Z})$ is the Kummer isomorphism, and

$$\cup : H^1(k, \mathbb{Z}/2\mathbb{Z}) \times H^1(k, \mathbb{Z}/2\mathbb{Z}) \to H^2(k, \mathbb{Z}/2\mathbb{Z})$$

is the cup product. Therefore the case $n = 3$ (which is the most crucial to the proof) of the theorem is equivalent to the injectivity of the restriction map

$$H^2(\mathbb{Q}, \mathbb{Z}/2\mathbb{Z}) \to \bigoplus_{p \in P_{\mathbb{Q}}} H^2(\mathbb{Q}_p, \mathbb{Z}/2\mathbb{Z}) \ \oplus \ H^2(\mathbb{R}, \mathbb{Z}/2\mathbb{Z})$$

where $P_{\mathbb{Q}}$ is the set of the rational primes. Moreover we have the residue isomorphism for $p \in P_{\mathbb{Q}}$:

$$\partial_p : H^2(\mathbb{Q}_p, \mathbb{Z}/2\mathbb{Z}) \simeq H^1(\mathbb{F}_p, \mathbb{Z}/2\mathbb{Z}), \tag{3.1}$$

and Theorem 3.1 is equivalent to the injectivity of the residue map:

$$H^2(\mathbb{Q}, \mathbb{Z}/2\mathbb{Z}) \xrightarrow{\partial} \bigoplus_{p \in P_{\mathbb{Q}}} H^1(\mathbb{F}_p, \mathbb{Z}/2\mathbb{Z}) \ \oplus \ H^2(\mathbb{R}, \mathbb{Z}/2\mathbb{Z}). \tag{3.2}$$

This fact has been extended to the following.

Theorem 3.2. *(Brauer-Hasse-Noether and Witt) Let X be either* $\mathrm{Spec}(\mathcal{O}_K)$ *with $\mathcal{O}_K$ the ring of integers in a number field or in a local field, or a proper smooth curve over a finite field. Let K be the function field of X. For simplicity, in case K is a number field, we assume that n is odd or that $X(\mathbb{R}) = \emptyset$ (namely K is totally imaginary). Then the residue map*

$$H^2(K, \mathbb{Z}/n\mathbb{Z}(1)) \xrightarrow{\partial} \bigoplus_{x \in X_{(0)}} H^1(\kappa(x), \mathbb{Z}/n\mathbb{Z}) \tag{3.3}$$

is injective, where $X_{(0)}$ is the set of the closed points of X, $\kappa(x)$ is the residue field of $x \in X$, and $\mathbb{Z}/n\mathbb{Z}(1)$ is defined as in (2.2).

We remark that there is a natural isomorphism

$$H^2(K, \mathbb{Z}/n\mathbb{Z}(1)) \simeq Br(K)[n],$$

where $Br(K)$ is the Brauer group of K (the set of equivalence classes of central simple algebras over K endowed with a suitable group structure). Thus Theorem 3.2 in case K is a global field, is equivalent to the Hasse principle for the Brauer group of K, namely the following global-local principle for such an algebra A:

$$A \simeq M_n(K) \Leftrightarrow A \otimes_K K_x \simeq M_n(K_x) \quad (\forall x \in X_{(0)})$$

where $M_n(*)$ is the matrix algebra and K_x is the completion of K at x.

In 1985 K.Kato [K] formulated a fascinating framework of conjectures which generalize Theorem 3.2 to higher dimensional *arithmetic schemes X,* namely a scheme of finite type over a finite field or the ring of integers in a number field

or a local field. He defined a complex of abelian groups $KC_\bullet(X, \mathbb{Z}/n\mathbb{Z})$ (now called the Kato complex of X):

$$\cdots \xrightarrow{\partial} \bigoplus_{x \in X_{(a)}} H^{a+1}(x, \mathbb{Z}/n\mathbb{Z}(a)) \xrightarrow{\partial} \bigoplus_{x \in X_{(a-1)}} H^a(x, \mathbb{Z}/n\mathbb{Z}(a-1)) \xrightarrow{\partial} \cdots$$
$$\cdots \xrightarrow{\partial} \bigoplus_{x \in X_{(1)}} H^2(x, \mathbb{Z}/n\mathbb{Z}(1)) \xrightarrow{\partial} \bigoplus_{x \in X_{(0)}} H^1(x, \mathbb{Z}/n\mathbb{Z})$$

Here $H^*(x, \mathbb{Z}/n\mathbb{Z}(a))$ is the Galois cohomology of the residue fields $\kappa(x)$ of x and $\mathbb{Z}/n\mathbb{Z}(a)$ is defined as in (2.2). The term in degree a is the direct sum of the Galois cohomology group for $x \in X_{(a)}$, where

$$X_{(a)} = \{x \in X \mid \dim \overline{\{x\}} = a\},$$

the set of those points of X whose closure in X has dimension a. Note that $x \in X_{(a)}$ if and only if $\mathrm{trdeg}_{\mathbb{F}_p} \kappa(x) = a$ or $\mathrm{trdeg}_{\mathbb{Q}} \kappa(x) = a - 1$.

In case X is as in Theorem 3.2, $KC_\bullet(X, \mathbb{Z}/n\mathbb{Z})$ coincides with the complex (3.3), and the assertion of Theorem 3.2 is equivalent to the vanishing of the first homology group $H_1\big(KC_\bullet(X, \mathbb{Z}/n\mathbb{Z})\big)$.

We define Kato homology of an arithmetic scheme X as

$$KH_a(X, \mathbb{Z}/n\mathbb{Z}) = H_a\big(KC_\bullet(X, \mathbb{Z}/n\mathbb{Z})\big) \quad (a \geq 0). \tag{3.4}$$

We will also use

$$KH_a(X, \mathbb{Q}/\mathbb{Z}) = \varinjlim_n KH_a(X, \mathbb{Z}/n\mathbb{Z}),$$
$$KH_a(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell) = \varinjlim_n KH_a(X, \mathbb{Z}/\ell^n\mathbb{Z}),$$

where ℓ is a prime. Kato notices that Theorem 3.2 admits the following conjectural generalization.

Conjecture 3.3. *Let X be a proper smooth variety over a finite field. Then*

$$KH_a(X, \mathbb{Z}/n\mathbb{Z}) = 0 \quad (\forall a > 0).$$

We remark that Geisser [Ge2] defined Kato homology with integral coefficient and studied an integral version of Conjecture 3.3.

Conjecture 3.4. *Let X be a regular scheme proper flat over the ring $\mathcal{O}_k$ of integers in a number field. Assume*

($*$) *either n is odd or k is totally imaginary.*

Then

$$KH_a(X, \mathbb{Z}/n\mathbb{Z}) = 0 \quad (\forall a > 0).$$

We note that the assumption $(*)$ may be removed by modifying $KH_a(X, \mathbb{Q}/\mathbb{Z})$ (see [JS1] Conjecture C on page 482).

Conjecture 3.5. *Let X be a regular scheme proper and flat over* $\mathrm{Spec}(\mathcal{O}_k)$ *where $\mathcal{O}_k$ is the ring of integers in a local field. Then*

$$KH_a(X, \mathbb{Z}/n\mathbb{Z}) = 0 \quad \textit{for } a \geq 0.$$

The relationship of Kato homology of an arithmetic scheme to its motivic cohomology is explained in the following lemma (see [JS2], Lemma 6.2).

Lemma 3.6. *Let X be a connected regular scheme of finite type over a finite field or the ring $\mathcal{O}_k$ of integers in a number field or of a local field with $d =$* $\dim(X)$*. For an integer $i \geq 0$, assume $\mathbf{(BK)}^t_{X,\ell}$ with $t = 2d - i + 1$ (see §2). Then the cycle class map* (2.1)

$$\rho_X : H^i_M(X, \mathbb{Z}/n\mathbb{Z}(r)) \to H^i_{\acute{e}t}(X, \mathbb{Z}/n\mathbb{Z}(r))$$

is an isomorphism for $r > d :=$ $\dim(X)$*, and there is an exact sequence*

$$KH_{2d-i+2}(X, \mathbb{Z}/n\mathbb{Z}) \to H^i_M(X, \mathbb{Z}/n\mathbb{Z}(d)) \xrightarrow{\rho_X} H^i_{\acute{e}t}(X, \mathbb{Z}/n\mathbb{Z}(d)) \to KH_{2d-i+1}(X, \mathbb{Z}/n\mathbb{Z}).$$

4. Results on Cohomological Hasse Principle

In this section we state the known results on the Kato conjectures 3.3, 3.4 and 3.5. Let X be as in the conjectures. As explained, the Kato conjectures in case $\dim(X) = 1$ rephrase the classical fundamental facts on the Brauer group of a global field and a local field.

Kato [K] proved Conjectures 3.3, 3.4, and 3.5 in case $\dim(X) = 2$. He deduced it from higher class field theory for X proved in [KS2] and [Sa1]. For X of dimension 2 over a finite field, the vanishing of $KH_2(X, \mathbb{Z}/n\mathbb{Z})$ in Conjecture 3.3 had been earlier established in [CTSS] (prime-to-p-part), and by M. Gros [Gr] for the p-part.

The first result after [K] is the following:

Theorem 4.1. *(Saito [Sa2]) Let X be a smooth projective 3-fold over a finite field F. Then $KH_3(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell) = 0$ for any prime $\ell \neq \mathrm{ch}(F)$.*

This result was immediately generalized to the following:

Theorem 4.2. *(Colliot-Thélène [CT], Suwa [Sw]) Let X be a smooth projective variety over a finite field. Then*

$$KH_a(X, \mathbb{Q}/\mathbb{Z}) = 0 \quad \textit{for } 0 < a \leq 3$$

[CT] handled the prime-to-p part where $p = \mathrm{ch}(F)$, and Suwa [Sw] later adapted the technique of [CT] to handle the p-part. A tool in [Sa2] is a class field theory of surfaces over local fields, while the technique in [CT] is global and different from that in [Sa2].

The arithmetic version of the above theorem was established in the following:

Theorem 4.3. *(Jannsen-Saito [JS1]) Let X be a regular projective flat scheme over $S = \mathrm{Spec}(\mathcal{O}_k)$ where k is a number field or a local field. Fix a prime ℓ. Assume that for any closed point $v \in S$, the reduced part of $X_v = X \times_S v$ is a divisor with simple normal crossings on X and that X_v is reduced if $v|\ell$. For simplicity, if k is a number field, we assume that $\ell \neq 2$ or k is totally imaginary. Then*

$$KH_a(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell) = 0 \quad \textit{for } 0 < a \leq 3$$

The following theorem is a direct consequence of [J], Theorem 0.5. It reduces Conjecture 3.4 to Conjecture 3.5 for Kato homology with $\mathbb{Q}/\mathbb{Z}$-coefficient.

Theorem 4.4. *(Jannsen [J]) Let X be a regular projective flat scheme over $S = \mathrm{Spec}(\mathcal{O}_k)$ where k is a number field. For each closed point $v \in S$, let $S_v = \mathrm{Spec}(\mathcal{O}_{k_v})$ where k_v is the completion of k at v and write $X_{S_v} = X \times_S S_v$. Fix a prime ℓ and assume for simplicity that $\ell \neq 2$ or k is totally imaginary. Then we have a natural isomorphism*

$$KH_a(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell) \simeq \bigoplus_{v \in S_{(0)}} KH_a(X_v, \mathbb{Q}_\ell/\mathbb{Z}_\ell) \quad \textit{for } a > 0.$$

In the next theorem, Conjecture 3.3 is shown assuming resolution of singularities.

Theorem 4.5. *(Jannsen [J], Jannsen-Saito [JS2]) Let X be a projective smooth variety of dimension d over a finite field. Let $t \geq 1$ be an integer. Then we have*

$$KH_a(X, \mathbb{Q}/\mathbb{Z}) = 0 \quad \textit{for } 0 < a \leq t$$

if either $t \leq 4$ or $\mathbf{(RS)}_d$, or $\mathbf{(RES)}_{t-2}$ (see below) holds.

$\mathbf{(RS)}_d$: For any X integral and proper of dimension$\leq d$ over F, there exists a proper birational morphism $\pi : X' \to X$ such that X' is smooth over F. For any U smooth of dimension$\leq d$ over F, there is an open immersion $U \hookrightarrow X$ such that X is projective smooth over F with $X - U$ a divisor with simple normal crossings on X.

$\mathbf{(RES)}_t$: For any smooth projective variety X over F, any divisor Y with simple normal crossings on X with $U = X - Y$, and any integral closed subscheme $W \subset X$ of dimension$\leq t$ such that $W \cap U$ is regular, there exists a birational proper map $\pi : X' \to X$ such that X' is projective

smooth over F and $\pi^{-1}(U) \simeq U$, and that $Y' = X' - \pi^{-1}(U)$ is a divisor with simple normal crossings on X', and that the proper transform of W in X' is regular and intersects transversally with Y'.

We note that a proof of $(\mathbf{RES})_{\mathbf{2}}$ is given in [CJS] based on an idea of Hironaka, which enables us to obtain the unconditional vanishing of Kato homology in degree $a \leq 4$.

The above approach has been improved to remove the assumptions $(\mathbf{RS})_d$ and $(\mathbf{RES})_t$ on resolution of singularities, at least if we are restricted to the prime-to-ch(F) part:

Theorem 4.6. *(Kerz-Saito [KeS], [Sa3]) Let X be a proper smooth variety over a finite field F. For a prime $\ell \neq \mathrm{ch}(F)$, we have $KH_a(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell) = 0$ for $a > 0$.*

A key to the proof is the following refinement of de Jong's alteration theorem due to Gabber (see [Il2]).

Theorem 4.7. *(Gabber) Let F be a perfect field and X be a variety over F. Let $W \subset X$ be a proper closed subscheme. Let ℓ be a prime different from $\mathrm{ch}(F)$. Then there exists a projective morphism $\pi : X' \to X$ such that*

- *X' is smooth over F and the reduced part of $\pi^{-1}(W)$ is a divisor with simple normal crossings on X.*
- *π is generically finite of degree prime to ℓ,*

The same technique proves the following arithmetic version as well:

Theorem 4.8. *(Kerz-Saito [KeS], [Sa3]) Let X be a regular scheme, proper flat scheme over a henselian discrete valuation ring with finite residue field F. Then, for every prime $\ell \neq \mathrm{ch}(F)$, we have $KH_a(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell) = 0$ for $a \geq 0$.*

We remark that one can prove the above results with $\mathbb{Z}/\ell^n\mathbb{Z}$-coefficient instead of $\mathbb{Q}_\ell/\mathbb{Z}_\ell$-coefficient by using the Bloch-Kato conjecture:

Theorem 4.9. *Let X and ℓ be as in Theorem 4.6 or Theorem 4.8. Assume $(\mathbf{BK})^t_{X,\ell}$ holds. Then we have $KH_a(X, \mathbb{Z}/\ell^n\mathbb{Z}) = 0$ for $0 < a \leq t$.*

In the rest of this section we give a very rough sketch of the proof of Theorem 4.6. We fix a finite field F and work in the category $\mathcal{C}$ of schemes separated of finite type over F. We first recall the following:

Definition 4.10. Let $\mathcal{C}_*$ be the category with the same objects as $\mathcal{C}$, but morphisms are the proper maps in $\mathcal{C}$. Let Ab be the category of abelian group.

A homology theory $H = \{H_a\}_{a\in\mathbb{Z}}$ on $\mathcal{C}$ is a sequence of covariant functors:

$$H_a(-):\ \mathcal{C}_* \to Ab$$

satisfying the following conditions:

(i) For each open immersion $j : V \hookrightarrow X$ in $\mathcal{C}$, there is a map $j^* : H_a(X) \to H_a(V)$, associated to j in a functorial way.

(ii) If $i : Y \hookrightarrow X$ is a closed immersion in X, with open complement $j : V \hookrightarrow X$, there is a long exact sequence (called localization sequence)

$$\cdots \xrightarrow{\partial} H_a(Y) \xrightarrow{i_*} H_a(X) \xrightarrow{j^*} H_a(V) \xrightarrow{\partial} H_{a-1}(Y) \longrightarrow \cdots .$$

(The maps ∂ are called the connecting morphisms.) This sequence is functorial with respect to proper maps or open immersions, in an obvious way.

It is an easy exercise to check that Kato homology (3.4)

$$KH(-,\Lambda) = \{KH_a(-,\Lambda)\}_{a\in\mathbb{Z}} \quad (\Lambda = \mathbb{Z}/n\mathbb{Z},\ \mathbb{Q}/\mathbb{Z},\ \mathbb{Q}_\ell/\mathbb{Z}_\ell)$$

provides us with a homology theory on $\mathcal{C}$. Another homology theory which we use is the étale homology theory $H^{\text{ét}}(-,\Lambda)$ on $\mathcal{C}$ given by

$$H^{\text{ét}}_a(X,\Lambda) := H^{-a}(X_{\text{ét}}, R\,f^!\Lambda) \quad \text{for } f : X \to \mathrm{Spec}(F) \text{ in } \mathcal{C}.$$

where $Rf^!$ is the right adjoint of $Rf_!$ defined in [SGA 4], XVIII, 3.1.4. Using a result of [JSS], we can identify $KH_a(X,\Lambda)$ with an E^2-term of the niveau spectral sequence to get the following map as an edge homomorphism

$$\epsilon_a :\ H^{\text{ét}}_{a-1}(X,\Lambda) \to KH_a(X,\Lambda) \quad \text{for each } a \geq 1 \text{ and } X \in \mathcal{C}.$$

This gives rise to a natural transformation of homology theories

$$\epsilon :\ H^{\text{ét}}(-,\Lambda)[-1] \to KH(-,\Lambda).$$

We now keep our attention to the above homology theories in case $\Lambda = \mathbb{Q}_\ell/\mathbb{Z}_\ell$ with $\ell \neq \mathrm{ch}(F)$ and in this case we simply write $KH_a(X)$ and $H^{\text{ét}}_a(X)$. For each integer $d > 0$ consider the following condition:

KC(d): For any connected $X \in \mathcal{C}$ with $\dim(X) \leq d$ which is proper and smooth over F we have $KH_a(X) = 0$ for $a \geq 1$.

We prove **KC**(d) by induction on d. One of the basic ingredients in the proof is a result of Jannsen and Saito [JS2], Lemma 3.4 relying on weight arguments [D] which implies the following (see [Sa3], Lemma 3.10 for its proof):

Claim 4.11. *Assume* **KC**$(d-1)$. *Let $X \in \mathcal{C}$ be connected proper smooth over F, and let Y be a divisor with simple normal crossings on X such that one of the irreducible components of Y is ample. Put $U = X - Y$. Then the composite map*

$$\delta_a :\ H^{\acute{e}t}_{a-1}(U) \xrightarrow{\epsilon_a} KH_a(U) \xrightarrow{\partial} KH_{a-1}(Y)$$

is injective for $1 \leq a \leq d$ and surjective for $a \geq 2$.

Now we sketch a proof of $\mathbf{KC}(d-1) \Longrightarrow \mathbf{KC}(d)$. Let $X \in \mathcal{C}$ be a connected proper smooth over F with $\dim(X) = d$. Fix an element $\alpha \in KH_a(X)$ for $a \geq 1$. We have to show $\alpha = 0$. From the construction of ϵ_a, it is easy to see that there is a dense open subscheme $j : U \to X$ satisfying the condition

$$(*) \qquad j^*(\alpha) \text{ is in the image of } \epsilon_a : H^{\text{ét}}_{a-1}(U) \to KH_a(U).$$

Suppose for the moment that $Y = X - U$ is a divisor with simple normal crossings on X. Then one can use a Bertini argument to find a hypersurface section $H \hookrightarrow X$ such that $Y \cup H$ is a divisor with simple normal crossings. Replacing Y by $Y \cup H$ and U by $U - U \cap H$, the condition $(*)$ is preserved. Consider the commutative diagram

$$\begin{tikzcd}
KH_{a+1}(U) \arrow[r, "\partial"] & KH_a(Y) \arrow[r] & KH_a(X) \arrow[r, "j^*"] & KH_a(U) \arrow[r, "\partial"] & KH_{a-1}(Y) \\
H^{\text{ét}}_a(U) \arrow[u, "\epsilon_{a+1}"] \arrow[ur, "\delta_{a+1}"'] & & & H^{\text{ét}}_{a-1}(U) \arrow[u, "\epsilon_a"] \arrow[ur, "\delta_a"'] &
\end{tikzcd}$$

By the assumption $\mathbf{KC}(d-1)$, Claim 4.11 implies that the map δ_a is injective and the map δ_{a+1} is surjective. A simple diagram chase shows that $\alpha = 0$.

In the general case in which $Y \hookrightarrow X$ is not necessarily a divisor with simple normal crossings we use Theorem 4.7 to find an alteration $f : X' \to X$ of degree prime to ℓ such that $f^{-1}(Y)$ is a divisor with simple normal crossings. We then construct a pullback map

$$f^* : KH_a(X) \to KH_a(X')$$

which allows us to conduct the above argument for $f^*(\alpha) \in KH_a(X')$. This implies $f^*(\alpha) = 0$ and taking the pushforward gives $f_* f^*(\alpha) = \deg(f)\,\alpha = 0$. Since $\deg(f)$ is prime to ℓ we conclude $\alpha = 0$ and therefore we have finished the proof.

The construction of the necessary pullback map on Kato homology, especially in the arithmetic case, and its compatibility with the pullback map on etale homology are the most severe technical difficulties. This problem is solved using Rost's version of intersection theory and the method of deformation to normal cones [R].

5. Application: Finiteness of Motivic Cohomology

In the following sections we present some applications of the results on the cohomological Hasse principle of §4. The first apllication is on the finiteness conjecture for motivic cohomology.

Theorem 5.1. *Let X be a quasi-projective scheme over either a finite field F or a henselian discrete valuation ring with finite residue field F. Let $n > 0$ be an integer prime to* $\mathrm{ch}(F)$ *and assume* $\mathbf{(BK)}^t_{X,\ell}$ *for all primes $\ell | n$ and integers $t \geq 0$. Then* $\mathrm{CH}^r(X, q; \mathbb{Z}/n\mathbb{Z})$ *is finite for all $r \geq \dim(X)$ and $q \geq 0$.*

Proof When X is regular and projective over the base, the assertion follows from Theorem 4.9 and Lemma 3.6. The general case is reduced to the special case by using the localization sequence for $\mathrm{CH}^r(X, q; \mathbb{Z}/n\mathbb{Z})$ and Gabbers's theorem 4.7 (and its variant for schemes over a discrete valuation ring). For simplicity we only treat the case over a finite field F. We may assume $n = \ell^m$ for a prime $\ell \neq \mathrm{ch}(F)$. We proceed by induction on $\dim(X)$. First we remark that the localization sequence for higher Chow groups implies that for a dense open subscheme $U \subset X$, the finiteness of $\mathrm{CH}^r(X, q; \mathbb{Z}/n\mathbb{Z})$ for all $r \geq \dim(X)$ and q is equivalent to that of $\mathrm{CH}^r(U, q; \mathbb{Z}/n\mathbb{Z})$. Thus it suffices to show the assertion for any smooth variety U over F. If U is an open subscheme of a smooth projective variety X over F, we have already seen that the assertion holds for X and hence for U by the above remark. In general Gabbers's theorem 4.7 implies that there exist an open subscheme V of a smooth projective variety X over F, an open subscheme W of U, and a finite étale morphism $\pi : V \to W$ of degree prime to ℓ. We know that the assertion holds for V so that it holds for W by a standard norm argument. This completes the proof by the above remark. □

We note that the above theorem implies an affirmative result on the Bass conjecture. Let $K'_i(X, \mathbb{Z}/n\mathbb{Z})$ be Quillen's higher K-groups with finite coefficients constructed from the category of coherent sheaves on X (which coincide with the algebraic K-groups with finite coefficients constructed from the category of vector bundles when X is regular).

Corollary 5.2. *Under the assumption of Theorem 5.1, $K'_i(X, \mathbb{Z}/n\mathbb{Z})$ is finite for $i \geq \dim(X) - 2$.*

Proof Theorem 2.7 implies that $\mathrm{CH}^r(X, q; \mathbb{Z}/n\mathbb{Z})$ is finite for $r \leq q+1$. Hence the assertion follows from Theorem 5.1 and the Atiyah-Hirzebruch spectral sequence (see [Le] for its construction in the most general case):

$$E_2^{p,q} = \mathrm{CH}^{-q/2}(X, -p-q; \mathbb{Z}/n\mathbb{Z}) \Rightarrow K'_{-p-q}(X, \mathbb{Z}/n\mathbb{Z})$$

(note $E_2^{p,q}$ can be nonzero only if $q \leq 0$ and $p + q \leq 0$). □

6. Application: Special Values of Zeta Functions

Let X be a smooth projective variety over a finite field F. We consider the zeta function

$$\zeta(X, s) = \prod_{x \in X_{(0)}} \frac{1}{1 - N(x)^{-s}} \quad (s \in \mathbb{C})$$

where $N(x)$ is the cardinality of the residue field $\kappa(x)$ of x. The infinite product converges absolutely in the region $\{s \in \mathbb{C} \mid \Re(s) > \dim(X)\}$ and can be continued to the whole s-plane as a meromorphic function. Indeed the fundamental results of Grothendieck and Deligne imply that

$$\zeta(X,s) = \prod_{0 \le i \le 2d} P_X^i(q^{-s})^{(-1)^{i+1}},$$

where $P_X^i(t) \in \mathbb{Z}[t]$, and that for every integer r

$$\zeta(X,r)^* := \lim_{s \to r} \zeta(X,s) \cdot (1 - q^{r-s})^{\rho_r}$$

is a rational number, where $\rho_r = -\mathrm{ord}_{s=r}\zeta(X,s)$. The problem is to express these values in terms of arithmetic invariants associated to X. It has been studied by Milne [Mil] (who used étale cohomology) and Lichtenbaum [Li] (who used (conjectural) étale motivic complexes) and Geisser [Ge1] (who used Weil-étale cohomology). As an application of Theorem 4.9, we get the following new result on the problem.

Theorem 6.1. *Let X be a smooth projective variety over a finite field F. Let $p = \mathrm{ch}(F)$ and $d = \dim(X)$.*

(1) For all integers j, the torsion part $H_M^j(X, \mathbb{Z}(d))_{tors}$ of $H_M^j(X, \mathbb{Z}(d))$ is finite modulo the p-primary torsion subgroup. Moreover, $H_M^j(X, \mathbb{Z}(d))_{tors}$ is finite if $d \le 4$.

(2) We have the equality up to a power of p:

$$\zeta(X,0)^* = \prod_{0 \le j \le 2d} |H_M^j(X, \mathbb{Z}(d))_{tors}|^{(-1)^j} \tag{6.1}$$

The equality holds also for the p-part if $d \le 4$.

Remark 6.2. Let $X = \mathrm{Spec}(\mathcal{O}_K)$ where $\mathcal{O}_K$ is the ring of integers in a number field. The formula (6.1) should be compared with the formula

$$\lim_{s \to 0} \zeta(X,s) \cdot s^{-\rho_0} = -\frac{|H_M^2(X, \mathbb{Z}(1))_{\mathrm{tors}}|}{|H_M^1(X, \mathbb{Z}(1))_{\mathrm{tors}}|} \cdot R_K$$

which is obtained by rewriting the class number formula (1.1) using motivic cohomology. Thus (6.1) may be viewed as a geometric analogue of the class number formula. Note that the regulator R_K does not appear in (6.1) since $H_M^j(X, \mathbb{Z}(d))$ is (conjecturally) finite for $j \ne 2d$.

Proof of Theorem: For simplicity we treat only the case $\ell \neq \mathrm{ch}(F)$. Put

$$H^j_M(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d)) = \varinjlim_n H^j_M(X, \mathbb{Z}/\ell^n\mathbb{Z}(d)),$$

$$H^j_{\text{ét}}(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d)) = \varinjlim_n H^j_{\text{ét}}(X, \mathbb{Z}/\ell^n\mathbb{Z}(d)).$$

By Theorem 4.6 and Lemma 3.6 we have an isomorphism

$$H^j_M(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d)) \simeq H^j_{\text{ét}}(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d)). \tag{6.2}$$

By (1.4) we have an exact sequence

$$0 \to H^j_M(X, \mathbb{Z}(d)) \otimes \mathbb{Q}_\ell/\mathbb{Z}_\ell \to H^j_M(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d)) \to H^{j+1}_M(X, \mathbb{Z}(r))\{\ell\} \to 0 \tag{6.3}$$

where $M\{\ell\}$ denotes the ℓ-primary torsion part for an abelian group M. Assuming $j \neq 2d$, one can show using Deligne's proof of the Weil conjecture [D] and a theorem of Gabber [Ga] that $H^j_{\text{ét}}(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d))$ is finite and trivial for almost all ℓ (see [CTSS], Theorem 2). Thus (6.2) and (6.3) imply $H^j_M(X, \mathbb{Z}(d)) \otimes \mathbb{Q}_\ell/\mathbb{Z}_\ell = 0$ and we get an isomorphism of finite groups

$$H^{j+1}_M(X, \mathbb{Z}(r))\{\ell\} \simeq H^j_{\text{ét}}(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d)). \tag{6.4}$$

This shows the first assertion (1). For the proof of (2), we use the formula

$$\zeta(X, 0)^* = \frac{[H^0_{\text{ét}}(X, \mathbb{Z})_{\text{tors}}][H^2_{\text{ét}}(X, \mathbb{Z})_{\text{cotor}}][H^4_{\text{ét}}(X, \mathbb{Z})] \cdots}{[H^1_{\text{ét}}(X, \mathbb{Z})][H^3_{\text{ét}}(X, \mathbb{Z})][H^5_{\text{ét}}(X, \mathbb{Z})] \cdots} \tag{6.5}$$

due to Milne [Mil], Theorem 0.4. Here $H^0_{\text{ét}}(X, \mathbb{Z}) = \mathbb{Z}$, $H^1_{\text{ét}}(X, \mathbb{Z}) = 0$, and $H^j_{\text{ét}}(X, \mathbb{Z})$ is finite for $j \geq 3$, and the cotorsion part $H^2_{\text{ét}}(X, \mathbb{Z})_{\text{cotor}}$ of $H^2_{\text{ét}}(X, \mathbb{Z})$ is finite. By arithmetic Poincaré duality we have

$$H^{2d-i}_{\text{ét}}(X, \mathbb{Q}_\ell/\mathbb{Z}_\ell(d)) \simeq \mathrm{Hom}(H^{i+1}_{\text{ét}}(X, \mathbb{Z}_\ell), \mathbb{Q}_\ell/\mathbb{Z}_\ell),$$

where $H^{i+1}_{\text{ét}}(X, \mathbb{Z}_\ell) = \varprojlim_n H^{i+1}_{\text{ét}}(X, \mathbb{Z}/\ell^n\mathbb{Z})$, and this group is finite for $i \geq 1$. Thus the desired assertion follows from the following isomorphisms

$$H^j_{\text{ét}}(X, \mathbb{Z}_\ell) \simeq H^j_{\text{ét}}(X, \mathbb{Z})\{\ell\} \quad \text{for } j \geq 3,$$

$$H^2_{\text{ét}}(X, \mathbb{Z}_\ell) \simeq H^2_{\text{ét}}(X, \mathbb{Z})_{\text{cotor}}\{\ell\},$$

which can be easily shown by using the exact sequence of étale sheaves

$$0 \to \mathbb{Z} \xrightarrow{\ell^n} \mathbb{Z} \to \mathbb{Z}/\ell^n\mathbb{Z} \to 0. \qquad \square$$

7. Application: Higher Class Field Theory

Another application of Theorem 4.9 is a generalization of the higher dimensional class field theory by Schmidt-Spiess [ScSp] which describes the abelian fundamental group of a smooth scheme over a finite field by using its Suslin homology of degree 0 (see (7.2) below). The generalization is its higher-degree variant and establishes an isomorphism between the Suslin homlogy of higher degree and the dual of étale cohomology (see Theorem 7.2 below).

We start with a brief review of higher dimensional class field theory. Let X be a regular scheme of finite type over $\mathbb{F}_p$ or $\mathbb{Z}$. Higher dimensional class field theory aims at describing all relations among the Frobenius elements

$$\sigma_x \in \pi_1^{ab}(X)$$

associated to closed points of X. Here $\pi_1^{ab}(X)$ is the abelian fundamental group of X, which classifies the abelian finite étale coverings of X. To be more precise, let $X_{(0)}$ be the set of the closed points of X. For $x \in X_{(0)}$, the residue field $\kappa(x)$ of x is finite. The closed immersion $x \to X$ induces $\rho_x : \pi_1^{ab}(x) \to \pi_1^{ab}(X)$ and $\pi_1^{ab}(x)$ is the absolute Galois group of $\kappa(x)$ which is topologically generated by the q-th power map where $q = |\kappa(x)|$. The Frobenius element $\sigma_x \in \pi_1^{ab}(X)$ is defined as its image under ρ_x. This defines the map

$$\rho_X : Z_0(X) \to \pi_1^{ab}(X)\ ;\ \big(n_x\big)_{x\in X_{(0)}} \to \prod_{x\in X_{(0)}} (\sigma_x)^{n_x}$$

where $Z_0(X) = \bigoplus_{x\in X_{(0)}} \mathbb{Z}$ is the group of zero cycles on X. It was shown by Lang that the image of ρ_X is dense in $\pi_1^{ab}(X)$ and the problem is to determine its kernel.

The question was first answered by Kato-Saito [KS2], which used the higher idele class group of X defined as the cohomolgy group of the sheaf of relative Milnor K-group with respect to the Nisnevich topology. Unfortunately, the description of the kernel of ρ_X in this formulation is not direct and does not give a clear answer to the above question except in the case where X is proper over the base. It is the higher unramified class field theory stated as an (almost) isomorphism:

$$\rho_X\ :\ \mathrm{CH}_0(X) \to \pi_1^{ab}(X)$$

(see Theorem 2.6). Recall

$$\mathrm{CH}_0(X) = \mathrm{Coker}\left(\bigoplus_{y\in X_{(1)}} \kappa(y)^\times \xrightarrow{\delta} \bigoplus_{x\in X_{(0)}} \mathbb{Z}\right)$$

where $X_{(1)}$ is the set of the generic points of the integral curves on X and δ is given by taking the divisors of functions on those curves.

An essential improvement has been given by the following theorem due to Schmidt-Spiess [ScSp] and Kerz-Schmidt-Wiesend [W], [KeSc] (see also [Sz]).

Theorem 7.1. *Let X be a connected regular scheme of finite type over $\mathbb{F}_p$ or $\mathbb{Z}$. Let $n > 0$ be an integer prime to the characteristic of the function field of X. Then $\rho_X : Z_0(X) \to \pi_1^{ab}(X)$ induces an isomorphism*

$$\mathrm{Coker}\left(\bigoplus_{y\in X_{(1)}} \kappa(y)_{\Sigma,n} \xrightarrow{\delta} \bigoplus_{x\in X_{(0)}} \mathbb{Z}/n\mathbb{Z}\right) \simeq \pi_1^{ab}(X)/n$$

For $y \in X_{(1)}$, $\kappa(y)_{\Sigma,n}$ is a subgroup of $\kappa(y)^\times$ defined as follows. For simplicity we restrict to the case X is over $\mathbb{F}_p$. Let $C \subset X$ be the closure of y in X, $\widetilde{C}$ be its normalization, and $\overline{C}$ be the smooth compactification of $\widetilde{C}$, and put $\Sigma_y = \overline{C} - \widetilde{C}$. Then

$$\kappa(y)_{\Sigma,n} = \left\{a \in \kappa(y)^\times \;\middle|\; a \in \left(\kappa(y)_x^\times\right)^n \quad \text{for all } x \in \Sigma_y\right\}$$

where $\kappa(y)_x$ is the completion of $\kappa(y)$ at x.

In what follows we assume that X is smooth over a finite field. Schmidt-Spiess [ScSp] have established a canonical isomorphism

$$\mathrm{Coker}\left(\bigoplus_{y\in X_{(1)}} \kappa(y)_{\Sigma,n} \xrightarrow{\delta} \bigoplus_{x\in X_{(0)}} \mathbb{Z}/n\mathbb{Z}\right) \simeq H_0^S(X, \mathbb{Z}/n\mathbb{Z}). \tag{7.1}$$

(A similar isomorphism was shown by Schmidt [Sc] when X is flat over $\mathbb{Z}$ under a certain tameness condition). Thus Theorem 7.1 can be rephrased as a canonical isomorphism

$$H_0^S(X, \mathbb{Z}/n\mathbb{Z}) \simeq \pi_1^{ab}(X)/n. \tag{7.2}$$

As an application of Theorem 4.9, we can extend this to the following.

Theorem 7.2. ([KeS]) *Let X be a connected smooth scheme over a finite field $\mathbb{F}_q$ and let $n > 0$ be an integer prime to* $\mathrm{ch}(\mathbb{F}_q)$. *Then there exists a canonical isomorphism for all integers $i \geq 0$*

$$H_i^S(X, \mathbb{Z}/n\mathbb{Z}) \simeq H_{\acute{e}t}^{i+1}(X, \mathbb{Z}/n\mathbb{Z})^* := \mathrm{Hom}\big(H_{\acute{e}t}^{i+1}(X, \mathbb{Z}/n\mathbb{Z}), \mathbb{Z}/n\mathbb{Z}\big),$$

where $H_i^S(X, \mathbb{Z}/n\mathbb{Z})$ is the Suslin homology defined in §1. In particular $H_i^S(X, \mathbb{Z}/n\mathbb{Z})$ is finite.

The case $i = 0$ of Theorem 7.2 is reduced to the isomorphism (7.2) by the natural isomorphism $H_{\text{ét}}^1(X, \mathbb{Z}/n\mathbb{Z})^* \simeq \pi_1^{ab}(X)/n$.

Remark 7.3. Let X be separated of finite type over a finite field F. Assuming resolution of singularities over F, Geisser [Ge3] proved that $H_i^S(X, \mathbb{Z}/n\mathbb{Z})$ is finite for all integers i and n.

8. Application: Resolution of Quotient Singularities

We fix a field k and assume $\mathrm{ch}(k) = 0$. Let $\mathcal{C}$ be the category $\mathcal{C}$ of separated schemes of finite type over k. Let $\mathcal{S} \subset \mathcal{C}$ be the subcategory of smooth projective schemes over k. Fix an abelian group Λ. Based on work of Gillet and Soulé [GS], Jannsen ([J], Theorem 5.9) proved the following.

Theorem 8.1. *There exists a homology theory (cf. Definition 4.10)*

$$H^W(-, \Lambda) : \mathcal{C}_* \to \mathcal{A}b$$

such that for all $X \in \mathcal{S}$, we have

$$H_a^W(X, \Lambda) = \begin{cases} \Lambda^{\pi_0(X)} & a = 0 \\ 0 & a \neq 0, \end{cases}$$

where $\pi_0(X)$ is the set of connected components of X. We call $H_a^W(X, \Lambda)$ the weight homology group of X with coefficient Λ.

We briefly review the construction of [GS]. To a simplicial object in $\mathcal{S}$:

$$X_\bullet : \quad \cdots X_2 \begin{array}{c} \xrightarrow{\delta_0} \\ \xleftarrow{s_0} \\ \xrightarrow{\delta_1} \\ \xleftarrow{s_1} \\ \xrightarrow{\delta_2} \end{array} X_1 \begin{array}{c} \xrightarrow{\delta_0} \\ \xleftarrow{s_0} \\ \xrightarrow{\delta_1} \end{array} X_0,$$

we associate a chain complex of abelian groups

$$W(X_\bullet, \Lambda) : \ \cdots \to \Lambda^{\pi_0(X_n)} \xrightarrow{\partial} \Lambda^{\pi_0(X_{n-1})} \xrightarrow{\partial} \cdots \xrightarrow{\partial} \Lambda^{\pi_0(X_0)},$$

where $\partial : \Lambda^{\pi_0(X_n)} \to \Lambda^{\pi_0(X_{n-1})}$ is defined as $\partial = \sum_{a=0}^n (-1)^a \partial_a$ with

$$\partial_a : \Lambda^{\pi_0(X_n)} \to \Lambda^{\pi_0(X_{n-1})} \ ; \ (x_i)_{i \in \pi_0(X_n)} \to \left(\sum_{\delta_a(i)=j} x_i \right)_{j \in \pi_0(X_{n-1})}.$$

For $X \in \mathcal{C}$ choose an open immersion $j : X \to \overline{X}$ with $\overline{X} \in \mathcal{C}$ proper over k and let $i : Y = \overline{X} - X \to \overline{X}$ be the closed immersion for the complement. By

[GS] 1.4, one can find a diagram

$$\begin{array}{ccc} Y_\bullet & \xrightarrow{i_\bullet} & \overline{X}_\bullet \\ \downarrow \pi_Y & & \downarrow \pi_X \\ Y & \xrightarrow{i} & \overline{X} \end{array} \tag{8.1}$$

where $Y_\bullet$ and $\overline{X}_\bullet$ are simplicial objects in $\mathcal{S}$ and π_X and π_Y are hyperenvelopes. To this diagram one associates a complex

$$Cone\big(W(Y_\bullet, \Lambda) \xrightarrow{i_{\bullet *}} W(\overline{X}_\bullet, \Lambda)\big).$$

By [GS], 1.4, the image $W(X, \Lambda)$ of the above complex in the homotopy category of chain complexes of abelian groups depends only on X and not on a choice of the diagram (8.1). The homology theory in Theorem 8.1 is defined as

$$H_a^W(X, \Lambda) := H_a(W(X, \Lambda)) \quad \text{for } X \in \mathcal{C}.$$

For example, if X is a divisor with simple normal crossings on a smooth projective variety over k, $H_a^W(X, \Lambda)$ is a homology group of the complex:

$$\cdots \to \Lambda^{\pi_0(X^{(a)})} \xrightarrow{\partial} \Lambda^{\pi_0(X^{(a-1)})} \xrightarrow{\partial} \cdots \xrightarrow{\partial} \Lambda^{\pi_0(X^{(1)})}.$$

where $X_1, \ldots, X_N$ are the irreducible components of X and

$$X^{(a)} = \coprod_{1 \le i_i < \cdots < i_a \le N} X_{i_1, \ldots, i_a} \quad (X_{i_1, \ldots, i_a} = X_{i_1} \cap \cdots \cap X_{i_a}),$$

and the differentials ∂ are obvious ones.

Theorem 8.2. *([KeS2]) Let X be a quasi-projective smooth variety over k with action of a finite group G. Let X/G be the geometric quotient ([Mu], Ch.II §7, SGA 1 V §1).*

(1) Assume that X is projective. Then $H_a^W(X/G, \mathbb{Z}) = 0$ for all $a > 0$.

(2) Assume that the singular locus Z of X/G is proper over k. Let $\pi : Y \to X/G$ be a proper birational morphism such that Y is smooth over k and π is an isomorphism over outside Z. Let E be the reduced part of $\pi^{-1}(Z)$. Then $H_a^W(E, \mathbb{Z}) \simeq H_a^W(Z, \mathbb{Z})$ for all a. In particular, if Z is regular, then $H_a^W(E, \mathbb{Z}) = 0$ for all $a > 0$.

Here we explain an idea of the proof of Theorem 8.2(1). The second assertion (2) is an easy consequence of (1). Since $H_a^W(Y, \mathbb{Z})$ for $Y \in \mathcal{C}$ is finitely generated, it suffices to show the assertion for the weight homology group with coefficient $\Lambda = \mathbb{Q}/\mathbb{Z}$. Without loss of generality, we assume that k is finitely generated over $\mathbb{Q}$. Then the basic idea of the proof is to introduce an arithmetic invariant

$KH_a(Y)$ for $Y \in \mathcal{C}$ which is defined without referring to desingularizations (or hyperenvelopes) and to show the following facts:

($*1$) $H_a^W(Y, \mathbb{Q}/\mathbb{Z}) \simeq KH_a(Y)$ for all $Y \in \mathcal{C}$.

($*2$) $KH_a(X/G) = 0$ for $a \neq 0$, where X/G is as in Theorem 8.2(1).

To define such an invariant we consider

$$H_a(X) := \varinjlim_n \operatorname{Hom}\big(H_c^a(X_{\text{ét}}, \mathbb{Z}/n\mathbb{Z}), \mathbb{Q}/\mathbb{Z}\big) \quad \text{for } X \in Ob(\mathcal{C}). \tag{8.2}$$

where $H_c^a(X_{\text{ét}}, \mathbb{Z}/n\mathbb{Z})$ is the étale cohomology with compact support (cf. [JS2], Example 2.5). It provides a homology theory on $\mathcal{C}$ and gives rise to the niveau spectral sequence:

$$E^1_{p,q}(X) = \bigoplus_{x \in X_{(p)}} H_{p+q}(x) \quad \Rightarrow \quad H_{p+q}(X) \quad \text{with } H_a(x) = \varinjlim_{V \subseteq \overline{\{x\}}} H_a(V). \tag{8.3}$$

Here the limit is over all non-empty open subschemes $V \subseteq \overline{\{x\}}$. The affine Lefschetz theorem implies $E^1_{p,q}(X) = 0$ for $q < 0$ and the desired arithmetic invariant $KH_a(X)$ is defined as $E^2_{a,0}(X)$, an E^2 term of the spectral sequence. By the same techniques as the proof of Theorems 4.5 and 4.6, one can prove the following.

Theorem 8.3. *For $X \in \mathcal{S}$, we have*

$$KH_a(X) = \begin{cases} (\mathbb{Q}/\mathbb{Z})^{\pi_0(X)} & a = 0 \\ 0 & a \neq 0, \end{cases}$$

The assertion ($*1$) follows from Theorem 8.3 and a result of Jannsen [J], Theorem 5.13. We note that the proof of Theorem 8.3 uses the weight argument ([D]) and requires the assumption that k is finitely generated. In order to show the assertion ($*2$), we apply the same argument to the equivariant version of (8.2). We fix a finite group G and let $\mathcal{C}_G$ be the category of quasi-projective schemes over k with a G-action. We consider

$$H_a^G(X) := \varinjlim_n \operatorname{Hom}\big(H_c^a(G; X_{\text{ét}}, \mathbb{Z}/n\mathbb{Z}), \mathbb{Q}/\mathbb{Z}\big) \quad \text{for } X \in Ob(\mathcal{C}). \tag{8.4}$$

Here

$$H_c^a(G; X_{\text{ét}}, \mathbb{Z}/n\mathbb{Z}) := \mathbb{R}\Gamma(G, \mathbb{R}\Gamma(\overline{X}_{\text{ét}}, j_!\mathbb{Z}/n\mathbb{Z}))),$$

is the equivariant étale cohomology with compact support, where $j : X \hookrightarrow \overline{X}$ is any equivariant compactification of X, and $R\Gamma(G, -)$ is the derived functor of taking G-invariants. This provides a homology theory on $\mathcal{C}_G$ and the equivariant version $KH_a^G(X)$ of $KH_a(X)$ is defined as an E^2-term of the associated niveau spectral sequence. Then ($*2$) follows from the following.

Theorem 8.4. *([KeS2]) Let $X \in \mathcal{C}_G$ be smooth over k.*

(1) We have a natural isomorphism $KH_a^G(X) \simeq KH_a(X/G)$ for all a.

(2) If X is projective, we have

$$KH_a^G(X) = \begin{cases} (\mathbb{Q}/\mathbb{Z})^{\pi_0(X/G)} & a = 0 \\ 0 & a \neq 0, \end{cases}$$

References

[B1] S. Bloch, *Higher Algebraic K-theory and class field theory for arithmetic surfaces*, Ann. of Math. **114** (1981), 229–265.

[B2] S. Bloch, *Algebraic cycles and higher algebraic K-theory*, Adv. Math. **61** (1986), 267–304.

[BK] S. Bloch and K. Kato, *p-adic étale cohomology*, Publ. Math. IHES **63** (1986), 107–152.

[CJS] V. Cossart, U. Jannsen and S. Saito, *Resolution of singularities for embedded surfaces*, in preparation (see www.mathematik.uni-regensburg.de/Jannsen).

[CT] J.-L. Colliot-Thélène, *On the reciprocity sequence in the higher class field theory of function fields*, Algebraic K-Theory and Algebraic Topology (Lake Louise, AB, 1991), (J.F. Jardine and V.P. Snaith, ed), 35–55, Kluwer Academic Publishers, 1993.

[CTSS] J.-L. Colliot-Thélène, J.-J. Sansuc and C. Soulé, *Torsion dans le groupe de Chow de codimension deux*, Duke Math. J. **50** (1983), 763–801.

[D] P. Deligne, *La conjecture de Weil II*, Publ. Math. IHES **52** (1981), 313–428.

[Ga] O. Gabber, *Sur la torsion dans la cohomologie l-adique d'une variété*, C. R. Acad. Sc. Paris Sér. I Math. **297** (1983), 179–182.

[Ge1] T. Geisser, *Weil-étale cohomology over finite fields*, Math. Ann. **330** (2004), 665–692.

[Ge2] T. Geisser, *Arithmetic homology and an integral version of Kato's conjecture*, To appear in J. Reine Angew.

[Ge3] T. Geisser, *On Suslin's singular homology and cohomology*, preprint.

[GL] T. Geisser and M. Levine, *The Bloch-Kato conjecture and a theorem of Suslin-Voevodsky*, J. Reine Angew. **530** (2001), 55–103.

[Gra1] D. Grayson, *Finite generation of K-groups of a curve over a finite field (after Quillen)*, Lecture Notes in Math. **966** (1982), Springer-Verlag, Berlin, 69–90.

[Gra2] D. Grayson, *The motivic spectral sequence*, in: Handbook of K-theory 1, eds. E. friedlander, D. Grayson, Springer (2005).

[Gr] M. Gros, *Sur la partie p-primaire du groupe de Chow de codimension deux*, Comm. Algebra **13** (1985), 2407–2420.

[GS] H. Gillet and C. Soulé, *Descent, motives and K-theory*, J. Reine Angew. **478** (1996), 127–176.

[HW] C. Weibel, *Axioms for the Norm Residue Isomorphism*, K-theory Preprint Archives, http://www.math.uiuc.edu/K-theory/0809/

[Il1] L. Illusie, *Complexe de De Rham-Witt et cohomologie cristalline*, Ann. Ec. Norm. Sup. 4 série **12** (1979), 501–661.

[Il2] L. Illusie, *On Gabber's refined uniformization*, http://www.math.u-psud.fr/∼illusie/

[J] U. Jannsen, *Hasse principles for higher dimensional fields*, http://arxiv.org/abs/0910.2803.

[JS1] U. Jannsen and S. Saito, *Kato homology of arithmetic schemes and higher class field theory*, Documenta Math. Extra Volume: Kazuya Kato's Fiftieth Birthday (2003), 479–538

[JS2] U. Jannsen and S. Saito, *Kato conjecture and motivic cohomology over finite fields*, http://arxiv.org/abs/0910.2815

[JSS] U. Jannsen, S. Saito and K. Sato, *Etale duality for constructible sheaves on arithmetic schemes*, http://arxiv.org/abs/0910.3759.

[K] K. Kato, *A Hasse principle for two dimensional global fields*, J. für die reine und angew. Math. **366** (1986), 142–183.

[KS1] K. Kato and S. Saito, *Unramified class field theory of arithmetic surfaces*, Ann. of Math. **118** (1985), 241–275.

[KS2] K. Kato and S. Saito, *Global class field theory of arithmetic schemes*, Contemporary Math. **55**(1986), 255–331.

[KeSc] M. Kerz and A. Schmidt *Covering data and higher dimensional global class field theory* J. of Number Theory **129** (2009), 2569–2599

[KeS] M. Kerz and S. Saito, *Kato conjecture and motivic cohomology for arithmetic schemes*, in preparation.

[KeS2] M. Kerz and S. Saito, *Equivariant Kato conjecture and the weight homology of quotient schemes*, in preparation.

[Le] M. Levine, *K-theory and motivic cohomology of schemes*, preprint.

[Li] S. Lichtenbaum, *Values of zeta functions at non-negative integers*, In: Number theory, Noordwijkerhout 1983, Lecture Notes in Math. —**1068**, 127–138.

[Mil] J.S. Milne, *Values of zeta functions of varieties over finite fields*, American J. of Math. **108** (1986), 297–360.

[MS] A.S. Merkurjev and A.A. Suslin, *K-cohomology of Severi-Brauer Varieties and the norm residue homomorphism*, Math. USSR Izvestiya **21** (1983), 307–340.

[Mu] D. Mumford, *Abelian varieties*, Tata Institute of Fundamental Research. Bombay, Oxford University Press (1970).

[MVW] A.S. Merkurjev and A.A. Suslin, *Lecture notes on motivic cohomology*, Clay Math. Monographs **2**, Amerian Math. Society, Clay Math. Institute

[P] B. Poonen, *Bertini theorems over finite fields*, Ann. of Math. **160** (2004), 1099–1127.

[Q] D. Quillen, *Finite generation of the group K_i of rings of algebraic integers*, Lecture Notes in Math. **341** (1973), Springer-Verlag, Berlin, 195–214.

[R] M. Rost, *Chow groups with coefficients*, Doc. Math. J. **1** (1996), 319–393.

[Sa1] S. Saito, *Unramified class field theory of arithmetic schemes*, Ann. of Math. **121** (1985), 251–281.

[Sa2] S. Saito *Cohomological Hasse principle for a threefold over a finite field*, in: Algebraic K-theory and Algebraic Topology, NATO ASI Series, **407** (1994), 229–241, Kluwer Academic Publishers

[Sa3] S. Saito, *Recent progress on the Kato conjecture*, http://www.lcv.ne.jp/∼smaki/en/preprints/index.html.

[Sat1] K. Sato, *p-adic étale Tate twists and duality of arithmetic schemes* (with an appendix by Hagihara, K.). Ann. Sci. Éc. Norm. Sup. (4) **40** (2007), 519–588

[Sat2] K. Sato, *Characteristic classes for p-adic étale Tate twists and the image of p-adic regulators*, in preparation

[SJ] A. Suslin and S. Joukhovitski, *Norm Varieties* J. Pure Appl. Alg. **206** (2006), 245–276.

[Sc] A. Schmidt, *Singular homology of arithmetic schemes* Algebra Number Theory **1**(2) (2007), 183–222.

[ScSp] A. Schmidt and M. Spiess, *Singular homology and class field thoery of varieties over finite fields* J. Reine Angew. Math. **527** (2000), 13–37.

[SV1] A. Suslin and V. Voevodsky, *Singular homology of abstract algebraic varieties* Invent. Math. **123** (1996), 61–94.

[SV2] A. Suslin and V. Voevodsky, *Bloch-Kato conjecture and motivic cohomology with finite coefficients*, in: Cycles, Transfer, and Motivic Homology Theories, Annals of Math. Studies **143**, Princeton University Press, 2000.

[Sw] N. Suwa, *A note on Gersten's conjecture for logarithmic Hodge-Witt sheaves*, K-theory **9** (1995), 245–271.

[Sz] T. Szamuely, *Corps de classes des schémas arithmétiques* http://www.renyi.hu/szamuely/publ.html.

[V1] V. Voevodsky, *Triangulated categories of motives over a field*, Cycles, Tansfers, and Motivic Homology Theories, Annals of Math. Studies, Princeton University Press

[V2] V. Voevodsky, *On motivic cohomology with $\mathbb{Z}/l$-coefficients*, K-theory Preprint Archives, http://www.math.uiuc.edu/K-theory/0639/

[V3] V. Voevodsky, *Motivic Eilenberg-MacLane spaces*, K-theory Preprint Archives, http://www.math.uiuc.edu/K-theory/0864/

[V4] V. Voevodsky, *Motivic cohomology are isomorphic to higher Chow groups*, Int. Math. Res. Not. **7** (2002), 351–355.

[W] G. Wiesend, *Class field theory for arithmetic schemes*, Math. Z. **256**(4) (2007), 717–729.

[We1] C. Weibel, *The norm residue isomorphism theorem*, K-theory Preprint Archives, http://www.math.uiuc.edu/K-theory/0934/

[We2] C. Weibel, *Patching the Norm Residue Isomorphism Theorem*, K-theory Preprint Archives, http://www.math.uiuc.edu/K-theory/0844/

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Betti Numbers of Syzygies and Cohomology of Coherent Sheaves

Frank-Olaf Schreyer* and David Eisenbud†

Abstract

The Betti numbers of a graded module over the polynomial ring form a table of numerical invariants that refines the Hilbert polynomial. A sequence of papers sparked by conjectures of Boij and Söderberg have led to the characterization of the possible Betti tables up to rational multiples—that is, to the rational cone generated by the Betti tables. We will summarize this work by describing the cone and the closely related cone of cohomology tables of vector bundles on projective space, and we will give new, simpler proofs of some of the main results. We also explain some of the applications of the theory, including the one that originally motivated the conjectures of Boij and Söderberg, a proof of the Multiplicity Conjecture of Herzog, Huneke and Srinivasan.

Mathematics Subject Classification (2010). Primary 13D02; Secondary 14F05.

Keywords. Betti numbers, free resolutions, syzygies, cohomology of coherent sheaves, multiplicity

1. Introduction

Hilbert's Syzygy theorem states that every finitely generated graded module over a polynomial ring $S = K[x_1, \ldots, x_n]$ has a finite free resolution of length at most n. Hilbert's motivation was to show that the *Hilbert function* $h_M(k) := \dim_K M_k$ is given by a polynomial $p_M(k)$ for large k, as follows: for a graded module $M = \oplus_k M_k$ over the standard graded polynomial ring S consider a finite free graded resolution, that is, an exact sequence

$$\mathbf{F}: \quad (0 \leftarrow M \leftarrow) F_0 \leftarrow F_1 \leftarrow \ldots \leftarrow F_r \leftarrow 0$$

*Fakultät für Mathematik und Informatik, E2 4, Universität des Saarlandes, D-66123 Saarbrücken, Germany. E-mail: schreyer@math.uni-sb.de

†Department of Mathematics, University of California, Berkeley, Berkeley CA 94720. E-mail: eisenbud@math.berkeley.edu.

with $F_i = \oplus_j S(-j)^{\beta_{i,j}}$ and $S(-j)$ the free cyclic S module with generator in degree j. The numbers $\beta_{i,j}$ are called the *graded Betti numbers* of the resolution. From the exactness of the resolution Hilbert deduced the formula

$$h_M(k) = \dim_K M_k = \sum_{i=0}^{r} (-1)^i \sum_j \beta_{i,j} \binom{k+n-j}{n}$$

for the Hilbert function h_M. Since the combinatorial binomial coefficient $\binom{m}{n}$ agrees with the polynomial

$$\frac{m(m-1)\cdots(m-n+1)}{n!} \in \mathbb{Q}[m]$$

when m is large enough, the same formula defines the Hilbert polynomial $p_M(k) \in \mathbb{Q}[k]$. The Hilbert polynomial $p_M(k)$ captures the most important properties of M. For example, $\deg p_M = \dim M$, and the leading coefficient of p_M times $(\dim M)!$ is the the multiplicity of M.

A minimal free resolution of a module M is a free resolution $\mathbf{F}$ such that no proper summand of F_{i+1} maps surjectively onto the kernel $\ker(F_i \to F_{i-1})$. It is determined up to isomorphism by M (see for example [6]), and thus the *graded Betti numbers* $\beta_{i,j} = \beta_{i,j}(M)$ of a minimal free resolution are invariants of M. We collect the Betti numbers of M as usual in a *Betti table* $\beta(M) = (\beta_{i,j}(M))$. Hence, $\beta(M)$ is a numerical invariant that refines the Hilbert polynomial.

There are many papers whose goal is to describe this invariant and its possible values in special cases. In 2006 Mats Boij and Jonas Söderberg suggested a relaxation of this problem that opened the door to a radically different approach. The set of Betti tables form a semigroup, since the direct sum of modules corresponds to the addition of Betti tables. Allowing multiplication by positive rational numbers instead of just positive integers, we get a rational convex cone, the cone of Betti tables. Boij and Söderberg conjectured that the *extremal rays* of this cone are spanned by Betti tables of so called *pure resolutions*, described below.

In [11] we showed that these conjectures were true. Besides proving the existence of pure resolutions this involved finding the equations of the facets of the cone of Betti tables. To describe how this was done, we must introduce another invariant of a finitely generated graded module M that also refines the Hilbert polynomial. Consider the coherent sheaf $\mathcal{E}$ on $\mathbb{P}^{n-1}$ represented by M. The value of the Hilbert polynomial $p_M(k)$ coincides with the Euler characteristic of the twisted sheaf: $\chi\mathcal{E}(k)$. Since the Euler characteristic is the alternating sum of the dimensions of the cohomology groups, we can also think of the *cohomology table* $\gamma(\mathcal{E}) = (h^j\mathcal{E}(k))$ of $\mathcal{E}$ as a refinement of the Hilbert polynomial. And, just as with Betti tables, it is natural to consider the convex rational cone generated by the cohomology tables.

The key idea in our proof of the Boij-Söderberg conjecture was to show that the facets of the cone of Betti tables of finite length modules come from

the extremal rays in the closed cone of cohomology tables of vector bundles on projective space. We identified these extremal rays and showed that vector bundles with such extremal cohomology tables exist, so that the cone of cohomology tables, like the cone of Betti tables, is closed. Using our results, Boij and Söderberg extended the theory to arbitrary modules [2]. On the cohomology side, we generalized the results to arbitrary coherent sheaves [12].

A flurry of other papers and preprints including [9], [24], [14] [15] and [7] have added to the basic picture and its applications. In this note we give a new and simpler proof of our main result on the cone of Betti tables, and we give a simpler treatment of the theorem of Boij and Söderberg on Betti tables of arbitrary finitely generated modules. We also explain some applications, and survey what is known in some other cases.

2. Betti Tables

As above, let $S = K[x_1, \ldots, x_n]$ be a polynomial ring over a field K, graded with each x_i of degree 1. For a finitely generated graded S-module M we may regard its Betti table as an integral point in an infinite dimensional $\mathbb{Q}$-vector space:

$$\beta(M) = (\beta_{i,j}(M)) \in \bigoplus_{i=0}^{n} \bigoplus_{j\in\mathbb{Z}} \mathbb{Q}.$$

The cone generated by finite positive rational linear combinations of Betti tables is called the *Boij-Söderberg cone*. Our main result on Betti tables is a description of this cone in terms of its extremal rays.

Definition 1. A finitely generated graded S-module M is called *pure* of type $d := (d_0, \ldots, d_r)$ if

1. M is Cohen-Macaulay of codimension r; that is, $F_i = 0$ for $i > r$ and $\dim M = n - r$
2. in a minimal free resolution of M as above, the free module F_i is generated by elements of degree d_i; that is, $\beta_{i,j} = 0$ when $j \neq d_i$.

Proposition 2.1. *If M is a pure module of type d, then $\beta(M)$ is determined up to a rational factor by d. Further, $\beta(M)$ spans an extremal ray in the Boij-Söderberg cone of Betti tables.*

Proof. Suppose $\beta(M) = \sum_{\ell=1}^{N} q_\ell \beta(M_\ell)$ with rational numbers $q_\ell > 0$. We have to prove that the Betti tables $\beta(M_\ell)$ all lie in the same ray as $\beta(M)$. Each of the modules M_ℓ has $\dim M_\ell \leq \dim M$, because otherwise the Hilbert polynomial of M would have larger degree. Since by the Auslander-Buchsbaum-Serre formula, the length of a free resolution is at least the codimension, the equality of Betti tables implies that each M_ℓ is Cohen-Macaulay with the same codimension r as M and that each M_ℓ is pure with the same type $(d_1, \ldots, d_r)$ as M. The proof

that each $\beta(M_\ell)$ lies in the same ray as $\beta(M)$ follows by an argument of Herzog and Kühl [17]: Consider the Hilbert series of M defined as

$$H_M(t) = \sum_{d\in\mathbb{Z}} \dim M_d t^d = \frac{\sum_{i,j}(-1)^i\beta_{i,j}t^j}{(1-t)^n} \in \mathbb{Q}[[t]][t^{-1}].$$

This rational function has a pole of order $\dim M$ at $t = 1$, or equivalently, the numerator $\sum_{i,j}(-1)^i\beta_{i,j}t^j$ has a zero of order $\operatorname{codim} M$. In case of a pure module the numerator simplifies to

$$\sum_{i=0}^{r}(-1)^i\beta_{i,d_i}t^{d_i}.$$

Hence, the $r+1$ numbers $\beta_{0,d_0},\dots,\beta_{r,d_r}$ satisfy a system of r linear equations

$$\sum_{i=0}^{r}\beta_{i,d_i}d_i^s = 0 \text{ for } s = 0,\dots r-1$$

of Vandermond type. Hence by Cramer's rule,

$$\beta_{i,d_i} = q \prod_{t>s,\ t,s\neq i}(d_t - d_s)$$

are determined by the type up to a common rational factor q.

□

With $\beta(d)$ we denote the rational table on the ray of type d normalized such that

$$\beta_{i,d_i}(d) = \prod_{j\neq i}\frac{1}{|d_j - d_i|}.$$

To formulate our main result one more preparation is necessary: we order the strictly increasing sequences d as follows:

$$d = (d_0,\dots,d_r) \le d' = (d'_0,\dots,d'_{r'})$$

if $r \ge r'$ and $d_i \le d'_i$ for $i = 1,\dots,r'$. One can think of this as the termwise order if one simply extends each sequence $d = (d_0,\dots,d_r)$ to $(d_0,\dots,d_r,\infty,\infty,\dots)$.

We can now state the main result of the theory concerning the cone of Betti tables:

Theorem 2.2 ([2, 11]). *Let $S = k[x_1,\dots,x_n]$ be as above.*

1. *For every strictly increasing sequence of integers $d = (d_0,\dots,d_r)$ with $r \le n$, there exist pure S-modules of type d.*
2. *The Betti table of any finitely generated graded S-module may be written uniquely as a positive rational linear combination of the Betti tables of a set of pure modules whose types form a totally ordered sequence.*

The second statement of the theorem has two nice interpretations that may help to clarify its meaning. First, geometrically, it really says that the cone of Betti tables is a *simplicial fan*, that is, it is the union of simplicial cones, meeting along facets. The maximal simplicial cones in the fan correspond to maximal chains (totally ordered subsets) in the partially ordered set of degree sequences: the simplicial cone is the set of finite positive rational combinations of Betti tables whose degree sequences lie in the chain. These simplices and cones are thus infinite ascending unions of finite-dimensional cones, corresponding to Betti tables with finite support, that is, resolutions where the free modules are generated in a given bounded range of degrees.

Second, algorithmically, the theorem implies that there is a greedy algorithm that gives the decomposition. Rather than trying to specify this formally, we give an example with $n = 3$. To describe it compactly, we will write the Betti table of a module M as an array whose entries in the i-th column are the $\beta_{i,j}$—that is, the i-th column corresponds to the free module F_i. For reasons of efficiency and tradition, we put $\beta_{i,j}$ in the $(j-i)$-th row.

Consider the $K[x,y,z]$-module $M = S/(x^2, xy, xz^2)$. The minimal free resolution of M has the form

$$S \leftarrow S(-2)^2 \oplus S(-3) \leftarrow S(-3) \oplus S(-4)^2 \leftarrow S(-5) \leftarrow 0$$

and is represented by an array

$$\beta(M) = \begin{pmatrix} 1 & & & \\ & 2 & 1 & \\ & 1 & 2 & 1 \end{pmatrix}$$

where all the entries not shown are equal to zero.

To write this as a positive rational linear combination of pure diagrams, we first consider the "top row", corresponding to the generators of lowest degree in the free modules of the resolution. These are in the positions

$$\begin{pmatrix} * & & & \\ & * & * & \\ & & & * \end{pmatrix}$$

corresponding to the degree sequence $(0, 2, 3, 5)$. There is in fact a pure module $M_1 = S/I_1$ with resolution

$$\beta(M_1) = \begin{pmatrix} 1 & & & \\ & 5 & 5 & \\ & & & 1 \end{pmatrix}.$$

The greedy algorithm now instructs us to subtract *the largest possible q_1 that will leave the resulting table $\beta(M) - q_1\beta(M_1)$ having only non-negative terms.* We see at once that $q_1 = 1/5$.

We now repeat this process starting from $\beta(M) - q_1\beta(M_1)$; the theorem guarantees that there will always be a pure resolution whose degree sequence matches the top row of the successive remainders. In this case we arrive at the expression

$$\beta(M) = \begin{pmatrix} 1 & & & \\ & 2 & 1 & \\ & 1 & 2 & 1 \end{pmatrix} = 1/5 \begin{pmatrix} 1 & & & \\ & 5 & 5 & \\ & & & 1 \end{pmatrix} + 1/10 \begin{pmatrix} 3 & & & \\ & 10 & & \\ & & 15 & 8 \end{pmatrix}$$
$$+1/6 \begin{pmatrix} 1 & & & \\ & & & \\ & 4 & 3 & \end{pmatrix} + 1/3 \begin{pmatrix} 1 & & & \\ & & & \\ & 1 & & \end{pmatrix}.$$

All the fractions and tables that occur are of course invariants—apparently new invariants—of M.

3. Facets of the Cone and Cohomology Tables

We next focus on the facets of the cone of Betti tables (compare with [1, 2]). Consider a finite chain of degree sequences. Since the rays corresponding to the degree sequences are linearly independent, these rays generate a simplicial cone. A facet (maximal face) of this cone is generated by all but one of the rays in our chain. If this ray corresponds to the degree sequence b, then we may assume that b is neither the largest nor the smallest degree sequence by replacing our chain with a longer chain of degree sequences. Consider the degree sequences $a > b$ and $c < b$ immediately above and below in our chain. By inserting a finite number of degree sequences between a and b if necessary, we can achieve that a and b differ in at most one position. Similarly, we can achieve that b and c differ in at most one position. Then the facet of this simplicial cone obtained by deleting b is an *outer face*—that is, it will lie on the boundary of the cone of Betti tables—if either a and c differ in precisely one position τ, or a and c differ in precisely two consecutive positions τ and $\tau+1$ and $a_\tau \geq c_{\tau+1}$, compare [1], Proposition 2.2. Indeed, suppose that a and c differ in the position τ and k with $k > \tau$ and, moreover, $k > \tau + 1$ or $k = \tau + 1$ and $a_\tau < c_{\tau+1}$. Then both sequences $b = (\ldots, a_\tau, \ldots, c_k, \ldots)$ and $b' = (\ldots, c_\tau, \ldots, a_k, \ldots)$ are increasing, hence valid degree sequences between with a and c. For $a_k < \infty$ we have the numerical identity

$$(a_k - a_\tau)\beta(a) + (c_k - c_\tau)\beta(c) = (c_k - a_\tau)\beta(b) + (a_k - c_\tau)\beta(b'),$$

which becomes $\beta(a) + (c_k - c_\tau)\beta(c) = (c_k - a\tau)\beta(b) + \beta(b')$, in case $a_k = \infty$. Hence, once we know that pure modules for arbitrary degree sequences exist, we can deduce that the facet of the simplicial cone obtained by dropping b (or b') lies in the interior of the cone of Betti tables.

Hence, apart from the existence of pure resolution, we have to show that every potential outer face as above is indeed an outer face. A typical example which could lead to an outer face is the chain

$$a = (0,3,4) > b = (0,2,4) > c = (0,1,4)$$

for the case that a and c differ in only one position and

$$a = (0,2,3,4) > b = (0,1,3,4) > c = (0,1,2,4)$$

in case a and c differ in two positions. In the first case, the linear function $\beta \mapsto \beta_{\tau,b_\tau}$ is positive on the ray corresponding to b and vanishes on all other rays in any simplex corresponding to a chain of degree sequence containing a, b, c. Clearly, this functional is also non-negative on Betti tables of arbitrary module. So these are indeed outer faces.

The second case is more complicated. We start by replacing a, b and c by $(\ldots, a_{\tau-1}, c_{\tau+1}, c_{\tau+1} + 1, \ldots), (\ldots, a_{\tau-1}, c_{\tau+1} - 1, c_{\tau+1} + 1, \ldots)$ and $(\ldots, a_{\tau-1}, c_{\tau+1} - 1, c_{\tau+1}, \ldots)$. For example, we will replace the triple

$$(0,\infty,\infty) > (0,1,\infty) > (0,1,3)$$

by

$$(0,3,4) > (0,2,4) > (0,2,3).$$

We will see that the equation, which we will derive below in this new situation, works for the face obtained by deleting the original b as well.

Now take a complete chain extending $a > b > c$. We can compute the coefficient $\delta_{i,j}$ of a functional $\delta : \beta \mapsto \sum_{ij} \delta_{i,j}\beta_{i,j}$ vanishing on the facet opposite to b recursively. We start by taking $\delta_{i,a_i} = 0$ and work our way up and down in the chain. The condition $\delta(\beta(c)) = 0$ determines the ratio of δ_{τ,c_τ} while $\delta_{\tau+1,c_{\tau+1}}$ and $\delta(\beta(b)) > 0$ determines the sign. Moving one step down from c in the chain, determines one more coefficient of δ. In the example above we can look at the degree sequence

$$(0,1,3,4) > (0,1,2,4) > (0,1,2,3) = (0,1,2,3,\infty) > \ldots > (0,1,2,3,6) >$$

$$(0,1,2,3,5) > (0,1,2,3,4) > (-1,1,2,3,4) > \ldots$$

whose Betti tables are

$$\begin{pmatrix} 2 & 4 & & \\ & & 4 & 2 \\ & & & \end{pmatrix} \begin{pmatrix} 3 & \mathbf{8} & \mathbf{6} & & \\ & & & 1 & \\ & & & & \end{pmatrix} \begin{pmatrix} 1 & 3 & 3 & \mathbf{1} & \\ & & & & \\ & & & & \end{pmatrix} \ldots \begin{pmatrix} 10 & 36 & 45 & 20 & \\ & & & & \\ & & & & \mathbf{1} \end{pmatrix}$$

$$\begin{pmatrix} 4 & 15 & 20 & 10 & \\ & & & & \mathbf{1} \\ & & & & \end{pmatrix} \begin{pmatrix} 1 & 4 & 6 & 4 & \mathbf{1} \\ & & & & \\ & & & & \end{pmatrix} \begin{pmatrix} \mathbf{1} & & & & \\ & 10 & 20 & 15 & 4 \\ & & & & \end{pmatrix} \ldots$$

and obtain

$$\delta = (\delta_{i,j}) = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ 21 & -12 & 5 & 0 & -3 \\ 12 & -5 & 0 & 3 & -4 \\ 5 & 0 & -3 & 4 & -3 \\ \mathbf{0} & 3 & -4 & 3 & 0 \\ 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & 5 \\ 0 & 0 & 0 & 0 & 12 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Of course, moving up in degree from a will always yield zero coefficients. Note, that the results of these computations apparently do not depend on the specific choice of the complete chain extending $a > b > c$.

To prove Theorem 2.2, we have to show that each such δ is nonnegative on the Betti table $\beta(M)$ of an arbitrary module. Our key observation is that the numbers appearing are dimensions of cohomology groups of what we call supernatural vector bundles on $\mathbb{P}^{r-1}$.

Definition 2. A vector bundle $\mathcal{E}$ on $\mathbb{P}^m$ has *natural cohomology* [16] if for each k at most one of the groups

$$H^i(\mathcal{E}(k)) \neq 0.$$

It has *supernatural cohomology* if in addition the Hilbert polynomial

$$\chi(\mathcal{E}(k)) = \frac{\text{rank } \mathcal{E}}{m!} \prod_{j=1}^{m} (k - z_j)$$

has m distinct integral roots $z_1 > z_2 > \ldots > z_m$.

Note that a supernatural vector bundle $\mathcal{E}$ has non-vanishing cohomology in the following range (see [11]):

$$\begin{cases} H^0(\mathcal{E}(k)) \neq 0 \\ H^i(\mathcal{E}(k)) \neq 0 \\ H^m(\mathcal{E}(k)) \neq 0 \end{cases} \text{ if and only if } \begin{cases} k > z_1 \\ z_i > k > z_{i+1} \\ z_m > k \end{cases}.$$

For a coherent sheaf $\mathcal{E}$ on $\mathbb{P}^m$ we denote by

$$\gamma(\mathcal{E}) = (\gamma_{j,k}) \in \bigoplus_{j=0}^{m} \prod_{k \in \mathbb{Z}} \mathbb{Q} \text{ with } \gamma_{j,k} = h^j(\mathcal{E}(k))$$

its cohomology table. Analogous to the theorem on free resolutions we have

Theorem 3.1 ([11]). *The extremal rays of the rational cone of cohomology tables of vector bundles on $\mathbb{P}^m$ are generated by cohomology tables of supernatural vector bundles.*

More precisely: Every cohomology table of a vector bundle is a unique positive rational combination of cohomology tables of supernatural vector bundles, whose root sequences form a chain.

Here we order the root sequences component wise.

The crucial new concept is the following pairing between Betti tables of modules and cohomology tables of coherent sheaves. We define $\langle \beta, \gamma \rangle$ for a Betti table $\beta = (\beta_{i,k})$ and a cohomology table $\gamma = (\gamma_{j,k})$ by

$$\langle \beta, \gamma \rangle = \sum_{i \geq j} (-1)^{i-j} \sum_k \beta_{i,k} \gamma_{j,-k}$$

Theorem 3.2 (Positivity 1,[11, 12]). *For $\mathbf{F}$ any free resolution of a finitely generated graded $K[x_0, \ldots, x_m]$-module M and $\mathcal{E}$ any coherent sheaf on $\mathbb{P}^m$, we have*

$$\langle \beta(\mathbf{F}), \gamma(\mathcal{E}) \rangle \geq 0.$$

Moreover, if M has finite length and $H^{i+1}(\widetilde{F}_i \otimes \mathcal{E}) = 0$ for all $i \geq 0$, then

$$\langle \beta(\mathbf{F}), \gamma(\mathcal{E}) \rangle = 0.$$

Note that if $\widetilde{F}_i = \oplus_{j \in \mathbb{Z}} \mathcal{O}(-j)^{\beta_{i,j}}$ then

$$\langle \beta(\mathbf{F}), \gamma(\mathcal{E}) \rangle = \sum_{i \geq j} (-1)^{i-j} h^j(\widetilde{F}_i \otimes \mathcal{E})$$

Proof. We first treat the case where $\mathcal{E}$ is a vector bundle. In this case we have an exact complex

$$0 \leftarrow \mathcal{M}_0 \leftarrow \widetilde{F}_0 \otimes \mathcal{E} \rightarrow \widetilde{F}_1 \otimes \mathcal{E} \leftarrow \ldots \leftarrow \widetilde{F}_r \otimes \mathcal{E} \leftarrow 0$$

with $\mathcal{M}_0 = \widetilde{M} \otimes \mathcal{E}$. Breaking it up in short exact sequences

$$\begin{array}{ccccccccc}
0 & \leftarrow & \mathcal{M}_0 & \leftarrow & \widetilde{F}_0 \otimes \mathcal{E} & \leftarrow & \mathcal{M}_1 & \leftarrow & 0 \\
0 & \leftarrow & \mathcal{M}_1 & \leftarrow & \widetilde{F}_1 \otimes \mathcal{E} & \leftarrow & \mathcal{M}_2 & \leftarrow & 0 \\
0 & \leftarrow & \mathcal{M}_2 & \leftarrow & \widetilde{F}_2 \otimes \mathcal{E} & \leftarrow & \mathcal{M}_3 & \leftarrow & 0 \\
 & & & & \vdots & & & &
\end{array}$$

we get the desired functional by taking the alternating sum of the Euler characteristics of initial parts of the corresponding long exact sequences in coho-

mology:

$$\begin{array}{ccccccc}
 & & H^0(\widetilde{F}_0 \otimes \mathcal{E}) & \leftarrow & H^0(\mathcal{M}_1) & \leftarrow & 0 \\
 & & & & & & \\
 & & H^1(\widetilde{F}_1 \otimes \mathcal{E}) & \leftarrow & H^1(\mathcal{M}_2) & \leftarrow & \\
H^0(\mathcal{M}_1) & \leftarrow & H^0(\widetilde{F}_1 \otimes \mathcal{E}) & \leftarrow & H^0(\mathcal{M}_2) & \leftarrow & 0 \\
 & & & & & & \\
 & & H^2(\widetilde{F}_2 \otimes \mathcal{E}) & \leftarrow & H^2(\mathcal{M}_3) & \leftarrow & \\
H^1(\mathcal{M}_2) & \leftarrow & H^1(\widetilde{F}_2 \otimes \mathcal{E}) & \leftarrow & H^1(\mathcal{M}_3) & \leftarrow & \\
H^0(\mathcal{M}_2) & \leftarrow & H^0(\widetilde{F}_2 \otimes \mathcal{E}) & \leftarrow & H^0(\mathcal{M}_3) & \leftarrow & 0 \\
 & & \vdots & & & &
\end{array}$$

Hence, $\langle \beta(\mathbf{F}), \gamma(\mathcal{E}) \rangle = \sum_{j=0}^{m} \dim \operatorname{coker} H^j(\mathcal{M}_{j+1}) \to H^j(\widetilde{F}_j \otimes \mathcal{E})) \geq 0$.

In the general case, where $\mathcal{E}$ is not necessarily locally free, the complex at the beginning of the proof may not be exact. However, we note that what we need to prove depends only on the cohomology table of $\mathcal{E}$, not on the sheaf itself. Hence, we can replace $\mathcal{E}$ with a translate $g^*\mathcal{E}$ for any $g \in PGL(m+1)$. When g is a general element, [23] shows that the sheaves $Tor_i(\tilde{M}, \mathcal{E}) = 0$ for $i > 0$; that is, the complex

$$0 \leftarrow \tilde{M} \otimes g^*\mathcal{E} \leftarrow \widetilde{F}_0 \otimes g^*\mathcal{E} \to \widetilde{F}_1 \otimes \mathcal{E} \leftarrow \ldots \leftarrow \widetilde{F}_r \otimes g^*\mathcal{E} \leftarrow 0$$

is exact, and the same argument applies.

For the vanishing statement, we note that in this case $\widetilde{\mathbf{F}} \otimes \mathcal{E}$ is exact as well, and $\mathcal{M}_0 = 0$. By induction we obtain $H^i(\mathcal{M}_i) = 0$ from $H^i(\widetilde{F}_{i-1} \otimes \mathcal{E}) = 0$, and all the Euler characteristics are in fact zero. $\square$

The facet equation in the example above is obtained from the vector bundle $\mathcal{E}$ on $\mathbb{P}^2 \overset{\iota}{\hookrightarrow} \mathbb{P}^3$, that is the kernel of a general map $\mathcal{O}^5_{\mathbb{P}^2}(-1) \to \mathcal{O}^3_{\mathbb{P}^2}$. The coefficients of the functional $\langle -, \gamma(\iota_* \mathcal{E}) \rangle$ are

$$\begin{pmatrix}
\vdots & \vdots & \vdots & \vdots & \vdots \\
21 & -12 & 5 & 0 & -3 \\
12 & -5 & 0 & 3 & -4 \\
5 & 0 & -3 & 4 & -3 \\
0 & 3 & -4 & 3 & 0 \\
0 & 4 & -3 & 0 & 5 \\
0 & 3 & 0 & -5 & 12 \\
0 & 0 & 5 & -12 & 21 \\
0 & 0 & 12 & -21 & 32 \\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}$$

This is not quite the functional we wanted, which had zeros in place of some of the nonzero values. To correct this, we define "truncated" functionals $\langle -, \gamma \rangle_{\tau,\kappa}$

by putting zero coefficients in the appropriate spots:

$$\begin{aligned}\langle\beta,\gamma\rangle_{\tau,\kappa} \quad = \quad & \textstyle\sum_{k\le\kappa}\beta_{\tau,k}\gamma_{\tau,-k}+\sum_{j<\tau}\sum_k\beta_{j,k}\gamma_{j,-k}\\ & -\textstyle\sum_{k\le\kappa+1}\beta_{\tau+1,k}\gamma_{\tau,-k}-\sum_{j<\tau}\sum_k\beta_{j+1,k}\gamma_{j,-k}\\ & +\textstyle\sum_{i>j+1}(-1)^{i-j}\sum_k\beta_{i,k}\gamma_{j,-k}\end{aligned}$$

Theorem 3.3 (Positivity 2, [11, 12]). *For* $\mathbf{F}$ *the* <u>*minimal*</u> *free resolution of a finitely generated graded* $K[x_0,\dots,x_m]$*-module and* $\mathcal{E}$ *any coherent sheaf on* $\mathbb{P}^m$*, we have*

$$\langle\beta(\mathbf{F}),\gamma(\mathcal{E})\rangle_{\tau,\kappa}\ge 0.$$

Proof. We replace $\mathcal{E}$ by a general translate as above, to achieve homological transversaltity to $\mathbf{F}$. Let E be a graded module representing the sheaf $\mathcal{E}$. Consider the Čech resolution $0\to C^0\to C^1\to C^2\to\dots$ of E with $C^p=\oplus_{i_0<\dots,i_p}E[x_{i_0}^{-1},\dots,x_{i_p}^{-1}]$ and the tensor product

$$\begin{array}{ccccccccc}
 & & \vdots & & \vdots & & \vdots & & \\
 & & \uparrow & & \uparrow & & \uparrow & & \\
0 & \leftarrow & F_0\otimes C^2 & \leftarrow & F_1\otimes C^2 & \leftarrow & F_2\otimes C^2 & \leftarrow & \dots\\
 & & \uparrow & & \uparrow & & \uparrow & & \\
0 & \leftarrow & F_0\otimes C^1 & \leftarrow & F_1\otimes C^1 & \leftarrow & F_2\otimes C^1 & \leftarrow & \dots\\
 & & \uparrow & & \uparrow & & \uparrow & & \\
0 & \leftarrow & F_0\otimes C^0 & \leftarrow & F_1\otimes C^0 & \leftarrow & F_2\otimes C^0 & \leftarrow & \dots\\
 & & \uparrow & & \uparrow & & \uparrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

of complexes.

By the homological transversality the horizontal cohomology is concentrated in the F_0-column. Hence, the total complex has homology only in non-negative cohomological degrees. The vertical cohomology in internal degree 0 on the diagonal or below are the groups

$$\begin{array}{ccc}
 & & H^2(\widetilde{F}_2\otimes\mathcal{E})\\
 & H^1(\widetilde{F}_1\otimes\mathcal{E}) & H^1(\widetilde{F}_2\otimes\mathcal{E})\\
H^0(\widetilde{F}_0\otimes\mathcal{E}) & H^0(\widetilde{F}_1\otimes\mathcal{E}) & H^0(\widetilde{F}_2\otimes\mathcal{E})
\end{array}$$

The Euler characteristic of this diagram is again $\langle\beta(\mathbf{F}),\gamma(\mathcal{E})\rangle$. If we split the internal degree 0 part of the spectral sequence $H_{vert}(C\otimes F)\Rightarrow H_{tot}(C\otimes F)$ as a sequence of K-vector spaces, then we obtain a complex

$$\dots\leftarrow A_p=\bigoplus_{i-j=p}H^j(\widetilde{F}_i\otimes\mathcal{E})\quad\leftarrow\bigoplus_{i-j=p+1}A_{p+1}=H^j(\widetilde{F}_i\otimes\mathcal{E})\leftarrow\dots$$

that is exact in negative cohomological degrees. Note that this gives a different proof (essentially our original proof) of part of the positivity result of Theorem 3.2, since $\langle\beta(\mathbf{F}),\gamma(\mathcal{E})\rangle=\dim\operatorname{coker}(A_1\to A_0)\ge 0$.

Consider the submodules

$$B_0 = \bigoplus_{j<\tau} H^j(\widetilde{F}_j \otimes \mathcal{E}) \oplus \bigoplus_{k\leq\kappa} H^\tau(\mathcal{O}(-k)^{\beta_{\tau,k}} \otimes \mathcal{E}) \subset A_0$$

and

$$B_1 = \bigoplus_{j<\tau} H^j(\widetilde{F}_{j+1} \otimes \mathcal{E}) \oplus \bigoplus_{k\leq\kappa+1} H^\tau(\mathcal{O}(-k)^{\beta_{\tau+1,k}} \otimes \mathcal{E}) \subset A_1$$

corresponding to the truncation. The diagram

$$\begin{array}{ccc} A_0 & \leftarrow & A_1 \\ \uparrow & & \uparrow \\ B_0 & \leftarrow & B_1 \end{array}$$

commutes, because $\mathbf{F}$ is minimal. Hence,

$$\begin{array}{rcl} \langle\beta(\mathbf{F}),\gamma(\mathcal{E})\rangle_{\tau,\kappa} & = & \langle\beta(\mathbf{F}),\gamma(\mathcal{E})\rangle - \dim A_0 + \dim B_0 + \dim A_1 - \dim B_1 \\ & = & \dim\ker(A_1 \to A_0) + \dim\operatorname{coker}(B_1 \to B_0) \\ & & - \dim\ker(B_1 \to B_0) \geq 0, \end{array}$$

because $\ker(B_1 \to B_0) \subset \ker(A_1 \to A_0)$. □

These stronger versions of the vanishing results of [11] allow us to give a direct proof of the main theorem of [2]:

Final part of the Proof of Theorem 2.2. The facet equation, which cuts out the desired face corresponding to a degree sequence $a > b > c$ with c of length r that only differ in positions τ and $\tau+1 \leq r$ and satisfies $a_\tau \geq c_{\tau+1}$, is given by taking a supernatural vector bundle $\mathcal{E}$ on $\mathbb{P}^{r-1} \subset \mathbb{P}^{n-1}$ with root sequence $(z_1 > z_2 > \ldots > z_{r-1}) = (-b_0 > \ldots > -b_{\tau-1} > -b_{\tau+2} > \ldots > -b_r)$, $\kappa = c_{\tau+1} - 1$ and the functional

$$\langle -,\gamma(\mathcal{E})\rangle_{\tau,\kappa}.$$

Indeed, for a pure module M with a degree sequence $d \leq c$, we have

$$\langle\beta(M),\gamma(\mathcal{E})\rangle_{\tau,\kappa} = \langle\beta(M),\gamma(\mathcal{E})\rangle,$$

and the vanishing follows from Theorem 3.2: Since the length of d is at least the length of c we can reduce to the case where M has finite length, because the Betti numbers of M and M/xM as an S/xS-module for a linear nonzero divisor x of M coincide. Furthermore, $H^{i+1}(\mathcal{E}(-d_i)) = 0$ because $-d_i \geq -c_i \geq z_{i+1}$ or $i+1 \geq r$.

The vanishing $\langle\beta(M),\gamma(\mathcal{E})\rangle_{\tau,\kappa} = 0$ for all pure M with a degree sequence $d \geq a$ is trivially true by our choice of $\mathcal{E}$ and the truncation.

Finally, $\langle\beta(b),\gamma(\mathcal{E})\rangle_{\tau,\kappa} > 0$, because $z_\tau = -b_{\tau-1} > -b_\tau = -c_\tau \geq -\kappa = -c_{\tau+1} + 1 > -c_{\tau+1} > -c_{\tau+2} = -b_{\tau+2} = z_{\tau+1}$ and hence $H^\tau(\mathcal{E}(-b_\tau)) \neq 0$. Thus $\langle -,\gamma(\mathcal{E})\rangle_{\tau,\kappa} = 0$ cuts out the desired face. □

Conversely, the essential facet equations of the cone of cohomology tables of vector bundles are of type $\langle \mathbf{F}(M), -\rangle_{\tau,\kappa}$ for an appropriate finite length pure module M, see [11].

4. Existence

To complete the proof of both Boij-Söderberg decompositions, it is now enough to establish the existence of supernatural vector bundles and pure resolutions for arbitrary root or degree sequences. In each case there are two methods known. For equivariant resolutions or homogeneous vector bundles in characteristic 0 one can use Schur functors [9, 10, 24]. For arbitrary fields, one can use a push down method [11]. For bundles this is a simple application of the Künneth formula applied to $\mathcal{E} = \pi_*\mathcal{O}(a_1, \ldots, a_m)$, where π is a finite linear projection $\pi : \mathbb{P}^1 \times \ldots \times \mathbb{P}^1 \to \mathbb{P}^m$ and $\mathcal{O}(a_1, \ldots, a_m)$ is a suitable line bundle on the product.

For resolutions this is an iteration of the Lascoux method [22] to get the Buchsbaum-Eisenbud family of complexes associated to generic matrices [4]: We start with $\mathcal{K}$, a Koszul complex on $\mathbb{P}^{r-1} \times \mathbb{P}^{m_1} \times \ldots \times \mathbb{P}^{m_s}$ of $r + \sum_{i=1}^{s} m_i$ forms of multidegree $(1, \ldots, 1)$ tensored with $\mathcal{O}(-d_0, a_1, \ldots, a_s)$. Here s is the number of desired non-linear maps and m_j+1 is the desired degree of the j-th non-linear map. The spectral sequence for $R\pi_*\mathcal{K}$ of the projection $\pi : \mathbb{P}^{r-1} \times \mathbb{P}^{m_1} \times \ldots \times \mathbb{P}^{m_s} \to \mathbb{P}^{r-1}$ gives rise to the desired complex if we choose $a_1, \ldots, a_s$ suitably. Indeed, we may apply Proposition 4.1 s-times: For any product $X_1 \times X_2$ with projections $p : X_1 \times X_2 \to X_1$ and $q : X_1 \times X_2 \to X_2$ and sheaves $\mathcal{L}_i$ on X_i, we set

$$\mathcal{L}_1 \boxtimes \mathcal{L}_2 := p^*\mathcal{L}_1 \otimes q^*\mathcal{L}_2.$$

Proposition 4.1. *Let $\mathcal{F}$ be a sheaf on $X \times \mathbb{P}^m$, and let $p : X \times \mathbb{P}^m \to X$ be the projection. Suppose that $\mathcal{F}$ has a resolution of the form*

$$0 \to \mathcal{G}_N \boxtimes \mathcal{O}(-e_N) \to \cdots \to \mathcal{G}_0 \boxtimes \mathcal{O}(-e_0) \to \mathcal{F} \to 0$$

with degrees $e_0 < \cdots < e_N$. If this sequence contains the subsequence $(e_{k+1}, \ldots, e_{k+m})$
$= (1, 2, \ldots, m)$ for some $k \geq 0$ then

$$R^\ell p_*\mathcal{F} = 0 \text{ for } \ell > 0$$

and $p_*\mathcal{F}$ has a resolution on X of the form

$$\begin{aligned} 0 \to \mathcal{G}_N \otimes H^m\mathcal{O}(-e_N) \to \cdots \\ \to \mathcal{G}_{k+m+1} \otimes H^m\mathcal{O}(-e_{k+m+1}) \xrightarrow{\phi} \mathcal{G}_k \otimes H^0\mathcal{O}(-e_k) \to \\ \cdots \to \mathcal{G}_0 \otimes H^0\mathcal{O}(-e_0) \end{aligned} \quad (1)$$

Proof. From the numerical hypotheses we see that $e_i \leq 0$ for $i \leq k$ and $e_i \geq m+1$ for $i \geq k+m+1$. Consider the spectral sequence

$$E_1^{i,-j} = R^i p_*(\mathcal{G}_j \boxtimes \mathcal{O}(-e_j)) \Rightarrow R^{i-j} p_* \mathcal{F}.$$

By the projection formula, the terms of the E_1 page are

$$R^i p_*(\mathcal{G}_j \boxtimes \mathcal{O}(-e_j)) = \begin{cases} \mathcal{G}_j \otimes H^m(\mathbb{P}^m, \mathcal{O}(-e_j)) & \text{if } j \geq k+m+1 \text{ and } i = m \\ \mathcal{G}_j \otimes H^0(\mathbb{P}^m, \mathcal{O}(-e_j)) & \text{if } j \leq k \text{ and } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the spectral sequence degenerates to the complex (1), where ϕ is a differential from the m-th page and the other maps are differentials from the first page. In particular, only terms $E_\infty^{i,-j}$ with $i \leq j$ can be nonzero. On the other hand, the terms $R^{i-j} p_* \mathcal{F}$ can be nonzero only for $i \geq j$. Hence, the complex (1) is exact and resolves $\oplus_{i \geq 0} E_\infty^{i,-i} = E_\infty^{0,0} = p_* \mathcal{F}$, while the higher direct images of $\mathcal{F}$ vanish. □

5. Applications, Extensions of the Basic Theory and Open Questions

The application that motivated Boij and Söderberg to make their Conjecture, was the following sharp version of the Multiplicity Conjecture of Huneke and Srinivasan [18].

Theorem 5.1 ([2]). *For any finitely generated module M of projective dimension r and codimension s generated in degree 0, we have the following bounds for the Hilbert series:*

$$\left(\prod_{i=1}^{r} a_i\right) H(\beta(a), t) \leq \frac{H(M,t)}{\beta_{0,0}(M)} \leq \left(\prod_{i=1}^{s} b_i\right) H(\beta(b), t) \ ,$$

where $a = (0, a_1, a_2, ..., a_r)$ are the minimal shifts and $b = (0, b_1, b_2, ..., b_s)$ are the maximal shifts in a minimal free resolution of M. Equality on either side implies that the resolution is pure. In particular, the right hand inequality implies the Multiplicity Conjecture, that is, the multiplicity of M is bounded by

$$\text{mult}\,(M) \leq \beta_{0,0}(M)\, \frac{b_1 \cdot \ldots \cdot b_s}{s!}$$

with equality if and only if M is Cohen-Macaulay with a pure resolution.

Sketch. If a is a degree sequence of length r with $a_0 = 0$, then the Betti table $(\prod_{i=1}^{r} a_i)\, \beta(a)$ is normalized such that $\beta_{0,0} = 1$. Given two degree sequences $a < b$ with $a_0 = b_0 = 0$, then the Hilbert series of the normalized tables satisfy

$$H\left(\left(\prod_{i=1}^{r} a_i\right) \beta(a), t\right) < H\left(\left(\prod_{i=1}^{s} b_i\right) \beta(b), t\right) .$$

The result follows because the normalized Boij-Söderberg decomposition is a convex combination. □

Let $a < b$ be two degree sequences of equal length. The part of the Boij-Söderberg cone of tables $\beta = (\beta_{i,j})$ with $\beta_{i,j} = 0$ unless $a_i \leq j \leq b_i$ is a finite, equi-dimensional simplical fan.

Turning to the monoid of Betti tables of modules, we have:

Theorem 5.2 (Erman, [14]). *The monoid of Betti tables of Cohen-Macaulay modules with Betti tables bounded by the degree sequences $a < b$ is finitely generated.*

Note that the index of actual Betti tables among the integral points on a ray of the Boij-Söderberg cone can be arbitrary large [7]. However, along the extremal rays, Eisenbud and Weyman conjecture that any sufficiently large integral point is the Betti table of a module.

To understand the monoid is substantially more difficult than understanding the cone. For example, unlike the Boij-Söderberg cone the monoid of Betti tables depends on the characteristic of the ground field [21]. A case where we understand the monoid completely can be found in [7].

We believe that the most important next step in trying to understand the monoid better, would be a proof of the Eisenbud-Buchsbaum-Horrocks rank conjecture [5]. Daniel Erman [15] uses the Boij-Söderberg decomposition to prove the rank conjecture for $M = S/I$ a cyclic module, provided that the minimal generators of the ideal I are sufficiently large compared to the regularity.

Turning to coherent sheaves, it is no longer true that the cohomology table of any sheaf is a finite sum of tables of supernatural sheaves of various dimensions. What remains true is that a cohomology table of an arbitrary coherent sheaf on $\mathbb{P}^m$ is an infinite (convergent) sum with non-negative coefficients of cohomology tables of supernatural sheaves, whose zero sequences form a chain, see [12]. This result, however, does not characterize the Boij-Söderberg cone of coherent sheaves on $\mathbb{P}^m$ but only its closure in $\bigoplus_{i=0}^m \prod_{j\in\mathbb{Z}} \mathbb{R}$.

If $X \subset \mathbb{P}^m$ is a subvariety of dimension d we can ask about the Boij-Söderberg cone of its coherent sheaves. Consider a linear Noether normalization $\pi : X \to \mathbb{P}^d$. Then, for any sheaf $\mathcal{F}$ on X, the cohomology table of $\mathcal{F}$ and $\pi_*\mathcal{F}$ coincide. Hence the Boij-Söderberg cone of $(X, \mathcal{O}_X(1))$ is a subcone of the one of $(\mathbb{P}^d, \mathcal{O}(1))$. If they coincide then there exists a sheaf $\mathcal{U}$ on X, whose cohomology table coincides with that of $\mathcal{O}_{\mathbb{P}^d}$ up to a multiple. By Horrocks criterion [19], this implies $\pi_*\mathcal{U} \cong \mathcal{O}^r_{\mathbb{P}^d}$. By the very definition [10], this means that $\mathcal{U}$ is an Ulrich sheaf on X.

Conversely, if an Ulrich sheaf exists then for a sheaf $\mathcal{G}$ on $\mathbb{P}^d$, the cohomology table of $\mathcal{U} \otimes \pi^*\mathcal{G}$ and $\mathcal{G}$ coincide up to the factor r. This proves

Theorem 5.3 ([13]). *The Boij-Söderberg cone of the coherent sheaves on a variety X of dimension d with respect to a very ample polarization $\mathcal{O}_X(1)$ coincides with the Boij-Söderberg cone of $(\mathbb{P}^d, \mathcal{O}(1))$ if and only if X carries an Ulrich sheaf.*

Varieties that have an Ulrich sheaf include curves and hypersurfaces. They are closed under Segre products, Veronese re-embeddings and transversal intersections. In [10] we conjecture that every variety has an Ulrich sheaf.

Very little is known for the extension of this theory to the multi-graded setting. I believe that there will be beautiful results ahead in this direction.

References

[1] M. Boij and J. Söderberg. Graded Betti numbers of Cohen-Macaulay modules and the multiplicity conjecture. J. Lond. Math. Soc. **78** (2008) 85–106.

[2] M. Boij and J. Söderberg. Betti numbers of graded modules and the Multiplicity Conjecture in the non-Cohen-Macaulay case. Preprint: arXiv:0803.1645.

[3] M. Boij and G. Fløystad. The cone of Betti diagrams of bigraded artinian modules of codimension two. Preprint: arXiv:1001.3238.

[4] D. Buchsbaum and D. Eisenbud. Generic free resolutions and a family of generically perfect ideals. Advances in Math. **18** (1975) 245–301.

[5] D. Buchsbaum and D. Eisenbud. Algebra structures for finite free resolutions, and some structure theorems for ideals of codimension 3. Amer. J. Math. **99** (1977), no. 3, 447485.

[6] D. Eisenbud. *Commutative algebra. With a view toward algebraic geometry.* Graduate Texts in Mathematics, 150. Springer-Verlag, New York, 1995. xvi+785 pp.

[7] D. Eisenbud, D. Erman and F.-O.Schreyer. Beyond Numerics: The Existence of Pure Filtrations. Preprint: arXiv:1001.0585

[8] D. Eisenbud, G. Fløystad and F.-O. Schreyer. Sheaf cohomology and free resolutions over exterior algebras. Trans. Amer. Math. Soc. **355** (2003) 4397–4426.

[9] D. Eisenbud, G. Fløystad and J. Weyman. The existence of pure free resolutions. Annales de l'Inst. Fourier. To appear. arXiv:0709.1529

[10] D. Eisenbud and F.-O. Schreyer with an Appendix by J. Weyman. Resultants and Chow forms via exterior syzygies. J. Amer. Math. Soc. **16** (2003), 537–579.

[11] D. Eisenbud and F.-O. Schreyer. Betti Numbers of Graded Modules and Cohomology of Vectos Bundles. J. Amer. Math. Soc. **22** (2009), 859–888.

[12] D. Eisenbud and F.-O. Schreyer. Cohomology of Coherent Sheaves and Series of Supernatural Bundles. To appear in Jour. Euro. Math. Soc. Preprint: arXiv:0902.1594

[13] D. Eisenbud and F.-O. Schreyer. Boij-Söderberg theory. To appear in: G. Fløystad, T. Johnsen and A.L. Knudson (editors). Combinatorial aspects of commutative algebra and algebraic geometry, Proceeding of the Abel Symposium, 2009.

[14] D. Erman. The Semigroup of Betti Diagrams. Algebra and Number Theory 3 (2009) 341–365.

[15] D. Erman. A special case of the Buchsbaum-Eisenbud-Horrocks rank conjecture. Preprint: arXiv:0902.0316

[16] R. Hartshorne and A. Hirschowitz. Cohomology of a general instanton bundle. Ann. Sci. de l'École Normale Sup. (1982) 365–390.

[17] J. Herzog and M. Kühl. On the Betti numbers of finite pure and linear resolutions. Comm. in Alg. **13** (1984) 1627–1646.

[18] J. Herzog and H. Srinivasan. Bounds for Multiplicities. Trans. Am. Math. Soc. (1998) 2879–2902.

[19] G. Horrocks. Vector bundles on the punctured spectrum of a local ring. Proc. London Math. Soc. (3) **14** (1964) 689–713.

[20] C. Huneke and M. Miller. A note on the multiplicity of Cohen-Macaulay algebras with pure resolutions. Canad. J. Math. 37 (1985), 1149–1162.

[21] M. Kunte. Gorenstein modules of finite length. Thesis, Universität des Saarlandes (2008). Preprint: arXiv:0807.2956

[22] A. Lascoux. Syzygies des variétés déterminantales. Adv. in Math. **30** (1978) 202–237.

[23] E. Miller and D. Speyer. A Kleiman-Bertini theorem for sheaf tensor products. J. Algebraic Geom. **17** (2008), 335–340.

[24] S. Sam, and J. Weyman. Pieri resolutions for classical groups. Preprint: arXiv:0907.4505

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Algebraic Cycles on Singular Varieties

Vasudevan Srinivas*

Abstract

We discuss algebraic cycles on singular varieties, in relation to the Grothendieck group of vector bundles. This theory, which is still not fully worked out, seems to admit some surprises. On the other hand, conjectured aspects of the refined structure of cycle groups of nonsingular varieties, predicted by motivic considerations, seem to have plausible extensions to singular varieties, which can be verified in some nontrivial examples.

Mathematics Subject Classification (2010). 14C17, 14C30, 14B05.

Keywords. Chow ring, singular varieties.

1. Introduction

The aim of this article is to discuss algebraic cycles on (quasi-)projective algebraic varieties, which may have singularities. Our goal is to see to what extent the theory of the Chow ring of a smooth variety can be generalized to the singular case, keeping in mind an expected relation with the Grothendieck group of vector bundles, and the analogy with the even degree singular cohomology ring in topology.

We will work over an algebraically closed ground field k. Eventually we will restrict to the case when k has characteristic 0.

Now suppose X is a non-singular projective variety. We first review properties of the Chow ring, as developed in Fulton's book [1]. We begin by recalling a definition.

Definition 1.

$Z^p(X) =$ Group of *algebraic cycles of codimension* p in X
$=$ Free abelian group on subvarieties of X which are irreducible
of codimension p.

*Supported by J.C. Bose Fellowship of the Department of Science and Technology, India.
School of Mathematics, Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai-400005, India. E-mail: srinivas@math.tifr.res.in.

It is standard to impose various *equivalence relations* on algebraic cycles, and pass to the quotient groups, to get more "reasonable" invariants; a basic one is *rational equivalence*.

Definition 2. The group $R^p(X) \subset Z^p(X)$ of cycles *rationally equivalent to 0* is generated by cycles of the form

$$\sum_i n_i A_i - \sum_j m_j B_j$$

where, for some suitable irreducible $W \subset X \times \mathbb{P}^1$ of codimension p, we have

$$\{A_1, A_2, \ldots\} = \text{irreducible components of } W \cap X \times \{0\},$$

$$\{B_1, B_2, \ldots\} = \text{irreducible components of } W \cap X \times \{\infty\}.$$

Here $0, \infty$ denote any two distinct points of $\mathbb{P}^1_k$; the coefficients n_i, m_j are certain *intersection multiplicities*.

We think of the cycles obtained from the intersections

$$W \cap X \times \{t\}, \quad \forall\ t \in \mathbb{P}^1_k,$$

as a *rational 1-parameter family* of codimension p algebraic cycles in X. Since $\operatorname{Aut} \mathbb{P}^1_k = \mathrm{PGL}_2(k)$ acts transitively on pairs of distinct points of $\mathbb{P}^1$, the equivalence relation above does not depend on which pair of points $\{0, \infty\} \subset \mathbb{P}^1_k$ is chosen, and all the cycles in such a rational 1-parameter family are rationally equivalent to each other.

Definition 3.

$$CH^p(X) = \frac{Z^p(X)}{R^p(X)}$$

$$= \text{Chow group of codim. } p \text{ cycles modulo rational equivalence.}$$

For example, if X is a curve, $CH^0(X) = \mathbb{Z}$, $CH^1(X) = \mathbb{Z} \oplus J(X)$, where $J(X)$ is the *Jacobian variety* of X, studied classically (Riemann, Abel, Jacobi ...).

One of the important results proved in [1] is the following.

Theorem 1.1. *For a nonsingular variety X,*

$$CH^*(X) = \bigoplus_{p=0}^{\dim X} CH^p(X)$$

has a multiplication, defined by intersecting suitable pairs of irreducible subvarieties of X, making it into a commutative graded ring, the Chow ring *of X.*

The Chow ring $CH^*(X)$ is an important and subtle invariant of the variety X, which is usually very difficult to compute. But some structural results are known about it, and there are important conjectures related to it, including the famous *Hodge Conjecture* (we return to this later).

A first reason to study the Chow ring is that it is an algebraic version of the *even degree cohomology ring* defined in algebraic topology. Thus, if X is a complex projective variety of dimension n, we have the ring

$$H^{2*}(X,\mathbb{Z}) = \bigoplus_{i=0}^{\dim X} H^{2i}(X,\mathbb{Z})$$

which has the following properties.

- It is a *commutative, graded ring*, which is additively a finitely generated abelian group. (If we also include the cohomology of odd degree, we obtain a $\mathbb{Z}/2\mathbb{Z}$-graded algebra over the even degree cohomology.)
- It is *contravariant functorial* for arbitrary (continuous) maps, and in fact for homotopy classes of maps.
- There are *Chern classes*

  $$c_i(V) \in H^{2i}(X,\mathbb{Z})$$

  for complex vector bundles V, which are *functorial under pull-backs*: if $f : Y \to X$ is a morphism of complex varieties, then

  $$c_i(V) \mapsto c_i(f^*V) \text{ under } f^* : H^{2i}(X,\mathbb{Z}) \to H^{2i}(Y,\mathbb{Z}).$$

  In fact, if $K_0^{\text{top}}(X)$ denotes the Grothendieck ring of (continuous) complex vector bundles on X, where the multiplication is induced by tensor products of bundles, then by a result of Atiyah and Hirzebruch, the *Chern character* gives an isomorphism of rings

  $$K_0^{\text{top}}(X) \otimes \mathbb{Q} \to H^{2*}(X,\mathbb{Q})$$

  The individual cohomologies $H^{2i}(X,\mathbb{Q})$ correspond to distinct *eigenspaces for Adams operations*, which are certain ring endomorphisms of $K_0^{\text{top}}(X)$ defined using the tensor and exterior power operations on vector bundles.
- (Cycle classes) If $Z \subset X$ is an irreducible subvariety of codimension p *which satisfies* $Z \cap X_{sing} = \emptyset$, then one finds that

  $$H^j(X, X \setminus Z, \mathbb{Z}) = \begin{cases} 0 & \text{if } i < 2p \\ \mathbb{Z} & \text{for } i = 2p \end{cases}$$

  (excision, together with the "Thom isomorphism theorem"). The image of 1 under the natural map

  $$\mathbb{Z} = H^{2p}(X, X \setminus Z, \mathbb{Z}) \to H^{2p}(X,\mathbb{Z})$$

  thus defines a *cycle class* of the subvariety Z.

For *nonsingular* X, we get an induced *homomorphism of graded rings*

$$CH^*(X) \to H^{2*}(X, \mathbb{Z}),$$

so that the intersection product of algebraic cycles is compatible with the cup product in cohomology.

- If X is a non-singular complex projective variety, then its cohomology satisfies *Poincaré duality*, an isomorphism of graded groups

$$H^{2\dim X - *}(X, \mathbb{Z}) \to H_*(X, \mathbb{Z})$$

induced by the *cap product* with the *fundamental homology class*

$$\mu[X] \in H_{2\dim X}(X, \mathbb{Z}).$$

However, if X has singularities, then although there is a fundamental class, the cap product with it need not be an isomorphism.

- The cohomology of an algebraic variety carries "extra structure", reflecting the very special nature of algebraic varieties among topological spaces. Grothendieck's theory of *motives* and it generalizations are an attempt to elucidate this extra structure.

One aspect of this "extra structure" comes from *Hodge theory*. If $X \subset \mathbb{P}^n_{\mathbb{C}}$ is a nonsingular complex projective variety, then its cohomology has a *Hodge decomposition* for each $0 \le i \le 2\dim X$,

$$H^i(X, \mathbb{Z}) \otimes \mathbb{C} = \bigoplus_{0 \le r,s \le \dim X,\ r+s=i} H^{r,s}(X)$$

where $H^{r,s}(X)$ are complex vector spaces, satisfying

$$\overline{H^{r,s}(X)} = H^{s,r}(X) \subset H^i(X, \mathbb{Z}) \otimes \mathbb{C},$$

where the overbar denotes the natural complex conjugation on $H^i(X, \mathbb{Z}) \otimes \mathbb{C}$ obtained from that on $\mathbb{C}$, which maps any $\mathbb{C}$-subspace into another such; thus $\operatorname{rank} H^i(X, \mathbb{Z})$ is *even* if i is odd.

Define

$$Hg^p(X) = \text{ inverse image of } H^{p,p}(X) \text{ under } H^{2p}(X, \mathbb{Z}) \to H^{2p}(X, \mathbb{Z}) \otimes \mathbb{C}$$

$$= \text{group of } \textit{Hodge cycles} \text{ on } X \text{ of codimension } p.$$

It can be shown that the cycle map defined above has the property that

$$\text{image}\left(CH^p(X) \to H^{2p}(X, \mathbb{Z})\right) \subset Hg^p(X).$$

The *Lefschetz* $(1,1)$ *Theorem* is the statement that

$$\text{image}\left(CH^1(X) \to H^2(X,\mathbb{Z})\right) = Hg^1(X)$$

$$= \{\alpha \in H^2(X,\mathbb{Z}) \mid \alpha_{\mathbb{C}} \in H^{1,1}(X) \subset H^2(X,\mathbb{Z}) \otimes \mathbb{C}\}.$$

The *Hodge Conjecture* asserts that

$$\text{image}\left(CH^p(X) \to H^{2p}(X,\mathbb{Z})\right) \otimes \mathbb{Q} = Hg^p(X) \otimes \mathbb{Q}.$$

It is natural to ask if there are generalizations/conjectures for singular varieties; we will return to this theme later.

2. The Singular Case: First Steps

From analogy between the Chow ring and cohomology, we might look for a "Chow ring" for a singular projective variety X satisfying the following properties.

- It should be a commutative, graded ring, contravariant in X for arbitrary morphisms.
- We should have a natural isomorphism $CH^1(X) \cong \text{Pic}\,(X)$, the Picard group of line bundles.
- $CH^*(X)$ should admit a suitable theory of Chern classes for vector bundles.
- $CH^p(X)$ should admit cycle classes for arbitrary subvarieties of codimension p which are disjoint from the singular locus.
- The Chern Character should give an isomorphism of rings

  $$K_0(X) \otimes \mathbb{Q} \cong CH^*(X) \otimes \mathbb{Q}$$

  where $K_0(X)$ is the Grothendieck group of *algebraic vector bundles* on X, such that the individual $CH^i(X) \otimes \mathbb{Q}$ are identified with the appropriate eigenspaces for Adams operators on $K_0(X) \otimes \mathbb{Q}$.
- When $k = \mathbb{C}$, there should be a (functorial) "cycle class" ring homomorphism

  $$CH^*(X) \to H^{2*}(X,\mathbb{Z})$$

  which coincides with the earlier cycle class map on irreducible $Z \subset X$ with $Z \cap X_{sing} = \emptyset$. The algebraic and topological Chern classes of (algebraic) vector bundles should be compatible with this homomorphism.
- The Chow ring should carry "extra structure" corresponding to the "extra structure" on cohomology.

In his book [1], Fulton develops the theory of the Chow ring of a non-singular variety, and in particular, its intersection product, using an analogy with Poincaré duality. Here are some features of his theory, from our perspective.

- The Chow group defined earlier

  $$CH_F^p(X) = Z^p(X)/R^p(X)$$

  is seen as analogous to the homology group

  $$H_{2\dim X - 2p}(X, \mathbb{Z}),$$

  even if X has singularities. We use the notation $CH_F^p(X)$ to indicate it is Fulton's definition.

- Chern classes are defined "directly" for vector bundles as certain *operators on these "homology groups"*, satisfying suitable functorial properties.

- A "Poincaré dualtiy theorem" is proved for non-singular varieties, identifying the underlying additive group of a suitable ring of such operators with the additive "homology group", through a sort of "cap-product".

Even for projective varieties with singularities, the whole machinery makes sense, except that we do not have "Poincaré duality" any more. This leads to a "Fulton-Chow ring of operators", even for singular projective varieties.

One property of the Fulton Chow groups $CH_F^*(X)$ is that, if

$$G_0(X) = \text{the Grothendieck group of coherent sheaves on } X,$$

then there is an isomorphism of rational vector spaces

$$G_0(X) \otimes \mathbb{Q} \cong \bigoplus_{p=0}^{\dim X} CH_F^p(X) \otimes \mathbb{Q}$$

(this is a version of the *singular Riemann-Roch theorem*, proved in [1]). Note that $G_0(X)$ is also covariantly functorial, for projective varieties, and so has a "homology like" character.

However, in general the Grothendieck groups $K_0(X)$ and $G_0(X)$ of vector bundles and coherent sheaves, respectively, are *not* isomorphic, even tensored with $\mathbb{Q}$. This is somehow analogous to the fact that Poincaré dualty fails for singular X. They even have different functoriality: K_0 is contravariant functorial for arbitrary morphisms, using the pull-back operation on vector bundles.

The Fulton-Chow ring of operators, with rational coefficients, *does not* coincide with $K_0(X) \otimes \mathbb{Q}$, for some projective varieties X.

For example, if $\pi : Y \to X$ is a resolution of singularities, then it is easy to see that with rational coefficients, the Fulton-Chow ring of operators of X is a

subring of $CH^*(Y) \otimes \mathbb{Q}$. This is a simple consequence of the *surjectivity* of the corresponding pushforward map $\pi_* : CH_F^*(Y) \otimes \mathbb{Q} \to CH_F^*(X) \otimes \mathbb{Q}$. However $K_0(X) \otimes \mathbb{Q} \to K_0(Y) \otimes \mathbb{Q}$ need not be injective in general, even for curves.

Similarly, we do not expect to have a cycle class homomorphism from the Fulton-Chow ring of operators to cohomology, in general.

Hence *the Fulton-Chow ring of operators is not the ring we are looking for.*

3. The Singular Case: Continued

3.1. A Chow ring. In fact, a theory of the Chow ring for singular varieties, with all of the above desired features, has *not yet been constructed* in the published literature, to the best of our knowledge. But there are partial results.

Of course, we first postulate that

$$CH^1(X) \cong \mathrm{Pic}\,(X),$$

where $\mathrm{Pic}\,(X)$ is the group of line bundles, and the isomorphism determines the 1st Chern class.

For $d = \dim X$, Levine and Weibel [2] gave the following definition, in 1985,

$$CH^d(X) = \frac{\text{Free abelian group on points of } X \setminus X_{sing}}{\text{Modified rational equivalence}}$$

where "modified rational equivalence" is generated by divisors of rational functions on curves, where the curves meet the singular locus "correctly", and the rational functions are invertible at the points of intersection. For example, if $d = 2$, we consider only curves which are *reduced Cartier divisors*; if X_{sing} has codimension ≥ 2, we consider only *curves disjoint from* X_{sing}. Some further properties are:

- for nonsingular X, we recover the usual definition of $CH^d(X)$
- $CH^d(X) \otimes \mathbb{Q} \cong F^d K_0(X) \otimes \mathbb{Q}$, where $F^d K_0(X)$ is the subgroup generated by classes of smooth points
- if $k = \mathbb{C}$, then $CH^d(X)_{tors} \cong H^{2d-1}(X, \mathbb{Q}/\mathbb{Z})$ ("Roitman Theorem", [3]), originally due to Roitman in the smooth case
- there is a good "Albanese theory" associated to $CH^d(X)$ (see [4]).

Marc Levine has an unpublished preprint (*circa* 1984), with a construction of a Chow ring for quasi-projective varieties X, satisfying all the desired properties, except possibly functoriality for arbitrary morphisms (see [5] for an overview; the detailed construction is in the preprint cited there, entitled "A Geometric Theory of the Chow ring for singular varieties"; the surface case is worked out in [18]). Levine's technique did not yield a proof of functoriality, though there is no reason to think it is wrong for his Chow ring.

So we may reasonably conjecture that such a theory of the "Chow cohomology ring" exists, and possibly equals the ring defined by Levine.

3.2. A cohomological formula? One may try to guess a cohomological formula for the desired Chow ring. One plausible guess is

$$\bigoplus_{p=0}^{\dim X} H^p(X, \mathcal{K}_{p,X}),$$

where $\mathcal{K}_{p,X}$ are the sheafified Quillen K-groups; this is a true formula in the non-singular case (Bloch's Formula, proved in general by Quillen [6], see also [7]), and Levine [8] extended it to *singular surfaces*. But Levine and I (2002, unpublished) found a counterexample in $\dim 3$.

Another guess, where we use the Milnor K-theory sheaves (which again gives the Chow groups for smooth X), instead of the Quillen sheaves, also turns out to be wrong in the singular case.

Our counterexample, for both formulas, is the "boundary of a 4-simplex", which is a singular union of copies of affine 3-spaces $\mathbb{A}^3_{\mathbb{C}}$, given by an equation

$$xyzw(1 - x - y - z - w) = 0$$

in $\mathbb{A}^4_{\mathbb{C}}$, with coordinates x, y, z, w, for which the Grothendieck group of vector bundles can be shown to be $\mathbb{Z} \oplus K_3(\mathbb{C})$. Thus, $K_0(X) \otimes \mathbb{Q}$ has 3 nontrivial Adams weight subspaces: of weight 0, corresponding to the rank map on vector bundles (the summand $\mathbb{Z}$ in K_0), of weight 2, given by $K_3^{\mathrm{ind}}(\mathbb{C}) \otimes \mathbb{Q}$, the indecomposable part of K_3, and of weight 3, given by $K_3^M(\mathbb{C}) \otimes \mathbb{Q}$, the Milnor K_3. Hence, these must be the rational Chow groups $CH^i(X) \otimes \mathbb{Q}$, $i = 0, 2, 3$, and they are known to be nontrivial.

The group $H^3(X, \mathcal{K}_3) \otimes \mathbb{Q}$ is seen to account for the whole of $K_3(\mathbb{C}) \otimes \mathbb{Q}$, including the indecomposable part. This means that $H^2(\mathcal{K}_2) \otimes \mathbb{Q}$ vanishes, while at the same time, $CH^2(X) \otimes \mathbb{Q} \neq 0$. Thus we end up having a counterexample to both possible cohomological formulas for the "Chow ring", even with rational coefficients.

3.3. Codimension of support. We comment next on our desired *cycle class property*:

> $CH^p(X)$ should admit cycle classes for arbitrary subvarieties of codimension p which are disjoint from the singular locus.

At first sight, this seems formulated in *too weak a form*: we might well expect that if $\mathcal{F}$ is a coherent sheaf on a projective variety X, which has a finite resolution by vector bundles (i.e., whose stalks have *finite homological dimension* as modules over the respective local rings), then there is an associated "cycle class" in $CH^p(X)$, where p is the codimension of the support of $\mathcal{F}$. With $\mathbb{Q}$-coeffcients, we might further expect to be able to express this cycle class using the p-th graded component of the Chern character of $\mathcal{F}$, and this Chern character should vanish in graded degrees $< p$.

In particular, if X has dimension d, and $\mathcal{F}$ is supported at a finite set of points, we should thus expect the corresponding cycle to lie in $CH^d(X)$, and to be a rational multiple of the Chern character of $\mathcal{F}$.

However, it turns out that this is *false* in general. A counterexample comes from suitably interpreting the work of Dutta,Hochster and Maclaughlin [9]. They showed that if $X \subset \mathbb{P}^4_k$ is the 3-dimensional quadric cone defined by

$$X = \{x_1x_2 - x_3x_4 = 0\} \subset \mathbb{P}^4_k,$$

which has a unique singular point (the vertex of the cone)

$$P = \{x_0 = 1, x_1 = x_2 = x_3 = x_4 = 0\},$$

then there is a sheaf $\mathcal{F}$ of finite homological dimension, supported only at the point P, such that if $L = \{x_1 = x_3 = 0\} \cong \mathbb{P}^2_k$ is a plane contained in X, then

$$\chi(\mathcal{F}, \mathcal{O}_L) = \sum_{j\geq 0}(-1)^j \ell_P \left(\mathcal{T}or_j^{\mathcal{O}_X}(\mathcal{F}, \mathcal{O}_L)\right) = -1$$

where ℓ_P denotes the length of the stalk of a sheaf supported at P.

$\chi(\mathcal{F}, \mathcal{O}_L)$ is the K-theoretic expression of the "intersection product" of $ch_X(\mathcal{F})$ and L, which corresponds to Fulton's cap product, where ch_X denotes the Chern character.

This means that *the Chern character of $\mathcal{F}$ must have a component in* $CH^2(X) \otimes \mathbb{Q}$, even though we would naively think it must lie solely in $CH^3(X) \otimes \mathbb{Q}$!

The example $\mathcal{F}$ is given by writing down explicitly a module of length 15 over the local ring

$$\frac{k[x_1, x_2, x_3, x_4]_{(x_1,x_2,x_3,x_4)}}{(x_1x_2 - x_3x_4)}.$$

They construct 15×15-matricies A, B, C, D of numbers, which are nilpotent, mutually commute, and satisfy $AB = CD$, and explicitly construct a free resolution of the resulting module. They also explicitly compute all the Tor modules, which also have finite length, and finally the alternating sum of the lengths, to obtain the answer -1.

It is natural to ask: what is the "conceptual meaning" of this example, and can we construct others? It turns out that *Thomason's localization theorem* in algebraic K-theory leads to an explanation of this phenomenon.

Given a local ring of (say) an isolated singularity $R = \mathcal{O}_{X,P}$ (where X is a projective variety, $P \in X_{sing}$ an isolated point), there is a Grothendieck group $K_0^P(R)$ of bounded complexes of finite rank, free R-modules whose homology has finite length.

One may ask:

- what is the structure of $K_0^P(R)$?

- what linear functionals can be defined on $CH^*_F(X)\otimes\mathbb{Q}$ using cap products with Chern characters of elements of $K_0^P(X)\otimes\mathbb{Q}$?

Let $\widetilde{K}_0^P(R)$ be the kernel of the obvious map $K_0^P(R)\to\mathbb{Z}$ given by

$$F^\bullet \mapsto \chi(F^\bullet) = \sum_i (-1)^i \ell(H^i(F^\bullet)).$$

Then Thomason's theorem implies a formula

$$\widetilde{K}_0^P(R) \cong K_1(U)/K_1(\operatorname{Spec} R) = K_1(U)/R^*,$$

where $U = \operatorname{Spec} R \setminus \{P\}$ is the *punctured spectrum* of the local ring, and K_1 is the Quillen algebraic K_1-group (special cases of this formula were proved earlier, from works of Levine, and myself).

Now U is a regular scheme of Krull dimension $d-1$, where $d = \dim X$, and its K-theory in general has a nontrivial eigenspace decomposition under the Adams operations, with $d-1$ pieces, one of which is $\Gamma(U,\mathcal{O}_U^*)$, which is "usually" R^* (see, for example, [10] for a discussion of Adams operations for higher K-theory). If $d\geq 2$, we may well have other nontrivial pieces in this eigenspace decomposition. This is a (crude) answer to the first question.

Paul Roberts and I gave an answer in [11] to the second question, on the possible functionals on the Chow group obtained by "intersection products", in the case when $X\subset\mathbb{P}^n_k$ is the projective cone over $Y\subset\mathbb{P}^{n-1}_k$, a smooth projective variety, and $P\in X$ is the vertex (so that $X_{sing}=\{P\}$).

We found the following precise answer: if $h\in CH^1(Y)$ is the class of a hyperplane section, let V be the kernel in $CH^*(Y)\otimes\mathbb{Q}$ of multiplication by h, and let W be the image of V in the quotient of $CH^*(Y)\otimes\mathbb{Q}$ by *numerical equivalence* (this quotient of the Chow ring is known to be a finite dimensional vector space over $\mathbb{Q}$). Then the desired space of functionals is naturally isomorphic to W.

This precise answer is rather difficult to compute. However, suppose k has characteristic 0, and Y satisfies the following properties:

- Y is defined over the field $\overline{\mathbb{Q}}$ of algebraic numbers
- Y satisfies *Grothendieck's Standard Conjectures* (in particular, that the inverse of the Lefschetz operator is given by an algebraic correspondence)
- Y satisfies the *Bloch-Beilinson Conjecture*, that the cycle map to rational Deligne-Beilinson cohomology is injective on $CH^*(Y_{\overline{\mathbb{Q}}})\otimes\mathbb{Q}$.

Then: the dimension of the desired space of functionals on $CH^*_F(X)\otimes\mathbb{Q}$ equals the rank of the *algebraic primitive cohomology* of Y.

Of course $Y=\mathbb{P}^1\times\mathbb{P}^1\subset\mathbb{P}^3$ is known to satisfy all the above properties, and has a primitive cohomology class in H^2, so the Dutta-Hochster-Maclaughlin example is "explained". Other related calculations, trying to make the "interesting" functionals more explicit, may be found in [12].

For arbitrary (smooth projective) Y, the dimension of algebraic primitive cohomology gives an *upper bound* for the dimension of the space of functionals, which can in general be strict, for varieties not defined over $\overline{\mathbb{Q}}$.

For general local rings, we presumably cannot expect to find such a detailed answer. However, Kurano has introduced the notion of numerical equivalence on the rational (Fulton) Chow groups of a Noetherian local ring $(R, \mathfrak{m})$, or equivalently of the rational Grothendieck group $G_0(R) \otimes \mathbb{Q}$. An element $\alpha \in G_0(R)$ is called numerically equivalent to 0 if for any bounded complex $F_\bullet$ of finitely generated free R-modules with homology of finite length, the class (defined using the tensor product of modules)

$$[F_\bullet] \cdot \alpha \in G_0(R/\mathfrak{m}) = \mathbb{Z}$$

vanishes. In terms of the rational Chow group, this is expressed in terms of vanishing of the local Chern character (in the sense of [1]) of the perfect complex on the element of the Chow group.

Using localization techniques, as well as other tools (etale cohomology, de Jong's alterations, etc.) Kurano has shown [13] that for an excellent local ring, under a mild hypothesis[1], the rational Chow groups modulo numerical equivalence of local ring are finite dimensional vector spaces. These are rather mysterious invariants of local rings; for example, I believe we do not understand much about even the "Neron-Severi group" of a local ring, obtained as a quotient of its divisor class group.

3.4. Hodge conjecture. We now turn to the question: is there a reasonable Hodge Conjecture for singular complex projective varieties?

If X is a smooth projective variety, we can interpret the Hodge cycle groups as

$$Hg^p(X) = \mathrm{Hom}_{HS}\left(\mathbb{Z}(-p), H^{2p}(X, \mathbb{Z})\right).$$

Here $\mathbb{Z}(-p)$ is the abelian group $\mathbb{Z}$ with a "Hodge decomposition" $\mathbb{Z} \otimes \mathbb{C} = (\mathbb{Z} \otimes \mathbb{C})^{p,p}$, the symbol Hom_{HS} denotes the Hom group in the *category of Hodge structures*, which is an abelian category.

The homology and cohomology groups of a complex algebraic variety carry *mixed Hodge Structures*, and these form an abelian category MHS, from the work of Deligne. One possibility in the singular case is to look at the homology groups instead (and use Poincaré duality to relate to the non-singular case).

In fact there is a cycle map $CH^p_F(X) \to H_{2\dim X - 2p}(X, \mathbb{Z})$, and one can ask if

$$\begin{aligned}\text{image}\, CH^p_F(X) \otimes \mathbb{Q} \to H_{2\dim X-2p}(X, \mathbb{Q}) \\ = \mathrm{Hom}_{MHS}\left(\mathbb{Z}(\dim X - p), H_{2\dim X-2p}(X, \mathbb{Z})\right) \otimes \mathbb{Q},\end{aligned}$$

[1] Assume the ring contains $\mathbb{Q}$, or is essentially of finite type over a field, $\mathbb{Z}$ or a complete discrete valuation ring.

(with the "twist" $\mathbb{Z}(\dim X - p)$ instead of $\mathbb{Z}(-p)$, because we work with homology).

In the smooth case, this is actually a restatement of the usual Hodge Conjecture. However, one can see using Hironaka's resolution of singularities [14], and Deligne's foundational work [15], that this "singular homology Hodge conjecture" is a *consequence* of the usual Hodge conjecture for smooth varieties; this was pointed out by U. Jannsen (see [16]).

However, we can instead ask: what is the "Hodge theoretic" characterization, if any, of the image of

$$CH^*(X) \to H^{2*}(X, \mathbb{Z}),$$

or of this image tensored with $\mathbb{Q}$ (which we may define using the Chern character on $K_0(X)$)?

The first step here is to ask: what is the analogue, if any, of the Lefschetz $(1,1)$ theorem?

Before proceeding further, we remark that if X is a singular complex variety, it may be possible to find a proper morphism $f : Y \to X$ which is a bijection on points, but not an isomorphism. It will then be a homeomorphism on the underlying topological spaces, and so induce an isomorphism on singular cohomology, which will be compatible with the underlying Mixed Hodge structures. However, we need not have that $K_0(X) \to K_0(Y)$ is an isomorphism, and can even have a cokernel of positive rank (see [17] for an example).

So, for statements related to the cycle map into singular cohomology, it makes sense to restrict to the class of *seminormal varieties*. For our purposes, a variety X (in characteristic 0) is seminormal if, for any proper morphism $f : Y \to X$ which is bijective on points, f must be an isomorphism. This can be expressed intrinsically in terms of the local rings of X. In char. $p > 0$ there is also a definition, taking inseparability into account (see [19], [20] for more on this notion).

Now, even for a hypersurface $X \subset \mathbb{P}^3_{\mathbb{C}}$ with an isolated singularity (which is thus normal, Cohen-Macaulay, etc.), it is *false* in general that

$$\text{image}\left(CH^1(X) \to H^2(X, \mathbb{Z})\right) = \text{Hom}_{MHS}\left(\mathbb{Z}(-1), H^2(X, \mathbb{Z})\right).$$

A simple counterexample is the projective surface

$$X = \{w(x^2z - y^3) + (x^4 + y^4 + z^4) = 0\} \subset \mathbb{P}^3_{\mathbb{C}}.$$

This is a rational quartic surface with a triple point, resolved by 1 blow-up, whose exceptional divisor is an irreducible rational curve with a cusp. From this, one can show (see [21]) that

$$\operatorname{rank} \operatorname{Pic}(X) < \operatorname{rank} H^2(X, \mathbb{Z}) = \operatorname{rank} \operatorname{Hom}_{MHS}(\mathbb{Z}(-1), H^2(X, \mathbb{Z})).$$

To remedy this, for any projective variety X, we may consider the subspace

$$L^1H^2(X, \mathbb{Z}) = \text{Zariski locally trivial classes in } H^2(X, Z)$$

$$= \{\alpha \in H^2(X,\mathbb{Z}) \mid \alpha \in \ker\left(H^2(X,\mathbb{Z}) \to \oplus_i H^2(U_i,\mathbb{Z})\right)$$
$$\text{for some Zariski open cover } \{U_i\} \text{ of } X\}.$$

The choice of notation $L^1H^2(X,\mathbb{Z})$ is explained later.

From Deligne's results [15], $L^1H^2(X,\mathbb{Z}) \subset H^2(X,\mathbb{Z})$ is a "sub-Mixed-Hodge structure" (it should in fact correspond to a "submotive"), and clearly

$$\text{image}\left(CH^1(X) \to H^2(X,\mathbb{Z})\right) \subset L^1H^2(X,\mathbb{Z})$$

since $CH^1(X) = \text{Pic}\,(X)$, and any algebraic line bundle on X is, by definition, locally trivial for the Zariski topology. Hence we in fact have that for any projective complex variety X,

$$\text{image}\left(CH^1(X) \to H^2(X,\mathbb{Z})\right) \subset \text{Hom}_{MHS}\left(\mathbb{Z}(-1), L^1H^2(X,\mathbb{Z})\right).$$

Now one can show:

Theorem 3.1. *(Singular Lefschetz (1,1) Theorem) If X is a seminormal complex projective variety, , then*

$$\text{image}\left(CH^1(X) \to H^2(X,\mathbb{Z})\right) = \text{Hom}_{MHS}\left(\mathbb{Z}(-1), L^1H^2(X,\mathbb{Z})\right).$$

This was proved for normal varieties by J. Biswas and myself [17] in 2000, and more recently by Barbieri, Rosenschon and myself in general (see [22]) by somewhat different arguments (the argument with Biswas seems to also yield a similar version of the Tate Conjecture, in the normal case; the analogue of the Tate conjecture in the seminormal case seems to be an interesting open question).

Now we may ask: can we also similarly formulate a version of the Hodge Conjecture?

An attempt to do this goes as follows. We may regard the identity map on a complex variety X as a continuous map $\pi^X : X \to X_{Zar}$ between the analytic and Zariski sites, and so have a *Leray spectral sequence*

$$E_2^{p,q} = H^p(X_{Zar}, R^q\pi_*^X\mathbb{Z}) \Rightarrow H^{p+q}(X,\mathbb{Z}),$$

where on the abutment, we are working with the complex topology, and singular cohomology is thought of as cohomology of the constant sheaf $\mathbb{Z}$. In particular, the spectral sequence determines a *decreasing filtration* on each singular cohomology group

$$H^i(X,\mathbb{Z}) \supset L^1H^i(X,\mathbb{Z}) \supset \cdots \supset L^iH^i(X,\mathbb{Z}) \supset L^{i+1}H^i(X,\mathbb{Z}) = 0.$$

It is clear that

$$L^1H^i(X,\mathbb{Z}) = \ker\left(H^i(X,\mathbb{Z}) \to H^0(X_{Zar}, R^i\pi_*^X\mathbb{Z})\right)$$

consists of the elements which are locally trivial in the Zariski topology, in the same sense as before.

One can also show that

- the image of the p-th component of the topological Chern character from $K_0(X) \otimes \mathbb{Q}$ to $H^{2*}(X, \mathbb{Q})$ has image contained in $L^p H^{2p}(X, \mathbb{Q})$
- if X is nonsingular, then in fact

$$\text{image } \big(CH^p(X) \to H^{2p}(X, \mathbb{Z})\big) \;=\; L^p H^{2p}(X, \mathbb{Z}),$$

  and $L^{p+1} H^{2p}(X, \mathbb{Z}) = 0$. Both these assertions are part of what was called *Washnitzer's Conjecture*, proved by Bloch and Ogus [23] (see also [24]) that the filtration L^* coincides with Grothendieck's "coniveau filtration" N^*, sometimes also called the "arithmetic filtration".
- It should be remarked that the naive definition of the filtration N^i in the singular case is not the "correct" one, for example, one has $L^1 H^2(X, \mathbb{Z}) \subsetneq N^1_{naive} H^2(X, \mathbb{Z})$ for certain singular projective X.

This suggests the following questions:
(i) is $\{L^p H^i(X, \mathbb{Z})\}$ a filtration by Mixed Hodge structures (or "submotives")?
(ii) assume this holds; then if ch^p_X is the p-th component of the topological Chern character on $K_0(X)$, is

$$\text{image}\, ch^p_X = \text{Hom}_{MHS}(\mathbb{Z}(-p), L^p H^{2p}(X, \mathbb{Z})) \otimes \mathbb{Q}?$$

Assuming a positive answer to the first question, we may regard the second question as an analogue of the Hodge Conjecture.

Note that if X is non-singular, both questions have a positive answer, independent of conjectures, from the results of Bloch and Ogus. Results of Collino [25] imply that if X has exactly 1 singular point, then the above questions have a positive answer.

I expect that
(i) always holds, and
(ii) holds if X is seminormal, and defined over $\overline{\mathbb{Q}}$, while it might be false in general.

3.5. Fine structure. Finally, I turn to the "extra structure" of the Chow ring, first in the smooth case, and then discuss the singular case.

For smooth complex varieties, one important aspect of such "extra structure" is the Bloch Conjecture, and its refinement and generalization by Beilinson, which we may formulate as the existence of "good filtrations" on the Chow rings with rational coefficients of smooth varieties, which are compatible with pull-backs, and push-forwards under proper maps, in the natural way (see [26], [27], [28] for more on this theme). Key additional properties are that

- $\ker \big(CH^*(X) \otimes \mathbb{Q} \to H^{2*}(X, \mathbb{Q})\big)$ is the first level of this filtration
- the filtration induced on $CH^i(X) \otimes \mathbb{Q}$ has at most $i + 1$ steps

- each graded piece of $CH^i(X) \otimes \mathbb{Q}$ is "governed" by a certain graded piece of the cohomology $H^*(X, \mathbb{Q})$.

We do not elaborate on the vague last condition above. Instead, a concrete assertion, with $d = \dim X$, is:
$CH^d(X \setminus Y) \otimes \mathbb{Q} = 0$ for some i-dimensional subvariety $Y \iff H^0(X, \Omega^j) = 0$ for all $j > d - i$.
Here $\implies$ is in fact an old theorem of Mumford, extended by Roitman (see [29], [30]).

Another aspect of the "extra structure" on Chow groups is the system of Bloch-Beilinson Conjectures (see [31], [32]), relating Chow and motivic cohomology groups to the behaviour of L-functions, and refinements (like the Bloch-Kato conjecture [33]). A consequence of these conjectures is the following statement:

> if X is a smooth projective variety over $\overline{\mathbb{Q}}$, and $X_{\mathbb{C}}$ is the corresponding complex variety, then the composite cycle map
>
> $$CH^p(X) \otimes \mathbb{Q} \to CH^p(X_{\mathbb{C}}) \otimes \mathbb{Q} \to H^{2p}(X_{\mathbb{C}}, \mathbb{Z}(p)_{\mathcal{D}}) \otimes \mathbb{Q}$$
>
> is injective, for all p.

Here, $H^{2p}(X_{\mathbb{C}}, \mathbb{Z}(p)_{\mathcal{D}})$ is the *Deligne cohomology*, which fits into an exact sequence

$$0 \to J^p(X_{\mathbb{C}}) \to H^{2p}(X_{\mathbb{C}}, \mathbb{Z}(p)_{\mathcal{D}}) \to Hg^p(X_{\mathbb{C}}) \to 0$$

where $Hg^p(X_{\mathbb{C}})$ is the group of Hodge cycles, which appeared while formulating the Hodge Conjecture, and $J^p(X_{\mathbb{C}})$ is the *p-th Intermediate Jacobian*, first defined by Griffiths, and given in terms of the Hodge decomposition of $H^{2p-1}(X_{\mathbb{C}}, \mathbb{Z}) \otimes \mathbb{C}$ by the formula

$$J^p(X_{\mathbb{C}}) = \frac{H^{0,2p-1}(X_{\mathbb{C}}) \oplus H^{1,2p-2}(X_{\mathbb{C}}) \oplus \cdots \oplus H^{p,p-1}(X_{\mathbb{C}})}{H^{2p-1}(X_{\mathbb{C}}, \mathbb{Z})}.$$

A special case of the above, when $p = d = 2$, when combined with the Lefschetz hyperplane theorem, yields the following Conjecture:

> let $X = \operatorname{Spec} A$ be an affine smooth variety of dimension $d > 1$ over $\overline{\mathbb{Q}}$, then $CH^d(X) = 0$.

A remarkable feature of this conjecture is that *the geometry of $X_{\mathbb{C}}$ plays no role in the conjecture*. It is thus a deep, arithmetical conjecture about $\overline{\mathbb{Q}}$-algebras.

The Bloch Conjecture for complex affine surfaces says, on the other hand:

> let $X = \operatorname{Spec} A$ be a smooth affine surface over $\mathbb{C}$; then $CH^2(X) = 0$ $\iff$ X has a smooth compactification $\overline{X}$ which has no non-zero global 2-forms (or equivalently satisfies $H^2(\overline{X}, \mathcal{O}_{\overline{X}}) = 0$).

The equivalence of the 2 conditions follows from Serre duality.

At present, we do not have a single example of (say) a complex smooth affine surface, which is obtained by complexification of a surface defined over $\overline{\mathbb{Q}}$, for which we can verify that the class of any $\overline{\mathbb{Q}}$-rational point is 0, but where there exist $\mathbb{C}$-points with non-trivial class (e.g., by the Mumford-Roitman result, if $H^0(\overline{X}, \Omega^2_{\overline{X}}) \neq 0$).

For example, the conjecture predicts that the class of any $\overline{\mathbb{Q}}$-point of the affine Fermat hypersurface

$$\{x^4 + y^4 + z^4 = 1\} \subset \mathbb{A}^3_{\mathbb{C}}$$

is 0 in CH^2; this is not known. However, there exist "many" $\mathbb{C}$-points whose classes are not mutually rationally equivalent, since the projective Fermat hypersurface

$$\{x^4 + y^4 + z^4 = w^4\} \subset \mathbb{P}^3_{\mathbb{C}}$$

has a non-zero global 2-form (in fact it is a K3 surface).

Now I can state a result obtained jointly with Amalendu Krishna (see [35], [36]). We have more general results (on normal surfaces, and on graded $\overline{\mathbb{Q}}$-algebras, and he has further generalizations [37], [38], but I will not go into all that here).

Theorem 3.2. *Let $Y \subset \mathbb{P}^n_k$ be a smooth projective curve, and let X be the affine cone over Y. Let $\overline{X} \subset \mathbb{P}^{n+1}_{\mathbb{C}}$ be the projective cone over Y.*

(i) if $k = \mathbb{C}$, then $CH^2(X) = 0 \iff H^2(X, \mathcal{O}_{\overline{X}}) = 0 \iff H^1(Y, \mathcal{O}_Y(1)) = 0$.

(ii) if $k = \overline{\mathbb{Q}}$, then $CH^2(X) = 0$.

Thus, one finds that if

$$A = \frac{\overline{\mathbb{Q}}[x, y, z]}{(x^4 + y^4 + z^4)}$$

then

$$CH^2(\operatorname{Spec} A) = 0,$$

while

$$CH^2(\operatorname{Spec}(A \otimes \mathbb{C})) \cong \mathbb{C}\text{-vector space of uncountable dimension.}$$

This suggests that at least in some form, the "fine structure" present/conjectured on the Chow groups of smooth varieties might have extensions to the singular case. Furthermore, in some cases, one can find nontrivial singular examples verifying the conjectures, which are still elusive in the smooth case. These ideas need to be developed further, to get a more precise general picture.

4. Some Additional Remarks

4.1. Some new tools from K-theory. We now report briefly on some techniques from K-theory that may help shed light on the study of cycles in the singular case.

At the level of singular cohomology, an important tool which relates the structure of the cohomology of a singular variety to that of smooth varieties is the notion of *chomological descent*. Thus, if X is a singular projective complex variety, and $X_\bullet \to X$ is a smooth, proper hypercovering, then cohomological descent identifes $H^*(X, \mathbb{Z})$ with the simplicial cohomology $H^*(X_\bullet, \mathbb{Z}_{X_\bullet})$, which in turn is the abutment of two spectral sequences; this relationship is the basis of Deligne's definition of the Mixed Hodge structure on the cohomology of a singular projective variety, and many other constructions.

However, at the level of Grothendieck groups of vector bundles, there is no simple relationship between $K_0(X)$ and (say) K_0 of locally free sheaves on $X_\bullet$; thus we do not have any direct technique to relate the Grothendieck group of X with the K-theory of the (nonsingular) components X_n in the simplicial scheme $X_\bullet$.

This is related to the phenomenon that K-theory for singular schemes is not "homotopy invariant", i.e., that $K_i(X \times \mathbb{A}^1) \neq K_i(X)$ in general, for singular schemes, though we do have such an identification for regular schemes. Thus, such a smooth proper hypercover somehow "cannot see" the non-homotopy invariant part of K-theory.

One way this has been addressed in the literature is to consider a new definition of K-theory, called "homotopy invariant K-theory", which agress with the Quillen K-theory of vector bundles for regular schemes, does satisfy "homotopy invariance" in the above sense, and (in general) gives different results than Quillen K-theory for non-regular schemes. This is a reasonable approach from the perspective of, say, $\mathbb{A}^1$-homotopy theory.

However, even for $\mathrm{Pic}\,(X)$, the homotopy invariant K-theory seems to give in general a different result than the usual one, even if X is a hypersurface with isolated singularities. So, from our perspective of trying to understand K_0 of vector bundles in the presence of singularities, homotopy invariant K-theory does not provide a satisfactory substitute.

Some new techniques have emerged, which have made progress in the direction of relating the K-theory of a singular variety with (say) the K-theory of the components of a smooth proper hypercover. These rely on two things: (i) the notion of the cdh topology, a Grothendieck topology on schemes which is a variant of the h-topology of Voevodsky, and (ii) Cortiñas' *infinitesimal K-theory* groups, which are homotopy groups of a suitable homotopy fiber, and so fit into an exact sequence involving the Jones-Goodwillie trace maps between K-theory and negative cyclic homology (see [39], [40].

This allows one, in principle, to understand (in some fashion) the difference between the K-theory of a singular scheme and that of a smooth proper

hypercover, in terms of a similar difference between the corresponding negative cyclic homologies, which in turn are closely related to Kahler differentials. It is technically difficult to extract information relevant to us from this package, but it offers new possibilities, and allows us to make new calculations, seemingly impossible with earlier techniques. For example, the interested reader can consult [41], to get an idea about these techniques, and their scope; other articles and preprints are available at [42].

4.2. Intersection Chow groups? In a direction somewhat different from our theme so far, we comment on what we might mean by "intersection Chow groups", or at least "intersection K_0". This is motivated by the notion of intersection cohomology, introduced by Goresky and MacPherson, and developed further by Deligne, Beilinson, Bernstein and Gabber [43]. One important result in [43] is the decomposition theorem, which in principle identifies the intersection cohomology of a singular projective variety as a subquotient of the cohomology of a resolution of singularities, and thus determines a "pure motive", whose Chow groups (say, with rational coefficients) might be thought of as "intersection Chow groups". As far as I understand it, this is basically the approach of Hanamura and Corti in [44], [45].

A different idea might be to look for subcategories of the derived category of coherent sheaves of a projective variety, which are closed under Grothendieck-Serre duality. This is analogous to the defining conditions for perverse sheaves, which involve Grothendieck-Verdier duality. In an unpublished work, Deligne studied such categories, and an account of this was given by Bezrukavnikov [46]. The resulting notions of "perverse coherent sheaves" seem to be of interest in representation theory (see for example [47]).

From a K-theory perspective, it may be interesting to study the Grothendieck groups, and perhaps the "K-theory" of these categories, which might yield interesting new invariants for singular varieties. At this point, this is little more than speculation; thinking of "perverse coherent sheaves", we might ask if there is some "direct" definition of an "interesting" group which is "in-between" the Picard group and the divisor class group. We leave this poser to the reader!

References

[1] W. Fulton, *Intersection Theory*, Ergebnisse Math. 3. Folge Band 2, Springer-Verlag (1984).

[2] M. Levine, C. A. Weibel, *Zero cycles and complete intersections on singular varieties*, J. Reine Angew. Math. **359** (1985) 106–120.

[3] J. Biswas, V. Srinivas, *Roitman's theorem for singular projective varieties*, Compositio Math. **119** (1999) 213–237.

[4] H. Esnault, V. Srinivas, E. Viehweg, *The universal regular quotient of the Chow group of points on projective varieties*, Invent. Math. **135** (1999) 595–664.

[5] M. Levine, *The Chow ring of a singular variety*, Rend. Sem. Mat. Univ. Politec. Torino **42** (1984), no. 3, 1–14.

[6] D. Quillen, *Higher Algebraic K-theory*, Lecture Notes in Math. **341**, Springer-Verlag (1973).

[7] V. Srinivas, *Algebraic K-Theory. Second Edition*, Progress in Math. **90**, Birkhäuser (1995).

[8] M. Levine, *Bloch's formula for singular surfaces*, Topology **24** (1985) 165–174.

[9] S. P. Dutta, M. Hochster, J. E. McLaughlin, *Modules of finite projective dimension with negative intersection multiplicities*, Invent. math. **79** (1985) 253–291.

[10] M. Levine, *Lambda operations, K-theory and motivic cohomology*, in *Algebraic K-Theory (Toronto, ON, 1996)*, Fields Inst. Comm. **16**, Amer. Math. Soc. (1997) 131–184.

[11] Paul C. Roberts, V. Srinivas, *Modules of finite length and finite projective dimension*, Invent. Math. **151** (2003) 1–27.

[12] G. Piepmeyer, P. Roberts, *Constructing modules of finite projective dimension with prescribed intersection multiplicities*, J. Algebra **294** (2005) 569–589.

[13] K. Kurano, *Numerical equivalence defined on Chow groups of local rings*, Invent. Math. **157** (2004) 575–619.

[14] H. Hironaka, *Resoluion of singularities of an algebraic variety over a field of characteristic zero, I, II*, Ann. Math.(2) **79** (1964) 109–203, 205–326.

[15] P. Deligne, *Théorie de Hodge II, III*, Publ. Math. IHES **40** (1971) 5–57 and **44** (1974) 5–77.

[16] U. Jannsen, *Mixed Motives and Algebraic K-Theory. With appendices by S. Bloch and C. Schoen.*, Lecture Notes in Math. **1400**, Springer-Verlag (1990).

[17] J. Biswas, V. Srinivas, *A Lefschetz (1,1) theorem for normal projective complex varities*, Duke Math. J. **101** (2000) 427–458.

[18] J. G. Biswas, V. Srinivas, *The Chow ring of a singular surface*, Proc. Indian Acad. Sci. (Math. Sci.) **108** (1998) 227–249.

[19] S. Greco, C. Traverso, *On seminormal schemes*, Compositio Math. **40** (1980) 325–365.

[20] R. Swan, *On seminormality*, J. Alg. **67**(1) (1980 210–229.

[21] L. Barbieri-Viale, V. Srinivas, *The Néron-Severi group and the mixed Hodge structure on H^2*, J. Reine Ang. Math. **450** (1994) 37–42.

[22] Luca Barbieri-Viale, Andreas Rosenschon, V. Srinivas, *The Néron-Severi group of a proper seminormal complex variety*, Math. Zeit. **261** (2009) 261–276.

[23] S. Bloch, A. Ogus, *Gersten's conjecture and the homology of schemes*, Ann. Sci. E.N.S. (4) **7** (1974) 181–201.

[24] Colliot-Thélène, R. Hoobler, B. Kahn, *The Bloch-Ogus-Gabber Theorem* in *Algebraic K-theory (Toronto, ON, 1996)*, Fields Inst. Commun. **16**, Amer. Math. Soc. (1997) 31–94.

[25] A. Collino, *Washnitzer's conjecture and the cohomology of a variety with a single isolated singularity*, Ill. J. Math. **29** (1985) 353–364.

[26] S. Saito, *Motives and filtrations on Chow groups*, Invent. Math. **125** (1996) 149–196.

[27] S. Saito, *Motives and filtrations on Chow groups. II*, in *The arithmetic and geometry of algebraic cycles (Banff, AB, 1998)*, NATO Sci. Ser. C Math. Phy. Sci. **548**, Kluwer (2000) 321–346.

[28] U. Jannsen, *Motivic sheaves and filtrations on Chow groups*, in *Motives (Seattle, WA, 1991)*, Proc. Symp. Pure Math. **55**, Part 1, Amer. Math. Soc. (1994) 245–302.

[29] D. Mumford, *Rational equivalence of 0-cycles on surfaces.* J. Math. Kyoto Univ. **9** (1968) 195–204.

[30] A. A. Roitman, *Rational equivalence of zero-dimensional cycles*, Math. Sbornik (N.S.) **89 (131)** (1972) 569–585, 671.

[31] D. Ramakrishnan, *Regulators, algebraic cycles and values of L-functions*, in *Algebraic K-theory and algebraic number theory (Honolulu, HI, 1987)*, Contemp. Math. **83**, Amer. Math. Soc. (1989) 183–310.

[32] M. Rapoport, N. Schappacher, P. Schneider (eds.), *Beilinson's conjectures on special values of L-functions*, Perspectives in Math. **4**, Academic Proess (1988).

[33] S. Bloch, K. Kato, *L-functions and the Tamagawa numbers of motives*, in *The Grothendieck Festschrift, Vol. I*, Progress in Math. **86**, Birkhüser (1990).

[34] V. Srinivas, *Zero cycles on singular varieties*, in *The arithmetic and geometry of algebraic cycles (Banff, AB, 1998)*, NATO Sci. Ser. C. Math. Phys. Sci. **548**, Kluwer, Dordrecht (2000) 347–382.

[35] A. Krishna, V. Srinivas, *Zero cycles and K-theory on normal surfaces*, Ann. Math. **(2) 156** (2002) 155–195.

[36] A. Krishna, V. Srinivas, *Zero cycles on singular varieties* in *Algebraic cycles and motives*, Vol. 1, London Math. Soc. Lect. Notes Ser. **343**, Cambridge (2007) 264–277.

[37] A. Krishna, *Zero cycles on singular surfaces*, J. K-Theory **4** (2009) 101–143.

[38] A. Krishna, *Zero cycles on a threefold with isolated singularities*, J. Reine Ang. Math. **594** (2006) 93–115.

[39] G. Cortiñas, *Infinitesimal K-theory*, J. Reine Ang. Math. **503** (1998) 129–160.

[40] G. Cortiñas, *The obstruction to excision in K-theory and cyclic homology*, Invent. Math. **164** (2006) 143–173.

[41] G. Cortiñas, C. Haesemeyer, M. Schlichting, C. Weibel, *Cyclic homology, cdh-cohomology and negative K-theory*, Ann. Math. **167** (2008) 549–573.

[42] C. Weibel, *Papers using cdh techniques*, at link http://www.math.rutgers.edu/~weibel/papers.html

[43] A. A. Beilinson, J. Bernstein, P. Deligne, *Faisceaux pervers*, in *Analysis and topology on singular spaces, I (Luminy, 1981)*, Asterisque, **100**, Soc. Math. France, Paris (1982), 5–171.

[44] M. Hanamura, A. Corti, *Motivic decomposition and Intersection Chow groups.I.*, Duke Math. J. **103** (2000) 459–522.

[45] M. Hanamura, A. Corti, *Motivic decomposition and Intersection Chow groups. II.*, Pure Appl. Math. Q. **3** (2007) 181–203.

[46] R. Bezrukavnikov, *Perverse coherent sheaves (after P. Deligne)*, preprint: arXiv:math.AG/0005152.

[47] P. Achar, D. Sage, *Perverse coherent sheaves and the geometry of special pieces in the unipotent variety*, Adv. Math. **220** (2009) 1265–1296.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

An Exercise in Mirror Symmetry

Richard P. Thomas*

Abstract

This expository article is an attempt to illustrate the power of Kontsevich's homological mirror symmetry conjecture through one example, the heuristics of which lead to an algebro-geometric construction of knot invariants.

Mathematics Subject Classification (2010). Primary 14J33; Secondary 53D37, 57M27, 53C26.

Keywords. Mirror symmetry, Khovanov cohomology.

1. Introduction

This paper can be thought of as a companion to the paper [32], giving the background, mirror symmetric motivation, and helpful pictures that are missing there. Along the way we give a geometric description of Manolescu's isomorphism [18] between an open subset of a Hilbert scheme of points on an ALE space and the Slodowy slice to a nilpotent matrix with two equal Jordan blocks considered by Seidel and Smith, along the lines of the construction in [11]. We use a description of these ALE spaces as blow ups which is probably well known to experts but was new to me, giving maps between them that are crucial to our construction.

Heuristics. We treat mirror symmetry as a heuristic device to motivate constructions on one side of the mirror that reflect better known constructions on the other. We make no rigorous claims for our putative mirrors; for instance we are not claiming that a hyperkähler resolution of a singularity is mirror to a hyperkähler smoothing. Though we will use examples where this ansatz works well, in the key example it *fails* (see Section 5.1) and has to be augmented with a deformation.

*Department of Mathematics, Imperial College, London, UK.
E-mail: richard.thomas@imperial.ac.uk.

Acknowledgements. I would like to thank Gordon Brown, Chris Murphy, Wilson Sutherland, Simon Donaldson, Paul Seidel, Mikhail Khovanov, Shing-Tung Yau, Daniel Huybrechts, Tom Bridgeland and Rahul Pandharipande for educating me, and my collaborator Ivan Smith with whom much of this work was done. Thanks also to Arthur Greenspoon and Ivan Smith for carefully reading the manuscript.

2. Symplectic Geometry

We begin by surveying some standard constructions in symplectic geometry. We skate over many technical issues, in particular Floer cohomology, gradings, the construction of the Fukaya category, and the difficulties in doing symplectic parallel transport in noncompact spaces. Most of these are dealt with manfully in the wonderful papers of Paul Seidel [23, 24, 25].

2.1. Parallel transport.

A family of projective manifolds

$$p\colon \mathcal{X} \to B$$

will not in general be locally trivial over its smooth locus $B^* \subset B$; the complex structure will vary. As Paul Seidel once taught me, symplectic geometry is what is left when you look for what *is* locally constant. (I liked this because it sounded like it might subordinate symplectic geometry to algebraic geometry.) Here the symplectic form ω is given by pulling back the Fubini-Study form via a projective embedding. Over B^* there is a connection on the family: take the annihilator of the fibrewise tangent bundle $T_{\mathcal{X}/B}$ under ω to define the horizontal subbundle of $T\mathcal{X}$. Parallel transport along this connection preserves the symplectic form, and so identifies smooth fibres $\mathcal{X}_{b_0}$, $\mathcal{X}_{b_1}$ by symplectomorphisms, once we pick a path between their images b_0, b_1 in the base B^*. In particular the monodromy around a loop in B^* can be taken to be a *symplectomorphism* of any such fibre $(X, \omega) \cong (\mathcal{X}_b, \omega|_{\mathcal{X}_b})$.

This connection is not flat. Any two tangent vectors $v_1, v_2 \in T_b B^*$ have unique horizontal lifts $\tilde{v}_i \in \Gamma(T_{\mathcal{X}}|_{\mathcal{X}_b})$. Thinking of

$$h := \omega(\tilde{v}_1, \tilde{v}_2)$$

as a Hamiltonian function on $\mathcal{X}_b$ it defines an infinitesimal symplectomorphism of $\mathcal{X}_b$ by the Hamiltonian vector field X_h whose contraction with $\omega|_{\mathcal{X}_b}$ is dh. This $\mathrm{Ham}(\mathcal{X}_b, \omega|_{\mathcal{X}_b})$-valued 2-form on B^* is the curvature of the connection. Therefore isotopic loops in B^* give rise to different but Hamiltonian isotopic monodromies. We get a homomorphism

$$\pi_1(B^*) \to \mathrm{Aut}(X, \omega)$$

to the group of symplectomorphisms modulo Hamiltonian isotopies.

Pick a singular point x_0 lying above a point $b_0 \in B$ in the discriminant locus, and a path in B^* to $b \in B^*$. The locus L of points of $\mathcal{X}_b$ that flow to x_0 by

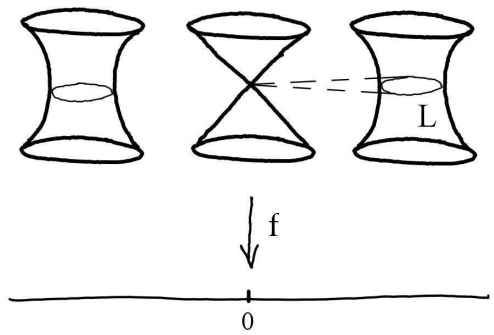


Figure 1. Vanishing cycle L of the family (2.1).

parallel transport along the path is called the *vanishing cycle* of the singularity x_0. Because the flow preserves the symplectic structure, L is isotropic (where it is smooth): $\omega|_{TL} \equiv 0$. If x_0 is an isolated critical point then L is in fact Lagrangian.

The curvature of the symplectic connection blows up as we approach such singular points. Taking smaller and smaller loops in B^* around b_0 the monodromy symplectomorphism approaches the identity away from the vanishing cycle.

2.2. The ordinary double point. We start with a basic affine local model. Consider the family

$$f\colon \mathbb{C}^{n+1} \longrightarrow \mathbb{C}, \qquad f(\mathbf{x}) = \sum_{i=1}^{n+1} x_i^2. \tag{2.1}$$

Over 0 we get the n-dimensional ordinary double point $\sum x_i^2 = 0$, while over $\epsilon \neq 0$ we find its smoothing $X_\epsilon = \{\sum x_i^2 = \epsilon\}$. We use the symplectic structure inherited from the standard Kähler form on $\mathbb{C}^{n+1}$.

Using the $O(n+1)$ symmetry it is easy to see that the vanishing cycle L over $\epsilon \neq 0$ along the straight line path to $0 \in \mathbb{C}$ is the real slice

$$x_i \in \sqrt{\epsilon}.\mathbb{R} \subset \mathbb{C}$$

of X. Scaling coordinates by $\epsilon^{-1/2}$ this is just the sphere

$$L = S^n = \left\{\sum x_i^2 = 1\right\} \subset \mathbb{R}^{n+1}.$$

In fact take $\epsilon \in (0, \infty)$, without loss of generality, and take real and imaginary parts: $x_j = a_j + ib_j$. Consider $\mathbf{a} = (a_i)$ and $\mathbf{b} = (b_i)$ as lying in $\mathbb{R}^{n+1}$ and $(\mathbb{R}^{n+1})^*$ respectively, and give $\mathbb{C}^{n+1} = \mathbb{R}^{n+1} \oplus (\mathbb{R}^{n+1})^* = T^*\mathbb{R}^{n+1}$ its canonical symplectic structure. Then the equation $f = \epsilon$ becomes

$$\sum a_i^2 - b_i^2 = \epsilon, \quad \sum a_i b_i = 0.$$

In particular $|\mathbf{a}|^2 = \epsilon + |\mathbf{b}|^2 > 0$ so we may divide $\mathbf{a}$ and multiply $\mathbf{b}$ by $|\mathbf{a}|$ to give a symplectomorphism of $f^{-1}(\epsilon)$ to

$$T^*S^n = \{(\mathbf{a}, \mathbf{b}) \in T^*\mathbb{R}^{n+1} \colon |\mathbf{a}| = 1,\ \mathbf{b}(\mathbf{a}) = 0\}.$$

The monodromy on going once anticlockwise around $\epsilon = 0$ is Seidel's generalised Dehn twist T_L [22] about L (first suggested by Arnol'd). This is (Hamiltonian isotopic to) the time π flow by the Hamiltonian $\phi(|\mathbf{b}|)$, where ϕ is a smooth monotonic function with $\phi(x) = x$ for small $x \geq 0$ and $\phi \equiv$const for large x. This flow is discontinuous across the vanishing cycle $\mathbf{b} = 0$, but after time π comes back to the antipodal map there and so becomes continuous again. (Alternatively use the standard metric to identify T^*S^n with TS^n. The latter has a canonical vector field which at a point $v \in T_pS^n$ is the horizontal lift $\tilde{v}$ of v to $T_{(p,v)}(TS^n)$. Flowing down $\tilde{v}/|\tilde{v}|$ is again discontinuous, cutting T^*S^n along its zero section then regluing after time π. Then use a bump function to glue this symplectomorphism to the identity away from the zero section.)

When $n = 1$ this reduces to the classical Dehn twist along an embedded S^1 in a Riemann surface: cut along S^1, rotate everything to one side of it through 2π, then reglue. Figure 2 shows its action on one of the cotangent fibres $\mathbb{R} \subset T^*S^1$. More generally given any middle dimensional cycle, the action of the Dehn

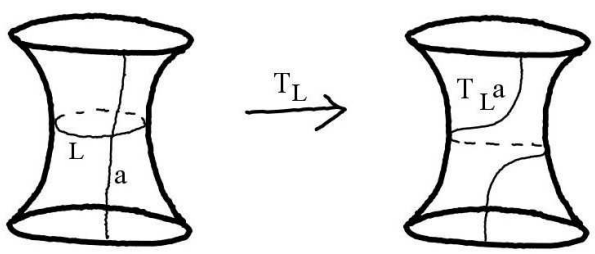


Figure 2. Action of the Dehn twist on a cotangent fibre a of T^*S^n.

twist, i.e. the monodromy around $\epsilon = 0$, can be described similarly: for every transverse intersection point with the vanishing cycle L, the cycle picks up a copy of L (connect summed to it at the intersection point). In particular in any projective family acquiring an ordinary double point we have the above local model near the vanishing cycle L (by Weinstein's theorem) and the action on middle degree homology H_n is given by the Picard-Lefschetz reflection

$$a \mapsto (a.[L])[L] + a. \tag{2.2}$$

One can keep more of the symplectic information by instead using the Fukaya A_∞-category [8, 25]. This has as objects certain Lagrangian submanifolds (with some extra decorations) and morphisms the Floer cochain complex $CF^*(L_1, L_2)$ whose generators are intersection points of generic Hamiltonian perturbations of L_1, L_2 with differential given by counting holomorphic discs running between the intersection points with boundary in the L_i. (The result is independent of

the choices of (almost) complex structure and Hamiltonian isotopy up to quasi-isomorphism.) The tautological evaluation map in this Fukaya category

$$CF^*(L, L') \otimes L \to L' \tag{2.3}$$

has a cone in the derived category $\mathcal{F}(X_\epsilon, \omega)$ of twisted complexes in the Fukaya category. Under certain conditions on the Maslov degree of the intersection points, this cone is equivalent to the (graded) Lagrangian connect sum of L' and L at its intersection points [7, 22, 23, 33]. The induced action of the Dehn twist on the derived Fukaya category indeed takes L' to the above cone [23], clearly categorifying the Picard-Lefschetz reflection (2.2) to which it reduces at the level of cohomology. Another way of saying this is that there is an exact triangle

$$HF^*(L, L') \otimes L \to L' \to T_L(L') \tag{2.4}$$

in $\mathcal{F}(X_\epsilon, \omega)$.

2.3. Families of quadrics. Another way of seeing the smoothing of the ordinary double point – i.e. a smooth fibre of (2.1) – is by fibring *it* over $\mathbb{C}$ using the last coordinate $x_{n+1} = t$:

$$\left\{ \sum_{i=1}^{n} x_i^2 = \epsilon - t^2 \right\} \subset \mathbb{C}^n_{x_i} \times \mathbb{C}_t \longrightarrow \mathbb{C}_t. \tag{2.5}$$

This expresses the n-dimensional affine quadric as a family of $(n-1)$-dimensional affine quadrics – the fibres $\sum_{i=1}^n x_i^2 = \text{const}$ where t is fixed. Each contains a canonical Lagrangian S^{n-1} real slice, except the two singular fibres where $\epsilon - t^2$ vanishes and the vanishing cycle collapses to a point. Picking a path between $t = \pm\epsilon^{1/2}$, the S^{n-1}-bundle over it (collapsing at the endpoints) gives a Lagrangian S^n as in Figure 3. This is the vanishing cycle of the degeneration of the total space given by tending $\epsilon \to 0$ (so that the path and the

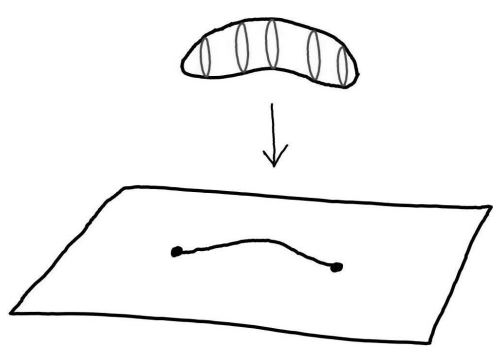


Figure 3. Lagrangian S^n fibred by S^{n-1}s over a matching path between the critical points of the fibration (2.5).

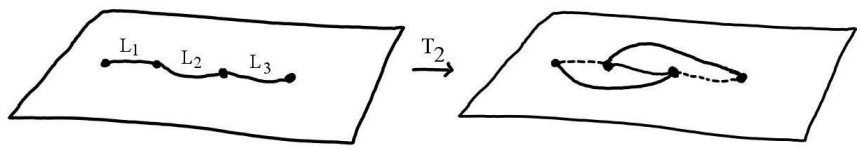


Figure 4. Action of the Dehn twist T_i on the A_{k-1}-chain.

vanishing cycle both collapse). Monodromy around this simply rotates the path anticlockwise through 180°, exchanging the endpoints and giving another way to view the Dehn twist.

This picture generalises by considering a degree k polynomial p on the right hand side of (2.5):

$$X = X_{\underline{\lambda}} := \left\{ \sum_{i=1}^{n} x_i^2 = p(t) \right\} \subset \mathbb{C}^n_{x_i} \times \mathbb{C}_t. \tag{2.6}$$

We fix p monic, with set $\underline{\lambda} = (\lambda_1, \ldots, \lambda_k)$ of *distinct, unordered* roots with centre of mass $0 \in \mathbb{C}$. Then $X_{\underline{\lambda}}$ is smooth (but acquires ordinary double points when p has double roots). By the same reasoning, paths between zeros λ_i of p give $O(n)$-invariant Lagrangian spheres in $X_{\underline{\lambda}}$. Such a sphere is the vanishing cycle of the degeneration given by bringing the two roots of p at its endpoints together along the path to produce an ordinary double point. We will be particularly interested in the $n = 2$ case of this construction, in which case the fibres are the *type-A_{k-1} ALE surfaces* $S_{\underline{\lambda}}$.

We get a smooth family of $X_{\underline{\lambda}}$ s over C_k^0, the configuration space of k distinct unordered points $\underline{\lambda}$ in the plane $\mathbb{C}$ with centre of mass the origin. Now $\pi_1(C_k^0) = B_k$, the braid group on k strands: a loop in C_k can be considered as a motion, as time runs from 0 to 1, of the k points through $\mathbb{C}$ (never touching, and starting and ending at the same set of points, possibly permuted); plotting the graph of this motion in $\mathbb{C} \times [0, 1]$ gives a braid. So the monodromy is a representation

$$B_k \to \mathrm{Aut}(X, \omega),$$

which is *faithful* [15]. Take as basepoint of C_k^0 a configuration of k points along the real line $\mathbb{R} \subset \mathbb{C}$, with the obvious A_{k-1}-chain of paths given by the intervals between them. Then the braid given by rotating the ith and $(i+1)$st points about each other in $\mathbb{C}$ while fixing the others gives the generator T_i of B_k. The corresponding automorphism $T_i \in \mathrm{Aut}(X, \omega)$ is the monodromy about the ordinary double point that $X_{\underline{\lambda}}$ acquires when the two points are brought together along the interval between them. Thus it is the Dehn twist in the Lagrangian sphere L_i fibring over that interval. It takes our A_{k-1}-chain of Lagrangian spheres to a different A_{k-1}-chain, as shown in Figure 4. The T_i satisfy the braid relations

$$\begin{array}{rcll} T_iT_jT_i & \cong & T_jT_iT_j, & |i-j| = 1, \\ T_iT_j & \cong & T_jT_i, & |i-j| > 1, \end{array} \tag{2.7}$$

in $\mathrm{Aut}(X,\omega)$ and so also in $\mathrm{Aut}(\mathcal{F}(X,\omega))$. (To be more careful one has to show that the T_i can be lifted to act on the decorations in the derived Fukaya category, in particular the grading.)

2.4. Spaces of matrices. The family (2.6) is a baby version of another natural family over C_k^0; the space M_k^0 of complex $k \times k$ trace-free matrices with distinct eigenvalues. This has a natural Kähler, and so symplectic, form ω inherited from $\mathbb{C}^{k^2}$. Consider the map

$$M_k^0 \to C_k^0 \tag{2.8}$$

taking a matrix to its set of eigenvalues $\underline{\lambda} \in C_k^0$. It has smooth fibre $M_{\underline{\lambda}}$, the $ad_{SL(k)}$-orbit of similar matrices with the same eigenvalues $\underline{\lambda}$. We get the monodromy representation

$$B_k \to \mathrm{Aut}(M_{\underline{\lambda}}, \omega). \tag{2.9}$$

In fact the family (2.1) for $n = 2$ is the above family (2.8) when $k = 2$, and (2.6) is also a Slodowy slice (at a nilpotent matrix with Jordan blocks of size $(1, k-1)$) of the fibration (2.8). The monodromies can also be described as coisotropic family Dehn twists modelled on relative versions of the 2-dimensional Dehn twist of Section 2.2 with $n = 2$; see [16, Section 3.4].

A different slice of the family (2.8) when $k = 2m$ is considered by Seidel and Smith [27]. Let SS_{2m} denote the space of trace-free matrices A with distinct eigenvalues and the following block form

$$A := \begin{pmatrix} A_1 & I_2 & 0 & 0 & 0 \\ A_2 & 0 & I_2 & 0 & 0 \\ \dots & & & \dots & \\ A_{m-1} & 0 & 0 & 0 & I_2 \\ A_m & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{2.10}$$

where A_i is any 2×2 matrix, A_1 is trace-free, and I_2 is the 2×2 identity matrix. Again the eigenvalue map makes this a smooth symplectic bundle

$$SS_{2m} \to C_{2m}^0, \tag{2.11}$$

with monodromy representation $B_{2m} \to \mathrm{Aut}(SS_{\underline{\lambda}}, \omega)$ on a fibre $SS_{\underline{\lambda}}$.

2.5. The Manolescu isomorphism. Manolescu [18] found another beautiful relationship between the Seidel-Smith family and the basic family (2.6) over C_{2m}^0. Namely, he showed that SS_{2m} can be identified with an explicit open subset of the relative Hilbert scheme of m points on the smooth fibres of the family of ALE surfaces given by (2.6) with $n = 2$ and $\deg p = 2m$.

Manolescu described his isomorphism by ingenious algebraic manipulation, but it is possible to describe it geometrically as follows. We fix m and work on

one fibre $SS_{\underline{\lambda}}$, *fixing* the degree $2m$ monic polynomial $p_{\underline{\lambda}}(x)$ with roots $\underline{\lambda}$ that is the characteristic polynomial of matrices in $SS_{\underline{\lambda}}$.

Since the A_i commute with the other 2×2 blocks in A (2.10), we can evaluate the determinant of $xI_{2m} - A$ blockwise to give the 2×2 matrix polynomial

$$A(x) := I_2 x^m - A_1 x^{m-1} - A_2 x^{m-2} - \ldots - A_m \tag{2.12}$$

with determinant $\det(A(x)) = p_{\underline{\lambda}}(x)$.

In fact it is convenient to work with the matrices

$$B(x) := A(x)J = Jx^m - (A_1 J)x^{m-1} - (A_2 J)x^{m-2} - \ldots - (A_m J), \tag{2.13}$$

where multiplication by

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad J^2 = -1,$$

is invertible, preserves determinants, and takes trace-free matrices to symmetric matrices. Therefore writing the polynomial-valued 2×2 matrices $B(x)$ in the form

$$B(x) = \begin{pmatrix} V(x) & U(x) \\ W(x) & X(x) \end{pmatrix}, \tag{2.14}$$

we have that U and $-W$ are monic of degree m, U and W have equal coefficients of x^{m-1} (the $\operatorname{tr} A_1 = 0$ condition), and V, X have degree $m-1$ and satisfy

$$\det(B(x)) = V(x)X(x) - U(x)W(x) = p_{\underline{\lambda}}(x). \tag{2.15}$$

Matrices $B(x)$ (2.14) satisfying these conditions are entirely equivalent to matrices $A \in SS_{\underline{\lambda}}$ (2.10).

Considering $B(x)$ to be an endomorphism of the trivial rank 2 bundle over $\mathbb{C}_x$, we study it via its spectral curve. Plotting the two eigenvalues $y_1(x), y_2(x)$ of $B(x)$ gives a curve

$$C_B := \{(x, y)\colon \det(yI_2 - B(x)) = 0\} \subset \mathbb{C}_x \times \mathbb{C}_y$$

double covering $\mathbb{C}_x$. Expanding out gives the equation of $C_B \subset \mathbb{C}_x \times \mathbb{C}_y$ as

$$y^2 - \operatorname{tr}(B(x))y + p_{\underline{\lambda}}(x) = 0. \tag{2.16}$$

Over this curve is the natural line subbundle $\mathrm{Eig} \to C_B$ of the trivial rank two bundle given by the corresponding eigenspace of $B(x)$. At $(x, y) \in C_B$, $yI_2 - B(x)$ has rank ≤ 1 and top row $(y - V(x) \quad -U(x))$, so an obvious element of the kernel is its perpendicular

$$\begin{pmatrix} U(x) \\ y - V(x) \end{pmatrix}.$$

This defines a generator of the eigenspace except when it vanishes, i.e. except at the points $(\alpha_i, V(\alpha_i))$, where α_i are the m roots of $U(x)$. (And from (2.14)

or (2.16) one sees that indeed $y = V(x)$ is on one branch of C_B at the roots of $U(x)$; the other branch being $y = X(x)$.)

So we have exhibited a section of Eig vanishing on the length-m divisor $D = \{(\alpha_i, V(\alpha_i))\}$, or, more precisely,

$$D = \{U(x) = 0 = y - V(x)\} \in \operatorname{Hilb}^m C_B. \tag{2.17}$$

In particular, at smooth points of C_B, we find that

$$\operatorname{Eig} \cong \mathcal{O}_{C_B}(D). \tag{2.18}$$

Write the equation (2.16) of the curve $C_B \subset \mathbb{C}_x \times \mathbb{C}_y$ as

$$y(\operatorname{tr}(B(x) - y) = p_{\underline{\lambda}}(x).$$

Plotting the graph of the *other* eigenvalue

$$Y = \operatorname{tr}(B(x)) - y$$

of $B(x)$ embeds C_B in

$$S_{\underline{\lambda}} := \{yY = p_{\underline{\lambda}}(x)\} \subseteq \mathbb{C}_x \times \mathbb{C}_y \times \mathbb{C}_Y. \tag{2.19}$$

This is the *affine blow up* of $\mathbb{C}^2_{xy}$ in the points $(\lambda_i, 0)$ defined by $y = 0 = p_{\underline{\lambda}}(x)$, and is isomorphic to the ALE surface (2.6) (with $n = 2$ and $k = 2m$). (The usual blow up is given by the same equation in $\mathbb{C}^2 \times \mathbb{P}^1_Y$ but we are removing the locus $Y = \infty$ – the proper transform of $y = 0$ – to get $S_{\underline{\lambda}}$. Since by (2.15) the curve $C_B \subset \mathbb{C}^2$ never hits $y = 0$ except at the roots of $p_{\underline{\lambda}}(x)$ it more naturally lies in the blow up (2.19) of $\mathbb{C}^2$ than in $\mathbb{C}^2$ itself. This will help us to invert the construction below. What is going on here is that a point $(x, y) \in C_B$ determines the other eigenvalue $Y = p_{\underline{\lambda}}(x)/y$ by (2.15) except when $y = 0$. At such points, i.e. when x is one of the roots λ_i, the fact that $y = 0$ tells us nothing as we already knew that C_B goes through $(\lambda_i, 0)$ by (2.15). To invert the construction we will need to know the gradient of C_B at this point instead, and this determines the other eigenvalue. The blow up (2.19) achieves this.)

Manolescu's map then maps A (2.10) (or equivalently $B(x)$ (2.13)) to the image of the divisor D (2.17) under the inclusion

$$\operatorname{Hilb}^m C_B \subset \operatorname{Hilb}^m S_{\underline{\lambda}}.$$

By its definition (2.17) we see that D projects to the length m subscheme $\{U(x) = 0\} \in \operatorname{Hilb}^m \mathbb{C}_x$ under the obvious projection $S_{\underline{\lambda}} \to \mathbb{C}_x$. In other words no part of D is tangent to the fibres of this projection and the restriction of the projection to D is an isomorphism. This proves one half of the following.

Theorem 2.20. [18, Prop 2.7] *The above construction gives an isomorphism between the space $SS_{\underline{\lambda}}$ and the open subset of* $\operatorname{Hilb}^m S_{\underline{\lambda}}$ *consisting of subschemes whose projection to $\mathbb{C}_x$ also have length m.*

The proof of the converse is now easy. Fix $D \in \mathrm{Hilb}^m S_{\underline{\lambda}}$ whose projection to $\mathbb{C}_x$ has length m. This defines a unique degree m monic polynomial $U(x)$ with those roots. The function $y|_D$ defines a function on the projection of D in $\mathbb{C}_x$, and there is a unique degree $m-1$ polynomial $V(x)$ on $\mathbb{C}_x$ whose restriction takes the same values. Similarly $Y|_D$ defines $X(x)$. Finally a degree m polynomial $W(x)$, with leading two coefficients -1 and the x^{m-1} coefficient of $U(x)$ respectively, is uniquely determined by comparing coefficients in the equation (2.15), using the fact that the coefficient of x^{2m-1} in $p_{\underline{\lambda}}$ is $\sum \lambda_i = \operatorname{tr} A = 0$. This determines $B(x)$ (2.14), as required.

More geometrically, we are saying that D determines the curve C_B through it, and (at least at smooth points of C_B) the eigensheaf $\mathrm{Eig} = \mathcal{O}_{C_B}(D)$ (2.18). Pushing this down gives the trivial rank two bundle, on $\mathbb{C}_x$, while the scalar endomorphism y descends to an endomorphism $B(x)$ of this trivial rank two bundle. This is the classical spectral curve construction for Higgs bundles [10]. I only recently discovered that the link to Hilbert schemes was discovered 15 years ago by Hurtubise [11].

2.6. Digression – fixed point locus. In [28] Seidel and Smith also consider the involution on $SS_{\underline{\lambda}}$ given by replacing each A_i by its transpose. The fixed point locus consists of those matrices $A(x)$ (2.12) which are symmetric; after multiplying by J we get those matrices $B(x)$ (2.13) which are trace-free.

This fixes the eigenvalues of $B(x)$ (since its determinant is also fixed (2.15)) and so the (smooth) spectral curve,

$$C_B := \{y^2 = p_{\underline{\lambda}}\}. \tag{2.21}$$

Restricted to this locus, the above gives a geometric description of the algebraic construction in [28] (a precursor [29] of Manolescu's construction). The result is an embedding of the fixed point locus of $SS_{\underline{\lambda}}$ in

$$\mathrm{Sym}^m C_B.$$

The image is the complement of the "hyperelliptic locus" of $\mathrm{Sym}^m C_B$ – i.e. it is the length-m subschemes of the hyperelliptic curve $C_B \to \mathbb{C}_x$ whose projection to $\mathbb{C}_x$ also have length m. In [28] Seidel and Smith use this to make a beautiful link between their construction of Khovanov cohomology (of Section 2.8) to Ozsváth-Szabó theory. So in this setting the passage from Ozsváth-Szabó theory to Khovanov cohomology is a form of complexification, replacing the Riemann surface (2.6 with $n = 1$) by the hyperkähler ALE surface (2.6 with $n = 2$) – i.e. replacing (2.21) by (2.19) – and taking Hilb^m of either.

2.7. ALE spaces as affine blow ups. Buried in the description of the Manolescu embedding we saw how to describe the ALE surfaces $S_{\underline{\lambda}}$ (2.6) as affine blow ups. Here we emphasise the construction and a consequence.

Fixing monic p with roots $\underline{\lambda}$, we consider the ALE surface

$$S_{\underline{\lambda}} = \{xy = p(t)\} \subset \mathbb{C}_x \times \mathbb{C}_y \times \mathbb{C}_t \tag{2.22}$$

with its obvious projection to $\mathbb{C}_x \times \mathbb{C}_t$. This is an isomorphism except over the points $x = 0 = p(t)$ of $\mathbb{C}^2$, where the fibre is an exceptional copy of $\mathbb{C}$. This is the affine blow up of $\mathbb{C}^2$ in $x = 0 = p(t)$: the usual blow up given by the same formula in $\mathbb{C}_x \times \mathbb{P}^1_y \times \mathbb{C}_t$ but with $y = \infty$ (the proper transform of the t-axis $x = 0$) removed.

The usual A_{k-1}-chain of Lagrangian S^2s in (2.22) can be seen as follows. Pick an A_{k-1}-chain of paths in $\mathbb{C}_t$ between the roots of $p(t)$. Multiplying by the radius ϵ circle about the origin in $\mathbb{C}_x$ gives k Lagrangian $S^1 \times [0,1]$ tubes in $\mathbb{C}^2$. Blow up $\mathbb{C}^2$ symplectically by removing balls of radius ϵ about each point of $x = 0 = p(t)$ and collapsing the Hopf fibration on the boundary S^3s. This collapses the tubes to Lagrangian S^2s forming our A_{k-1}-chain; see for instance [31].

As Ivan Smith explained to me, this can also be seen as a "spinning" ([26] is a good recent reference) of $\mathbb{C}_t$ over the roots $\underline{\lambda}$ of $p(t)$. The fibres of the projection to $\mathbb{C}_t$ are conics $\mathbb{C}^*$ (the fibres of $\mathbb{C}_x \times \mathbb{C}_t \to \mathbb{C}_t$ with the t-axis ($x = 0$) removed) except over the roots of $p(t)$ where we get the singular conics $\mathbb{C} \cup_0 \mathbb{C}$ (the exceptional fibre union the original fibre $\mathbb{C}_x$).

What is nice about the description as an affine blow up is that it demonstrates natural maps between the ALE spaces that are compatible with the A_k-chains. Ignoring the centre of mass condition for simplicity, let

$$S_{k-1} \subset \overline{S}_{k-1}$$

denote the ALE surface (2.22) with $\underline{\lambda} = (1, 2, \ldots, k)$ inside the full blow up of $\mathbb{C}^2$ in the points $(0,1), (0,2), \ldots, (0,k)$.

Then $\overline{S}_k$ is the blow up of $\overline{S}_{k-1}$ in the point $(0, \infty, k+1)$. On removing $y = \infty$ we get a projection $S_k \to S_{k-1}$. And since we have removed the blow up point $(0, \infty, k+1)$, we also get an inclusion $S_{k-1} \hookrightarrow S_k$ which is a right inverse. These maps are holomorphic; there are also maps preserving the real symplectic structure once we remove a ball about $(0, \infty, k+1)$ from $\overline{S}_{k-1}$, which will be sufficient for our needs in the next Section.

2.8. The Seidel-Smith construction. Seidel and Smith managed to produce an invariant of links using the space SS_{2m} (2.10). Via the Manolescu isomorphism, and using plait closure in place of braid closure, the construction should become the following. (Since the technical details have only been carried out carefully [27] in the open subset $SS_{\underline{\lambda}} \subset \mathrm{Hilb}^m S_{\underline{\lambda}}$, the following is partly conjectural, and should be thought of only as motivation for the mirror construction. In particular $\mathrm{Hilb}^m S_{\underline{\lambda}}$ is not an exact symplectic manifold, so the definition of Floer cohomology needs some care.)

We fix one of the ALE surfaces (2.22), writing it as

$$S_{2m-1} := \left\{ xy = \prod_{i=1}^{2m} (t - \lambda_i) \right\} \subset \mathbb{C}_{x,y} \times \mathbb{C}_t,$$

where $\underline{\lambda}$ is a collection of $2m$ distinct numbers $\lambda_i \in \mathbb{C}$ (with average zero). We also choose an A_{2m-1}-configuration of paths γ_i running between them, as in Figure 4, and so an A_{2m-1}-chain of Lagrangian spheres $L_i \subset S_{2m-1}$.

In turn this defines the Lagrangian $(S^2)^m$

$$\mathcal{L} = \mathcal{L}_m := L_1 \times L_3 \times \ldots \times L_{2m-3} \times L_{2m-1} \tag{2.23}$$

in the Hilbert scheme

$$H_m := \mathrm{Hilb}^m S_{2m-1},$$

via the map $L_1 \times \ldots \times L_{2m-1} \subset (S_{2m-1})^m \to \mathrm{Sym}^m S_{2m-1} \longrightarrow \mathrm{Hilb}^m S_{2m-1}$. (Since the L_{2i-1} are disjoint, the map's image lies in the complement of the large diagonal, over which $\mathrm{Hilb}^m S_{2m-1} \to \mathrm{Sym}^m S_{2m-1}$ is an isomorphism.)

The relative Hilbert schemes of the family of $S_{\underline{\lambda}}$s (2.6) gives a quasi-projective family over C^0_{2m}. Taking monodromy, we see that the braid group lifts to the symplectomorphism group of $\mathrm{Hilb}^m S_{2m-1}$. The Kähler form is the one pulled back via the resolution $\mathrm{Hilb}^m \to \mathrm{Sym}^m$, minus $\epsilon[E]$, where E is the exceptional divisor. By making $\epsilon \to 0$ we can ensure that the action of $\beta \in B_{2m}$ is arbitrarily close, away from the exceptional locus, to the action of $\beta \times \ldots \times \beta$ on $\mathrm{Sym}^m S_{2m-1}$.

Then for any $\beta \in B_m$ define the braid invariant

$$SS^*(\beta) := HF^{*+m+w}(\mathcal{L}, \beta\mathcal{L}) \tag{2.24}$$

to be the Floer cohomology of $\mathcal{L}$ and its image under β (assuming the technical details can be overcome to define this, and as a graded $\mathbb{C}$-vector space rather than a module over a Novikov ring). Here the writhe w is the number of positive minus the number of negative crossings in the braid β.

In fact $SS^*(\beta)$ should be an invariant of the isotopy class of the link given by the plait closure of β. By a result of Birman [3], modified slightly in [2], and the fact that Floer cohomology is functorial under (graded) symplectomorphisms (so that $HF^*(\mathcal{L}, \alpha\beta\mathcal{L}) = HF^*(\alpha^{-1}\mathcal{L}, \beta\mathcal{L})$, for instance), to deduce this it is sufficient to prove the following; see Figure 5 and the further explanation below.

1. $T_1\mathcal{L} \cong \mathcal{L}[-1]$,
2. $T_{2i-1}T_{2i}\mathcal{L} \cong T_{2i-1}^{-1}T_{2i}^{-1}\mathcal{L}$,
3. $T_{2i}T_{2i-1}T_{2i+1}T_{2i}\mathcal{L} \cong \mathcal{L}$, and
4. $HF^*(\mathcal{L}_m, \beta\mathcal{L}_m) \cong HF^{*+1}(T_{2m}^{\pm 1}\mathcal{L}_{m+1}, \beta\mathcal{L}_{m+1})$.

We now explain these relations, starting with (1). As we have already seen, T_1 simply flips the path running between the first two roots λ_1, λ_2. This preserves L_1 but shifts its grading by the $[-1]$ on the right hand side of (1). Since we are skating over the issue of grading we content ourselves with noting only that it reverses the orientation of L_1 (this is equivalent to the action on grading

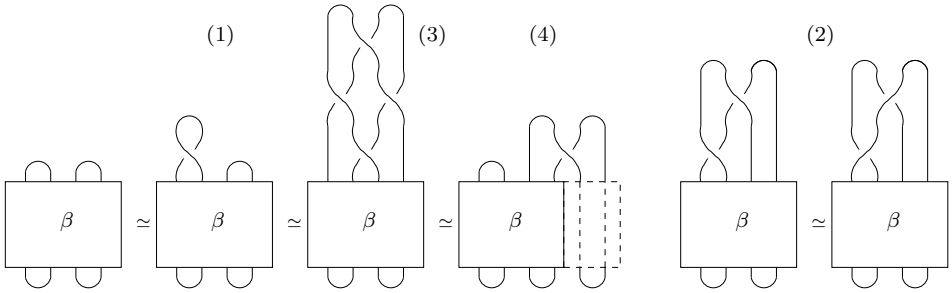


Figure 5. Equivalent plait closures of a braid $\beta \in B_4$.

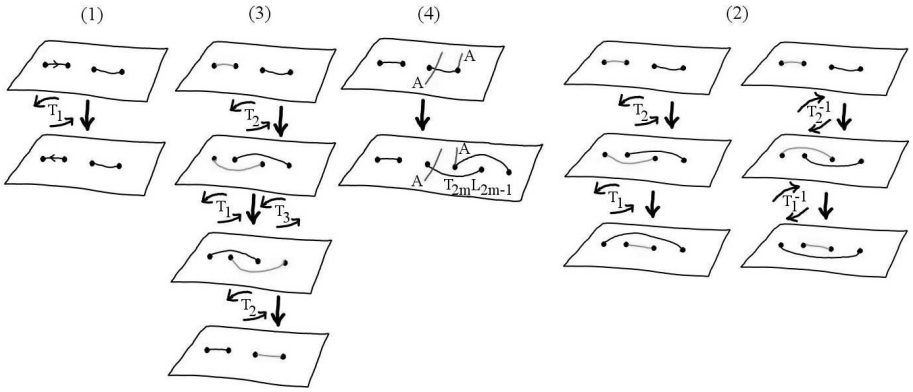


Figure 6. Action of the moves (1) – (4) on the Lagrangians L_i fibring over the paths shown. This gives the action on $\mathcal{L} \subset \mathrm{Hilb}^m S_{2m-1}$, which is a product (2.23) of L_i s.

mod 2). Since the other L_{2i-1}, $i \geq 2$ are untouched by T_1 the relation (1) follows.

Secondly we consider (3). As shown in Figure 6, $T_{2i}T_{2i-1}T_{2i+1}T_{2i}$ simply swaps $L_{2i\pm1}$ (and leaves the other L_{2j-1} alone). But in $\mathrm{Hilb}^m S_{2m-1}$ the order of the factors of $\mathcal{L}$ is unimportant, so (3) follows.

Relation (4) (stabilisation as we increase the number of strands in our braid, or Markov II as it is called in [27]) is slightly more involved. The left hand side is computed in H_m, with β an element of B_{2m}. The right hand side takes place in H_{m+1}, with β considered as an element of B_{2m+2} via the standard inclusion $B_{2m} \hookrightarrow B_{2m+2}$. Here we are using the inclusion of ALE spaces $S_{2m-1} \subset S_{2m+1}$ of Section 2.7.

In Figure 6 is drawn part of an arbitrary $O(2)$-invariant Lagrangian A which is generated in $\mathcal{F}(S_{2m-1})$ by L_i, $i \leq 2m-1$ ($\beta\mathcal{L}_m$ in (4) being a product of such things). We have drawn intersections of A with L_{2m-1} in either the root

λ_{2m} or elsewhere. This corresponds to a splitting

$$HF^*(L_{2m-1}, A) \;\cong\; HF^{*+1}(L_{2m}, A) \oplus HF^*(T_{2m}L_{2m-1}, A) \tag{2.25}$$

coming from the exact triangle (cf. (2.4))

$$L_{2m-1} \to T_{2m}L_{2m-1} \to L_{2m} \tag{2.26}$$

in $\mathcal{F}(S_{2m-1})$. (One can show that $HF^*(\,\cdot\,, A)$ applied to the second arrow vanishes for $A = L_i$, $i \leq 2m-1$, and so for any A, to give the splitting (2.25).) The first summand in (2.25) corresponds to the intersections at the root λ_{2m}; these come from intersections with the next Lagrangian L_{2m} along via cup product with the $HF^1(L_{2m}, L_{2m-1})$ class of the intersection of L_{2m-1} and L_{2m} (the extension class of the triangle (2.26)). The other intersection points are those which survive when L_{2m-1} is Dehn twisted about L_{2m}, as shown in Figure 6, and form the second summand of (2.25).

Since A has no intersections with L_{2m+1} the first summand is isomorphic to $HF^{*+1}(T_{2m}L_{2m+1}, A)$, as can also be seen from Figure 6. The upshot is that if $\mathcal{A}$ is a product of Lagrangians of the form A, the intersection points used to calculate $HF^*(\mathcal{L}_m, \mathcal{A})$ can be matched with intersection points used to calculate $HF^{*+1}(T_{2m}\mathcal{L}_{m+1}, \mathcal{A})$. More precisely their Floer cohomologies can be matched using (2.25). Applied to $\mathcal{A} = \beta\mathcal{L}_m$ this gives (4).

Finally we come to relation (2). We calculate on S_{2n-1} that both $T_{2i-1}T_{2i}$ and $T_{2i-1}^{-1}T_{2i}^{-1}$ leave L_{2j+1} alone for $j \neq i, i-1$, and take L_{2i-1} to L_{2i}. This is clear from Figure 6. Their actions on L_{2i+1} differ, however. They both take it to connect sums of L_{2i-1}, L_{2i} and L_{2i+1}, but in the opposite direction:

$$T_{2i-1}T_{2i}L_{2i+1} \;\cong\; L_{2i+1}\#L_{2i}\#L_{2i-1}, \tag{2.27}$$

$$T_{2i-1}^{-1}T_{2i}^{-1}L_{2i+1} \;\cong\; L_{2i-1}\#L_{2i}\#L_{2i+1}. \tag{2.28}$$

Here # is the *graded Lagrangian connect sum* [22, 33], and is *not* symmetric. It can be described in an $O(2)$-symmetric manner by the connect-summed paths in Figure 6 – with the connect sums in opposite directions corresponding to paths above and below their intersection point.

The two Lagrangians (2.27, 2.28) are certainly *not* Hamiltonian isotopic in S_{2m-1}, so that $T_{2i-1}T_{2i}\mathcal{L}$ and $T_{2i-1}^{-1}T_{2i}^{-1}\mathcal{L}$ are *not* Hamiltonian isotopic in either the product $(S_{2m-1})^m$ or symmetric product $\mathrm{Sym}^m S_{2m-1}$. However Seidel and Smith prove they *are* Hamiltonian isotopic in SS_{2m}, and therefore also in $\mathrm{Hilb}^m S_{2m-1}$. We want to think about this categorically as follows.

In the derived Fukaya category, we see $T_{2i-1}T_{2i}\mathcal{L}$ and $T_{2i-1}^{-1}T_{2i}^{-1}\mathcal{L}$ as extensions of the same objects in the opposite direction. On deforming the symplectic space $\mathrm{Sym}^m S_{2m-1}$ to $\mathrm{Hilb}^m S_{2m-1}$ (by "inflating" the exceptional divisor – subtracting a small amount of the class of the exceptional divisor from the degenerate symplectic form pulled back from $\mathrm{Sym}^m S_{2m-1}$) the Lagrangians $T_{2i-1}^{\pm1}T_{2i}^{\pm1}\mathcal{L}$ deform because both $\mathcal{L}$ and the symplectomorphisms T_i do. However the pieces $L_{2i\pm1} \times L_{2i}$ of the extensions do *not* deform as Lagrangians

– the class $[E]$ restricts to a nonzero class thereon (because $L_{2i\pm1}$ intersects L_{2i} inside S_{2m-1}). And then for general reasons, if two extensions of the same pieces deform while the pieces do not then the deformations of the extensions become isomorphic. The algebro-geometric analogue of this will be clearer to see in Section 5.

Using slightly different techniques in a fibre of SS_{2m}, Seidel and Smith prove carefully that they get an invariant of links up to isotopy. Conjecturally their invariant can be derived from the famous Khovanov cohomology $KH^{*,*}\otimes\mathbb{C}$ [14] by a certain collapse of the latter's bigrading. In the algebro-geometric mirror described later, it will in fact be possible to get the full bigrading and prove the isomorphism to $KH^{*,*}\otimes\mathbb{C}$.

3. Simultaneous Resolution

In each of the examples (2.1), (2.6), (2.8) and (2.11) – in the first two cases only in dimension $n=2$ – the families have a remarkable property. The complete family $\mathcal{X}\to B$ (*including* the singular fibres now) can be pulled back to a new family $\mathcal{X}'\to B'$ via a finite basechange $B'\to B$, such that $\mathcal{X}'$ admits a *simultaneous resolution*

$$\pi\colon \overline{\mathcal{X}}\to\mathcal{X}'.$$

This is a map which is birational, and a resolution of singularities on each fibre. In particular on each smooth fibre it restricts to an isomorphism. So the smooth fibres fit together with the *resolutions* of the singular fibres in a smooth family $\overline{\mathcal{X}}\to B'$. Thus the smoothings and resolutions of the singular fibres of $\mathcal{X}\to B$ are diffeomorphic (something which is obviously not true for the $n=1$ dimensional node (2.1), for instance) and is related to the fact that they are hyperkähler [12].

3.1. Surface ordinary double point. The simplest case is the smoothing of the surface ordinary double point,

$$\mathcal{X}=\{x^2+y^2+w^2=t\}\subset\mathbb{C}^3_{xyw}\times\mathbb{C}_t\to\mathbb{C}_t.$$

If we pull this back by the double cover $t\mapsto t^2$ of the base then the total space becomes singular itself, with the threefold ordinary double point singularity

$$\mathcal{X}'=\{x^2+y^2+w^2=t^2\}\subset\mathbb{C}^3_{xyw}\times\mathbb{C}_t\to\mathbb{C}_t. \tag{3.1}$$

Setting $X=x+iy, Y=x-iy, T=t+w, W=t-w$ this becomes

$$\mathcal{X}'=\{XY=TW\}\subset\mathbb{C}^4$$

fibring over $\mathbb{C}$ by the function $(T+W)/2$. Blowing up the Weil divisor $(X=0=T)$ gives a resolution $\overline{\mathcal{X}}\to\mathcal{X}'$ which is an isomorphism away from the origin.

More explicitly, $\overline{\mathcal{X}}$ is the graph of the rational function $X/T = W/Y : \mathcal{X}' \to \mathbb{P}^1$ in $\mathcal{X}' \times \mathbb{P}^1$:

$$\overline{\mathcal{X}} := \{(X, Y, T, W, [\lambda : \mu]) \in \mathbb{C}^4 \times \mathbb{P}^1 : XY = TW,\ \mu X = \lambda T,\ \mu W = \lambda Y\}.$$

Then $\overline{\mathcal{X}} \to \mathcal{X}'$ is an isomorphism on all of the smooth fibres of (3.1), and replaces the central fibre's surface ordinary double point by its minimal resolution – i.e. its blow up with a $\mathbb{P}^1$ exceptional set C. (So the exceptional set of the whole family is this $C \cong \mathbb{P}^1$, which is *not* a divisor: $\overline{\mathcal{X}} \to \mathcal{X}'$ is a *small* resolution).

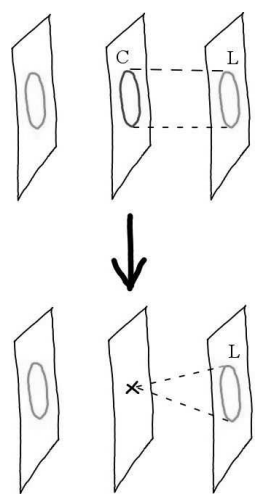


Figure 7. Simultaneous resolution $\overline{\mathcal{X}}$ of the family (3.1), with the Lagrangian vanishing cycles $L \cong S^2$ limiting to the holomorphic exceptional curve $C \cong \mathbb{P}^1$.

We picture this in Figure 7. By its definition as a vanishing cycle, under symplectic parallel transport the *Lagrangian* L limits to the *holomorphic* exceptional $\mathbb{P}^1 = C$. This is remarkable but no contradiction; the pull back of the standard Kähler form from $\mathcal{X}'$ is symplectic on the general fibre (and zero on restriction to L) but degenerate on the central fibre (it is precisely zero along C). One could perturb to get a nondegenerate Kähler form on $\overline{\mathcal{X}}$, giving nonzero area to C, but this would then also have nonzero area on the (homologous) L which would therefore cease to be Lagrangian.

One can also ask what the limit of the Dehn twists is on the central fibre. Consider the graph in $\mathcal{X}_\epsilon \times \mathcal{X}_\epsilon$ of the monodromy about the circle of radius ϵ. As $\epsilon \to 0$, this approaches the identity away from the vanishing cycle L. Arbitrarily close to L we can always find $\epsilon > 0$ and a point that the Dehn twist takes to any other given point. So in the limit we get all of $C \times C$ (since C is the limit of L_ϵ).

The upshot is that as $\epsilon \to 0$ the limit of the Dehn twists about the L_ϵ is the *holomorphic* correspondence

$$\Delta \cup (C \times C) \tag{3.2}$$

in $\overline{\mathcal{X}}_0 \times \overline{\mathcal{X}}_0$, where Δ is the diagonal.

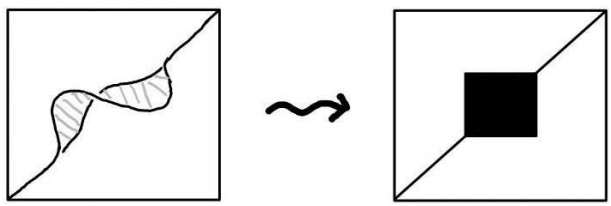


Figure 8. The graph of the Dehn twist limits to the correspondence $\Delta \cup (C \times C)$. (Despite the crude picture, the two irreducible components Δ and $C \times C$ have the same dimension 2.)

The family of $S_{\underline{\lambda}}$s over C_k^0 (2.22) also admits a simultaneous resolution after basechange, with the A_{k-1}-chain of Lagrangian S^2s limiting to the A_{k-1}-chain of holomorphic $\mathbb{P}^1$s in the minimal resolution. When $k = 2m$, taking the relative Hilbert scheme of this new family gives (a birational model of) a similar simultaneous resolution for the space SS_{2m} (2.11) via Manolescu's embedding. Instead of describing these examples in detail we pass straight to the final, and universal example. The previous examples can be obtained from this by taking slices.

3.2. Adjoint quotient and the Flag variety. We partially compactify the adjoint quotient (2.8) with the space of *all* trace-free $k \times k$ matrices, mapping via the roots of its characteristic polynomial to $\operatorname{Sym}^k \mathbb{C}$:

$$M_k \to \operatorname{Sym}^k \mathbb{C}. \tag{3.3}$$

We basechange by the projection $\mathbb{C}^k \to \operatorname{Sym}^k \mathbb{C}$ that forgets the order of k-tuples. In other words we consider the space of matrices with a chosen ordering of the roots (with multiplicities) of its characteristic polynomial:

$$M_k' \to \mathbb{C}^k.$$

At a point $(A, \lambda_1, \ldots, \lambda_{2k}) \in M_k'$ with distinct roots, so that the matrix has distinct eigenvalues λ_i with eigenspaces L_i, there is a canonical associated flag $0 < V_1 < \ldots V_{k-1} < V$ given by $V_i = \oplus_{j \le i} L_j$. This is preserved by A, and characterised by the property that A acts on V_i/V_{i-1} with weight λ_i. Therefore the space $\overline{M}_k$ defined as

$$\Big\{ \big(A, \underline{\lambda}, (0 < V_1 < \cdots < V_{k-1} < V)\big) \colon AV_i \subseteq V_i \ \forall i,\ A \text{ acts on } V_i/V_{i-1} \text{ as } \lambda_i \Big\} \tag{3.4}$$

has a forgetful map to M'_k which is an isomorphism over the good locus of matrices with distinct eigenvalues. In fact $\overline{M}_k \to M'_k$ is a simultaneous resolution, restricting over each fibre of $M'_k \to \mathbb{C}^k$ to a resolution of singularities. The central fibre is the cotangent bundle T^*Fl of the Flag variety, because its fibre over a point $(0 < V_1 < \cdots < V_{k-1} < V) \in Fl$ is

$$\{A : V \to V : AV_i \subseteq V_{i-1}\}.$$

It provides a resolution of the central fibre of $M_k \to \mathrm{Sym}^k \mathbb{C}$, i.e. of the nilpotent cone of matrices with no nonzero eigenvalues. The general fibre is diffeomorphic to it; in fact it is *symplectomorphic* to T^*Fl with its canonical real symplectic structure as the cotangent bundle of a real manifold.

A similar picture to Figure 7 holds. While Fl is a *holomorphic* subvariety of the central fibre, it is the limit of *Lagrangian* vanishing cycles $Fl \subset T^*Fl$ in the general fibre.

In the central fibre T^*Fl live the divisors

$$N_i := \pi_i^* T^*Fl_i \subset T^*Fl, \tag{3.5}$$

where $\pi_i : Fl \to Fl_i$ is the map to the partial flag variety that forgets the ith term V_i in the flag. In the general fibre (seen as symplectomorphic to T^*Fl) they are coisotropic with characteristic foliation $\pi_i|_{N_i}$ a fibration by isotropic S^2s. As λ_i and λ_{i+1} come together in the base $\mathrm{Sym}^k \mathbb{C} = \{\text{eigenvalues}\}$ (3.3), N_i is the relative vanishing cycle that collapses along this characteristic foliation to a family of surface ordinary double points. Doing the family generalised Dehn twist about N_i ([21, Section 1.4], [19, Section 2.3]) should give the braid group of symplectic monodromies of (2.9). The limit of the graphs of these symplectomorphisms is the subvariety

$$\Delta \cup (N_i \times_{Fl_i} N_i) \ \subset \ T^*Fl \times T^*Fl. \tag{3.6}$$

4. Homological Mirror Symmetry

Kontsevich's homological mirror symmetry conjecture [17] is an amazing categorical expression of Witten's formulation of mirror symmetry in terms of A- and B-models. It has become a vast subject that we will only touch on through our example.

Roughly speaking, Kontsevich says that two closed Calabi-Yau manifolds should be considered as mirror pairs when the derived Fukaya category of one is isomorphic to the derived category of coherent sheaves on the other. Symplectic geometry (the "A-model") on one side is equated with complex geometry (the "B-model") on the other side. In particular the plentiful automorphisms of a symplectic manifold should be mirrored not by holomorphic automorphisms of the mirror (of which there are few) but by autoequivalences of its derived category.

4.1. Surfaces. For the examples of the last section, passing from the general fibre to the resolution of the central fibre (using symplectic parallel transport and simultaneous resolution) gives a cheap way to swap complex and symplectic structures. As we have seen, Lagrangian submanifolds can become, in the limit, holomorphic (in fact complex Lagrangian, in the canonical holomorphic symplectic structure). Taking the structure sheaves of these limits means we have turned objects of the derived Fukaya category into objects of the derived category of coherent sheaves.

So it seems a reasonable guess that the mirror of the (symplectic) general fibre might be related to the (holomorphic) resolution of the central fibre. (That mirror symmetry is so simple here, not even changing the topology, is a feature of hyperkähler manifolds, with the mirror map being related to hyperkähler rotation. To make this more precise would involve complexifying our symplectic forms with B-fields, putting connections with curvature $B|_L$ on our Lagrangians L, introducing coisotropic branes, worrying about noncompactness, and working much harder. But we use mirror symmetry here only as a motivational guide.)

So in the simplest case we would like to think of the mirror of the symplectic manifold T^*S^2 (the smoothing of the surface ordinary double point) as something like the complex surface $S = T^*\mathbb{P}^1$ (the resolution of the surface ordinary double point). As usual we denote the Lagrangian S^2 by L and the holomorphic $\mathbb{P}^1$ by C, so we would like mirror symmetry to relate

$$L \in \mathcal{F}(T^*S^2) \quad \text{to} \quad \mathcal{O}_C(-1) \in D(S),$$

where D denotes the bounded derived category of coherent sheaves with compact support. (Work of Auroux and Seidel suggests one should remove certain loci from T^*S^2 and $T^*\mathbb{P}^1$ before they can sensibly be considered as mirror, but for our heuristic purposes we can ignore this.) The twist by the line bundle $\mathcal{O}(-1)$ is unimportant (since it defines an autoequivalence of $D(S)$) and is just for convenience.

Since the graph of the Dehn twist T_L about L limits (3.2) to the holomorphic subvariety

$$\Delta \cup (C \times C) \overset{\iota}{\hookrightarrow} S \times S, \tag{4.1}$$

it is natural to use this as a holomorphic correspondence on S. In fact we would like to lift this to an action on $D(S)$, mirror to the induced action of T_L on $\mathcal{F}(T^*S^2)$. So we might use the structure sheaf of (4.1) as a Fourier-Mukai kernel. For convenience we twist by the line bundle $\mathcal{L}$ which is $\mathcal{O}_S(C)$ on Δ glued to $\mathcal{O}(-1,-1)$ on $C \times C$ (both are isomorphic to $\mathcal{O}_C(-2)$ on Δ_C):

$$T_C := \pi_{2*}\big(\iota_*\mathcal{L} \otimes \pi_1^*(\ \cdot\ )\big) \colon D(S) \to D(S).$$

(Here $\pi_1, \pi_2 \colon S \times S \to S$ are the obvious projections, and the functors $\otimes$ and π_{2*} are derived. It turns out that using the untwisted structure sheaf gives the

inverse of the functor T_C; I don't know if this is significant or a coincidence.) Equivalently, the action of T_C on $E \in D(S)$ is

$$E \mapsto T_C E = \mathrm{Cone}\big(R\,\mathrm{Hom}(\mathcal{O}_C(-1), E) \otimes \mathcal{O}_C(-1) \to E\big), \tag{4.2}$$

where the arrow is the obvious evaluation map. (Taking E to be a complex of injectives, this map is canonical rather than defined up to homotopy, so the cone turns out to be functorial here [30].) Compare its mirror (2.3, 2.4).

More generally, the simultaneous resolution of the family of ALE surfaces (2.22) has central fibre the minimal resolution of

$$\{xy = t^k\} \subseteq \mathbb{C}^3. \tag{4.3}$$

Call this S, with its A_{k-1}-chain of exceptional -2-curves $C_i \subset S$ (the limit of an A_{k-1}-chain of Lagrangian vanishing cycles L_i on a general fibre). In fact the sheaves $A_i := \mathcal{O}_{C_i}(-1)$ satisfy the following homological definition of an A_{k-1}-chain in any derived category of coherent sheaves.

Definition 4.4. [30] *Objects $A_i \in D(S)$, $i = 1, \ldots, k-1$ form an A_{k-1}-chain of n-spherical objects if for all i, j,*

- $\mathrm{Ext}^*(A_i, A_i) \cong H^*(S^n, \mathbb{C})$,
- $A_i \otimes \omega_S \cong A_i$,
- $\bigoplus_p \mathrm{Ext}^p(A_i, A_j) = \begin{cases} \mathbb{C} & |i-j| = 1, \\ 0 & |i-j| > 1. \end{cases}$

For us $n = 2$, and the second, Calabi-Yau condition always holds since the canonical bundle ω_S of S is trivial. One can then define the Dehn twists about the A_i as in (4.2) by

$$T_{A_i} E := \mathrm{Cone}\big(R\,\mathrm{Hom}(A_i, E) \otimes A_i \to E\big), \tag{4.5}$$

or by Fourier-Mukai transform with the kernel

$$\mathrm{Cone}\big(A_i^\vee \boxtimes A_i \to \mathcal{O}_\Delta\big).$$

Here $^\vee$ denotes derived dual, and the arrow is restriction to the diagonal followed by evaluation (trace).

Theorem 4.6. [15, 30] *If the A_i form an A_{k-1}-chain then the $T_i = T_{A_i}$ define a (weak) faithful action of the braid group $B_k \hookrightarrow \mathrm{Aut}(D(S))$.*

In particular, the T_i are invertible and satisfy the braid relations

$$\begin{array}{rcll} T_i T_j T_i & \cong & T_j T_i T_j, & |i-j| = 1, \\ T_i T_j & \cong & T_j T_i, & |i-j| > 1. \end{array}$$

So our putative mirrors of Dehn twists really satisfy the same relations as the original twists (2.7). And we have put things in a more categorical framework, allowing twists around arbitrary spherical objects, as mirror symmetry suggests should be possible – according to Kontsevich's conjecture, all mirror symmetry needs to see is categorical properties, rather than specific geometry. For more on mirror symmetry for ALE surfaces see [13].

4.2. Higher dimensions. Our other examples of families over C_k^0 fit into a similar hyperkähler mirror symmetry picture. In fact they all follow from the case of the space of matrices of Section 3.2 by taking slices. In much the same way as described above, the family Dehn twists around the divisors N_i limit to the Fourier-Mukai transforms with kernels the structure sheaves of the limits $\Delta \cup (N_i \times_{Fl_i} N_i)$ (3.6) of the graphs of these symplectomorphisms. Up to twisting by a line bundle, these are the relative versions of the derived category Dehn twist (4.5), with action

$$E \mapsto \operatorname{Cone}\big(\iota_{i*}p_i^*p_{i*}\iota_i^! E \to E\big).$$

Here the arrow is evaluation, and p_i and ι_i are the obvious maps

$$\begin{array}{c} N_i \overset{\iota_i}{\hookrightarrow} T^*Fl \\ \downarrow{\scriptstyle p_i} \\ T^*Fl_i. \end{array} \tag{4.7}$$

Again these define autoequivalences $T_i\colon D(T^*Fl) \to D(T^*Fl)$ which satisfy the braid relations [1, 16]. In fact the T_i (both here and on the slices S of the last section) even admit natural transformations between them which satisfy the relations of the braid cobordism category, and these give rise to maps between the Khovanov cohomology groups of links of the next Section, when we fix a link cobordism. But we refer to [16] for this further extension of mirror symmetry.

The braid relations in this case are *much* harder than those in 2 dimensions. But Manolescu's isomorphism means that they follow from the simple two dimensional case for the spaces relevant to Khovanov cohomology.

5. Hilbert Schemes of ALE Spaces and Khovanov Cohomology

By now it should be clear how one would go about trying to mirror the Seidel-Smith construction to define Khovanov cohomology in a derived category of coherent sheaves. There is a slice of (3.4) that provides a simultaneous resolution of (the basechange of) SS_{2m}. The derived category of its central fibre carries a braid group action and a complex Lagrangian submanifold that $\mathcal{L}$ (2.23) limits to. Taking its structure sheaf (and possibly twisting by a line bundle) as an object of the derived category, one would like to show that the Exts from this object to its image under a braid give an invariant of the link closure of the braid.

Such a programme has been carried out in beautiful work of Cautis and Kamnitzer [5]. In fact they use a compactification of the above space related, via the geometric Satake correspondence, to the $sl(2)$ representations of the Reshetikhin-Turaev tangle calculus. This has the huge advantage of being generalisable to other Lie algebras [6]. However, as mentioned above, it is also hard work, involving calculations in high dimensions.

Manolescu's isomorphism suggests we might work with something like the Hilbert scheme of points on $S = S_{2m-1}$, the minimal resolution of the A_{2m-1}-singularity (4.3). This reduces most of the work to much simpler calculations with sheaves on the surfaces S_{2m-1}. In fact, by [4, 9] the category

$$D_m := D(\mathrm{Hilb}^m\, S_{2m-1})$$

has a canonical identification with the Σ_m-equivariant derived category of $(S_{2m-1})^m$, where the symmetric group Σ_m permutes the factors:

$$D(\mathrm{Hilb}^m\, S_{2m-1}) \;\cong\; D(S^m_{2m-1})^{\Sigma_m}. \tag{5.1}$$

5.1. However. One would expect the right hand side of (5.1) to be mirror to the Σ_m-equivariant Fukaya category of $S_{\underline{\lambda}}$ (2.19), which is *not* the Fukaya category of its Hilbert scheme, but can be thought of as playing the role of the Fukaya category of the singular symplectic space $\mathrm{Sym}^m\, S_{2m-1}$. Considering the Hilbert scheme as a symplectic deformation of this (subtracting a small multiple of the exceptional divisor of Hilb $\to$ Sym from the degenerate symplectic form one gets by pulling back from Sym) suggests the mirror might be a *deformation* of $\mathrm{Hilb}^m\, S_{2m-1}$. We will indeed use such a deformation related to the exceptional divisor.

This is an example where our naive description of mirror symmetry fails. The mirror of the smoothing $\mathrm{Hilb}^m\, S_{\underline{\lambda}}$ of a hyperkähler singularity $\mathrm{Hilb}^m\, S_{\underline{0}}$ appears not to be the obvious choice $\mathrm{Hilb}^m\, S_{2m-1}$ (which is birational to a resolution of $\mathrm{Hilb}^m\, S_{\underline{0}}$) but a *deformation* thereof.

5.2. The construction. Any $E \in D(S^m_{2m-1})$ defines an element

$$\Sigma_m.E := \bigoplus_{\sigma\in\Sigma_m} \sigma^* E \;\in\; D(S^m_{2m-1})^{\Sigma_m}, \tag{5.2}$$

with its obvious Σ_m-linearisation. Thus from the spherical objects $L_i := \mathcal{O}_{C_i}(-1) \in D(S_{2m-1})$ we define

$$\mathcal{L} = \mathcal{L}_m := \Sigma_m.(L_1 \boxtimes L_3 \boxtimes \ldots \boxtimes L_{2m-1}) \in D(S^m_{2m-1})^{\Sigma_m}. \tag{5.3}$$

Equivalently, the object $\mathcal{L} \in D(\mathrm{Hilb}^m\, S_{2m-1})$ can be described via the Haiman-BKR equivalence (5.1) as follows. As in (2.23), the composition

$$C_1 \times C_3 \times \ldots \times C_{2m-1} \hookrightarrow S^m_{2m-1} \to \mathrm{Sym}^m\, S_{2m-1} \dashrightarrow \mathrm{Hilb}^m\, S_{2m-1}$$

is an embedding since the C_{2i-1} do not intersect each other so their product avoids the diagonal locus over which the last map is not regular. Then

$$\mathcal{L} = \mathcal{O}_{C_1\times C_3\times\ldots\times C_{2m-1}}(-1,-1,\ldots,-1) \in D(\mathrm{Hilb}^m\, S_{2m-1}). \tag{5.4}$$

Any autoequivalence $T \in \mathrm{Aut}(D(S_{2m-1})$ induces a canonical autoequivalence $\Phi(T) \in \mathrm{Aut}(D(S^m_{2m-1})^{\Sigma_m})$ [20]. Its action on objects of the form (5.3) is the obvious one:

$$\Phi(T)\Big(\Sigma_m.(E_1 \boxtimes \ldots \boxtimes E_m)\Big) \;=\; \Sigma_m.\big(T(E_1) \boxtimes \ldots \boxtimes T(E_m)\big). \tag{5.5}$$

We apply this to the spherical twists $T_i := T_{\mathcal{O}_{C_i}(-1)}$:

$$\mathsf{T}_i := \Phi(T_i)[1] \;\in\; \mathrm{Aut}\left(D(S^m_{2m-1})^{\Sigma_m}\right). \tag{5.6}$$

Since Φ is a homomorphism, these define generators of a braid group action $B_{2m} \to \mathrm{Aut}(D_m)$. (The braid relations are homogeneous, so the extra shift [1] makes no difference.) Thus any $\beta \in B_{2m}$ gives an autoequivalence $\mathsf{T}_\beta \in \mathrm{Aut}(D_m)$. We define the braid invariant

$$kh^*(\beta) := \mathrm{Ext}^*_{D_m}(\mathcal{L}, \mathsf{T}_\beta\, \mathcal{L}[m]). \tag{5.7}$$

The shifts in the definitions (5.6, 5.7) match with the shift $w+m$ in the mirror Seidel-Smith construction (2.24).

5.3. Maps between ALE spaces. To study the dependence of (5.7) on m we will need the holomorphic analogue (or hyperkähler rotation) of the symplectic maps between ALE spaces of Section 2.7. So let S_{k-1} be the minimal resolution of $A_{k-1} := \{x^k = yz\} \subset \mathbb{C}^3$. We will exhibit a natural inclusion $S_{k-1} \subset S_k$ taking the A_{k-1}-chain of -2-curves $C_i \cong \mathbb{P}^1,\ i = 1, \ldots, k-1$ in the former to the first $k-1$ curves of the A_k-chain $C_1, \ldots, C_{k-1}, C_k$ in the latter.

Consider the blow up of $\mathbb{C}^2$ in the ideal (x^k, y). Call this $\overline{A}_{k-1}$. It can be constructed inductively via blow ups and a blow down in smooth centres:

1. Blow up the origin in $\mathbb{C}^2$, giving an exceptional divisor $E_1 \cong \mathbb{P}^1$.

2. Blow up the point $\infty \in E_1$ (its intersection with the proper transform of the x-axis). We get a new exceptional divisor E_2, and the proper transform of E_1 which is a -2-curve C_1.

(r) At the rth stage, blow up $\infty \in E_{r-1}$ to produce a new exceptional divisor E_r, and the proper transform of E_{r-1} is a -2-curve C_r.

After the kth step we get a surface $\overline{S}_{k-1}$ with an A_{k-1}-chain of -2-curves C_i and a -1-curve E_k; see Figure 9. Now blow down the $C_i,\ i = 1, \ldots, k-1$ to get $\overline{A}_{k-1}$.

Now $\overline{A}_{k-1} = \mathrm{Bl}_{(x^k,y)}\,\mathbb{C}^2 = \{\mu x^k = \lambda y\} \subset \mathbb{C}^2_{x,y} \times \mathbb{P}^1_{[\lambda:\mu]}$. Therefore if we remove the proper transform $\overline{\{y=0\}} = \{\mu = 0\}$ of the x-axis we can set $[\lambda : \mu] = [z : 1]$ to get the affine variety

$$\{x^k = yz\} \subset \mathbb{C}^2_{x,y} \times \mathbb{C}_z,$$

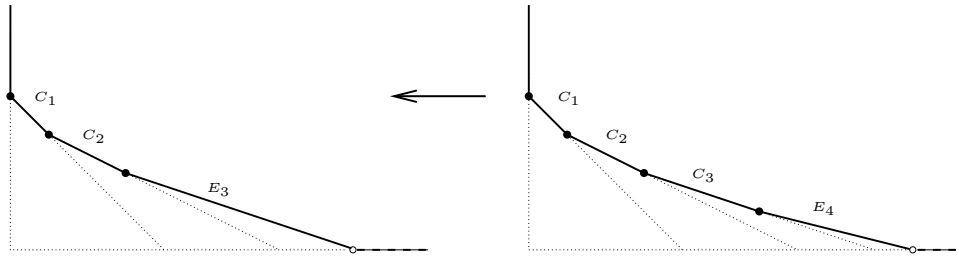


Figure 9. Newton polygon diagram of the blow up map $\overline{S}_2 \leftarrow \overline{S}_3$. On removing the divisors corresponding to the dashed lines (the proper transforms of the x-axis) we get an inclusion $S_2 \subset S_3$ in the opposite direction.

which is precisely A_{k-1}. Thus $\overline{A}_{k-1}$ and $\overline{S}_{k-1}$ are partial compactifications of A_{k-1} and S_{k-1} respectively (since $\overline{S}_{k-1}$ is the minimal resolution of S_{k-1}).

We obtained $\overline{S}_k$ from $\overline{S}_{k-1}$ by blowing up the latter in the point $\infty \in E_k$. But $\infty = \overline{\{y=0\}} \cap E_k$ lies in the divisor $\overline{\{y=0\}}$ that we remove from $\overline{S}_{k-1}$ to get S_{k-1}, so the inclusion $S_{k-1} \subset \overline{S}_{k-1}$ lifts to the blow up: $S_{k-1} \subset \overline{S}_k$. Its image is clearly contained in the open subset S_k, and maps the curves $C_i \subset S_{k-1}$ to the corresponding curves $C_i \subset S_k$, as claimed.

As in Section 2.8, to prove that kh^* is a link invariant under plait closure it is sufficient to prove the following; again see Figure 5.

1. $\mathsf{T}_1 \mathcal{L} \cong \mathcal{L}$,
2. $\mathsf{T}_{2i-1}\mathsf{T}_{2i}\,\mathcal{L} \cong \mathsf{T}_{2i-1}^{-1}\mathsf{T}_{2i}^{-1}\mathcal{L}$,
3. $\mathsf{T}_{2i}\mathsf{T}_{2i-1}\mathsf{T}_{2i+1}\mathsf{T}_{2i}\,\mathcal{L} \cong \mathcal{L}$, and
4. $\mathrm{Ext}^*_{D_m}(\mathcal{L}_m, \mathsf{T}_\beta\,\mathcal{L}_m[m]) \cong \mathrm{Ext}^*_{D_{m+1}}(\mathsf{T}_{2m}^{\pm 1}\mathcal{L}_{m+1}, \mathsf{T}_\beta\,\mathcal{L}_{m+1}[m+1])$.

In the last relation we use the inclusion $S_{2m-1} \hookrightarrow S_{2m+1}$ exhibited above.

Theorem 5.8. [32] *The relations (1), (3) and (4) hold in the categories D_m, but (2) does not.*

The proof is reduced by (5.5) to simple computations in $D(S_{2m-1})$ mirroring those of Section 2.8.

Firstly, $T_1 L_{2i+1} \cong L_{2i+1}$ for $i \geq 1$ by (4.2), because $\mathrm{Ext}^*(L_1, L_{2i+1}) = 0$. Since $\mathrm{Ext}^*(L_1, L_1) \cong H^*(S^2, \mathbb{C})$ we get the exact triangle in $D(S_{2m-1})$

$$L_1 \oplus L_1[-2] \longrightarrow L_1 \longrightarrow T_1 L_1.$$

The first map is the identity on the first factor, so $T_1 L_1 \cong L_1[-1]$. Therefore by (5.5), $\mathsf{T}_1\mathcal{L} \cong \mathcal{L}[-1][1] = \mathcal{L}$, which proves relation (1).

For (2) we note the following calculation on S_{2m-1}. If $A, B \cong \mathbb{P}^1$ are (possibly reducible) rational curves in S_{2m-1} intersecting in a single transverse point, then $\mathrm{Ext}^*(\mathcal{O}_A, \mathcal{O}_B) = \mathbb{C}[-1]$ and the resulting exact triangle

$$\mathcal{O}_B \to T_{\mathcal{O}_A}\mathcal{O}_B \to \mathcal{O}_A \tag{5.9}$$

expresses $T_{\mathcal{O}_A}\mathcal{O}_B$ as the nontrivial extension

$$T_{\mathcal{O}_A}\mathcal{O}_B \cong \mathcal{O}_{A\cup B}(1,0). \tag{5.10}$$

By $(1,0)$ we mean to twist by the line bundle which is the gluing of the trivial bundle on A and the degree 1 bundle $\mathcal{O}_B(A\cap B)$ on B. (A similar result to (5.10) holds when $\mathcal{O}_A, \mathcal{O}_B$ and $T_{\mathcal{O}_A}\mathcal{O}_B$ are all twisted by the same line bundle.) If we denote this extension by $\mathcal{O}_B\#\mathcal{O}_A$ it is the mirror of the Lagrangian connect sum of (2.27, 2.28). So, for instance, we picture

$$T_iL_{i-1} = L_{i-1}\#L_i = \mathcal{O}_{C_{i-1}\cup C_i}(0,-1)$$

as the path in $\mathbb{C}$, running from λ_{i-1} *over* λ_i to λ_{i+1}, over which its mirror is S^1-fibred. Similarly the connect sum in the opposite direction, which is $T_{i-1}L_i = \mathcal{O}_{C_{i-1}\cup C_i}(-1,0)$, corresponds to the path *under* λ_i. (See [34] for more on these pictures for objects of $D(S_{2m-1})$.)

Applying this twice we find that

$$T_{2i-1}T_{2i}L_{2i+1} = T_{2i-1}\mathcal{O}_{C_{2i}\cup C_{2i+1}}(-1,0) = \mathcal{O}_{C_{2i-1}\cup C_{2i}\cup C_{2i+1}}(-1,0,0), \tag{5.11}$$

the second equality following from (5.10) applied to $A = C_{2i-1}$ and $B = C_{2i}\cup C_{2i+1}$. Similarly,

$$T_{2i-1}^{-1}T_{2i}^{-1}L_{2i+1} \cong \mathcal{O}_{C_{2i-1}\cup C_{2i}\cup C_{2i+1}}(0,0,-1). \tag{5.12}$$

Finally $T_{2i-1}T_{2i}$ and $T_{2i-1}^{-1}T_{2i}^{-1}$ both take L_{2i-1} to L_{2i}, by similar calculations mirroring Figure 6, and they leave L_{2j+1} alone for $j \neq i, i-1$.

Since (5.11) and (5.12) are *not* isomorphic it follows from (5.5) that $\mathsf{T}_{2i-1}\mathsf{T}_{2i}\,\mathcal{L} \ncong \mathsf{T}_{2i-1}^{-1}\mathsf{T}_{2i}^{-1}\mathcal{L}$, i.e. (2) does not hold.

Repeated calculations with (5.10) on S_{2m-1} show that $T_{2i}T_{2i-1}T_{2i+1}T_{2i}$ also leaves L_{2j+1} alone for $j \neq i, i-1$, but swaps $L_{2i\pm1}$ (see Figure 6):

$$T_{2i}T_{2i-1}T_{2i+1}T_{2i}L_{2i\pm1} = L_{2i\mp1}.$$

Relation (3) then follows again from (5.5).

Finally (4) follows just as in the mirror situation of Section 2.8. The exact triangle (2.26) holds just as well in $D(S_{2m-1})$ – see (5.9) – giving the splitting

$$\begin{aligned}\mathrm{Ext}^*(L_{2m-1}, A) &\cong \mathrm{Ext}^{*+1}(L_{2m}, A) \oplus \mathrm{Ext}^*(T_{2m}L_{2m-1}, A)\\ &\cong \mathrm{Ext}^{*+1}(T_{2m}L_{2m+1}, A) \oplus \mathrm{Ext}^*(T_{2m}L_{2m-1}, A)\end{aligned}$$

that replaces (2.25) for any $A \in D(S_{2m-1})$ generated by L_i, $i \leq 2m-1$. Relation (4) follows easily; see [32] for full details.

5.4. Deformation. As suggested in Section 5.1, to get something which acts as a better mirror of $\mathrm{Hilb}^m S_{\underline{\lambda}}$ in which relation (2) holds, we should deform by something concentrated on the diagonal.

The exceptional divisor E of $H_m := \mathrm{Hilb}^m(S_{2m-1}) \to \mathrm{Sym}^m(S_{2m-1})$ has a class $[E] \in H^1(\Omega_{H_m})$ despite the noncompactness. (For instance the exact sequence $0 \to \Omega_{H_m} \to \Omega_{H_m}(\log D) \to \mathcal{O}_D \to 0$ has extension class in $\mathrm{Ext}^1(\mathcal{O}_D, \Omega_{H_m})$; its image in $\mathrm{Ext}^1(\mathcal{O}_{H_m}, \Omega_{H_m}) = H^1(\Omega_{H_m})$ is $[E]$.) Via the holomorphic symplectic form $\Omega_{H_m} \cong T_{H_m}$, and we get a canonical class $e \in H^1(T_{H_m})$, the space of first order deformations of H_m.

Using some twistor theory we get a canonical family $\mathcal{H} \to \mathbb{P}^1$ of holomorphic symplectic deformations $(\mathcal{H}_t, \sigma_t)$ of $H_m = \mathcal{H}_0$ in the direction of e; see [32]. The Lagrangian $\mathcal{L}$ (5.4) deforms along this deformation because it is disjoint from $[E]$. We show in [32] that the functors T_β also deform. Both sides of the relations (1), (3) and (4) therefore also deform along $\mathcal{H}$, and by rigidity of the complexes involved the equalities continue to hold.

Finally then we come to (2). As in (5.11, 5.12) we have (cf. (2.27, 2.28)),

$$\begin{aligned} T_{2i-1}T_{2i}L_{2i+1} &\cong L_{2i+1}\#L_{2i}\#L_{2i-1}, \\ T_{2i-1}^{-1}T_{2i}^{-1}L_{2i+1} &\cong L_{2i-1}\#L_{2i}\#L_{2i+1}, \end{aligned}$$

are extensions of the same objects but in opposite directions. On deforming $\mathrm{Hilb}^m S_{2m-1}$ along $\mathcal{H}$, the Lagrangians $T_{2i-1}^{\pm 1}T_{2i}^{\pm 1}\mathcal{L}$ deform because both $\mathcal{L}$ and the symplectomorphisms T_i do. However the pieces $L_{2i\pm 1} \times L_{2i} \times \ldots$ of the extensions do *not* deform (essentially because $[E]$ restricts to a nonzero class on their support since $L_{2i\pm 1}$ intersects L_{2i} inside S_{2m-1}). For general reasons, if two extensions of the same pieces deform while the pieces do not then the deformations of the extensions become isomorphic.

The baby model to keep in mind is to deform S itself so that $[C_1]$ and $[C_2]$ do not remain of type $(1,1)$, but their sum $[C_1] + [C_2]$ does. Then neither of $\mathcal{O}_{C_1}(-1)$ or $\mathcal{O}_{C_2}(-1)$ deform, but their extensions in different directions,

$$\mathcal{O}_{C_1 \cup C_2}(0,-1) \quad \text{and} \quad \mathcal{O}_{C_1 \cup C_2}(-1,0)$$

both deform and become isomorphic to $\mathcal{O}_C(-1)$, where C is the unique (smooth) rational curve that degenerates back to $C_1 \cup C_2$ on the central fibre.

5.5. Bigrading and Khovanov cohomology. There are also $\mathbb{C}^*$-actions on the spaces S_i with respect to which the inclusion maps $S_{k-1} \subset S_k$ are equivariant [32]. Since the constructions of this paper are equivariant with respect to this $\mathbb{C}^*$-action, we get extra $\mathbb{C}^*$-action, and so a bigrading, on the link invariant kh^*.

Finally, using the method of [5], one can show that the resulting $kh^{*,*}$ is in fact Khovanov cohomology $KH^{*,*} \otimes \mathbb{C}$ (up to a shift in bigrading). Building up a link from standard cobordisms one presents both $kh^{*,*}$ and $KH^{*,*} \otimes \mathbb{C}$

as iterated cones on the same standard pieces. (For $KH^{*,*}$ this is Khovanov's famous "cube of resolutions".) Because of some vanishing of Ext groups, this iterated cone is unique.

References

[1] R. Bezrukavnikov, I. Mirković, and D. Rumynin, *Singular localization and intertwining functors for reductive Lie algebras in prime characteristic.* Nagoya Math. Jour. **184**, 1–55, 2006. math.RT/0602075.

[2] S. Bigelow, *A homological definition of the Jones polynomial* Geom. & Top. Monogr. **4**, 29–41, 2002. math.GT/0201221.

[3] J. Birman, *On the stable equivalence of plat representations of knots and links*, Canad. Jour. Math. **28**, 264–290, 1976.

[4] T. Bridgeland, A. King and M. Reid, *The McKay correspondence as an equivalence of derived categories*, Jour. A.M.S. **14**, 535–554, 2001. math.AG/9908027.

[5] S. Cautis and J. Kamnitzer, *Knot homology via derived categories of coherent sheaves I, $sl(2)$-case*, Duke Math. Jour. **142**, 511–588, 2008. math.AG/0701194.

[6] S. Cautis and J. Kamnitzer, *Knot homology via derived categories of coherent sheaves II, $sl(m)$ case*, Invent. Math. **174**, 165–232, 2008. arXiv:0710.3216.

[7] K. Fukaya, *Mirror symmetry of Abelian variety and multi theta functions*, Jour. Alg. Geom. **11**, 393–512, 2002.

[8] K. Fukaya, Y.-G. Oh, H. Ohta, and K. Ono, *Lagrangian intersection Floer theory: anomaly and obstruction*, AMS/IP Stud. in Adv. Math., **46**, 2009.

[9] M. Haiman, *Hilbert schemes, polygraphs and the Macdonald positivity conjecture*, Jour. A.M.S. **14**, 941–1006, 2001. math.AG/0010246.

[10] N. Hitchin, *The self-duality equations on a Riemann surface*, Proc. London Math. Soc. **55**, 59–126, 1987.

[11] J. Hurtubise, *Integrable systems and algebraic surfaces*, Duke Math. Jour. **83**, 19-50, 1996.

[12] D. Huybrechts, *Compact hyper-Kähler manifolds: basic results*, Invent. Math. **135**, 63–113, 1999.

[13] A. Ishii, K. Ueda and H. Uehara, *Stability conditions on A_n-singularities*, math.AG/0609551.

[14] M. Khovanov, *A categorification of the Jones polynomial.* Duke Math. Jour., **101**, 359-426, 2000. math.QA/9908171.

[15] M. Khovanov and P. Seidel, *Quivers, Floer cohomology, and braid group actions*, J. Amer. Math. Soc. **15**, 203–271, 2002. math.QA/0006056.

[16] M. Khovanov and R. P. Thomas, *Braid cobordisms, triangulated categories, and flag varieties*, Homology, Homotopy and Applications **9**, 19–94, 2007. math.AG/0609335.

[17] M. Kontsevich, *Homological Algebra of Mirror Symmetry*, International Congress of Mathematicians, Zürich 1994. Birkhäuser, 1995. alg-geom/9411018.

[18] C. Manolescu, *Nilpotent slices, Hilbert schemes, and the Jones polynomial*, Duke Math. Jour. **132**, 311–369, 2006. math.SG/0411015.

[19] T. Perutz, *Lagrangian matching invariants for fibred four-maniolds I*, Geom. & Top. **11**, 759–828, 2007. math.SG/0606061.

[20] D. Ploog, *Equivariant autoequivalences for finite group actions*, Adv. in Math. **216**, 62–74, 2007. math.AG/0508625.

[21] P. Seidel, unpublished notes, 1997.

[22] P. Seidel, *Graded Lagrangian submanifolds*, Bull. Soc. Math. France. **128**, 103–146, 2000. math.SG/9903049.

[23] P. Seidel, *A long exact sequence for symplectic Floer cohomology*. Topology **42**, 1003–1063, 2003. math.SG/0105186.

[24] P. Seidel, *Lectures on four-dimensional Dehn twists*, In "Symplectic 4-manifolds and algebraic surfaces", 231–267, LNM **1938**, Springer, 2008. math.SG/0309012.

[25] P. Seidel, *Fukaya categories and Picard-Lefschetz theory*, Zürich Lectures in Advanced Mathematics, EMS, 2008.

[26] P. Seidel, *Suspending Lefschetz fibrations, with an application to local mirror symmetry*, to appear in Comm. Math. Phys., 2010. arXiv:0907.2063.

[27] P. Seidel and I. Smith, *A link invariant from the symplectic geometry of nilpotent slices*, Duke Math. Jour. **134**, 453–514, 2006. math.SG/0405089.

[28] P. Seidel and I. Smith, *Localization for involutions in Floer cohomology*, arXiv:1002.2648.

[29] P. Seidel and I. Smith, *Symplectic geometry of the adjoint quotient, I & II*, Lectures, MSRI, April 2004.

[30] P. Seidel and R. P. Thomas, *Braid group actions on derived categories of sheaves*, Duke Math. Jour. **108**, 37–108, 2001. math.AG/0001043.

[31] I. Smith, *Quadrics, quilts and representation varieties*, in preparation.

[32] I. Smith and R. P. Thomas, *Khovanov homology from ALE spaces*, preprint.

[33] R. P. Thomas, *Moment maps, monodromy and mirror manifolds*, Symplectic geometry and mirror symmetry (Seoul, 2000), 467–498, World Sci. Publishing, 2001.

[34] R. P. Thomas, *Stability conditions and the braid group*, Comm. Anal. Geom. **14**, 135–161, 2006. math.AG/0212214.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Invariants Entiers en Géométrie Énumérative Réelle

Jean-Yves Welschinger *

Résumé

Je rappelle les divers problèmes de géométrie énumérative réelle desquels j'ai pu extraire des invariants à valeurs entières, fournissant un pendant réel aux invariants de Gromov-Witten. Je discute l'optimalité des bornes inférieures fournies par ces invariants ainsi que certaines de leurs propriétés arithmétiques. Je présente enfin davantage de résultats garantissant la présence ou l'absence de disques pseudo-holomorphes à bord dans une sous-variété lagrangienne d'une variété symplectique donnée.

Mathematics Subject Classification (2010). Primary 53D45 ; Secondary 14N35.

Keywords. Enumerative geometry, rational curve, real algebraic variety, holomorphic discs.

Introduction

Le nombre de racines complexes d'un polynôme générique à une variable de degré d ne dépend pas du choix du polynôme et vaut d, tandis que lorsque ce polynôme est à coefficients réels, le nombre de ses racines réelles peut prendre toutes les valeurs de même parité que d comprises entre 0 et d. Ceci tient au fait que le corps des nombres complexes est algébriquement clos au contraire du corps des nombres réels. Bien plus généralement, le nombre de solutions d'un « système de n équations génériques »sur une variété projective complexe lisse de dimension n ne dépend que du degré de ces équations, alors qu'il dépend fortement du choix, même générique, de ces équations lorsqu'elles sont à coefficients réels et considérées sur le lieu réel d'une variété algébrique réelle. (En fait d'équations, il conviendrait plutôt de parler de sections génériques de n fibrés

*Je remercie le Centre national de la recherche scientifique ainsi que l'Agence nationale de la recherche pour leurs soutiens sans lesquels je n'aurais pu réaliser ces travaux.
Université de Lyon; CNRS; Université Lyon 1; Institut Camille Jordan.
E-mail: welschinger@math.univ-lyon1.fr.

en droites holomorphes disons très amples). Chaque problème de géométrie énumérative réelle peut en principe s'interpréter de cette manière. La variété projective réelle est l'espace des modules des objets géométriques que l'on veut compter et les équations proviennent des conditions d'incidences que l'on impose à ces objets.

Le principal phénomène présenté dans cet article de synthèse est le suivant : il est parfois possible de compter ces objets géométriques réels en fonction d'un signe $\pm$ de manière à extraire un entier indépendant du choix générique des conditions d'incidence. Dans le premier paragraphe, nous observons ce phénomène en comptant les courbes J-holomorphes rationnelles réelles dans une variété symplectique réelle de dimension quatre en fixant leur classe d'homologie et leur imposant de passer par un nombre adéquat de points réels ou bien complexes conjugués. Nous utilisons en effet le langage de la géométrie symplectique pour étudier ces problèmes énumératifs, tenant compte des résultats de M. Gromov [9] selon lesquels le caractère algébrique des variétés ne joue aucun rôle dans ces problèmes énumératifs, seule l'ellipticité de l'opérateur de Cauchy-Riemann sous-jacent intervient. Les entiers que l'on extrait de ce problème énumératif fournissent un invariant par déformation des variétés symplectiques réelles de dimension quatre (X, ω, c_X), qui prend la forme d'une fonction $\chi : d \in H_2(X;\mathbb{Z}) \mapsto \chi^d[T] \in \mathbb{Z}[T_1, \ldots, T_N]$ où N désigne le nombre de composantes connexes du lieu réel $\mathbb{R}X$ de la variété. On définit des invariants analogues pour les variétés symplectiques de dimension six « fortement semi-positives », par exemple positives, dans le troisième paragraphe et en incluant des conditions de tangence à une courbe réelle dans le deuxième. Ces derniers résultats s'appliquent en particulier à un problème classique de géométrie énumérative, le comptage des coniques tangentes à cinq coniques génériques données. Le nombre de solutions complexes vaut 3264, un résultat établi par de Joncquières au milieu du dix-neuvième siècle. On montre que le nombre de solutions réelles se trouve minoré par trente-deux lorsque les coniques réelles bordent cinq disques disjoints par exemple. En effet, la valeur absolue des invariants entiers que l'on introduit dans cet exposé borne inférieurement le nombre de solutions réelles du problème énumératif que l'on considère.

Un deuxième phénomène apparaît dans cet article, l'optimalité de ces bornes inférieures. On montre en effet dans le premier paragraphe que dans le cas des variétés symplectiques réelles de dimension quatre, lorsque le lieu réel possède une sphère, un tore ou bien, sous des conditions plus restrictives, un plan projectif réel et lorsqu'au plus un point est choisi réel et dans cette composante, il existe une structure presque complexe générique J pour laquelle le nombre de courbes J-holomorphes rationnelles réelles satisfaisant nos conditions d'incidence vaut exactement la valeur absolue de notre invariant, ceci quelle que soit la classe d'homologie de ces courbes rationnelles. Ce résultat vaut également pour la quadrique ellipsoïde de dimension trois, comme établi dans le troisième paragraphe. Cette optimalité est établie à l'aide de méthodes issues de la théorie symplectique des champs, méthodes qui nous permettent également parfois de

calculer le signe de notre invariant, d'établir des congruences satisfaites par ce dernier ainsi que de fournir des formules le calculant dans certains cas, calculs que l'on mène explicitement en bas degrés. Tous ces résultats font l'objet du premier paragraphe de cet article. En utilisant la notion d'involution antibirationnelle sur une variété symplectique de dimension quatre, on montre de la même manière dans le quatrième paragraphe l'existence de disques J-holomorphes à bords dans le tore de Clifford et satisfaisant des conditions d'incidences ponctuelles. Dans le cas d'une sphère lagrangienne dans une variété symplectique négative ou nulle, on montre au contraire dans ce même paragraphe, pour tout $E > 0$, l'existence de structures presque-complexes J pour lesquelles aucun disque ou membrane J-holomorphe d'énergie inférieure à E ne repose sur cette sphère, un résultat analogue à nos résultats d'optimalités puisqu'on atteint ainsi le minimum possible du nombre de disques ou membranes J-holomorphes. Remarquons à propos que l'obtention d'invariants entiers ou rationnels à partir du comptage des disques J-holomophes à bords dans une sous-variété lagrangienne est un problème classique de géométrie symplectique (et de la théorie des cordes ouvertes en physique théorique) pour lequel peu de solutions existent. Notre approche en fournit une lorsque la lagrangienne est fixée par une involution antiholomorphe. Remarquons également que l'absence de disques J-holomorphes pour certaines structures permet de définir l'homologie de Floer pour des sphères lagrangiennes dans les variétés symplectiques à première classe de Chern nulle, un autre problème classique de géométrie symplectique (et de symétrie miroir en physique théorique).

Le présent article est largement issu de mon mémoire d'habilitation à diriger des recherches, laquelle fut soutenue à l'École normale supérieure de Lyon en mars 2008.

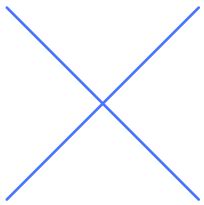
1. Invariants Énumératifs des Variétés Symplectiques Réelles de Dimension Quatre

1.1. Définition des invariants. Soit (X, ω, c_X) une *variété symplectique réelle* fermée de dimension quatre, par quoi on entend une variété symplectique fermée de dimension quatre (X, ω) équipée d'une involution c_X satisfaisant la relation $c_X^*\omega = -\omega$. Le lieu fixe $\mathbb{R}X$ de cette involution est supposé ici non-vide, c'est le *lieu réel* de la variété, lequel a la propriété d'être lagrangien. Ses composantes connexes sont étiquetées $(\mathbb{R}X)_1, \ldots, (\mathbb{R}X)_N$. On note $\mathcal{J}_\omega$ l'espace des structures presque-complexes ω-positives de (X, ω) de classe C^l, $l \gg 1$ et $\mathbb{R}\mathcal{J}_\omega \subset \mathcal{J}_\omega$ le sous-espace des structures J qui rendent l'involution c_X J-antiholomorphe. Ce sont tous deux des variétés de Banach séparables non-vides et contractiles.

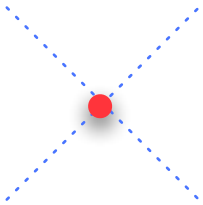
Soit $d \in H_2(X; \mathbb{Z})$ une classe d'homologie satisfaisant la relation $(c_X)_* d = -d$ et $J \in \mathbb{R}\mathcal{J}_\omega$ une structure presque-complexe générique. Les *courbes J-holomorphes rationnelles réelles* homologues à d, c'est-à-dire les sphères J-

holomorphes invariantes par c_X et homologues à d, forment alors un espace de dimension $c_1(X)d-1$, où $c_1(X)$ désigne la première classe de Chern de la variété (X,ω). Nous supposons cette dimension positive ou nulle, puisque le cas contraire signifie que l'espace en question est vide, puis faisons chuter cette dimension à zéro en imposant quelques contraintes à ces courbes, à savoir de passer par une collection $\underline{x}$ de $c_1(X)d-1$ points distincts. Ces derniers peuvent être choisis réels, c'est-à-dire fixés par c_X, ou bien complexes conjugués, c'est-à-dire échangés par c_X ; nous noterons r_i le nombre de points réels choisis dans $(\mathbb{R}X)_i$, $i \in \{1,\ldots,N\}$, et r_X le nombre de paires de points complexes conjugués, de sorte que $2r_X + \sum_{i=1}^{N} r_i = c_1(X)d-1$. L'ensemble $\mathcal{R}_d(\underline{x},J)$ des courbes J-holomorphes rationnelles réelles homologues à d qui satisfont ces contraintes supplémentaires est fini. Ces courbes sont de plus toutes irréductibles, immergées et n'ont que des points doubles transverses comme singularités. Remarquons que le cardinal $R_d(\underline{x},J) = \#\mathcal{R}_d(\underline{x},J)$ dépend en général des choix auxiliaires de la structure presque complexe et de la configuration de points, essentiellement parce que le corps des réels n'est pas algébriquement clos. Nous allons montrer qu'il en devient indépendant lorsque l'on compte ces courbes en fonction d'un signe convenablement choisi.

Soit $C \in \mathcal{R}_d(\underline{x},J)$, le nombre total de points doubles de C se calcule par la formule d'adjonction et vaut $\delta = \frac{1}{2}(d^2 - c_1(X)d + 2)$. Les points doubles réels de C sont de deux natures différentes. Ils peuvent être l'intersection locale de deux branches réelles ou bien l'intersection locale de deux branches complexes conjuguées. Ces points doubles réels sont dits *non-isolés* dans le premier cas et *isolés* dans le second



Point double réel non isolé



Point double réel isolé

Notons $m(C)$ le nombre de points doubles réels isolés de C, c'est la *masse* de C ; elle est majorée par δ. Pour tout entier m compris entre 0 et δ, on note $n_d(m)$ le nombre de courbes $C \in \mathcal{R}_d(\underline{x},J)$ de masse m. Posons finalement

$$\chi_r^d(\underline{x},J) = \sum_{m=0}^{\delta} (-1)^m n_d(m),$$

où $r = (r_1,\ldots,r_N)$.

Théorème 1.1 ([23], [25]). *Soient (X, ω, c_X) une variété symplectique réelle fermée de dimension quatre, N le nombre de composantes connexes de son lieu réel et $d \in H_2(X;\mathbb{Z})$ satisfaisant $c_1(X)d > 0$. Soient $\underline{x} \subset X$ une configuration réelle de $c_1(X)d - 1$ points distincts et $r = (r_1, \dots, r_N)$ le N-uplet associé. L'entier $\chi_r^d(\underline{x}, J)$ est indépendant du choix de $\underline{x}$ et du choix générique de $J \in \mathbb{R}\mathcal{J}_\omega$.*

Le Théorème 1.1 permet de noter cet entier χ_r^d sans ambiguïté. Lorsque $\sum_{i=1}^N r_i$ n'a pas la même parité que $c_1(X)d - 1$, on pose $\chi_r^d = 0$. On note alors $\chi^d[T]$ la fonction génératrice $\sum_{|r|=0}^{c_1(X)d-1} \chi_r^d T^r \in \mathbb{Z}[T_1, \dots, T_N]$, où $T^r = T_1^{r_1} \dots T_N^{r_N}$ et $|r| = r_1 + \dots + r_N$. Cette fonction est de même parité que $c_1(X)d-1$ et tous ses monômes ne dépendent en fait que d'une indéterminée. En effet, la partie réelle d'une sphère holomorphe réelle étant connexe, l'invariant χ_r^d est contraint de s'annuler lorsque les points réels de $\underline{x}$ ne sont pas tous choisis dans une même composante L du lieu réel. On adoptera la notation $\chi_r^d(L)$ pour indiquer que les $|r|$ points réels sont choisis dans L. On renvoie le lecteur à [25] pour une étude de la dépendance de χ_r^d en fonction de r.

Ainsi, la fonction $\chi : d \in H_2(X;\mathbb{Z}) \mapsto \chi^d[T] \in \mathbb{Z}[T_1, \dots, T_N]$ ne dépend que de la variété symplectique réelle fermée de dimension quatre (X, ω, c_X) et est invariante par déformation de cette dernière. Ceci signifie que si ω_t est une famille continue de formes symplectiques satisfaisant $c_X^* \omega_t = -\omega_t$, alors la fonction χ est la même pour tous les triplets (X, ω_t, c_X). Existe-t-il des invariants énumératifs analogues à ceux qui ressortent du Theorème 1.1 en genre quelconque, en dimension quelconque et avec des conditions d'incidence quelconques ? Nous n'avons que des débuts de réponses à ces questions.

1.2. Bornes inférieures et optimalité. Le nombre $R_d(\underline{x}, J)$ de courbes J-holomorphes rationnelles réelles homologues à d qui contiennent l'ensemble $\underline{x}$ de points que l'on s'est donné se retrouve ainsi borné inférieurement par la valeur absolue de l'invariant χ_r^d. Ce nombre est par ailleurs toujours majoré par le nombre total de courbes J-holomorphes rationnelles homologues à d et contenant $\underline{x}$, lequel nombre N_d ne dépend ni de J générique, ni de $\underline{x}$; c'est un invariant de Gromov-Witten de genre zéro de la variété (X, ω). Ainsi,

Corollaire ([25]). *Sous les hypothèses du Théorème 1.1, l'encadrement*

$$|\chi_r^d| \leq R_d(\underline{x}, J) \leq N_d$$

vaut pour tout choix de $\underline{x}$ et tout choix générique de $J \in \mathbb{R}\mathcal{J}_\omega$. $\square$

Les bornes inférieures apparaissant dans ce Corollaire 1.2 se trouvent être parfois optimales. C'est-à-dire qu'il est parfois possible d'exhiber une configuration $\underline{x}$ et une structure générique $J \in \mathbb{R}\mathcal{J}_\omega$ telles que toutes les courbes J-holomorphes rationnelles réelles comptées par χ_r^d le sont en fonction d'un unique et même signe. Nous présentons dans ce paragraphe les situations dans lesquelles nous avons été en mesure de montrer cette optimalité.

Théorème 1.2 ([29], [30]). *Soit (X, ω, c_X) une variété symplectique réelle fermée de dimension quatre et soit $d \in H_2(X;\mathbb{Z})$ une classe d'homologie satisfaisant $(c_X)_* d = -d$. Supposons que le lieu réel de cette variété possède une sphère ou un plan projectif réel L. Dans ce dernier cas, supposons que (X, ω, c_X) est elle-même symplectomorphe au plan projectif complexe éclaté en six boules complexes conjuguées au maximum. Les bornes inférieures apparues dans le Corollaire 1.2 sont sous ces hypothèses optimales dès que $0 \leq r \leq 1$. Le signe de l'invariant $\chi_r^d(L)$ est en outre dans ce cas déterminé par l'inégalité $(-1)^{\frac{1}{2}(d^2 - c_1(X)d+2)}\chi_r^d(L) \geq 0$.*

Remarque 1. La dernière partie du Théorème 1.2 signifie que le signe du coefficient de plus bas degré du polynôme $\chi^d(T)$ introduit au paragraphe 1.1 s'interprète comme la parité du genre lisse de la classe d. Le fait que ce signe puisse être négatif en degrés congrus à trois ou quatre modulo quatre dans le plan projectif complexe met en défaut la Conjecture 6 de [14].

Corollaire ([30]). *Soit d une classe d'homologie de dimension deux du plan projectif complexe ou de la quadrique ellipsoïde et $0 \leq r \leq 1$. Les bornes inférieures (1.2) sont atteintes pour la structure complexe standard lorsque les points complexes conjugués sont choisis très proches d'une conique imaginaire pure dans le premier cas et d'une section hyperplane réelle disjointe de L dans le second.* □

Théorème 1.3 ([30]). *Soit (X, ω, c_X) une variété symplectique réelle fermée de dimension quatre dont le lieu réel possède un tore L et soit $d \in H_2(X;\mathbb{Z})$ une classe d'homologie satisfaisant $(c_X)_* d = -d$. Les bornes inférieures du Corollaire 1.2 sont optimales lorsque $r = 1$. Lorsque le lieu réel est connexe -réduit au tore L-, l'invariant $\chi_1^d(L)$ est en outre positif. Dans le cas général, le signe de l'invariant $\chi_1^d(L)$ est déterminé par l'inégalité $(-1)^{\frac{1}{2}(d^2 - c_1(X)d+2)}\chi_1^d(L) \geq 0$ lorsque le lieu réel des courbes rationnelles ne s'annule pas dans $H_1(L;\mathbb{Z}/2\mathbb{Z})$, tandis qu'il est déterminé par l'inégalité $(-1)^{\frac{1}{2}(d^2 - c_1(X)d+2)}\chi_1^d(L) \leq 0$ lorsque ce dernier s'annule.*

Remarque 2. Dans le cas particulier de la quadrique hyperboloïde, la positivité de $\chi_1^d(L)$ avait été observée dans [14] par d'autres méthodes.

De savoir si les bornes supérieures apparues dans le Corollaire 1.2 sont optimales est un problème classique de géométrie énumérative réelle pour lequel on ne sait presque rien. Le seule chose que je puisse signaler est le critère suivant.

Corollaire ([25]). *Sous les hypothèses du Théorème 1.1, supposons que χ_r^d est positif (resp. négatif). Supposons qu'il existe une configuration réelle de points $\underline{x}$ et une structure générique $J \in \mathbb{R}\mathcal{J}_\omega$ telles qu'il existe $\frac{1}{2}(N_d - |\chi_r^d|)$ courbes J-holomorphes rationnelles réelles de masses impaires (resp. paires) homologues à d et passant par $\underline{x}$. Alors, toutes les courbes J-holomorphes rationnelles homologues à d et passant par $\underline{x}$ sont réelles, de sorte que les bornes supérieures du Corollaire 1.2 sont optimales.* □

Les bornes inférieures fournies par ces invariants sont-elles optimales en général ? La question se pose déjà dans le cas du plan projectif (ou de l'espace projectif de dimension trois, voir le §3.1).

1.3. Congruences. Étant donnée une classe d'homologie $d \in H_2(X;\mathbb{Z})$ d'une variété symplectique réelle de dimension quatre (X, ω, c_X), nous noterons $g_d = \frac{1}{2}(d^2 - c_1(X)d + 2)$ le genre lisse de d et $c_d = c_1(X)d - 1$ le degré attendu du polynôme $\chi^d(T)$ défini au §1.1.

Théorème 1.4 ([30]). *Soit (X, ω, c_X) une variété symplectique réelle fermée de dimension quatre dont le lieu réel possède une composante connexe L homéomorphe à une sphère. Soient $d \in H_2(X;\mathbb{Z})$ et $r \in \mathbb{N}$. Lorsque $2r+1 < c_d$, la puissance $2^{\frac{1}{2}(c_d-2r-1)}$ divise $\chi^d_r(L)$.*

Exemple : Le Théorème 1.4 s'applique à l'ellipsoïde de dimension deux lorsque d est un multiple positif, disons $\delta > 0$, d'une section plane réelle. Dans ce cas, $c_d = 4\delta - 1$ et $g_d = \delta^2 - 2\delta + 1 = \delta + 1 \mod (2)$. Par conséquent, $2^{2\delta-r-1}$ divise $\chi^d_r(L)$ lorsque $r < 2\delta - 1$. Nous avons également montré dans [30] que $2^{2\delta-r}$ divise $\chi^d_r(L)$ lorsque de plus $r = 2\delta + 1 \mod (4)$ ainsi que la congruence $\chi^d_{2\delta-3}(L) = 0 \mod (16)$.

Théorème 1.5 ([30]). *Soit (X, ω, c_X) une variété symplectomorphe au plan projectif complexe éclaté en six boules complexes conjuguées au maximum. Soit $d \in H_2(X;\mathbb{Z})$ une classe satisfaisant $c_d = c_1(X)d - 1 \geq 0$ et soient r, r_X des entiers naturels satisfaisant la relation $r + 2r_X = c_d$. Lorsque $r + 1 < r_X$, la puissance 2^{r_X-r-1} divise $\chi^d_r(L)$.*

Exemple : Le Théorème 1.5 s'applique au plan projectif complexe où d est un multiple positif, disons $\delta > 0$, d'une droite complexe. Dans ce cas, $8^{\frac{1}{2}(\delta-r-1)}$ divise χ^d_r lorsque $r + 1 < \delta$. Nous avons également montré dans [30] que $2^{\frac{1}{2}(3\delta-3r-1)}$ divise χ^d_r lorsque de plus $r = \delta+1 \mod (4)$ et $\chi^d_{\delta-3} = 0 \mod (64)$.

1.4. Calculs. L'invariant χ^d_r qui ressort du Theorème 1.1 fut rapidement estimé après que je l'ai introduit. G. Mikhalkin a proposé dans [16] un algorithme permettant, dans le cas des surfaces toriques réelles, le calcul de cet invariant lorsque le nombre r de points choisis réels est maximal. Cet algorithme a été plus tard étendu par E. Shustin [20] pour un choix quelconque de points réels. Il a été utilisé par I. Itenberg, V. Kharlamov et E. Shustin [12] pour estimer cet invariant, fournissant notamment la minoration $\chi^d_{3d-1} \geq \frac{1}{2}d!$ dans le cas du plan projectif, le calcul en degré inférieur ou égal à cinq, puis l'asymptotique $\log|\chi^d_{c_1(X)d-1}| \cong \log N_d$ dans le cas des surfaces de Del Pezzo réelles X, voir [14]. Ces derniers ont également plus récemment obtenu une formule de type Caporaso-Harris tropicale [13] pour le calcul de χ^d_{3d-1} dans le plan, après que A. Gathmann et H. Markwig [8] ont obtenus cette formule pour le calcul tropical de N_d. J. Solomon a également annoncé une formule calculant ces invariants χ^d_r dans le plan. E. Shustin [21] a adapté ces méthodes

tropicales pour obtenir des résultats analogues dans le cas de la quadrique ellipsoïde. Les méthodes de théorie symplectique des champs que j'ai pour ma part utilisé ([29], [30]) m'ont également permis d'obtenir des formules de type Caporaso-Harris mais avec des conditions de tangence imaginaires conjuguées. Les invariants relatifs qui interviennent dans ces formules sont introduits au §2.1.1. Je ne rappelle pas ici les formules générales qui se trouvent dans [30], mais simplement quelques calculs explicites qui en découlent facilement.

Corollaire ([30]). *Soit* (X, ω, c_X) *une variété symplectomorphe au plan projectif complexe. Alors,* $\chi^4(T) = o(T^2)$, $\chi^5(T) = 64 + 64T^2 + o(T^3)$, $\chi^6(T) = 1024T + 1536T^3 + o(T^4)$, $\chi^7(T) = -14336 + 11776T^2 + o(T^3)$ *et* $\chi^8(T) = -280576T + o(T^2)$.

Remarquons que $\chi^3(T) = 2T^2 + 4T^4 + 8T^8$; ce calcul de $\chi^d(T)$ en degré trois et les phénomènes discutés ici s'obtiennent simplement en éclatant les neuf points base d'un pinceau de cubiques planes et en calculant la caractéristique d'Euler du lieu réel de la surface obtenue, comme observé par V. Kharlamov [15] déjà dans les années 90. Toutefois, même l'existence d'une quartique rationnelle réelle plane passant par onze points réels en position générale n'était pas connue avant l'introduction de ces invariants χ_r^d. Les valeurs explicites de χ_r^d pour $d \leq 9$ et tout r furent entre temps obtenues dans [1] comme conséquence d'une formule de type Caporaso-Harris tropicale. Ces résultats mirent en défaut la conjecture de monotonie de [14], de sorte que la fonction $r \mapsto \chi_r^d$ n'est en général ni positive, ni monotone.

Corollaire ([30]). *Soit* (X, ω, c_X) *une variété symplectomorphe à la quadrique ellipsoïde de dimension deux. On note* h *la classe d'une section plane réelle de bidegré* $(1,1)$. *Alors,* $\chi^{2h}(T) = 2T^3 + 4T^5 + 6T^7$, $\chi^{3h}(T) = 16T + 16T^2 + o(T^3)$, $\chi^{4h}(T) = -256T + 320T^3 + o(T^4)$ *et* $\chi^{5h}(T) = 26880T + o(T^2)$.

Remarque 3. Cet invariant χ_r^d peut se définir purement en termes de fractions rationnelles complexes. Lorsque $r = 4d - 1$ par exemple, il compte algébriquement le nombre de fractions rationnelles $u = P/Q$, $P, Q \in \mathbb{C}[X]$ de degrés d, modulo reparamétrage par les homographies réelles de $PGL_2(\mathbb{R})$, telles que l'image $u(\mathbb{R}P^1)$ interpole un ensemble donné générique de $4d - 1$ points de la sphère de Riemann. Le signe en fonction duquel il convient de compter ces fractions rationnelles u est pair si u possède un nombre pair de points critiques dans chaque hémisphère $\mathbb{C}P^1 \setminus \mathbb{R}P^1$ et impair sinon. Il serait intéressant d'étudier cet invariant de la quadrique ellipsoïde en travaillant uniquement avec des fractions rationnelles.

Quelle est l'asymptotique de l'invariant χ_r^d, $r \leq 1$, calculé ici ? Nos formules calculent l'invariant en fonction d'une somme sur des arbres décorés. Quels sont les arbres qui sont asymptotiquement dominants/négligeables ? De plus, dans le cas du plan projectif par exemple, lorsque $r = 3d - 1$, notre formule calcule l'invariant comme une somme sur des arbres dont certains contribuent positivement et d'autres négativement. Ceci garantit l'existence de structures

presque-complexes pour lesquelles davantage de courbes rationnelles réelles satisfont nos conditions d'incidences que le nombre imposé par l'invariant χ^d_{3d-1}. Combien de courbes réelles a-t-on ainsi construit ? Enfin, notre méthode de calcul suivie dans la première section s'applique à toute variété symplectique de dimension quatre et calcule l'invariant χ en fonction d'invariants de Gromov-Witten de surfaces rationnelles relatifs à des courbes de carré -2 ou -4 lorsque L est une sphère ou un plan projectif réel. Que sait-on de ces invariants et qu'en déduire pour l'invariant χ ? Cette direction de recherche reste à développer. Par ailleurs, j'ignore dans quelles situations exactement il est possible de calculer l'invariant χ^d_r en fonction d'invariants relatifs imaginaires.

2. Invariants Relatifs des Variétés Symplectiques Réelles de Dimension Quatre

Les invariants χ^d_r introduits au §1.1 sont définis par un comptage de courbes J-holomophes rationnelles réelles soumises à des conditions d'incidence ponctuelles. Ils forment ainsi un analogue réel aux invariants de Gromov-Witten de genre zéro ponctuels. J'ai également défini de tels invariants en admettant que les courbes soient soumises à des conditions de tangence avec une courbe donnée, dans l'esprit de la théorie des invariants relatifs. Ces conditions de tangence peuvent être réelles ou bien complexe conjuguées. Dans le cas de conditions réelles, je n'ai pu définir de tels invariants relatifs qu'en admettant une seule condition de tangence et encore m'a-t-il fallu faire intervenir plusieurs types de courbes singulières. J'expose ces résultats dans le §2.1.1. J'ai pu en déduire des bornes inférieures pour le nombre de coniques réelles tangentes à cinq coniques données, un problème classique de géométrie énumérative. Dans le cas de conditions de tangence complexes conjuguées, la situation est bien meilleure et de tels invariants peuvent s'obtenir avec les mêmes méthodes que celles utilisées au §1.1. Je n'ai en fait introduit et utilisé ces invariants que dans des cas très particuliers, en utilisant le langage de la théorie symplectique des champs. Ils m'ont été utiles pour mener les calculs présentés au §1.4. J'expose ces résultats dans le §2.2.

2.1. Invariants relatifs réels

2.1.1. Définition des invariants. Soient (X, ω, c_X) une variété symplectique réelle de dimension quatre, $d \in H_2(X;\mathbb{Z})$ une classe d'homologie telle que $c_1(X)d \geq 2$ et $\underline{x}$ une configuration réelle de $c_1(X)d-2$ points distincts. Comme au §1.1, on note $\mathbb{R}X_1, \ldots, \mathbb{R}X_N$ les composantes connexes de $\mathbb{R}X$ et r_i le cardinal de $\underline{x} \cap \mathbb{R}X_i$, $i \in \{1, \ldots, N\}$. Soit $B \subset \mathbb{R}X$ une surface à bord lisse. En chaque point réel x_i de $\underline{x}$, on choisit une droite vectorielle T_i dans le plan tangent $T_{x_i}\mathbb{R}X$. Pour toute structure presque complexe $J \in \mathbb{R}\mathcal{J}_\omega$ suffisamment générique, on définit l'entier $\Gamma^{d,B}_r(J, \underline{x})$ comme la somme des nombres de

courbes J-holomorphes rationnelles réelles qui réalisent la classe d'homologie d, passent par la configuration $\underline{x}$ et qui proviennent des quatre familles suivantes :

- Les courbes tangentes au bord de B, elles sont comptées en fonction de leurs masses et de leur contact intérieur ou extérieur à B au point de tangence.
- Les courbes non-immergées, qui sont comptées en fonction de leurs masses et de la position du point de rebroussement par rapport à B.
- Les courbes possédant une des droites T_i comme tangente, qui sont comptées en fonction de leurs masses et de la position du point x_i correspondant à T_i par rapport à B.
- Les courbes réductibles, qui sont comptées en fonction de leurs masses et d'une multiplicité qui est le nombre de points réels d'intersection entre les deux composantes irréductibles de la courbe, chacun de ces points devant être compté positivement ou négativement selon qu'il est intérieur ou extérieur à B.

Ainsi, en notant respectivement $\mathcal{T}an^d_B(J,\underline{x})$, $\mathcal{C}usp^d(J,\underline{x})$, $\mathcal{T}an^d(J,\underline{x})$ et $\mathcal{R}ed^d(J,\underline{x})$ ces quatre ensembles finis de courbes J-holomorphes, l'entier $\Gamma^{d,B}_r(J,\underline{x})$ s'écrit

$$\sum_{C\in\cup\mathcal{T}an^d_L(J,\underline{x})\cup\mathcal{T}an^d(J,\underline{x})\cup\mathcal{C}usp^d(J,\underline{x})} (-1)^{m(C)}\langle C,B\rangle - \sum_{C\in\mathcal{R}ed^d(J,\underline{x})} (-1)^{m(C)}\,\mathrm{mult}_B(C).$$

Dans cette somme, *l'indice de contact* $\langle C,B\rangle$ vaut -1 (resp. $+1$) si $C\in\mathcal{T}an^d_L(J,\underline{x})$ et $\mathbb{R}C$ se trouve localement incluse dans (resp. en dehors de) B au voisinage du point de tangence y avec ∂B. Si $C\in\mathcal{C}usp^d(J,\underline{x})$ (resp. $C\in\mathcal{T}an^d(J,\underline{x})$), le point de rebroussement (resp. la droite T_i, $i\in I$) est unique et l'indice de contact $\langle C,B\rangle$ vaut -1 si ce point se situe en-dehors de B et $+1$ sinon. Si C est réductible, elle n'a que deux composantes irréductibles C_1, C_2, toutes deux réelles et

$$\mathrm{mult}_B(C)=\sum_{y\in\mathbb{R}C_1\cap\mathbb{R}C_2}\langle y,B\rangle,$$

où $\langle y,B\rangle$ vaut -1 lorsque y est extérieur à B et $+1$ s'il est intérieur.

Théorème 2.1 ([28]). *Soient (X,ω,c_X) une variété symplectique réelle fermée de dimension quatre et $B\subset\mathbb{R}X$ une surface à bord lisse. Soient N le nombre de composantes connexes de $\mathbb{R}X$ et $d\in H_2(X;\mathbb{Z})$ satisfaisant $c_1(X)d>1$, $c_1(X)d\neq 4$. Soient $\underline{x}\subset X\setminus\partial B$ une configuration réelle de $c_1(X)d-2$ points distincts et $r=(r_1,\dots,r_N)$ le N-uplet associé, supposé non nul. L'entier $\Gamma^{d,B}_r(J,\underline{x})$ est indépendant du choix de $\underline{x}$ et du choix générique de $J\in\mathbb{R}\mathcal{J}_\omega$.*

Le Théorème 2.1 permet sans ambiguïté de noter $\Gamma^{d,B}_r$ cet entier. Lorsque $\sum_{i=1}^N r_i$ n'a pas la même parité que $c_1(X)d$, on pose $\Gamma^{d,B}_r=0$. Comme au §1.1,

on note alors $\Gamma^{d,B}[T]$ la fonction génératrice $\sum_{|r|=0}^{c_1(X)d-2} \Gamma_r^{d,B} T^r \in \mathbb{Z}[T_1, \ldots, T_N]$. Cette fonction est de même parité que $c_1(X)d$ et tous ses monômes ne dépendent en fait que d'une indéterminée.

Ainsi, la fonction $\Gamma^B : d \in H_2(X;\mathbb{Z}) \mapsto \Gamma^{d,B}[T] \in \mathbb{Z}[T_1, \ldots, T_N]$ ne dépend que du quadruplet (X, ω, c_X, B). Elle est en outre invariante par déformation de ce quadruplet au sens où si ω_t est une famille continue de formes symplectiques satisfaisant $c_X^* \omega_t = -\omega_t$ et $B_t \subset \mathbb{R}X$ une isotopie de surfaces compactes, alors cette fonction est la même pour tout (X, ω_t, c_X, B_t).

Remarquons qu'en particulier $\Gamma_r^{d,B}(J, \underline{x})$ ne dépend pas de la position relative de $\underline{x}$ par rapport à B, que $\Gamma_r^{d,B} = -\Gamma_r^{d,\mathbb{R}X \setminus B}$ et que le cas particulier où B est vide est admissible et fournit un invariant que l'on a préalablement introduit dans [27]. Montrer l'invariance de $\Gamma_r^{d,\emptyset}(J, \underline{x})$ se trouve être une étape importante dans la démonstration de l'invariance de $\Gamma_r^{d,B}(J, \underline{x})$.

Théorème 2.2 ([28]). *Sous les hypothèses du Théorème 2.1, si B est un disque, $2\chi_{r+1}^d = \Gamma_r^{d,B} - \Gamma_r^{d,\emptyset}$. Si de plus, (X, ω, c_X) est symplectomorphe au plan projectif complexe, $\Gamma_r^{d,B} = -\Gamma_r^{d,\emptyset}$, tandis que si elle est symplectomorphe à la quadrique hyperboloïde de dimension deux, $\Gamma_r^{d,B} = 2\chi_{r+1}^d - \Gamma_r^{d,\emptyset}$.*

Corollaire ([28]). *Sous les hypothèses du Théorème 2.2, $\chi_{r+1}^d = -\Gamma_r^{d,\emptyset} = \Gamma_r^{d,B}$ dans le cas du plan projectif complexe et $\Gamma_r^{d,B} = 2\chi_{r+1}^d$, $\Gamma_r^{d,\emptyset} = 0$ dans le cas de la quadrique hyperboloïde de dimension deux.* □

2.1.2. Sur les 3264 coniques tangentes à cinq coniques génériques. Il est possible d'étendre les résultats du §2.1.1 à davantage de conditions de tangence avec le bord de B, au moins dans le cas de coniques. J'ai illustré ce phénomène en m'intéressant au problème ancien du comptage des coniques tangentes à cinq coniques génériques données. Le nombre de solutions complexes est 3264 comme l'a démontré de Joncquières en 1859 mais le nombre de solutions réelles dépend du choix des cinq coniques génériques. Soient $B_1, \ldots, B_5$ cinq disques plongés dans $\mathbb{R}P^2$ de sorte que leurs bords soient transverses deux à deux, et $J \in \mathbb{R}\mathcal{J}_\omega$. Notons Γ^B le nombre de coniques J-holomorphes réelles qui sont soit :

- irréductibles, tangentes à $B_1, \ldots, B_5$, et comptées positivement si elles sont tangentes intérieurement à B_i pour un nombre pair de $i \in \{1, \ldots, 5\}$, négativement sinon.
- réductibles, tangentes à quatre des cinq disques $B_1, \ldots, B_5$, et chacune comptée en fonction de la parité du nombre de disques en lesquelles elle est tangente intérieurement et de la position de son unique point singulier par rapport au cinquième disque en lequel elle n'est pas tangente.

Ainsi, en notant respectivement $\mathcal{C}on(J)$ et $\mathcal{C}on_{red}(J)$ ces deux ensembles finis de coniques J-holomorphes, on obtient

$$\Gamma^B(J) = \sum_{C\in\mathcal{C}on(J)} \langle C, B\rangle - \sum_{C\in\mathcal{C}on_{red}(J)} \langle C, B\rangle \,\mathrm{mult}_B(C) \in \mathbb{Z},$$

où lorsque $C \in \mathcal{C}on(J)$, *l'indice de contact* $\langle C, B\rangle$ vaut $\Pi_{i=1}^5 \langle C, B_i\rangle$; tandis que lorsque $C \in \mathcal{C}on_{red}(J)$ et $i_1, \ldots, i_4 \in \{1, \ldots, 5\}$ sont les entiers tels que C soit tangent aux bords de $B_{i_1}, \ldots, B_{i_4}$, $\langle C, B\rangle = \Pi_{j=1}^4 \langle C, B_{i_j}\rangle$ et $\mathrm{mult}_B(C) = +1$ si le point singulier de C appartient à B_{i_5} et -1 sinon.

Théorème 2.3 ([28]). *1) Cet entier $\Gamma^B(J)$ ne dépend pas du choix générique de la structure presque-complexe $J \in \mathbb{R}\mathcal{J}_\omega$ et est invariant par isotopie de $B = B_1 \cup \cdots \cup B_5$.*

2) Si $B_1, \ldots, B_5$ sont cinq disques disjoints, alors $\Gamma^B = 272$. Il en est de même si $B_1, \ldots, B_5$ sont proches de cinq droites doubles génériques.

Un disque est dit proche d'une droite double d'équation $y^2 = 0$ s'il a une équation de la forme $\{y^2 \leq \epsilon^2 x^2 - \delta\}$ pour ϵ et δ petits.

Corollaire ([28]). *Si $C_1, \ldots, C_5$ sont cinq coniques dont la classe d'isotopie est donnée par la deuxième partie du Théorème 2.3, alors le nombre de coniques réelles qui leur sont tangentes est minoré par 32 indépendamment du choix de $C_1, \ldots, C_5$ dans la classe d'isotopie.*

Démonstration : Le nombre de droites tangentes à deux coniques génériques vaut quatre, elles sont codées par les points d'intersection entre les deux coniques duales. Le nombre de coniques tangentes à quatre des cinq coniques $C_1, \ldots, C_5$ se trouve donc majoré par $240 = 5 * 3 * 4 * 4$, de sorte que le résultat découle de la définition de Γ^B et de la deuxième partie du Théorème 2.3. $\square$

Remarquons que le fait qu'il existe une configuration de cinq coniques réelles pour lesquelles les 3264 coniques tangentes à ces cinq coniques sont toutes réelles a été établi par F. Ronga, A. Tognoli et T. Vust [19]. Le théorème 2.1 montre la difficulté à définir des invariants relatifs avec conditions de tangence réelles. Dans le « monde tropical », la situation est parfois bien meilleure, voir [13]. Il en est de même avec des conditions de tangence complexes conjuguées, voir le §2.2.

2.2. Invariants relatifs imaginaires. Soit L une sphère, un tore ou un espace projectif réel de dimension $n = 2$ ou 3. Le fibré cotangent de L est équipé de sa forme de Liouville λ et de l'involution c_L définie par $(q, p) \in T^*L \mapsto (q, -p) \in T^*L$. Cette dernière satisfait $c_L^*\lambda = -\lambda$ de sorte que $(T^*L, d\lambda, c_L)$

est une variété symplectique réelle. Soit g une métrique à courbure constante sur L, U^*L l'ensemble des couples $(q,p) \in T^*L$ tels que $g(p,p) \leq 1$ et S^*L le bord de U^*L. La restriction de λ à S^*L est une forme de contact et l'on note R_λ le champ de Reeb associé. Le flot engendré par R_λ n'est autre que le flot géodésique. Notons $\mathcal{J}_\lambda$ l'espace des structures presque-complexes positives pour $d\lambda$ et asymptotiquement cylindriques sur une structure CR de S^*L. Plus précisément, le champ radial de T^*L identifie le complémentaire de la section nulle avec la symplectisation $(\mathbb{R}\times S^*L, d(e^\rho\lambda))$ de (S^*L,λ). On note $\mathcal{J}_\lambda$ l'espace des structures presque-complexes J positives pour $d\lambda$, de classe C^l, $l \gg 1$, qui satisfont $J(\frac{\partial}{\partial\rho}) = R_\lambda$ et préservent le noyau de λ pour $\rho \gg 1$ et qui enfin sont invariantes par translation par ρ au-delà d'un certain rang ρ_0. Nous notons alors $\mathbb{R}\mathcal{J}_\lambda \subset \mathcal{J}_\lambda$ le sous-espace des structures presque-complexes pour lesquelles c_L est J-antiholomorphe. Ces espaces $\mathcal{J}_\lambda$ et $\mathbb{R}\mathcal{J}_\lambda$ sont tous deux des variétés de Banach séparables non-vides et contractiles. Nous allons compter les courbes J-holomorphes rationnelles réelles pointées d'énergie de Hofer finie proprement immergées dans T^*L en fonction d'un signe ± 1 de façon à obtenir un invariant associé à T^*L. Rappelons que d'après le Théorème 1.2 de [11] et d'après [3], ces courbes rationnelles pointées convergent en leurs pointes vers des orbites de Reeb parcourues un nombre entier de fois, que l'on appelle multiplicité. La dimension de l'espace des modules de telles courbes dépend du nombre de pointes et des multiplicités associées. Afin d'obtenir un nombre fini de courbes, nous allons soumettre ces courbes à quelques contraintes, soit en les forçant à converger vers des orbites de Reeb prescrites, soit en les forçant à passer par des points de L ou des paires de points complexes conjuguées de $T^*L \setminus L$.

Soit e_i, $i \geq 1$, la suite d'entiers partout nulle sauf au i-ème rang où elle vaut un. Soient $\alpha = \sum_{i\in\mathbb{N}^*} \alpha_i e_i$ et $\beta = \sum_{i\in\mathbb{N}^*} \beta_i e_i$ deux suites d'entiers positifs qui s'annulent à partir d'un certain rang. Ces deux suites codent respectivement le nombre de paires d'orbites de Reeb complexes conjuguées limites prescrites et non prescrites de nos courbes, avec leur multiplicités $i \in \mathbb{N}^*$. Le nombre de pointes de nos courbes vaut donc $2v = 2\sum_{i\in\mathbb{N}^*}(\alpha_i + \beta_i)$ et nous choisissons un ensemble Γ de $\sum_{i\in\mathbb{N}^*} \alpha_i$ géodésiques fermées disjointes de L pour prescrire nos paires d'orbites de Reeb limites. À présent, afin de fixer nos contraintes ponctuelles, soient $r \in \mathbb{N}$ et $x_1, \ldots, x_r$ des points distincts de L. De même, soient $r_L \in \mathbb{N}$ et $\xi_1, \overline{\xi}_1, \ldots, \xi_{r_L}, \overline{\xi}_{r_L}$ des paires distinctes de points complexes conjugués de $T^*L \setminus L$, c'est-à-dire satisfaisant $c_L(\xi_i) = \overline{\xi}_i$. Nous supposons que

$$(n-1)r + 2(n-1)r_L + 2(n-1)\#\Gamma = 2v + \epsilon(n-1)\sum_{i\in\mathbb{N}^*} i(\alpha_i + \beta_i) + n - 3, \quad (1)$$

où $\epsilon = 2$ si L est homéomorphe à une sphère et $\epsilon = 1$ si L est homéomorphe à un espace projectif réel, tandis que nous supposons

$$(n-1)r + 2(n-1)r_L = 2v + n - 3 \text{ et } \alpha = 0 \quad (2)$$

si L est homéomorphe à un tore.

Alors, lorsque la structure presque-complexe $J \in \mathbb{R}\mathcal{J}_\lambda$ est générique, il n'y a qu'un nombre fini de courbes J-holomorphes rationnelles réelles d'énergie de Hofer finie, proprement immergées dans T^*L et ayant $2v$ pointes qui passent par $\underline{x}$, par chaque paire $\{\xi_i, \overline{\xi}_i\}$ et qui convergent vers les orbites de Reeb relevant les éléments de Γ ainsi que vers β_j autres paires d'orbites, $j \in \mathbb{N}^*$, chacune avec multiplicité j ou de classe d'homologie donnée si L est un tore. En effet, si L est un tore, il y a une infinité de géodésiques fermées primitives non homologues et la dimension (2) ne dépend pas du choix des classes d'homologies de sorte qu'il y a une infinité d'espaces de modules ayant la même dimension. Pour garantir la finitude, nous imposons les classes d'homologies des orbites de Reeb limites. Notons $\mathcal{R}(\alpha, \beta, \Gamma, \underline{x}, \underline{\xi}, J)$ cet ensemble fini de courbes, la généricité de J garantit qu'elles sont toutes immergées. Si L est de dimension deux, on pose

$$F_{(r,r_L)}(\alpha, \beta, \Gamma, \underline{x}, \underline{\xi}, J) = \sum_{C \in \mathcal{R}(\alpha,\beta,\Gamma,\underline{x},\underline{\xi},J)} (-1)^{m(C)} \in \mathbb{Z}.$$

Si L est de dimension trois, on l'équipe d'une structure spin. Ceci permet d'associer un état spinoriel $\mathrm{sp}(C)$ à chaque courbe $C \in \mathcal{R}(\alpha, \beta, \Gamma, \underline{x}, \underline{\xi}, J)$ comme expliqué au §3.1 et on pose

$$F_{(r,r_L)}(\alpha, \beta, \Gamma, \underline{x}, \underline{\xi}, J) = \sum_{C \in \mathcal{R}(\alpha,\beta,\Gamma,\underline{x},\underline{\xi},J)} \mathrm{sp}(C) \in \mathbb{Z}.$$

Théorème 2.4 ([30]). *Soit L une sphère, un tore ou un espace projectif réel de dimension $n = 2$ ou 3 muni d'une métrique à courbure constante. Soient α, β deux suites d'entiers positifs qui s'annulent à partir d'un certain rang. On choisit comme ci-dessus un ensemble Γ de géodésiques fermées et des ensembles $\underline{x}$, $\underline{\xi}$ de r et r_L points dans L et $T^*L \setminus L$ respectivement de sorte que ces nombres satisfassent (2) dans le cas du tore et (1) sinon. Lorsque $n = 3$, on suppose $r \neq 0$ et lorsque de plus $L \in \{S^3, \mathbb{R}P^3\}$, on suppose que J est invariante par le flot de Reeb pour $\rho \gg 1$. Alors, l'entier $F_{(r,r_L)}(\alpha, \beta, \Gamma, \underline{x}, \underline{\xi}, J)$ défini ci-dessus ne dépend ni du choix des contraintes $\Gamma, \underline{x}, \underline{\xi}$, ni du choix générique de la structure presque-complexe $J \in \mathbb{R}\mathcal{J}_\lambda$.*

L'entier $F_{(r,r_L)}(\alpha, \beta, \Gamma, \underline{x}, \underline{\xi}, J)$ étant indépendant de $\Gamma, \underline{x}, \underline{\xi}, J$ d'après le Théorème 2.4, nous le noterons $F_{(r,r_L)}(\alpha, \beta)$. Afin d'alléger encore cette notation, nous noterons cet entier $F(\alpha, \beta)$ lorsque $r_L = 0$, puisque la valeur de r est alors définie sans ambiguïté par les calculs de dimensions (1) et (2). Les Lemmes 2.5, 2.6 et 2.7 fournissent quelques calculs que l'on a pu mener. Les résultats du §1.4 reposent sur ces calculs.

Lemme 2.5 ([30]). *Si L est homéomorphe à une sphère de dimension deux et $r_L = 0$, on a $F(e_1, 0) = F(0, e_1) = 1$, $F(e_2, 0) = 2$, $F(0, e_2) = 8$, $F(2e_1, 0) = 2$, $F(e_1, e_1) = 4$ et $F(0, 2e_1) = 6$.*

Lemme 2.6 ([30]). *Si L est homéomorphe à un plan projectif réel et $r_L = 0$, on a $F(e_1, 0) = F(0, e_1) = F(e_2, 0) = F(2e_1, 0) = F(e_1, e_1) = F(0, 2e_1) = 1$ et $F(0, e_2) = 4$.*

Lemme 2.7 ([30]). *Si L est homéomorphe à un plan projectif réel et $r_L = 0$, on a $F(e_3, 0) = 2$, $F(0, e_3) = 12$, $F(e_1+e_2, 0) = 2$, $F(e_1, e_2) = 8$, $F(e_2, e_1) = 4$, $F(0, e_1 + e_2) = 24$, $F(3e_1, 0) = 2$, $F(2e_1, e_1) = 4$, $F(e_1, 2e_1) = 6$ et $F(0, 3e_1) = 8$.*

Toutefois, la valeur de l'invariant F qui ressort du Théorème 2.4 n'est pas connue en général. Il serait intéressant de développer des méthodes permettant son calcul.

3. Invariants En Dimension Six

Nous exposons dans ce paragraphe les résultats analogues à ceux présentés au §1 que l'on a pu établir en dimension six.

3.1. Définition des invariants dans les variétés algébriques réelles convexes. Rappelons qu'une variété projective lisse est dite convexe lorsque le groupe $H^1(\mathbb{C}P^1; u^*TX)$ s'annule pour tout morphisme $u : \mathbb{C}P^1 \to X$. Les principaux exemples que je connaisse sont les espaces homogènes projectifs, citons les produits d'espaces projectifs, la quadrique de $\mathbb{C}P^4$ ou encore la variété des drapeaux de $\mathbb{C}^3$. Il est à nouveau possible de définir un invariant en comptant algébriquement le nombre de courbes rationnelles réelles qui réalisent une classe d'homologie d donnée et passent par une configuration réelle de points $\underline{x}$ de cardinal $\frac{1}{2}c_1(X)d$, où $c_1(X)$ désigne la première classe de Chern de la variété et $c_1(X)d$ est supposé pair. Toutefois, le signe ± 1 en fonction duquel il convient de compter les courbes rationnelles réelles est plus délicat à définir. Le lieu réel $\mathbb{R}X = \text{fixe}(c_X)$ de X est une variété lisse de dimension réelle trois que l'on suppose orientable pour simplifier. Munissons-la d'une orientation ainsi que d'une métrique riemannienne auxiliaire. Son $SO_3(\mathbb{R})$-fibré principal des repères orthonormés directs s'étend alors en un $Spin_3$-fibré principal. En effet, l'obstruction à l'existence d'une telle extension est donnée en général par la classe caractéristique $w_2(\mathbb{R}X)$ et cette obstruction s'annule en dimension trois comme il découle de la formule de Wu. Lorsque la configuration réelle de points est suffisamment générique et possède au moins un point réel, d'une part les courbes rationnelles réelles qui passent par $\underline{x}$ et réalisent d sont toutes immergées (même lisses en général) et de partie réelle non vide, et d'autre part elle sont équilibrées. Ce dernier point signifie que le fibré normal de ces courbes se décompose sur $\mathbb{C}$ comme la somme directe de deux fibrés en droite isomorphes L et M, fibrés qui de plus peuvent être choisis réels. Notons $\mathcal{R}_d(\underline{x})$ cet ensemble fini de courbes rationnelles réelles. Chaque lieu réel de ces courbes fournit donc un nœud immergé dans la variété de dimension trois $\mathbb{R}X$, et ce nœud est de plus canoniquement équipé d'un repère mobile ou plutôt d'axes mobiles donnés par la tangente au nœud et les lieux réels des fibrés en droites L et M (en fait, seule la classe d'homotopie de ces axes mobiles est canoniquement définie, puisque la décomposition du fibré normal en somme

$L \oplus M$ n'est pas uniquement définie, mais c'est amplement suffisant pour nos besoins). Lorsque les lieux réels de L et M sont orientables, c'est-à-dire lorsque ces fibrés sont de degré pair, ces axes mobiles peuvent être enrichis de repères orthonormés. Ainsi, les nœuds définis par les courbes rationnelles réelles sont tous équipés de repères orthonormés mobiles qui fournissent des lacets dans le $SO_3(\mathbb{R})$-fibré principal des repères orthonormés, lacets qui relèvent les nœuds de $\mathbb{R}X$. Vient alors l'alternative suivante pour chaque lacet : soit ce lacet du $SO_3(\mathbb{R})$-fibré principal des repères se relève en un lacet du $Spin_3$-fibré principal donné par la structure *Spin* $\mathfrak{s}$, soit non. Ceci permet de définir *l'état spinoriel* $sp(C)$ de chaque courbe rationnelle réelle C comme valant $+1$ dans le premier cas, et -1 dans le second. Lorsque les lieux réels de L et M ne sont pas orientables, on modifie ces axes mobiles à l'aide d'un demi-tour à droite donné par l'orientation de $\mathbb{R}X$, ce qui permet de se ramener au cas précédent et de définir l'état spinoriel également dans ce cas. L'entier $\chi_r^{d,\mathfrak{s}}(\underline{x})$ n'est alors autre que le nombre de courbes rationnelles réelles qui réalisent la classe d'homologie d et passent par $\underline{x}$, ces courbes étant comptées en fonction de leur état spinoriel, de sorte que

$$\chi_r^{d,\mathfrak{s}}(\underline{x}) = \sum_{C \in \mathcal{R}_d(\underline{x})} sp(C) \in \mathbb{Z}.$$

On a noté $r = (r_1, \ldots, r_N)$ le N-uplet associé à $\underline{x}$; c'est-à-dire que N désigne le nombre de composantes connexes de $\mathbb{R}X$ et en notant $(\mathbb{R}X)_1, \ldots, (\mathbb{R}X)_N$ ces composantes, $r_i = \#(\underline{x} \cap (\mathbb{R}X)_i)$.

Théorème 3.1 ([26]). *Soient (X, c_X) une variété algébrique réelle convexe lisse de dimension trois et $\mathfrak{s}$ une structure $Spin_3$ sur son lieu réel $\mathbb{R}X$ supposé orientable. Soit $d \in H_2(X;\mathbb{Z})$ telle que $c_1(X)d$ soit pair et différent de quatre et soit $k_d = \frac{1}{2}c_1(X)d \in \mathbb{N}^*$. Soient $\underline{x} = (x_1, \ldots, x_{k_d})$ une configuration réelle de k_d points distinct dont au moins un réel et r le N-uplet associé. L'entier $\chi_{\underline{r}}^{d,\mathfrak{s}}(\underline{x})$ ne dépend alors pas du choix générique de $\underline{x}$.*

Ce résultat est valable aussi pour les variétés dont le lieu réel n'est pas orientable, moyennant le choix d'une structure Pin_3^- sur le lieu réel ; l'invariant s'annule alors lorsque k_d est pair, voir [26].

Le Théorème 3.1 permet sans ambiguïté de noter cet entier $\chi_r^{d,\mathfrak{s}}$. Lorsque $\sum_{i=1}^N r_i$ n'a pas la même parité que $k_d = \frac{1}{2}c_1(X)d$, on pose $\chi_r^{d,\mathfrak{s}} = 0$. On note alors $\chi^{d,\mathfrak{s}}[T]$ la fonction génératrice $\sum_{|r|=0}^{k_d} \chi_r^{d,\mathfrak{s}} T^r \in \mathbb{Z}[T_1, \ldots, T_N]$. Cette fonction est de même parité que $\frac{1}{2}c_1(X)d$ et tous ses monômes ne dépendent en fait que d'une indéterminée.

Ainsi, la fonction $\chi^{\mathfrak{s}} : d \in H_2(X;\mathbb{Z}) \mapsto \chi^{d,\mathfrak{s}}(T) \in \mathbb{Z}[T]$ est invariante par isomorphisme de la variété algébrique réelle convexe lisse de dimension trois (X, c_X). On en déduit à nouveau les bornes inférieures suivantes en géométrie énumérative réelle.

Corollaire ([26]). *Sous les hypothèses du Théorème 3.1, notons $R_d(\underline{x})$ le nombre de courbe rationnelles réelles connexes homologues à d qui passent par*

$\underline{x}$ et N_d l'invariant de Gromov-Witten de genre zéro associé. Alors, $|\chi_r^{d,\mathfrak{s}}| \leq R_d(\underline{x}) \leq N_d$. □

Finissons ce paragraphe par une interprétation topologique de nos résultats. Les singularités de l'espace $\mathbb{R}_\tau\overline{\mathcal{M}}^d_{k_d}(X)$ sont de codimension au moins deux, de sorte que cet espace possède une première classe de Stiefel-Whitney. Étant donné $D \in H_{3k_d-1}(\mathbb{R}_\tau\overline{\mathcal{M}}^d_{k_d}(X);\mathbb{Z}/2\mathbb{Z})$, on note $D^\vee$ son image sous le morphisme $H_{3k_d-1}(\mathbb{R}_\tau\overline{\mathcal{M}}^d_{k_d}(X);\mathbb{Z}/2\mathbb{Z}) \to H^1(\mathbb{R}_\tau\overline{\mathcal{M}}^d_{k_d}(X);\mathbb{Z}/2\mathbb{Z})$.

Proposition 3.2 ([24]). *La première classe de Stiefel-Whitney de toute composante $\mathbb{R}\mathcal{M}^*$ de $\mathbb{R}_\tau\mathcal{M}^d_{k_d}(X)$ qui contient une courbe équilibrée s'écrit*

$$w_1(\mathbb{R}\mathcal{M}^*) = (\mathbb{R}_\tau ev^d)^* w_1(\mathbb{R}_\tau X^{k_d}) + \sum_{D\subset Red'} \epsilon(D)D^\vee \in H^1(\mathbb{R}\mathcal{M}^*;\mathbb{Z}/2\mathbb{Z}),$$

où $\epsilon(D) \in \{0,1\}$ et lorsque $\epsilon(D) = 1$, la composante irréductible D de Red se trouve contractée par l'application d'évaluation $\mathbb{R}_\tau ev^d$. □

On a noté ici Red' la réunion du diviseur des courbes réductibles Red et de l'éventuel diviseur des courbes non-équilibrées $(u, C, \underline{z})$ telles que $\dim H^1(C; N_u \otimes \mathcal{O}_C(-\underline{z})) \geq 2$, si un tel diviseur existe. On note Red_1 la réunion des composantes irréductibles D de Red' pour lesquelles $\epsilon(D) = 1$. En dimension deux, cet ensemble a été déterminé dans [18]. Équipons $\mathbb{R}_\tau X^{k_d}$ d'un système de coefficients tordus entiers $\mathcal{Z}$ et notons $[\mathbb{R}_\tau X^{k_d}] \in H_{3k_d}(\mathbb{R}_\tau X^{k_d};\mathcal{Z})$ sa classe fondamentale. Notons $\mathcal{Z}^*$ le système de coefficients locaux induit sur $\mathbb{R}\mathcal{M}^*$, tiré en arrière de $\mathcal{Z}$ par $\mathbb{R}_\tau ev^d$.

Proposition 3.3 ([24]). *Sous les hypothèses de la Proposition 3.2, il existe une unique classe fondamentale $[\mathbb{R}\mathcal{M}^*] \in H_{3k_d}(\mathbb{R}\mathcal{M}^*, Red_1;\mathcal{Z}^*)$ telle qu'en toute courbe équilibrée $(u, C, \underline{z}) \in \mathbb{R}\mathcal{M}^*$, le morphisme*

$$(\mathbb{R}_\tau ev^d)_* : H_{3k_d}(\mathbb{R}\mathcal{M}^*, \mathbb{R}\mathcal{M}^* \setminus \{(u, C, \underline{z})\};\mathcal{Z}^*) \to H_{3k_d}(\mathbb{R}_\tau X^{k_d}, \mathbb{R}_\tau X^{k_d} \setminus \{u(\underline{z})\};\mathcal{Z})$$

envoie $[\mathbb{R}\mathcal{M}^*]$ sur $sp(u, C, \underline{z})[\mathbb{R}_\tau X^{k_d}]$. □

Comme $\mathbb{R}_\tau ev^d(\mathrm{Red}_1)$ est de codimension deux, le groupe $H_{3k_d}(\mathbb{R}_\tau X^{k_d}, \mathbb{R}_\tau ev^d(\mathrm{Red}_1);\mathcal{Z})$ est cyclique, engendré par $[\mathbb{R}_\tau X^{k_d}]$. L'entier $\chi_r^{d,\mathfrak{s}}$ n'est autre que celui défini par la relation $(\mathbb{R}_\tau ev^d)_*[\mathbb{R}_\tau\overline{\mathcal{M}}^d_{k_d}(X)] = \chi_r^{d,\mathfrak{s}}[\mathbb{R}_\tau X^{k_d}]$, où la classe fondamentale $[\mathbb{R}_\tau\overline{\mathcal{M}}^d_{k_d}(X)]$ est donnée par la Proposition 3.3.

3.2. Extension aux variétés symplectiques réelles fortement semi-positives. L'extension des résultats du §3.1 aux variétés symplectiques n'est pas immédiate, en partie parce-que le thérorème de Grothendieck [10] selon lequel les fibrés holomorphes sur la sphère de Riemann

sont entièrement décomposables n'est plus valable pour les fibrés normaux des courbes pseudo-holomorphes. Ces derniers sont des fibrés vectoriels complexes munis d'un opérateur de Cauchy-Riemann qui n'est que $\mathbb{R}$-linéaire et non $\mathbb{C}$-linéaire comme dans le cas de fibrés holomophes. Ces premiers sont des perturbations d'ordre zéro de ces derniers par des opérateurs $\mathbb{C}$-antilinéaires et sont parfois appelés « opérateurs de Cauchy-Riemann généralisés ». J'ai étendu dans [24] la notion d'état spinoriel pour un opérateur de Cauchy-Riemann généralisé surjectif.

La stratégie est la suivante. L'espace des opérateurs de Cauchy-Riemann généralisés réels sur un fibré vectoriel complexe réel donné est un espace de Banach affine, il contient les opérateurs de Cauchy-Riemann $\mathbb{C}$-linéaires comme sous-espace de Banach. Or chaque opérateur de Cauchy-Riemann surjectif définit une structure de fibré vectoriel holomorphe équilibré et possède donc un état spinoriel d'après les résultats du §3.1. Étant donné un opérateur de Cauchy-Riemann généralisé surjectif, on le relie à opérateur de Cauchy-Riemann surjectif par un chemin générique et on définit son état spinoriel comme celui de l'opérateur de Cauchy-Riemann si le chemin traverse un nombre pair de fois le mur des opérateurs non-surjectifs et son opposé sinon.

Soit alors (X, ω, c_X) une variété symplectique réelle fortement semi-positive de dimension six, c'est-à-dire pour laquelle toute classe sphérique $d \in H_2(X;\mathbb{Z})$ positive contre ω satisfait l'implication $c_1(X)d \geq 2-n \implies c_1(X)d \geq 1$. Les variétés symplectiques réelle positives, par exemple de Fano, satisfont cette condition. On suppose à nouveau pour simplifier le lieu réel de cette variété orientable et on l'équipe d'une structure spin $\mathfrak{s}$. Soit, comme au §3.1, $d \in H_2(X;\mathbb{Z})$ telle que $(c_X)_* d = -d$, $c_1(X)d$ soit pair et strictement plus grand que deux. Soient $k_d = \frac{1}{2}c_1(X)d$ et $\underline{x} = (x_1, \ldots, x_{k_d}) \in X^{k_d}$ une configuration réelle de k_d points distincts, dont au moins un réel. Lorsque $J \in \mathbb{R}\mathcal{J}_\omega$ est suffisamment générique, il n'y a qu'un nombre fini de courbes J-holomorphes rationnelles réelles connexes homologues à d et contenant $\underline{x}$. Ces courbes sont toutes irréductibles, lisses et de partie réelle non-vide. On note $\mathcal{R}_d(\underline{x}, J)$ cet ensemble fini de courbes. Le fibré normal de chacune de ces courbes $C \in \mathcal{R}_d(\underline{x}, J)$ est équipé d'un opérateur de Cauchy-Riemann généralisé surjectif D_C qui possède donc un état spinoriel $sp(C)$ d'après ce qui précède. On pose

$$\chi_r^{d,\mathfrak{s}}(\underline{x}, J) = \sum_{C \in \mathcal{R}_d(\underline{x}, J)} sp(C) \in \mathbb{Z}.$$

Théorème 3.4 ([24]). *Soit (X, ω, c_X) une variété symplectique réelle fortement semi-positive de dimension six, de lieu réel orientable muni d'une structure spin $\mathfrak{s}$. Soit $d \in H_2(X;\mathbb{Z})$ telle que $(c_X)_* d = -d$, $c_1(X)d$ est pair et strictement plus grand que deux. Soient $k_d = \frac{1}{2}c_1(X)d$ et $\underline{x}$ une configuration réelle de k_d points distincts, dont au moins un réel. et $r = (r_1, \ldots, r_N)$ le N-uplet associé. Alors, l'entier $\chi_r^{d,\mathfrak{s}}(\underline{x}, J)$ ne dépend ni du choix de $\underline{x}$, ni du choix générique de $J \in \mathbb{R}\mathcal{J}_\omega$.*

Remarquons que ce résultat permet de noter sans ambiguïté l'invariant $\chi_r^{d,\mathfrak{s}}$, c'est un invariant par déformation fortement semipositive de (X, ω, c_X). On en déduit les bornes inférieures suivantes.

Corollaire ([24]). *Sous les hypothèses du Théorème 3.4,* $|\chi_r^{d,\mathfrak{s}}| \leq \#\mathcal{R}_d(\underline{x}, J)$, *pour tout choix de configuration réelle* $\underline{x} \in X^{k_d}$ *telle que* $\underline{x} \cap \mathbb{R}X = r$, *et tout choix générique de* $J \in \mathbb{R}\mathcal{J}_\omega$. □

Les invariants qui ressortent des Théorèmes 1.1 et 3.1 ont été interprétés par C.-H. Cho [5] et J. Solomon [22]. Leur approche consiste à d'abord définir la classe fondamentale $[\mathbb{R}_\tau \overline{\mathcal{M}}_{k_d}^d(X)]$ donnée par la Proposition 3.3 en utilisant les travaux de K. Fukaya, Y.-G. Oh, H. Ohta et K. Ono [6], [7], puis à en déduire l'existence des invariants grâce à la relation entre classes fondamentales donnée à la suite de cette proposition. J. Solomon a étendu ces invariants aux courbes de genre strictement positifs mais de structure conforme fixée et aux variétés symplectiques de dimension six, notamment de Calabi-Yau. Dans le cas des quintiques de $\mathbb{C}P^4$, l'invariant a été calculé par R. Pandharipande, J. Solomon et J. Walcher [17].

3.3. Optimalité, congruences et calculs dans le cas de l'ellipsoïde de dimension trois

Théorème 3.5 ([30]). *Soient* (X, c_X) *la quadrique ellipsoïde de dimension trois et* $d \in H_2(X; \mathbb{Z})$ *satisfaisant* $c_1(X)d = 2 \mod (4)$. *L'invariant* χ_1^d *est alors négatif et les bornes inférieures apparues dans le Corollaire 3.1 sont optimales, atteintes lorsque les conditions d'incidence non réelles sont choisies suffisamment proches d'une section hyperplane réelle disjointe du lieu réel* $\mathbb{R}X$.

Remarque 4. La condition $c_1(X)d = 2 \mod (4)$ garantit la parité de l'entier k_d de sorte que l'on peut effectivement choisir un point réel. Lorsque $c_1(X)d = 0 \mod (4)$, et $r = 0$, l'invariant χ_r^d n'est pas défini. Toutefois, on a montré dans ce cas là qu'il existe une structure presque-complexe générique $J \in \mathbb{R}\mathcal{J}_\omega$ et k_d points complexes conjugués pour lesquels aucune courbe J-holomorphe rationnelle réelle homologue à d contient ces k_d points, voir le Théorème 4.2.

Théorème 3.6 ([30]). *Soient* (X, c_X) *la quadrique ellipsoïde de dimension trois et* d *un multiple positif, disons* $\delta > 0$, *d'une section hyperplane réelle. Lorsque* $6r + 1 \leq 3\delta$, *la puissance* $2^{\frac{3}{4}(\delta - 2r)}$ *divise* χ_r^d.

Corollaire ([30]). *Soit* (X, ω, c_X) *une variété symplectomorphe à la quadrique ellipsoïde de dimension trois. Alors,* $\chi_1^2 = -1$, $\chi_1^6 = 0$ *et* $\chi_1^{10} = -896$.

Dans le cas de l'espace projectif de dimension trois, une formule calculant $\chi_{2d}^d(\mathbb{C}P^3)$ pour tout degré d est annoncée par E. Brugallé et G. Mikhalkin dans [4]. En particulier, $\chi_{10}^5 = 45$, $\chi_{14}^7 = -14589$, tandis qu'en degré pair l'invariant s'annule pour des raisons de symétrie.

4. Sur la Présence et L'absence de Membranes J-holomorphes

4.1. Absence de membranes J-holomorphes. Soit C une membrane J-holomorphe à bord dans une sous-variété lagrangienne L d'une variété symplectique fermée (X, ω). Notons χ la caractéristique d'Euler de cette membrane, $d \in H_2(X, L; \mathbb{Z})$ sa classe d'homologie relative et $\mu_{TX} \in H^2(X, L; \mathbb{Z})$ la classe de Maslov de la paire (X, L). La dimension attendue de l'espace des déformations de C s'écrit $\langle \mu_{TX}, d\rangle + (n-3)\chi$. Cette dimension chute lorsque l'on impose à C des contraintes supplémentaires. Si l'on impose par exemple à cette membrane de rencontrer p cycles de codimensions $2 + q_1, \ldots, 2 + q_p$, cette dimension attendue chute de la somme $q = q_1 + \cdots + q_p$. Deux problèmes généraux sous-tendent nos résultats. Il s'agit d'une part de compter les membranes J-holomorphes homologues à d soumises à de telles conditions d'incidence de sorte que ce comptage ne dépende pas de J et ne dépende des conditions d'incidence qu'à homologie près. Il s'agit d'autre part de minimiser ce nombre de membranes. Si nous ne pouvons répondre au premier problème dans ce degré de généralité, il nous est par contre parfois possible de répondre au second sans même supposer l'égalité $q = \langle \mu_{TX}, d\rangle + (n-3)\chi$, lorsque le minimum en question est nul. Le présent paragraphe est consacré aux résultats que l'on a pu obtenir dans cette direction. Ici encore le minimum est atteint en allongeant le cou d'une structure presque complexe générale.

4.1.1. En dimension supérieure

Théorème 4.1 ([30], [31]). *Soit L une sphère lagrangienne dans une variété symplectique fermée (X, ω) satisfaisant $c_1(X) = \lambda\omega$, $\lambda \leq 0$ et soit $E > 0$. Supposons la dimension de X supérieure à cinq. Pour toute structure presque-complexe J générale ayant un cou suffisamment long au voisinage de L, cette variété ne possède ni membrane J-holomorphe reposant sur L ni courbe J-holomorphe rencontrant L qui soit d'énergie inférieure à E. Ce résultat reste valable en dimension quatre pour les courbes ou membranes de genre nul.*

Rappelons que l'énergie d'une courbe C est par définition l'intégrale de la forme ω sur cette courbe. Les variétés projectives à fibré canonique nul ou ample, par exemple les intersections complètes de multidegrés $(d_1, \ldots, d_k)$ de l'espace projectif de dimension N dès lors que $\sum_{i=1}^{k} d_i \geq N + 1$, satisfont les hypothèses du Théorème 4.1. Remarquons qu'une modification de ce dernier s'applique également aux variétés dont le fibré canonique est le produit d'un fibré ample et d'un fibré porté par un diviseur effectif disjoint de L. Le Théorème 4.1 permet de définir la cohomologie de Floer de sphères lagrangiennes dans les variétés symplectiques dont la première classe de Chern s'annule, voir [31] et [6], [7] pour une théorie de l'obstruction à définir en général une telle homologie.

Théorème 4.2 ([30]). *Soit L une sphère lagrangienne dans une variété symplectique fermée semipositive (X, ω) de dimension $2n \geq 6$ et soit $d \in$*

$H_2(X, L; \mathbb{Z})$. Écrivons $\langle \mu_{TX}, d\rangle + (n-3)\chi = q + r$ avec $q \in \mathbb{Z}$, $0 \leq r < 2 + (n-3)\chi$ et $\chi \leq 2$. Lorsque $q \geq 0$, choisissons p cycles de $X \setminus L$ de codimensions $2 + q_1, \ldots, 2 + q_p$ de sorte que $q = q_1 + \cdots + q_p$. Dès que la structure presque complexe générale J possède un cou suffisamment long au voisinage de L, cette variété ne contient aucune membrane J-holomorphe homologue à d, de caractéristique d'Euler χ qui rencontre ces p cycles et repose sur L. Ce résultat reste valable pour des membranes de genre nul lorsque $n = 2$.

Exemple : la quadrique ellipsoïde. Soit X la quadrique ellipsoïde de dimension complexe $n \geq 3$ et H une section hyperplane disjointe de L. Le groupe $H_2(X, L; \mathbb{Z})$ est monogène, engendré par la classe d_0 satisfaisant $\langle H, d_0\rangle = +1$. La première classe de Chern de X vaut nH, d'où l'on déduit le calcul $\langle \mu_{TX}, ld_0\rangle = 2ln$ quel que soit l'entier l. Écrivons $l = (n-1)a + b$, le Théorème 4.2 s'applique par exemple lorsque $n + 1 \leq 2b < 2n$, les membranes sont des disques et lorsque toutes les conditions d'incidence sont ponctuelles. Rappelons que le Théorème 3.5 traite du cas $r = n - 1$ et montre ainsi en un sens l'optimalité des hypothèses faites dans ce Théorème 4.2.

4.1.2. En dimension quatre. Nous noterons $\mathcal{M}_{g,b}$ l'espace des modules des structures complexes de la surface compacte connexe orientée de genre g ayant b composantes de bord.

Proposition 4.3 ([30]). *Soit L une sphère lagrangienne dans une variété symplectique fermée de dimension quatre (X, ω). On suppose que cette dernière ne possède pas de sphère symplectique S satisfaisant $\langle c_1(X), [S]\rangle > 0$. Soit $(d, g, b) \in H_2(X, L; \mathbb{Z}) \times \mathbb{N} \times \mathbb{N}^*$ et K un compact de $\mathcal{M}_{g,b}$. Alors, pour toute structure presque-complexe générale ayant un cou suffisamment long au voisinage de L, la variété ne possède pas de membrane J-holomorphe homologue à d à bord dans L et conforme à un élément de K.*

Proposition 4.4 ([30]). *Soit L une surface lagrangienne orientable hyperbolique dans une variété symplectique fermée de dimension quatre (X, ω) et soit $d \in H_2(X, L; \mathbb{Z})$. On note $N_d^g(\underline{x}, J)$ le nombre de courbes J-holomorphes homologues à d à bords dans L, de topologie et de structure conforme données et qui passent par une configuration $\underline{x}$ de points distincts de (X, ω) de cardinal adéquat, pour $J \in \mathcal{J}_\omega$ générique. Ce nombre $N_d^g(\underline{x}, J)$ s'annule pour toute structure presque-complexe générale ayant un cou suffisamment long au voisinage de L.*

Proposition 4.5 ([30]). *Soit (X, ω, c_X) une variété symplectique réelle fermée de dimension quatre dont le lieu réel possède un tore lagrangien ou bien une surface hyperbolique lagrangienne L, orientable ou non. On suppose que (X, ω, c_X) ne possède pas de sphère symplectique réelle S satisfaisant $\langle c_1(X), [S]\rangle > 1$ si L est orientable et $\langle c_1(X), [S]\rangle > 0$ sinon. Soit $(d, g, b) \in H_2(X, L; \mathbb{Z}) \times \mathbb{N} \times \mathbb{N}^*$ et K un compact de $\mathcal{M}_{g,b}$. Alors, pour toute structure presque-complexe générale ayant un cou suffisamment long au voisinage de L, la variété ne possède pas*

de membrane J-holomorphe homologue à d à bord dans L et conforme à un élément de K.

4.2. Présence de membranes J-holomorphes. Les résultats présentés aux §§1.1 et 3.1 permettent de garantir l'existence de disques J-holomorphes reposant sur une sous-variété lagrangienne d'une variété symplectique donnée, lorsque cette lagrangienne se trouve dans le lieu fixe d'une involution antisymplectique, laquelle est J-antiholomorphe et à condition que l'invariant que l'on a défini n'est pas nul. Nous souhaitons montrer ici qu'il est possible d'obtenir ces résultats pour une classe plus large de sous-variété lagrangiennes, en faisant intervenir la notion d'involutions antibirationnelles sur les variétés symplectiques.

4.2.1. Involutions antibirationnelles des variétés symplectiques de dimension quatre. Une involution c_X de la variété symplectique de dimension quatre (X, ω) qui est définie en-dehors d'un nombre fini de points $x_1, \ldots, x_k$ de X est dite *antibirationnelle* lorsqu'il existe un diagramme commutatif de la forme suivante :

$$\begin{array}{ccc} (Y, J_Y) & \xrightarrow{c_Y} & (Y, J_Y) \\ \pi \downarrow & & \downarrow \pi \\ (X, J_X) & \xrightarrow{c_X} & (X, J_X) \end{array}$$

où Y est une variété compacte de dimension quatre obtenue à partir de X en réalisant un nombre fini d'éclatements topologiques au-dessus des points x_i, $i \in \{1, \ldots, k\}$, J_X, J_Y sont des structures presque-complexes lisses et c_Y une involution J_Y-antiholomorphe sur Y toute entière. De plus, J_X est supposée ω-positive, c_X est J_X-antiholomorphe sur son lieu de définition et π est (J_Y, J_X)-holomorphe.

Les involutions antibirationnelles classiques sur les surfaces compactes de Kähler fournissent des exemples de telles surfaces. Remarquons que pour tout $i \in \{1, \ldots, k\}$, $\pi^{-1}(x_i)$ est un arbre de sphères J_Y-holomorphes n'ayant que des points doubles transverses comme singularités.

Lemme 4.6. *Supposons que pour tout $i \in \{1, \ldots, k\}$ et toute composante irréductible C de l'arbre $\pi^{-1}(x_i)$, $c_Y(C)$ ne soit pas contractée par π sur $x_1, \ldots, x_k$. Alors, le diagramme ci-dessus est unique à équivalence près, une fois donnée (X, ω, c_X).*

Soient (X, ω, J_X, c_X) satisfaisant les hypothèses du Lemme 4.6 et (Y, J_Y, c_Y) la variété de dimension quatre associée. Soit $\underline{y}$ l'ensemble fini $\big(\cup_{i=1}^k \pi^{-1}(x_i)\big) \cap c_Y\big(\cup_{i=1}^k \pi^{-1}(x_i)\big)$. L'involution antibirationnelle c_X est dite *simple* lorsqu'elle satisfait les hypothèses du Lemme 4.6 et lorsque $\underline{y}$ se trouve en-dehors des points doubles de $\cup_{i=1}^k \pi^{-1}(x_i)$.

Lemme 4.7. *Soit c_X une involution antibirationnelle simple de (X, ω) et (Y, c_Y) la variété de dimension quatre donnée par le Lemme 4.6. Alors, la*

deux-forme $\omega_Y = \pi^*\omega - (\pi \circ c_Y)^*\omega$ *est fermée et non-dégénérée en tout point de* $Y \setminus \underline{y}$. *Elle est également non-dégénérée en tout point d'intersection transverse de* $\big(\cup_{i=1}^k \pi^{-1}(x_i)\big) \cap c_Y\big(\cup_{i=1}^k \pi^{-1}(x_i)\big) \subset \underline{y}$.

Une telle deux-forme qui n'a qu'un nombre fini de noyaux de dimension deux sera dite *quasi-symplectique*. Remarquons qu'en particulier, lorsque l'intersection $\big(\cup_{i=1}^k \pi^{-1}(x_i)\big) \cap c_Y\big(\cup_{i=1}^k \pi^{-1}(x_i)\big)$ est transverse, la deux-forme ω_Y est symplectique.

La structure presque-complexe J_Y est ω_Y-positive dans le sens que pour tous $y \in Y$ et $v \in T_yY \setminus \{0\}$, soit v et $J_Y(v)$ engendrent le noyau de $\omega_Y|_y$, soit $\omega_Y(v, J_Y(v)) > 0$. Notons $\mathcal{J}_{\omega_Y}$ l'espace des structures presque-complexes de classe C^l qui sont ω_Y-positives. Si $J \in \mathcal{J}_{\omega_Y}$, alors $\overline{c}_Y^*(J) = -dc_Y \circ J \circ dc_Y$ appartient également à $\mathcal{J}_{\omega_Y}$. Notons $\mathbb{R}\mathcal{J}_{\omega_Y}$ le lieu fixe de cette action de $\mathbb{Z}/2\mathbb{Z}$ sur $\mathcal{J}_{\omega_Y}$.

Lemme 4.8. *Soient* c_X *une involution antibirationnelle simple sur* (X, ω) *et* (Y, c_Y, ω_Y) *la variété de dimension quatre donnée par les Lemmes 4.6, 4.7. Il existe* $J_0 \in \mathbb{R}\mathcal{J}_{\omega_Y}$ *tel que* $\omega_Y(J_0, J_0) = \omega_Y$ *et* $g_Y = \omega_Y(., J_0)$ *soit un deux-tenseur symétrique positif sur* Y, *défini en-dehors de* $\underline{y}$.

Pour tout voisinage U de $\underline{y}$ et tout $J_0 \in \mathbb{R}\mathcal{J}_{\omega_Y}$ donné par le Lemme 4.9, notons $\mathcal{J}_{\omega_Y}^{U,J_0}$ (resp. $\mathbb{R}\mathcal{J}_{\omega_Y}^{U,J_0}$) le sous-espace des $J \in \mathcal{J}_{\omega_Y}$ (resp. $J \in \mathbb{R}\mathcal{J}_{\omega_Y}$) telles que $J = J_0$ sur U.

Lemme 4.9. *Pour tous* U, J_0, *l'espace* $\mathcal{J}_{\omega_Y}^{U,J_0}$ *est une variété de Banach séparable non-vide et contractile. Le sous-espace* $\mathbb{R}\mathcal{J}_{\omega_Y}^{U,J_0}$ *en est une sous-variété de Banach séparable non-vide et contractile.* □

Remarque 5. La deux-forme $\pi^*\omega$ est limite d'une suite de formes symplectiques sur Y obtenues après un nombre fini d'éclatements de boules symplectiques dont les rayons convergent vers zéro. Par suite, la deux-forme ω_Y est limite d'une suite de formes symplectiques $(\omega_Y^n)_{n\in\mathbb{N}}$. Alors, $J \in \mathcal{J}_{\omega_Y}$ est ω_Y^n-positif pour n assez grand, principalement parce-que les noyaux de ω_Y deviennent symplectiques pour ω_Y^n.

4.2.2. Invariants énumératifs des involutions antibirationnelles simples. Soient c_X une involution antibirationnelle simple sur (X, ω) et $x_1, \ldots, x_k \in X$ les points où elle n'est pas définie. Soit (Y, ω_Y, c_Y) la variété quasi-symplectique de dimension quatre associée, voir le Lemme 4.7. Soient π la projection $Y \to X$ et $\underline{y}$ l'ensemble fini $\big(\cup_{i=1}^k \pi^{-1}(x_i)\big) \cap c_Y\big(\cup_{i=1}^k \pi^{-1}(x_i)\big)$. Soit $\mathbb{R}Y$ le lieu fixe de c_Y, on étiquette ses composantes connexes $(\mathbb{R}Y)_1, \ldots, (\mathbb{R}Y)_N$. Remarquons que la courbe $\big(\cup_{i=1}^k \pi^{-1}(x_i)\big) \cup c_Y\big(\cup_{i=1}^k \pi^{-1}(x_i)\big)$ n'intersecte $\mathbb{R}Y$ qu'en un nombre fini de points, de sorte qu'elle ne déconnecte aucune des courbes $(\mathbb{R}Y)_i$, $i \in \{1, \ldots, N\}$. Soient $d_Y \in H_2(Y; \mathbb{Z})$ tel que $(c_Y)_* d_Y = -d_Y$, $c_1(Y)d_Y > 0$ et $y = (y_1, \ldots y_{c_1(Y)d_Y - 1})$ une configuration réelle de $c_1(Y)d_Y - 1$

points distincts de $Y \setminus \big(\cup_{i=1}^k \pi^{-1}(x_i) \cup c_Y\big(\cup_{i=1}^k \pi^{-1}(x_i)\big)\big)$. Pour tout $i \in \{1, \dots, N\}$, notons $r_i = \#(y \cap (\mathbb{R}Y)_i)$ puis $r = (r_1, \dots, r_N)$. Soient U, voisinage de $\underline{y}$ et $J_0 \in \mathbb{R}\mathcal{J}_{\omega_Y}$ donnés par le Lemme 4.9. Alors, dès que U est suffisamment petit, pour tout $J \in \mathbb{R}\mathcal{J}_{\omega_Y}^{U,J_0}$ générique, il n'y a qu'un nombre fini de courbes J-holomorphes rationnelles réelles homologue à d_Y dans Y qui contiennent y. Ces courbes sont toutes irréductibles, immergées et n'ont que des points doubles transverses comme singularités. Le nombre total de leurs points doubles vaut $\delta_Y = \frac{1}{2}(d_Y^2 - c_1(Y)d_Y + 2)$. Pour tout entier m compris entre 0 et δ_Y, notons $n_d(m)$ le nombre de ces courbes qui sont de masse m. On pose alors

$$\chi_r^{d_Y}(y, J, U, J_0) = \sum_{m=0}^{\delta_Y} (-1)^m n_d(m).$$

Théorème 4.10. *L'entier $\chi_r^{d_Y}(y, J, U, J_0)$ ne dépend pas des choix de y, J, U et J_0.* □

Remarquons que l'entier $\chi_r^{d_Y}$ fourni par le Théorème 4.10 est un invariant par déformation du triplet (X, ω, c_X), puisque le triplet (Y, ω_Y, c_Y) lui est canoniquement associé.

4.2.3. Exemple : les tores isotopes au tore de Clifford. Soient $a, b \in \mathbb{R}_+^*$ et $\mathbb{T}_{a,b} \subset \mathbb{C}P^2$ le tore lagrangien défini par les équations $|x| = a$, $|y| = b$ dans les coordonnées affines $(x, y) \in (\mathbb{C})^2 \subset \mathbb{C}P^2$. Ce tore est le lieu fixe de l'involution antibirationnelle de Cremona $c^{a,b} : (x, y, z) \in \mathbb{C}P^2 \setminus \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\} \mapsto (a^2\overline{yz}, b^2\overline{xz}, \overline{xy}) \in \mathbb{C}P^2$. Cette involution antibirationnelle $c^{a,b}$, $a, b \in \mathbb{R}_+^*$, est simple. En effet, soit Y le plan projectif éclaté aux trois points $(1, 0, 0), (0, 1, 0), (0, 0, 1)$ et $\pi : Y \to \mathbb{C}P^2$ la projection associée. L'involution $c^{a,b}$ se relève en une involution antiholomorphe $c_Y^{a,b}$ définie partout, soit une structure réelle. Cette dernière envoie les trois diviseurs exceptionnels sur les transformées strictes des côtés du triangle $(1, 0, 0), (0, 1, 0), (0, 0, 1)$, d'où la simplicité de $c^{a,b}$. Ainsi, le Théorème 4.10 s'applique et fournit des invariants $\chi_r^{d_Y}$ par déformation du triplet $(\mathbb{C}P^2, \omega, c^{a,b})$. Le deuxième groupe d'homologie de Y est engendré par une droite générique et les diviseurs exceptionnels $E_1, \dots E_3$ de nos éclatements. La classe d'homologie d_Y de nos courbes rationnelles réelles de Y est déterminée par quatre entiers $d, d_1, \dots, d_3$ satisfaisant la relation $d = d_1 + d_2 + d_3$. Si l'on contracte $E_1, \dots E_3$, ces courbes se contractent sur des courbes rationnelles de degré d du plan qui ont un point de multiplicité d_1, d_2, d_3 en $(1, 0, 0), (0, 1, 0), (0, 0, 1)$ respectivement. Ces courbes rationnelles immergées ont en outre la propriété de rencontrer le tore $\mathbb{T}_{a,b}$ en une collection de points isolés et en un cercle immergé, elles consistent en fait en une paire de disques J-holomorphes qui reposent sur $\mathbb{T}_{a,b}$ et sont échangés par $c^{a,b}$. Si l'on contracte plutôt un diviseur exceptionnel, disons E_3, ainsi que son image sous $c_Y^{a,b}$, alors on obtient des courbes rationnelles de bidegré $(d_1 + d_3, d_2 + d_3)$ sur l'hyperboloïde quadrique $(\mathbb{C}P^1 \times \mathbb{C}P^1, \mathrm{conj} \times \mathrm{conj})$, qui ont une paire de points de multiplicité d_3 en deux points complexes conjugués, à savoir les points où E_3

et son image se contractent. Lorsque $d_3 = 0$ ou 1, cet invariant $\chi_r^{d_Y}$ vaut l'invariant correspondant dans l'hyperboloïde quadrique $(\mathbb{C}P^1 \times \mathbb{C}P^1, \mathrm{conj} \times \mathrm{conj})$, à savoir $\chi_r^{(d_1,d_2)}$ et $\chi_r^{(d_1+1,d_2+1)}$ respectivement. Des estimations de ces derniers se trouvent dans [12].

Corollaire. *Soient $r, s, d \in \mathbb{N}$ tels que $r + 2s = 2d - 1$ et supposons donnée une collection de r points distincts dans $\mathbb{T}_{a,b} \subset \mathbb{C}P^2$, $a, b \in \mathbb{R}_+^*$ ainsi qu'une collection de s paires distinctes de points dans $\mathbb{C}P^2 \setminus \big(\mathbb{T}_{a,b} \cup \{(1,0,0),(0,1,0),(0,0,1)\}\big)$ échangées par l'involution antibirationnelle $c^{a,b}$. Alors, pour tous $d_1, d_2, d_3 \in \mathbb{N}$ tels que $d = d_1 + d_2 + d_3$, il y a au moins $|\chi_r^{d_Y}|$ paires de disques J_X-holomorphes reposant sur $\mathbb{T}_{a,b}$, échangés par $c^{a,b}$, passant par les r points donnés et intersectant chacune des s paires de points complexes conjugués, dès lors que J_X se relève en une structure J_Y appartenant à l'un des $\mathbb{R}\mathcal{J}_{\omega_Y}^{U,J_0}$ donné par le Lemme 4.9. La réunion de ces deux disques dans chacune de ces paires forme une courbe rationnelle plane de degré d ayant un point de multiplicité d_1, d_2, d_3 en $(1,0,0),(0,1,0),(0,0,1)$ respectivement.* $\square$

Remarquons que des invariants énumératifs portant sur des disques à bords dans le tore de Clifford ont été obtenus par P. Biran et O. Cornea [2]. Les disques holomorphes à bords dans le tores de Clifford ont par ailleurs été étudiés par C.-H. Cho dans sa thèse en termes de produits de Blaschke. En ce qui concerne nos résultats présentés dans ce paragraphe 4.2, il reste à s'affranchir de la notion de simplicité (des involutions antibirationnelles).

Références

[1] A. Arroyo, E. Brugallé, and L. López De Medrano. Recursive formulas for Welschinger invariants. *Prépublication math.arXiv :0809.1541*, 2008.

[2] P. Biran and O. Cornea. A Lagrangian quantum homology. In *New perspectives and challenges in symplectic field theory*, volume 49 of *CRM Proc. Lecture Notes*, pages 1–44. Amer. Math. Soc., Providence, RI, 2009.

[3] F. Bourgeois. A Morse-Bott approach to Contact Homology. *Ph.D dissertation, Stanford University*, 2002.

[4] E. Brugallé and G. Mikhalkin. Enumeration of curves via floor diagrams. *C. R. Math. Acad. Sci. Paris*, 345(6) :329–334, 2007.

[5] C.-H. Cho. Counting real J-holomorphic discs and spheres in dimension four and six. *J. Korean Math. Soc.*, 45(5) :1427–1442, 2008.

[6] K. Fukaya, Y.-G. Oh, H. Ohta, and K. Ono. *Lagrangian intersection Floer theory : anomaly and obstruction. Part I*, volume 46 of *AMS/IP Studies in Advanced Mathematics*. American Mathematical Society, Providence, RI, 2009.

[7] K. Fukaya, Y.-G. Oh, H. Ohta, and K. Ono. *Lagrangian intersection Floer theory : anomaly and obstruction. Part II*, volume 46 of *AMS/IP Studies in Advanced Mathematics*. American Mathematical Society, Providence, RI, 2009.

[8] A. Gathmann and H. Markwig. The Caporaso-Harris formula and plane relative Gromov-Witten invariants in tropical geometry. *Math. Ann.*, 338(4) :845–868, 2007.

[9] M. Gromov. Pseudoholomorphic curves in symplectic manifolds. *Invent. Math.*, 82(2) :307–347, 1985.

[10] A. Grothendieck. Sur la classification des fibrés holomorphes sur la sphère de Riemann. *Amer. J. Math.*, 79 :121–138, 1957.

[11] H. Hofer, K. Wysocki, and E. Zehnder. Properties of pseudoholomorphic curves in symplectisations. I. Asymptotics. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 13(3) :337–379, 1996.

[12] I. Itenberg, V. Kharlamov, and E. Shustin. Welschinger invariant and enumeration of real rational curves. *Int. Math. Res. Not.*, (49) :2639–2653, 2003.

[13] I. Itenberg, V. Kharlamov, and E. Shustin. A Caporaso-Harris type formula for Welschinger invariants of real toric del Pezzo surfaces. *Comment. Math. Helv.*, 84(1) :87–126, 2009.

[14] I. V. Itenberg, V. M. Kharlamov, and E. I. Shustin. Logarithmic equivalence of the Welschinger and the Gromov-Witten invariants. *Uspekhi Mat. Nauk*, 59(6(360)) :85–110, 2004.

[15] A. I. Degtyarev and V. M. Kharlamov. Topological properties of real algebraic varieties : Rokhlin's way. *Uspekhi Mat. Nauk*, 55(4(334)) :129–212, 2000.

[16] G. Mikhalkin. Enumerative tropical algebraic geometry in $\mathbb{R}^2$. *J. Amer. Math. Soc.*, 18(2) :313–377 (electronic), 2005.

[17] R. Pandharipande, J. Solomon, and J. Walcher. Disk enumeration on the quintic 3-fold. *J. Amer. Math. Soc.*, 21(4) :1169–1209, 2008.

[18] N. Puignau. Première classe de Stiefel-Whitney des espaces d'applications stables réelles en genre zéro vers une surface convexe. *J. Inst. Math. Jussieu*, 8(2) :383–414, 2009.

[19] F. Ronga, A. Tognoli, and T. Vust. The number of conics tangent to five given conics : the real case. *Rev. Mat. Univ. Complut. Madrid*, 10(2) :391–421, 1997.

[20] E. Shustin. A tropical calculation of the Welschinger invariants of real toric del Pezzo surfaces. *J. Algebraic Geom.*, 15(2) :285–322, 2006.

[21] E. Shustin. Welschinger invariants of toric del Pezzo surfaces with nonstandard real structures. *Tr. Mat. Inst. Steklova*, 258(Anal. i Osob. Ch. 1) :227–255, 2007.

[22] J. P. Solomon. Intersection theory on the moduli space of holomorphic curves with Lagrangian boundary conditions. *Prépublication math.SG/0606429*, 2006.

[23] J.-Y. Welschinger. Invariants of real rational symplectic 4-manifolds and lower bounds in real enumerative geometry. *C. R. Math. Acad. Sci. Paris*, 336(4) :341–344, 2003.

[24] J.-Y. Welschinger. Enumerative invariants of strongly semipositive real symplectic six-manifolds. *Prépublication math.AG/0509121*, 2005.

[25] J.-Y. Welschinger. Invariants of real symplectic 4-manifolds and lower bounds in real enumerative geometry. *Invent. Math.*, 162(1) :195–234, 2005.

[26] J.-Y. Welschinger. Spinor states of real rational curves in real algebraic convex 3-manifolds and enumerative invariants. *Duke Math. J.*, 127(1) :89–121, 2005.

[27] J.-Y. Welschinger. Invariants of real symplectic four-manifolds out of reducible and cuspidal curves. *Bull. Soc. Math. France*, 134(2) :287–325, 2006.

[28] J.-Y. Welschinger. Towards relative invariants of real symplectic four-manifolds. *Geom. Funct. Anal.*, 16(5) :1157–1182, 2006.

[29] J.-Y. Welschinger. Invariant count of holomorphic disks in the cotangent bundles of the two-sphere and real projective plane. *C. R. Math. Acad. Sci. Paris*, 344(5) :313–316, 2007.

[30] J.-Y. Welschinger. Optimalité, congruences et calculs d'invariants des variétés symplectiques réelles de dimension quatre. *Prépublication math.SG/0707.4317*, 2007.

[31] J.-Y. Welschinger. Open strings, Lagrangian conductors and Floer functor. *Prépublication math.arXiv :0812.0276*, 2008.

Section 5

Geometry

Anna Erschler
Poisson-Furstenberg Boundaries, Large-scale Geometry and Growth of Groups 681

Jixiang Fu
On non-Kähler Calabi-Yau Threefolds with Balanced Metrics 705

William M. Goldman
Locally Homogeneous Geometric Manifolds 717

Larry Guth
Metaphors in Systolic Geometry 745

Sergei Ivanov
Volume Comparison via Boundary Distances 769

Xiaonan Ma
Geometric Quantization on Kähler and Symplectic Manifolds 785

Fernando Codá Marques
Scalar Curvature, Conformal Geometry, and the Ricci Flow with Surgery 811

Isabel Fernández* **and Pablo Mira***
Constant Mean Curvature Surfaces in 3-dimensional Thurston Geometries 830

Alexander Nabutovsky
Morse Landscapes of Riemannian Functionals and Related Problems 862

Frank Pacard
Constant Scalar Curvature and Extremal Kähler Metrics on Blow ups 882

Takao Yamaguchi
Reconstruction of Collapsed Manifolds 899

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Poisson-Furstenberg Boundaries, Large-scale Geometry and Growth of Groups

Anna Erschler*

Abstract

We give a survey of recent results on the Poisson-Furstenberg boundaries of random walks on groups, and their applications. We describe sufficient conditions for random walk to have non-trivial boundary, or, on the contrary, to have trivial boundary. We review recent progress in description of the boundary for random walks on various groups, including wreath products. We describe how the Poisson-Furstenberg boundary can be used to obtain lower bounds for the growth function of the groups of intermediate growth. We also discuss relation between properties of the boundary with other asymptotic properties of groups, including isoperimetry and various characteristics of random walks.

Mathematics Subject Classification (2010). Primary 20F69, 60B15; Secondary 43A05, 43A07, 60G50, 60J50, 30F15.

Keywords. Random walks on groups, boundary, harmonic function, amenable groups, growth of groups.

1. Boundaries of Random Walks on Groups

The Poisson boundary is a probability space, defined by a Markov chain (Feller [41]). In the case when the Markov chain is a random walk on a group, this space is naturally endowed with the action of this group, and there are several equivalent ways to define it (see Furstenberg [43, 44, 45], Kaimanovich, Vershik [61]). If the group acts on a symmetric space, then this action induces an action on a naturally defined geometric boundary of this space. The Poisson-Furstenberg boundary can be viewed in such cases as a probability measure on the geometric boundary, and this measure adds essential information to the

*Université Paris Sud XI, Orsay, France. E-mail: anna.erschler@math.u-psud.fr.

understanding of both algebraic and geometric properties of the group. An important feature is that, unlike the geometric boundary and unlike some other notions of boundary such as the Martin boundary, the Poisson-Furstenberg boundary behaves functorially with respect to the group homomorphisms. This has far-reaching applications, such as Furstenberg's approach to superrigidity theorems (see Furstenberg, Margulis, [45, 74]). Besides superrigidity, measures on the geometric boundaries appear for (not necessarily symmetric) hyperbolic spaces. For example, the measure on the boundary appears in the proof of monotonicity of the hyperbolic volume, where the use of this measure is crucial for finite volume non-compact manifolds (Thurston, see the exposition in Gromov [50]).

In the more general context, it happens that there is no natural geometric boundary attached to the group. However, the Poisson-Furstenberg boundary is always well defined, in so far as we fix some probability measure on the group. This is the subject of the present paper. The Poisson-Furstenberg boundary, regarded as a measure space with the group action, is related to many natural questions in random walks and harmonic analysis, and in the last years it turned out that this space has also applications to the growth of groups.

There are several ways to define the Poisson-Furstenberg boundary for a random walk on a group. We recall some of the equivalent definitions.

Definition 1. Consider two infinite trajectories X and Y. We say that they are equivalent if they coincide after some instant, possibly up to the time shift. This means there exits $N, k \geq 0$ such that $X_i = Y_{i+k}$ for all $i > N$. Consider the measurable hull of this equivalence relation in the space of infinite trajectories. The quotient by this equivalence relation is called *the Poisson-Furstenberg boundary.*

If in this definition we do not allow the time shift, that is, if we say that X and Y are equivalent whenever $X_i = Y_i$ for all $i > N$, then the resulting quotient space is called *the tail boundary.* For a random walk on a group these two definitions give the same space, while in a more general context of random walks on graphs the tail boundary may happen to be larger than the Poisson-Furstenberg boundary. The Poisson boundary is an interesting notion to study in the more general contexts of Markov chains and random walks, but there is more additional structure on such spaces in the case of random walks on groups. Apart from the above mentioned fact about the tail boundaries, there are many other manifestations of such phenomena. The entropy criterion, for example, which we recall below, does not hold in a more general context of not necessarily homogeneous spaces, such as non vertex-transitive graphs.

A function $F : G \to \mathbb{R}$ is called *μ-harmonic*, if for all $g \in G$ it holds $F(g) = \sum_{h \in G} F(gh)\mu(h)$. The Poisson-Furstenberg boundary can be equivalently defined in terms of bounded harmonic functions (from the subgroup, generated by the support of μ to $\mathbb{R}$):

Definition 2. The space of all bounded μ-harmonic functions, can be endowed with multiplication: given two bounded harmonic functions f_1 and f_2, put

$$(f_1 \times f_2)(x) = \lim_{n\to\infty} \sum_x f_1(gx) f_2(gx) \mu^{*n}(x).$$

One can prove that the limit above exists, and that this product is associative. It is easy to check that then $f_1 \times f_2$ is harmonic. Since the limit of bounded harmonic function with respect to the supremum norm is again bounded and harmonic, one concludes that the space of bounded μ-harmonic functions forms a commutative Banach algebra. Its spectrum Π_μ is endowed with a probability measure ν, defined by the following equality, which holds for all f: $\int \hat{f}(x) d\nu(x) = f(e)$, where $\hat{f}$ is the Gelfand transform of f. The set Π_μ, as a measure G-space, is isomorphic to the Poisson-Furstenberg boundary.

In particular, the equivalence of the definitions implies that the group G admits nonconstant bounded harmonic functions with respect to some measure μ, with the support generating G, if and only if the Poisson-Furstenberg boundary of the random walk is non-trivial. To see one of the implications observe that for any subset of the boundary, according to the first definition, the probability to hit this set (that is, the probability that the equivalence class of the trajectory belongs to this set) is a harmonic function between 0 and 1, which is non-constant so far as the set, as well as its compliment in the boundary, both have positive probability.

For more on different definitions of the Poisson boundary see [61]. For more recent surveys see [42, 6].

A harmonic function on a group is a discrete counterpart of a harmonic function on a Riemannian manifold. The random walk is said to be *symmetric* if $\mu(g) = \mu(g^{-1})$ for any $g \in G$. Given a regular cover M, with deck transformation group G, there is a symmetric measure μ on G, such that the Poisson-Furstenberg boundary of G can be identified with that of M, in particular, G admits bounded μ-harmonic functions if and only if M does. This measure is called Furstenberg discretization or Furstenberg-Lyons-Sullivan discretization [71, 55]. This measure is in general infinitely supported. It has a rapid decay, exponential moments of this measure are finite.

(We do not touch in this paper the questions concerning unbounded harmonic functions, and their relation to the random walk, such as positive harmonic functions and the corresponding Martin boundary, see [85]).

Some of the questions one asks about boundaries of random walks are as follows:

- given a group G and a probability measure μ on G, can we say whether the boundary of (G, μ) is trivial or not?
- If the boundary is non-trivial, can we describe at least some μ-boundaries, that is, some non-trivial quotients of the Poisson-Furstenberg boundary?
- Can one provide a complete description of the boundary (G, μ)?

- Can one obtain some information on the large scale geometry of G, granting some information on μ and the boundary of (G, μ), such as the triviality/non-triviality of the boundary of this random walk? Such as the description of this boundary?

We recall that if the boundary of the random walk is trivial, then the group, generated by the support of μ, is amenable, so the first question is essentially about amenable groups. The random walk (G, μ) is said to be *non-degenerate*, if the support of μ generates G as a semi-group. It is also known that any amenable group admits a non-degenerate symmetric measure with trivial boundary (Rosenblatt; Kaimanovich, Vershik [77, 60, 61]). For application of this criterion, as well of the generalization of this criterion for the case of amenable extension see [58, 83, 17, 7].

In many groups symmetric non-degenerate measures, provided by Kaimanovich-Vershik-Rosenblatt criterion, can not be chosen to have finite support. In [31] it is shown that on some groups such measures can not be chosen even in the class of measures with finite entropy. For a probability measure μ, we denote by $H(\mu)$ the entropy of μ, that is, $H(\mu) = -\sum_g \mu(g)\log(\mu(g))$, and we recall the entropy criterion of boundary triviality. The notion of the entropy of a random walk on a group is due to Avez [3]. The *entropy of the random walk* (G, μ) is defined as the limit, as n tends to infinity, of $H(\mu^{*n})/n$. Here μ^{*n} denotes the n-th convolution of μ, and the limit exists in view of the subadditivity $H(\mu^{*(n+k)}) \leq H(\mu^{*n}) + H(\mu^{*k})$.

Entropy Criterion (Avez-Derriennic-Kaimanovich-Vershik). *([4, 23, 60, 61]) Let G be a countable group, μ be a probability measure on G such that its entropy $H(\mu)$ is finite. Under this assumption the Poisson-Furstenberg boundary is trivial if and only if entropy of the random walk $h(\mu)$ is equal to zero.*

This criterion shows that just one number, defined by the n-th convolution of μ, gives an answer whether the boundary is trivial or not. The assumption that the entropy is finite is essential. Even if we know exactly the distribution of these n-th convolutions (in other words, if we know exactly the abstract distribution after the n steps of the random walk), then we can not in general forget the underlying group structure and say whether the boundary is trivial or not. There exist examples of measures μ on G, such that the boundary of (G, μ) is non-trivial, while the boundary of the random walk defined by the inverse measure $(G, \breve{\mu})$ is trivial (see Kaimanovich [54] and example 6.5 in [61]). It is clear, however, that the distributions after n-th step of the random walk are the same for the random walk and the inverse one. The entropy criterion tells us that this kind of phenomena can not happen for measures of finite entropy.

Some applications of the entropy criterion are immediate: for example, it tells us that any finitely supported (and more generally, any finite first moment) measure on a group of sub-exponential growth has trivial boundary. In other examples the estimate of the entropy, even understanding whether the entropy

is zero or not, can be significantly harder. In the next section we review the recent progress in this direction. In some classes of groups it seems easier to use the part of the entropy criterion that tells us that if entropy is zero, then the boundary is trivial. Though one can in many cases estimate the entropy from below and show that it is positive, in various classes of groups that were studied previously, this was not the only way to see that the boundary is non-trivial. Indeed, in some of the examples one could see directly that there are some non-trivial μ-boundaries [61], in others one was able to construct bounded harmonic functions on the group or on some covers with a given deck transformation group [71]. However, recently one has discovered other classes of groups, where the entropy criterion is so far the only known way to show the non-triviality of the boundary. In the next section we review some examples of this kind, for which we can not answer so far the above mentioned question about μ-boundaries.

An important tool for the complete description of the boundary is conditional entropy criterion, due to Kaimanovich, which is analogous to the entropy criterion. This criterion tells that if we have some μ-boundary and if we want to check whether this boundary is trivial it is necessary and sufficient (provided we work with measures with finite entropies) to check whether the conditional entropy is zero. In some sense it seems that "the larger the group is", the easier is to use this strategy. In Section 4 we try to give this statement a more precise meaning and we review some known results.

Some applications of boundary are very well studied, for example the already mentioned relation of boundary to the space of harmonic functions. Applications to the growth of groups are more recent. Since entropy criterion was established, it has been known that there is a strong relation between the growth of a group and non-triviality of the boundary. It may seem that the growth of a space or of a group is much easier to determine than triviality/non-triviality of the boundary. However, one can use the boundary behavior of the random walk in some cases (see section 3) as a tool for establishing lower estimates on the growth.

We recall that the *word length* l_S with respect to a finite generating set S of G is defined as follows. For each $g \in G$ the word $l_S(g)$ is the minimum of m, such that g is equal to the product of m elements of S and of their inverses. The i-th moment ($i \in \mathbb{R}$) of a measure μ on G with respect to the word length l_S is $\sum_{g \in G} \mu(g) l_S^i(g)$.

Remark 1. *The rate of escape (or the drift) l of the random walk (G, μ) with respect to some word metric l_S is defined as the limit of $L(n)/n$, where $L(n)$ is the expectation of the distance to the origin after n steps of the random walk. If μ is symmetric and has finite first moment with respect to some (and hence to all) word metrics in G, then the entropy of the random walk is positive if and only if the rate of escape of the random walk (G, μ) is positive (It is easy to see that $h \leq vl$, where v is the exponential growth rate of the group, (see Guivarc'h [52]), and so it is clear that $l = 0$ implies $h = 0$. The converse was proved by Varopoulos in [81] for finitely supported measures and then by Karlsson,*

Ledrappier in the general case [65]. Furthermore, it is shown in [39] that if μ is symmetric and has finite second moment, then there exists $C > 0$ such that $H(\mu^{*n}) \geq C(L(n)/n)^2$ for all n. (It is shown by Ledrappier [69], that $h \geq l^2$ for the Brownian motion on the covering manifold).

Given two function $f_1, f_2 : \mathbb{N} \to \mathbb{R}_+$, the notation $f_1(n) \sim f_2(n)$ means that there exists $C > 0$ such that $f_1(n) \leq Cf_2(Cn)$ and $f_2(n) \leq Cf_1(Cn)$. The random walk (G, μ) is said to be *simple*, if μ is symmetric and if the support of μ is a finite generating set of G.

Remark 2. *The study of the asymptotics of $L(n)$ was initiated by Vershik and Guivarc'h. It turns out that $L(n)$ can have various asymptotics [28]. For example, $L(n) \sim n/\ln(n)$ for any finitely supported symmetric random walk on the wreath product of $\mathbb{Z}^2$ with a finite group; $L(n) \sim n^{(1-2^{-k})}$ for some simple random walk on the k times iterated wreath product $\mathbb{Z} \wr (\mathbb{Z} \wr (\mathbb{Z} \wr \cdots \wr \mathbb{Z}) \dots))$. For further examples of evaluation and estimates of $L(n)$ see [28], [39], Yadin [86], and also Corollaries 1 and 2 in Section 3 and the remark after Theorem 6 in Section 5.*

It is not known so far whether any function $f(n)$ between $\sqrt{n}$ and n, with some regularity on its growth, is asymptotically equivalent to $L(n)$ for a simple random walk on some finitely generated group G. We also mention that upper bounds on $H(n)$ can be relevant for Liouville type theorems on the growth of unbounded harmonic functions (see [39]).

2. Applications of Entropy Criterion

The entropy criterion can be used to show that simple random walks on polycyclic groups and on solvable Baumslag-Solitar groups have trivial boundaries. See also [27] for triviality of the boundary for some random walks on iterated wreath products of $\mathbb{Z}$ and $\mathbb{Z}^2$.

A very interesting class of examples was discovered recently by Bartholdi and Virag [11], who studied a group that was defined earlier by Grigorchuk and Żuk. Using a notion of a "self-similar" random walk, they have shown that this group admits a finitely supported measure with zero rate of escape. Originally in their paper they used some special metric on this group, which is not a word metric, and later the argument was simplified by Kaimanovich [62] who works with $H(\mu^{*n})$ instead of the rate of escape and shows that the entropy of the random walk is zero. It turned out that this argument can be applied to wider classes of groups acting on rooted trees. See Bartholdi, Kaimanovich, Nekrashevych [12] for the case of groups generated by bounded automata and Amir, Angel, Virag [2] for a more general case of groups generated by so called linear activity automata (it is shown by Sidki in [78] that a group generated by a polynomial activity automaton never contains a non-Abelian free subgroup, and it is an open question whether all such groups are amenable). An interesting

feature of the above mentioned examples is that the vanishing of the rate of escape or of the entropy of a non-degenerated random walk is used to show that the groups under consideration are amenable. Thus, random walks and Kesten criterion help to understand in these examples whether the group is amenable. In all previously known amenable finitely generated group there is some known sequence of Følner sets.

Now we recall some examples of random walks with non-trivial boundaries. The simplest class of examples of amenable groups such that the simple random walks have non-trivial boundaries are wreath products $\mathbb{Z}^d \wr A$ (that is, semidirect products of $\mathbb{Z}^d$ with $\sum_{\mathbb{Z}^d} A$, with $\mathbb{Z}^d$ acting by shift on the index set), $d \geq 3$ (Kaimanovich, Vershik, [61]). To see that the boundary of the simple random walk is non-trivial, one observes that the projection of the random walk to the base group $\mathbb{Z}^d$ is transient, and that therefore for all $x \in \mathbb{Z}^d$ the coordinate a_x stabilizes along infinite trajectories of the random walk. This argument has generalizations in several contexts, where its application is less straightforward. Quotients of the Poisson-Furstenberg boundaries for certain groups acting on rooted trees, that we describe in Section 3, are reminiscent of this "lamplighter boundary" for wreath products. We recall also, that Kaimanovich [63] has shown, that a simple random walk on the Thompson group F has non-trivial boundary. Kaimanovich has observed that if a group acts on a line by piecewise-linear mappings with finite number of pieces, then the boundary is non-trivial whenever the orbits of the action are transient (since for such actions the ratio of the left and right derivatives at a given point stabilizes along infinite trajectories), and he has proved that for the group F these orbits are indeed transient. It would be interesting to understand the boundary behavior of random walks on more general groups of diffeomorphisms of the interval for a) simple random walks; b) for not necessarily finitely supported random walks. It is a long standing question whether Richard Thompson group F is non-amenable. If it turns out to be amenable, these questions become especially interesting.

Now we return to the wreath products and recall some additional properties of the boundaries of random walks on these groups. A transience argument, similar to the argument used in the finitely-supported case in [61], shows that any non-degenerate random walk with finite first moment has non-trivial boundary [56]. Another argument, based on an entropy estimation, is introduced in [31]. It consists of subdividing the space of trajectories of length n into conditional subspaces, such that there exists a subset of measure at least p, ($p > 0$ is a constant not depending on n), with the following properties: the trajectories, belonging to the same conditional event inside this subset, all have the following form. There exits a sequence $n_1^{(n)}, n_2^{(n)}, \ldots, n_{k_n}^{(n)}$, depending on each conditional event, such that $k_n \geq Cn$, where C is a positive constant not depending on n. The increments of the trajectory at times t others than $n_i^{(n)}$, $1 \leq i \leq k_n$ are fixed for a given conditional event. For each time instant t, $t = n_i^{(n)}$ (for some i) the increments take two possible values. All 2^{n_k} trajectories in each given

conditional event visit at moment n distinct elements of the group. The time instants $n_1^{(n)}$, $n_2^{(n)}$ ... correspond to visits of distinct points by the projection of our random walk to some space. In the case of wreath products $\mathbb{Z}^d \wr A$ this space is $\mathbb{Z}^d$.

If the random walk admits such partition into conditional events, then the inequality between entropy and mean conditional entropy implies that the entropy of the random walk is positive. It might seem that the assumption is essentially stronger then the positivity of entropy, but it is shown in [31] that it can applied to many classes of groups. Moreover, it is not clear whether there are any obstructions for this type of entropy estimates: does there exist a simple random walk (G, μ), having non-trivial boundary, and not admitting families of conditional events, satisfying these properties?

Such question can be viewed as a probabilistic (entropic) counterpart of the following still open question, raised by Rosenblatt in [77]: does any group G of exponential growth admit a Lipschitz imbedding of the infinite binary tree?

The argument, applied to the wreath products, shows that any non-degenerate random walk of finite entropy on wreath products $\mathbb{Z}^d \wr A$, $d \geq 3$, $\#A \neq 1$, has non-trivial boundary. The same conclusion holds for the free metabelian group on d generator: $Met_d = \langle g_1, g_2, g_d | uw = wu, u, w \in [G, G]\rangle$. Another series of examples, studied in [31] is as follows. Consider a finitely presented group B_d, defined by the following generators and relations

$$B_d = \langle a, s_i, t_j | a^{t_i} = aa^{s_i}, [s_i, s_j] = [t_i, t_j] = [s_i, t_j] = e, [a^u, a^w] = e\rangle,$$

where i,j in the presentation take the values between 1 and d, and u and w are any words in s_i and t_j. The group B_d is a subgroup of $\mathrm{GL}(2, \mathbb{Z}(X_1, \ldots, X_d))$. It is a metabelian (that is, solvable of solvability length 2) group, and its subgroup generated by s_i and a is isomorphic to the wreath product of $\mathbb{Z}^d \wr \mathbb{Z}$. The groups B_d are particular cases of a more general construction due to Baumslag, that assures that any finitely generated metabelian group can be imbedded into finitely presented metabelian group. However, there is no known relation in general between triviality of the boundary of random walks on a subgroup and triviality of the boundary of random walks on the ambient group. There are particular cases, where such relation does exist. For example, this relation is well known in the case when the subgroup is *recurrent* for our random walk , that is, if the random walk (G, μ) returns to the subgroup H infinitely many times with probability one. In this case one considers the probability measure μ' on H, such that for any $h \in H$ the probability $\mu'(h)$ is equal to the probability that the random walk visits h at the instant of its first return to the subgroup. One shows that there is a canonical measure preserving bijection between the boundary of (G, μ) and that of (H, μ'). A recent result of Malyutin, Vershik [73] shows that for any group G, containing a free subgroup, and any simple random walk μ on G, the boundary of this free subgroup is a μ-boundary for the random walk (G, μ)). (These particular cases are not relevant to random walks we discuss). In general, it is not known whether non-triviality of some

(all) simple random walks on the subgroup implies the non-triviality of some (respectively all) simple random walks on a group, containing this subgroup. Thus the non-triviality of the boundary for simple random walks on the wreath products did not help to prove the non-triviality of the boundary of random walks on B_d. Entropy estimates from [31], applied to B_d, show that for $d \geq 3$ any simple random walk on B_d (and, more generally, any non-degenerate random walk of finite entropy) has non-trivial boundary.

Since Furstenberg discretization has finite entropy and since every finitely presented group, in particular, our B_d, serves as the fundamental group of a compact manifold M of a given dimension $d > 3$, the above implies the following

Theorem 1. *There exists a compact Riemannian manifold M, such that its fundamental group is amenable and such that its universal cover is not Liouville (that is, this universal cover admits non-constant bounded harmonic functions).*

Question 1. *What is the Poisson-Furstenberg boundary for the simple random walks on groups B_d?*

It is not clear even how to describe any non-trivial quotient of the boundary for these random walks.

3. Choquet-Deny Theorems. Groups of Intermediate Growth. Applications of Random Walks to Growth of groups

It is known that any random walk on a finitely generated group of polynomial growth has trivial Poisson-Furstenberg boundary. Indeed, it is shown by Dynkin and Malyutov in [26] that this statement, generalizing the classical Choquet-Deny theorem for Abelian groups (Blackwell [13]), holds for any finitely generated nilpotent group (see also Margulis [74] for description of all positive harmonic functions on nilpotent groups). By Polynomial Growth Theorem of Gromov [49] any group of polynomial growth is a finite extension of a nilpotent one. This can be used to show that the triviality of the boundary for nilpotent groups implies the triviality of the boundary for any measure on a group of polynomial growth. Now let G be a group of subexponential growth. That is, either G is of polynomial growth and is virtually nilpotent, or it has growth strictly between polynomial and exponential.

If we suppose that the measure μ on G has finite first moment, then in view of the entropy criterion the boundary is trivial. The question was whether a counterpart of Choquet-Deny theorem holds for any measure on a group of subexponential growth, that is, whether the condition to have finite first moment in the above mentioned statement is not essential. A negative answer is given in [32], where it is shown that some among Grigorchuk groups of intermediate growth admit a measure with non-trivial boundary. Moreover, on some of these groups this measure can be chosen to have finite entropy.

The idea of the construction of such measures and of the proof that the boundary is non-trivial is as follows. Given a group, acting on a rooted tree, we consider the action on the boundary of the tree and an orbit of a point x of the boundary under this action. In [32] we used an equivalent language of groups, acting by permutation of the interval (where the points of the interval $(0,1]$ are written as numbers in the k-ary numeral system, which correspond to the points of the k-regular rooted tree together with its boundary). The main focus in that paper is on Grigorchuk groups and their close generalizations. In this situation x could be chosen (in the terminology of actions on a rooted tree) to be the point of the boundary, corresponding to the right most ray of the tree. We say that the action of G on the interval $(0,1]$ verify *the strong condition* $(*)$ if the following holds. For any $g \in G$, $x, y \in (0,1]$ such that $g(x) = y$ and any $\delta > 0$ there exist $\epsilon > 0$ such that $g((x-\epsilon, x]) \subset (y-\delta, y]$. There exists a finite generating set S of G such that for any $s \in S$ and $x \in (0,1]$ satisfying $x \neq 1$ or $s(x) \neq 1$ there exist $a \in \mathbb{R}$ and $\epsilon > 0$ such that $s(y) = y + a$ for any $y \in (x-\epsilon, x]$. The standard action on the interval of any Grigorchuk group satisfies the strong condition (*).

Below we use the language of actions on trees, that seems slightly more adequate for some more general questions we want to address. Let G be a group acting on a rooted trees. For all y on the orbit of x one chooses a mapping T_{yx} defined from a left neighborhood of y to a left neighborhood of x in such a way that for all x, y, z the mapping $T_{zy}T_{yx}$ coincides with T_{zx} in some left neighborhood of y. Mappings T_{yx} allow us to multiply germs at different points, and we consider then the *group of germs*, generated by all $\mathrm{germ}(g, y)$, where $g \in G$ and y is on the orbit of x. We say that the action on a tree satisfies the strong condition (*) if the corresponding action on the interval satisfies this condition.

Theorem 2. *Let G be a group acting on the rooted tree, such that the action satisfies the strong condition (*) and suppose that there exists a subgroup H, such that the group of germs of H is not equal to the group of germs of G and such that the orbit of x under the action of H is infinite. Then G admits a measure with non-trivial boundary.*

In [32] it was assumed in Theorem 2 above that the group of germs of G is finite, but this condition can be easily dropped. One may check that the assumption of this theorem is verified for some of Grigorchuk groups. Moreover, on some of these groups one can additionally show that the constructed measure can be chosen to have finite entropy. Thus we get

Theorem 3. *i) There exist groups G of subexponential growth admitting probability measures μ with non-trivial Poisson boundary.*

ii) Moreover, there exist groups G of subexponential growth admitting probability measures μ of finite entropy such that the entropy $h(\mu)$ of the random walk (G, μ) is positive.

Theorem 2 can be applied to groups acting on rooted trees, the growth of which can be exponential or intermediate. It seems the most interesting that it can be applied to a large range of groups of intermediate growth. We want to stress however, that there are groups where it can not be applied and where we still do not know the answer to the question: does this group admit a measure with non-trivial boundary? In particular, this remains unknown for the first Grigorchuk group, which is the most well studied among groups of intermediate growth.

To prove Theorem 2, one constructs a measure μ such that its support belongs to the union of the subgroup H with some finite set in G and such that the induced random walk on the orbit is transient. One shows then that $\mathrm{germ}(g, x)$ modulo the group of germs of H stabilizes along infinite trajectories of the random walk.

The condition (*) in the way it is defined [32] is well suited for Grigorchuk groups, considered in that paper. In last years many new interesting examples of groups acting on rooted trees have been studied, for which this condition does not hold. It seems that this assumption in the theorem above can be very much weakened, and it is interesting to understand what is the optimal condition.

Our main motivation Theorem 2 is the construction of infinitely supported measures with non-trivial boundary (Theorem 3). However, a particular case of Theorem 2 above is when the orbital Schreier graph of H is transient. In this case the theorem shows that the simple random walk on G has non-trivial boundary. Recently Bondarenko [15] has shown that if G is generated by a bounded automaton, then the orbital Schreier graph of G is recurrent. It is known that such groups can be imbedded in a group, admitting a simple random walk with trivial Poisson-Furstenberg boundary. It seems that the assumption (*) in the corollary can be much weakened.

Question 2. *Can the the criterion from [32] be extended to provide a general criterion for recurrency/transiency of orbital Schreier graphes for groups acting on rooted trees?*

To have a sufficient condition for the Schreier graph being recurrent, we have to exclude cases such as $\mathbb{Z}^d$, $d \geq 3$, which act on a rooted trees and have trivial boundary, but it is seems that there could be criteria, much more general than those explained in [32], in terms of triviality of the boundary.

3.1. Application to growth. Let G be a finitely generated group and S be a finite generating set of G. The growth function $v_{G,S}(n)$ is the number of elements of G that can be written as a product of at most n elements of S and their inverses. It is shown in [32] that some measures with non-trivial boundaries on groups of intermediate growth can be used to obtain lower bounds on the growth of these groups. They are used in [32] to obtain the following bounds on the growth of certain Grigorchuk groups

$$\exp(n/\log^{2+\epsilon}(n)) \leq v_{G,S}(n) \leq \exp(n/\log^{1-\epsilon}(n)),$$

for all sufficiently large n. Here the upper bound is essentially due to Grigorchuk [46]. The lower bound follows from the fact, that the group G admits a measure μ with non-trivial boundary, with a certain control on the decay of μ. A generalization of this idea is introduced in [34], where we provide new lower bounds for the growth of groups of the form $\exp(n^\alpha)$.

Another application in [34] provides lower bounds for the escape $L(n)$ of random walks on certain groups, acting on rooted trees. The strategy is as follows. Given a group G, acting on a rooted tree, construct another auxiliary group G_2, $G \subset G_2$, such that the group of germs of G_2 is larger than the group of germs of G. In this situation the upper bounds on growth of G_2 provide lower bounds for the asymptotic behavior of $L(n)$ for random walks on G. We introduce in [34] the *critical constant* $c_{RT}(G,H)$ of a subgroup H in a group G. This constant is defined as $\sup \beta$, where the supremum is taken over all β, for which there exists a random walk on G, of finite β-moment, such that the induced random walk on G/H is transient. Suppose that the action of the auxiliary group G_2 satisfies the strong condition (*) and that the growth function of G_2 is bounded from above by $\exp(n^\gamma)$. One proves that in this case the critical constant of the stabilizer of 1 in G is at most γ. The proof uses the Poisson boundary argument similar to the proof of theorem 2. On the other hand, one observes that if $L(n) \leq Cn^\xi$, then $c_{RT}(G,H) \geq 1/(2\xi)$ for any finite index subgroup H in G. Moreover, if $c_{RT}(G,H) < 1/(2\xi)$, then $\sum_{n=1}^\infty L(n)n^{-(1+\epsilon+\xi)} = \infty$, for some $\epsilon > 0$. Applying this for H which is equal to the stabilizer of 1 in G, we conclude that the asymptotics of the escape of any simple random walk (G,μ) satisfies $L_{G,\mu}(n) \geq n^{1/(2\gamma)}$ for infinitely many n. For example, let G be the first Grigorchuk group. In this case one is able to construct the auxiliary group G_2, with the growth at most $\exp(n^\gamma)$, where $\gamma = \log(2)/\log(2/X)$ and X is the positive solution of the equation $X^3 + X^2 + X - 2$. (For the first Grigorchuk group such upper bound on the growth function is due to Bartholdi [10], and a similar argument works also for our group G_2). We have $\gamma < 0.768$. This implies

Corollary 1. *[34] For any simple random walk on the first Grigorchuk group* $\sum_{n=1}^\infty L_{G,\mu}(n)n^{-(1.65)} = \infty$, *and* $L_{G,\mu}(n) > n^{0.65}$ *for infinitely many* n.

It is proved by Grigorchuk that some of his groups are close to the first Grigorchuk group on one scale, and they are close to subgroups in direct sum of several copies of a solvable group H of exponential growth on the other scale. For this group H and for any symmetric finitely supported measure μ on H one can check that $L_{H,\mu} \leq C_1\sqrt{n}$, and this can be used to obtain the following corollary

Corollary 2. *There exists a Grigorchuk group G, such that a simple random walk on G satisfies* $\limsup(\log L_{G,\mu}(n))/n \geq 0.65$ *and* $\liminf(\log L_{G,\mu}(n))/n \leq 1/2$.

It is known (Lee, Peres [70]) that if G is an infinite finitely generated group, and μ is a symmetric finitely supported measure, such that its support generates G, then $L_{G,\mu}(n) \geq C\sqrt{n}$, for some $C > 0$ and all n.

Question 3. *Let G be a finitely generated group. Suppose that $L_{G,\mu}(n) \leq C\sqrt{n}$, where the measure μ is such that its support generates G. Can the growth of G be intermediate?*

There is no Grigorchuk group for which we know precisely the asymptotics of the growth function. And the estimates, obtained using random walks as explained above, provide in a sense the best known examples, where discrepancy between the upper and lower bounds is not too large. It would be very interesting to obtain more information on possible functions, that can be realized as the growth function of some groups. Grigorchuk has shown that there are groups with arbitrarily fast subexponential growth (more precisely, Grigorchuk shows in [46] that among his groups there are groups such that their the growth is minorized along a subsequence by a given subexponentially growing function, and essentially the same argument [33] shows that by taking a direct sum of two Grigorchuk groups we obtain a growth function, that is minorized by a given subexponential function for all sufficiently large values of n). A natural question would be: can any sufficiently fast growing subexponential function be realized as a growth function of some group? In particular, we want to know: does there exist $a < 1$ such any function $f \geq \exp(n^a)$ is equivalent to a growth function?

It is even more challenging to construct groups of super-polynomial growth with the smallest possible growth. A conjecture due to Grigorchuk [47] states that any super-polynomial growing function is bounded from below by $\exp(n^b)$, for some $b > 0$ (the strong form of this conjecture states that we can take $b = 1/2$). It is tempting to understand better the possible applications of the boundary theory for the class of groups of small intermediate growth. See [48] for other question concerning the growth of groups and, in particular, of Grigorchuk groups.

4. Complete Description of Poisson-Furstenberg Boundaries

The complete description of the Poisson-Furstenberg boundary has been known for the following finitely generated groups (under certain conditions on the decay of the probability measure defining the random walk):

- discrete subgroups in semi-simples Lie group (Furstenberg [45] for a particular case of an infinitely supported measure, "Furstenberg approximation", Ledrappier [68] for a more general class of measure on discrete

subgroups of $Sl(d, \mathbb{R})$, Kaimanovich [57] for a more general class of measures on discrete subgroup in an arbitrary semi-simple Lie group); see also Schapira [79] and Brofferio, Schapira [18];

- free groups (Dynkin, Malyutov [26] for random walks, with the defining measure supported on standard generators, Derriennic [24] for measures with finite support), more generally, for hyperbolic groups (Ancona [1] for measures with finite support, Kaimanovich [57] for measures of finite entropy and with finite logarithmic moments; see also Ballman Ledrappier [9]; for the question whether a given measure on the hyperbolic boundary can be realized as the hitting measure of a certain random walk see Connell, Muchnik [20]),

- Coxeter groups (follows from Karlsson, Margulis [66], see Theorem 6.1 in [64] for an explanation),

- groups with infinitely many ends (Woess [84] for finitely supported measures, [57] for a more general class of measures),

- the mapping class group (Kaimanovich, Masur [59]) and braid groups (Farb, Masur [40]).

We would like to stress that for some of the above mentioned groups, the boundary is described in terms of the space on which the group acts. It could be important and in some situations it seems to be harder to describe the boundary in more algebraic terms (see Vershik [82] for the statement of the problem and Malyutin, Vershik [73] for the results in this direction, including the stability of the so-called *Markov-Ivanovsky normal form* for random walks on braid groups).

- Wreath products of free groups with finite groups (Karlsson, Woess [67]),

- certain classes of groups acting by diffeomorphisms on a circle (Deroin, [22]).

For some classes of groups it is easier to identify the boundary for certain non-symmetric random walks, rather than for symmetric ones. It was done for

- random walks on the wreath product $\mathbb{Z}^d \wr B$, which have a non-zero drift of the projection on $\mathbb{Z}^d$ [57],

- random walks on solvable Baumslag-Solitar groups with a non-zero drift of the projection on $\mathbb{Z}$ [56], and, more generally, for such random walks on the group of rational affinities [16]. In the last two examples simple random walks have trivial boundary.

It was asked in [61] whether the "space of limit configurations", described in Section 2, provides a complete description of the Poisson-Furstenberg boundary in the wreath products $\mathbb{Z}^d \wr A$ ($d \geq 3$). The positive answer for $d \geq 5$ is given in [36], where we prove

Theorem 4. *Let $A = \mathbb{Z}^d$, $d \geq 5$, $\#B \geq 2$. If μ is a measure on $C = A \wr B$, such that the support of μ generates C as a group, the third moment of μ is finite and the projection of μ to $\mathbb{Z}^d$ is centered, then the Poisson-Furstenberg boundary is equal to the space of limit configurations.*

We hope that the argument in [36] can be extended also to the case of $d = 3$, $d = 4$.

In fact, Theorem 4 holds in general, without the assumption that the projection of μ to $\mathbb{Z}^d$ is centered. If this projection is not centered, then the projected random walk on $\mathbb{Z}^d$ has positive drift. As we have already mentioned, for measures such that the projection has positive drift, the result is due to Kaimanovich. Another special case of the theorem that was known previously is due to James and Peres, who have shown in [53] that the number of visits of points of the base provides a complete description of the Poisson-Furstenberg boundary of a certain measure on the semigroup $\mathbb{Z}^d \wr \mathbb{Z}_+$. The Poisson boundaries of certain random walks on wreath products of A with $\mathbb{Z}_+$ are equivalent to the *exchangeability* boundary of the projection random walk on A (see [14, 25, 56, 53, 36] for the definition of the exchangeability boundary, its properties and questions about this boundary).

A similar idea to that in the proof of Theorem 4 leads to description of the boundary for the free metabelian groups ([36]). It can be applied also to other groups with some resemblance to wreath products, such as extensions, by a finitely generated group A, of the finitary symmetric group on elements of A.

In the previous work that provided complete description of the boundary (see [45, 57, 68, 59, 66] and other above mentioned results), there was a natural candidate for the Poisson-Furstenberg boundary, and, moreover, there was a natural guess along which "directions" the trajectories converge to the limit point in this boundary. The main work was then to estimate conditional entropy (in many cases this can be done using Ray Criterion, though in some situations: modular group, a measure on a word-hyperbolic group without first moment, it is easier to work with Strip Criterion). One of the difficulties in proving Theorem 4 is that for wreath products, though there exists a natural and easy to describe candidate for the Poisson-Furstenberg boundary (lamplighter boundary), it is however not straightforward even to guess how the trajectories converge to the points of this boundary. The first step in the proof is to use the geometry and connectivity properties of the support of the limiting lamplighter configuration in order to "guess" approximatively which points the trajectory visits at certain time instants; the second step is to use this as a "ray approximation", to estimate conditional entropy and to prove, that "lamplighter boundary" is indeed the Poisson-Furstenberg boundary of the random walk under consideration.

5. Different Scales of Amenability. Asymptotic Invariants Related to Boundaries

Consider a symmetric non-degenerate probability finitely supported measure μ on G. As we have already mentioned, the boundary triviality of the random walk (G, μ) implies the amenability of G. For some amenable G the boundaries of (G, μ) is trivial, while for others such boundaries can be non-trivial. It is an open question whether the triviality of the boundary can depend on the choice of a simple random walk. For other questions related to dependence of entropy on the choice of defining measure see [37] and [38].

The fact that the Poisson-Furstenberg boundary of a simple random walk on G is trivial can be viewed as a strengthening of the fact that G is amenable.

Recall that a group G is said to be *amenable*, if it admits a finitely additive non-negative measure ν defined on all subsets of G, which is invariant under left translations and which has total mass one.

Kesten criterion of amenability says that a finitely generated group G is amenable if and only for some (and if and only if for all) non-degenerate finitely supported symmetric random walk on G the decay of the probability to return to the origin is subexponential. Another criterion is in terms of isoperimetric inequalities. Let S be a finitely generating set S. Given a subset $V \subset G$, its *boundary* $\partial_S V$ with respect to S is $\{v \in V : \exists s \in S : vs \notin V\}$. By *Følner criterion of amenability* a finitely generated group is amenable, if there exists a sequence of finite subsets V_n such that $|\partial_S V_n|/|V_n| \to 0$, as n tends to ∞. Here $|V|$ denotes the cardinality of the set V. The sequence V_n is called a *Følner sequence*, and the sets V_n are called *Følner sets*.

Given an amenable group G and a finite generating set S, the Følner function $Fol_{G,S}(n)$ is defined as the minimum of cardinality of V, where the minimum is taken over all finite subsets of V of G, such that the cardinality of the boundary of V with respect to the word metric $l_{G,S}$ is at least n times smaller than the cardinality of V. It is easy to see that if the group admits a sequence of Følner sets, then it admits an invariant mean: given a function on G, it suffices to consider the average value of this function for each Følner set, and then take the limit of this average, as n tends to ∞, along any non-principal ultrafilter. For a survey of equivalent definitions of amenability, see [19, 80].

Understanding the asymptotics of Følner function (in other words, understanding optimal *isoperimetric inequality*), in particular, obtaining lower bounds for Følner function is a question, related to large-scale geometry of groups. The study of Følner function was initiated by Vershik, who conjectured that $\mathbb{Z} \wr \mathbb{Z}$ provides an example of a group, with super-exponentially growing Følner function and asked whether the asymptotics of this function is n^n. Følner function of nilpotent groups were studied by Pansu, who proved the first asymptotically optimal isoperimetric inequality for a nilpotent, non virtually Abelian group. Later Varopoulos has shown that for virtually nilpotent group

of growth n^d the Følner function is asymptotically equivalent to n^d. His result was generalized by Coulhon and Saloff-Coste in [21], who have proved that for any group G the Følner function is asymptotically not less then the growth function of G: there exists C such that $Fol_{G,S}(Cn) \geq v_{G,S}(n)$.

Pittet and Saloff-Coste [75] have shown the the Følner function of $\mathbb{Z}^d \wr \mathbb{Z}/k\mathbb{Z}$, $d \geq 2$ is super-exponential (but their lower bound for the Følner function for these groups was not asymptotically optimal). The question of Vershik is answered in [29], where we prove the following more general

Theorem 5. *There exists $C > 0$ such that the following holds Let A and B be two finitely generated groups, B containing at least two elements. Let S_A and S_B be finite generating sets of A and B respectively, and S be the generating set of $A \wr B$, corresponding to the union of S_A and S_B. Then*

$$\text{Føl}_{A \wr B,S}(n) \geq C\text{Føl}_{B,S_B}(Cn)^{C\text{Føl}_{A,S_A}(Cn)}.$$

Under mild assumption on regularity of $\text{Føl}_A(n)$, the theorem provides asymptotically optimal lower bound for Følner function of the wreath product. Thus we obtain the first explicit examples of super-exponential asymptotics of $\text{Føl}_{G,S}(n)$, for example, it shows $\text{Føl}_{\mathbb{Z} \wr \mathbb{Z}} \sim n^n$, $\text{Føl}_{\mathbb{Z}^d \wr \mathbb{Z}/k\mathbb{Z}} \sim \exp(n^d)$. The theorem also implies that m times iterated exponent (for any $m \geq 1$) is a Følner function of some group.

Wreath products and groups resembling wreath products (see Gromov [51]) are so far the only known examples of groups with super-exponential Følner function, where we know the asymptotics of this function.

However, usually it is much easier to obtain a not necessarily optimal upper bound for the Følner function, that is, to produce a not necessarily optimal sequence of Følner sets in groups. For example, it is not difficult to see that if G is a group of subexponential growth, then some subsequence of balls $B_{G,s}(n_i)$ and corresponding spheres $Sph_{G,s}(n_i)$ satisfies $\#Sph_{G,s}(n_i)/\#B_{G,s}(n_i) \to 0$, that is, this subsequence of balls is a Følner sequence. This shows, that though asymptotic geometry and the forms of balls in the intermediate growth case are complicated and quite different from polynomial growth case, there are certain common properties of such groups and Abelian and nilpotent groups with respect to isoperimetry.

Conjecturably, there could be also some algebraic manifestation of the fact, that groups of intermediate growth, however intriguing they may be, share some common properties with nilpotent groups. We recall a question due to Grigorchuk: do all infinite simple groups have exponential growth? All Grigorchuk group act on rooted trees, and hence they are residually finite, and thus not simple. Not all groups of intermediate growth are residually finite: there exist central extensions (finite and infinite) of first Grigorchuk groups that have intermediate growth and that are not residually finite [30]. One can show (see Bajorska Macedonska [8]) that one of the two following statements hold: either any group of intermediate growth in an extension of a residually finite group

of intermediate growth; or there exist simple groups of intermediate growth. Indeed, let G be a group of intermediate growth and let R be the intersection of finite index subgroups in G. If G/R has super-polynomial growth, then this group is a residually finite group of intermediate growth. If the growth of G/R is polynomial, one proves that R is a finitely generated group, and concludes that R is a group of subexponential growth without subgroups of finite index. Take a simple quotient of R. It is clear that this quotient is an infinite group of subexponential, and hence of intermediate growth.

Another question is due to Grigorchuk and Pak: does an infinite group of subexponential growth always admit two infinite commuting subgroups? As to the first question, all infinite simple groups, known until now, are non-amenable. It is worth mentioning, that even the following weaker statement is unknown for the class of infinite groups of subexponential growth (and one might ask the same question for the larger class of groups, admitting simple random walk with trivial Poisson-Furstenberg boundary, and there are no known counterexamples even among amenable groups):

Question 4. *Let G be an infinite group of subexponential growth. Does there always exist an infinite subgroup H, which has infinite index in G?*

One could be inclined to say, that groups of subexponential growth are amenable in a very strong sense. For all Grigorchuk groups it is known additionally, that the Følner function of any of these groups is asymptotically bounded by $\exp(n^A)$. Moreover, A can be taken equal to 2 using to a self-similar random walk argument: see Kaimanovich [62], where it is explained that first Grigorchuk group has a self-similar measure with additional weight 1/2 at the identity. This mesure is supported on standard generators of the group, and it is symmetric. A similar argument shows that all Grigorchuk groups have a sequence of self-similar symmetric measures (that is, the measures are similar to the corresponding measure on a group of shifted measures), also with additional weight 1/2 at the identity. The latter fact can be used to show that for any Grigorchuk group, the entropy of the random walk, defined by the above mentioned measure, satisfies $H(n) \leq Cn^{1/2}$. This implies that for this (and hence also for any other simple random walk, see Pittet, Saloff-Coste [76]) on any Grigorchuk group, the probability to return to the origin satisfies $p_n(e,e) \leq \exp(-Cn^{1/2})$. The latter implies that Følner function of any of Grigorchuk groups is bounded from above by $\exp(Cn^2)$, for some $C > 0$. The main result of [35] shows, however, that Følner function of a group of subexponential growth can be arbitrarily large, that is

Theorem 6. *Given a function $f : \mathbb{N} \to \mathbb{R}$, there exists a group G_f of intermediate growth such that*

$$Fol_{G_f,S}(n) \geq f(n)$$

for all n.

The group G_f in this theorem can be chosen to be a torsion-group. Alternatively, it can be chosen to be a group without torsion. It would be interesting to understand in more detail growth and isoperimetry of such groups.

Another application of the construction from [35] is the existence of finitely generated group H of intermediate growth such that for any μ on G the escape satisfies $\limsup(\log L_{H,\mu}(n))/n = 1$, $\liminf(\log L_{H,\mu}(n))/n \leq \gamma$, for some $\gamma < 1$. (In terminology of [35] the group H is equal to an appropriate *"piecewise automatic group"* of the first Grigorchuk group with a non-amenable group). Moreover, one can use the construction from [35] to produce examples of groups H such that for any simple random walk on H it holds $\limsup(\log L_{H,\mu}(n))/n = 1$, $\liminf(\log L_{H,\mu}(n))/n = 1/2$.

The groups in Theorem 6 provide the first examples of groups with very large isoperimetry, such that simple random walks on these groups have trivial boundary.

Question 5. *Can such phenomenon occur for elementarily amenable group?*

Question 6. *What is the asymptotically largest possible Følner function for a solvable group, admitting a simple random walk with trivial boundary?*

As we have already mentioned, Theorem 6 shows that there is no upper bound for Følner function for groups with trivial boundary of simple random walks. If we suppose on the contrary, that the boundary of a simple random walk is non-trivial, Følner function of G can not be too small. Indeed, by the entropy criterion we know that the growth of G under this assumption is exponential, and by Coulhon Saloff-Coste isoperimetric inequality this implies that Følner function is asymptotically at least exponentially growing.

Question 7. *Suppose that a simple random walk on G has non-trivial boundary. What is the asymptotically smallest possible Følner function of G?*

I would like to thank Vadim Kaimanovich and Bruno Schapira for useful comments on this paper.

References

[1] A. Ancona, *Negatively curved manifolds, elliptic operators and the Martin boundary*, Ann. of Math., 125, 1987, 495–536.

[2] G. Amir, O. Angel, B. Virag, *Amenability of linear-activity automaton groups*, preprint, 2009, http://xxx.lanl.gov/abs/0905.2007

[3] A. Avez, *Entropie des groupes de type fini*, C.R.Acad.Sci.Paris, Sér. A, 275,(1972),1363–1366.

[4] A. Avez, *Harmonic functions on groups*, Differential Geometry and Relativity, Dordrecht-Holland, 1976, 27–32.

[5] R. Azencott, *Espaces de Poisson des groupes localement compacts*, Lecture Notes in Mathematics, Vol. 148. Springer-Verlag, Berlin-New York, 1970.

[6] M. Babillot, *An introduction to Poisson boundaries of Lie groups*, Probability measures on groups: recent directions and trends, 1–90, Tata Inst. Fund. Res., Mumbai, 2006.

[7] U. Bader, Y. Shalom, *Factor and normal subgroup theorems for lattices in products of groups*, Invent. Math. 163, No. 2, 415–454 (2006).

[8] B. Bajorska, O. Macedonska, *A note on groups of intermediate growth* Comm. Algebra 35 (2007), no. 12, 4112–4115.

[9] W. Ballman, F. Ledrappier, *The Poisson boundary for rank one manifolds and their cocompact lattices*, Forum Math. vol. 6 (1994) 301–313.

[10] L. Bartholdi, *The growth of Grigorchuk's torsion group*, Internat. Math. Res. Notices 1998, no. 20, 1049–1054.

[11] L. Bartholdi, B. Virág, *Amenability via random walks*, Duke Math. J. 130, No. 1, 39–56 (2005).

[12] L. Bartholdi, V.A. Kaimanovich, V.V. Nekrashevych, *On amenability of automata groups*, preprint, http://xxx.lanl.gov/abs/0802.2837

[13] D. Blackwell, *On transient Markov processes with a countable number of states and stationary transition probabilities*, Ann. Math. Stat., 26, 1955, 654–658.

[14] D. Blackwell, D. Freedman, *The tail σ-field of a Markov chain and a theorem of Orey*, Ann.Math.Statist. **35** 1964, 1291–1295.

[15] Ie. Bondarenko, *Groups generated by bounded automata and their Schreier graphs,* Phd Thesis, December 2007, http://txspace.tamu.edu/bitstream/handle/1969.1/85845/Bondarenko.pdf?sequence=1

[16] S. Brofferio, *Poisson boundary of random rational affinities*, Annales Inst. Fourier, 56, (2006), 499–515.

[17] M. Burger, N. Monod, *Continuous bounded cohomology and applications to rigidity theory*, Geom. Funct. Anal. 12 (2002), no. 2, 219–280.

[18] S. Brofferio, B. Schapira, *Poisson boundary of* $\mathrm{GL}_d(\mathbb{Q})$, to appear in Israel J. Math.

[19] T. Ceccherini-Silberstein, R.I. Grigorchuk,P. de la Harpe, *Amenability and paradoxical decompositions for pseudogroups and discrete metric spaces,* Proc. Steklov Inst. Math. 1999, no. 1 (224), 57–97

[20] Ch. Connell, R. Muchnik, *Harmonicity of Gibbs measures*, Duke Math. J. 137 (2007), no. 3, 461–509.

[21] Th. Coulhon, L. Saloff-Coste, *Isopérimétrie pour les groupes et les variétés*, Rev. Mat. Iberoam. 9, No.2, 293–314 (1993).

[22] B. Deroin, *Poisson boundary of a discrete group of diffeomorphisms of the circle*, preprint.

[23] Y. Derriennic, *Quelques applications du théoreme ergodique sous-additif*, Astérisque 74, (1980), 183–201.

[24] Y. Derriennic, *Marche aléatoire sur le groupe libre et frontiére de Martin*, Z. Wahrscheinlichkeitstheor. Verw. Geb. 32, 261–276 (1975).

[25] P. Diaconis, D. Freedman, *De Finetti's theorem for Markov chains*, Ann. Probab. 8, 115–130 (1980).

[26] E.B. Dynkin, M.B. Maliutov, *Random walks on groups with a finite number of generators*, Soviet. Math. Dokl. 2, 1961, 399–402.

[27] A. Erschler, *On the asymptotics of drift*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 283 (2001), 6, 251–257, 263; translation in J. Math. Sci. (N. Y.) 121 (2004), no. 3, 2437–2440.

[28] A. Erschler, On drift and entropy growth for random walks on groups. Ann. Probab. 31(3), 1193–1204.

[29] A. Erschler, *On isoperimetric profiles of finitely generated groups*, Geom. Dedicata 100 (2003), 157–171.

[30] A. Erschler, *Not residually finite groups of intermediate growth, commensurability and non-geometricity*, J. Algebra 272 (2004), no. 1, 154–172.

[31] A. Erschler, *Liouville property for groups and manifolds*, Inv. Mathematicae, 155, 2004, 55–80.

[32] A. Erschler, *Boundary behavior for groups of subexponential growth*, Ann. Math, 160, no 3, 2004, 1183–1210

[33] A. Erschler, *On Degrees of Growth of Finitely Generated Groups*, Funct.Anal. and its Appl., Vol. 39, No. 4, pp. 317320, (2005).

[34] A. Erschler, *Critical constants for recurrence of random walks on homogeneous G-spaces*, Ann. Inst. Fourier, 2005, 55, fas.2, 493–509

[35] A. Erschler, *Piecewise automatic groups*, Duke Math. J. 134 (2006), no. 3, 591–613.

[36] A. Erschler , *Poisson-Furstenberg boundary of random walks on wreath products and free metabelian groups*, to appear in Commentarii Math.Helvetici.

[37] A. Erschler, *On continuity of range, entropy and drift for random walks on groups*, to appear in proceedings of the workshop *Boundaries* in Graz.

[38] A. Erschler, V. Kaimanovich, *Continuity of entropy for random walks on hyperbolic groups*, in preparation.

[39] A. Erschler, A. Karlsson, *Homomorphisms to $\mathbb{R}$ constructed from random walks*, to appear in Ann. Inst. Fourier.

[40] B. Farb, H. Masur, *Superrigidity and mapping class groups*, Topology. 37 (1998), No. 6, 1169–1176.

[41] W. Feller, *Boundaries induced by non-negative mathrices*, Trans. Amer. Math. Soc., 83, 1956, 19–54.

[42] A. Furman, *Random walks on groups and random transformations*, Handbook of dynamical systems, Vol. 1A, 931–1014, North-Holland, Amsterdam, 2002.

[43] H. Furstenberg, *A Poisson formula for semi-simple Lie groups*, Ann.Math., 77, 335–386, 1963.

[44] H. Furstenberg, *Boundary theory and stochastic processes on homogeneous spaces*, Proc. Symp. Pure Math., 26, 193–229, 1974, Providence R.I.: Amer. Math. Soc.

[45] H. Furstenberg, *Random walks and discrete subgroups of Lie groups*, Advances Probab. Related Topics, Vol. 1 (1971), 1–63

[46] R.I. Grigorchuk, *Degrees of growth of finitely generated groups and the theory of invariant means*, Math. USSR-Izv. 25 (1985), 259–300. MR 0764305 592, 601

[47] R.I. Grigorchuk, *On growth in group theory*, Proc. of ICM, Vol. I, II (Kyoto, 1990), 325–338, Math. Soc. Japan, Tokyo, 1991.

[48] R.I. Grigorchuk, *Solved and unsolved problems around one group*, Infinite groups: geometric, combinatorial and dynamical aspects, 117–218, Progr. Math., 248, Birkhäuser, Basel, 2005.

[49] M. Gromov, *Groups of polynomial growth and expanding maps*, Publ. Math., Inst. Hautes Etud. Sci. 53, 53–78 (1981).

[50] M. Gromov, *Hyperbolic manifolds (according to Thurston and Jorgenson)*, Bourbaki Seminar, Vol. 1979/80, pp. 40–53, Lecture Notes in Math., 842, Springer, Berlin-New York, 1981.

[51] M. Gromov, *Entropy and isoperimetry for linear and non-linear group actions*, Groups Geom. Dyn. 2 (2008), no. 4, 499–593.

[52] Y. Guivarc'h, *Sur la loi des grands nombres et le rayon spectral d'une marche aléatoire*, Astérisque 74 (1980), 47–98.

[53] N. James, Yu. Peres, *Cutpoints and exchangeable events for random walks*, Theory Probab. Appl. 41, No.4, 666–677 (1996) and Teor. Veroyatn. Primen. 41, No.4, 854–868 (1996).

[54] V.A. Kaimanovich, *Examples of non-commutative discrete groups with non-trivial exit-boundary*, Journal of Sov.Math., 28, No.4, 579–591 (1985).

[55] V.A. Kaimanovich, *Discretisation of bounded harmonic functions on covering manifolds and entropy*, Potential Theory (Nagoya, 1990), 213–223, Berlin: de Gruyter, 1992.

[56] V.A. Kaimanovich, *Poisson boundaries of random walks on discrete solvable groups*, Heyer, Herbert (ed.), Probability measures on groups X. Proceedings of the tenth Oberwolfach conference, held November 4–10, 1990 in Oberwolfach, Germany. New York, NY: Plenum Publishing Corporation. 205–238 (1991).

[57] V.A. Kaimanovich, *The Poisson formula for groups with hyperbolic properties*, Annals of Math., 152, No.3, 659–692, 2000.

[58] V.A. Kaimanovich, *The Poisson boundary of amenable extensions*, Monatsh. Math. 136 (2002), no. 1, 9–15.

[59] V.A. Kaimanovich, H. Masur, *The Poisson boundary of the mapping class group*, Inv.Math, v. 125, n. 2 (1996), 221–264.

[60] V.A. Kaimanovich, A.M. Vershik, *Random walks on groups: boundary, entropy, uniform distribution,* Soviet. Math. Dokl. 20, 1170–1173, 1979.

[61] V.A. Kaimanovich, A.M. Vershik, *Random walks on discrete groups: Boundary and entropy*, Ann. Probab. 11, 457–490 (1983).

[62] V.A. Kaimanovich, *"Münchhausen trick" and amenability of self-similar groups*, Int. J. Algebra Comput. 15, No. 5–6, 907–937 (2005).

[63] V.A. Kaimanovich, announcement on the conference in Gaeta, 2003.

[64] A. Karlsson, F. Ledrappier, *On laws of large numbers for random walks*, Ann. Prob. 2006, Vol. 34, No. 5, 1693–1706.

[65] A. Karlsson, F. Ledrappier, *Linear drift and Poisson boundary for random walks*, Pure Appl. Math. Q. 3 (2007) 1027–1036

[66] A. Karlsson, G. Margulis, *A multiplicative ergodic theorem and nonpositively curved spaces*, Comm. Math. Phys. 208 (1999) 107–123.

[67] A. Karlsson, W. Woess, *The Poisson boundary of Lamplighter random walks on trees*, Geom. Dedicata, 124 (2007) 95–107

[68] F. Ledrappier, *Poisson boundaries of discrete groups of mathricies*, Isr.J. of Mathematics, 50, 1985, 319–336.

[69] F. Ledrappier, *Linear drift and entropy for regular covers*, Preprint 2009.

[70] J.R.Lee, Yu.Peres, *Harmonic maps on amenable groups and a diffusive lower bound for random walks*, http://xxx.lanl.gov/abs/0911.0274

[71] Th. Lyons, D. Sullivan, *Function theory, random paths and covering spaces,* J. Diff. Geom. 19, 299–323, 1984.

[72] G.A. Margulis, *Discrete subgroups of semisimple Lie groups*, Ergebnisse der Mathematik und ihrer Grenzgebiete, 3. Folge, 17. Berlin etc.: Springer-Verlag. ix, 388 p. DM 148.00 (1991).

[73] A.V. Malyutin, A.M. Vershik, *PF-boundary of the baid group and Markov-Ivanovsky normal form*, Izv. Ross. Akad. Nauk Ser. Mat. 72 (2008), no. 6, 105–132.

[74] G.A. Margulis, *Positive harmonic functions on nilpotent groups*, Soviet Math. Dokl. 7 1966 241–244.

[75] Ch. Pittet, L. Saloff-Coste, *Amenable groups, isoperimetric profiles and random walks*, In: J. Cossey et al. (eds.), Geometric Group Theory Down Under (Canberra, Australia, July 1419, 1996), de Gruyter, Berlin, 1999, pp. 293316.

[76] Ch. Pittet, L. Saloff-Coste, *On the stability of the behavior of random walks on groups*, J. Geom. Anal. 10 (2000), no. 4, 713–737.

[77] J. Rosenblatt, *Ergodic and mixing random walks on locally compast groups*, Math. Ann., 257 (1981), 31–42.

[78] S. Sidki, *Finite automata of polynomial growth do not generate a free subgroup*, Geom.Dedicata, 108:193–204, (2004).

[79] B. Schapira, *Poisson boundary of triangular matrices in a number field*, Ann. Inst. Fourier (Grenoble) 59 (2009), no. 2, 575–593.

[80] A. Valette, *Amenability and Margulis super-rigidity*, Representation theory and complex analysis, 235–258, Lecture Notes in Math., 1931, Springer, Berlin, 2008.

[81] N. Th. Varopoulos, *Long range estimates for Markov chains*, Bull. Sci. Math. 109 225–252, (1985).

[82] A.M. Vershik, *Dynamic theory of growth in groups: Entropy, boundaries, examples*, Russ. Math. Surv. 55, No.4, 667–733 (2000); translation from Usp. Mat. Nauk 55, No.4, 59–128 (2000).

[83] A.G. Willis, *Probability measures on groups and some related ideals in group algebras*, J. Funct. Anal. 92 (1990), no. 1, 202–263.

[84] W. Woess, *Boundaries of random walks on graphs and groups with infinitely many ends*, Israel J. Math. 68 (1989) 271–301.

[85] W. Woess, *Random walks on infinite graphs and groups*, Cambridge Tracts in Mathematics, 138, Cambridge University Press, 2000.

[86] A.Yadin, *Rate of escape of the mixer chain,* Electron. Commun. Probab. 14 (2009), 347–357.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

On non-Kähler Calabi-Yau Threefolds with Balanced Metrics

Jixiang Fu*

Abstract

The solution of the Strominger system can be viewed as a canonical structure on non-Kähler Calabi-Yau threefolds with balanced metrics. In this talk, we review the existence of balanced metrics on non-Kähler complex manifolds and the existence of solutions to the Strominger system.

Mathematics Subject Classification (2010). 53.

Keywords. Calabi-Yau manifold, Balanced metric, Strominger system, hermitian-Yang-Mills metric, Monge-Ampère equation, form-type Calabi-Yau equation.

1. Introduction

The principal concern of this paper is on non-Kähler Calabi-Yau threefolds with balanced metrics. Calabi-Yau manifolds are compact complex manifolds with trivial canonical line bundle. When the manifold is Kähler, Yau's theorem [34] on the Calabi conjecture provides a unique Ricci-flat Kähler metric in each Kähler class. Such metrics are called the Calabi-Yau metrics.

By the Clemens-Friedman construction, a large class of non-Kähler Calabi-Yau threefolds are obtained from Kähler Calabi-Yau threefolds by blowing down rational curves and smoothing the resulting singularities. For example, the connected sum of k copies of $S^3 \times S^3$ for any $k \geq 2$ can be given a complex structure in this way. Based on this construction, Reid speculated that any two projective Calabi-Yau threefolds can be connected by a sequence of deformations, contractions and smoothing through non-Kähler Calabi-Yau threefolds. This speculation demonstrates the potential role of non-Kähler complex manifolds.

It is therefore important to construct canonical metrics on non-Kähler Calabi-Yau manifolds. At first one should choose in general a good hermitian

*Partially supported by NSFC grants 10771037 and 10831008.
Institute of Mathematics, Fudan University, Shanghai 200433, China.
E-mail: majxfu@fudan.edu.cn

metric which is weaker than Kähler. One proposal is the balanced metric, which is also called the semi-Kähler metric in older references. Then one can consider how to construct a canonical metric in each "balanced class". In principle, we have a 1-1 correspondence between balanced metrics/forms and d-closed strictly positive definite $(n-1, n-1)$-forms, where n is the complex dimension of the manifold. In this sense, the balanced class of a balanced metric/form ω can be defined as

$$\mathcal{P}(\omega) = \{\omega^{n-1} + \sqrt{-1}\ \partial\bar{\partial}\varphi > 0 \mid \varphi \text{ is a real } (n-2, n-2)\text{-form}\}.$$

When the complex dimension of the manifold is three, the solution to the Strominger system can be viewed as a canonical structure on such manifolds. In 1986, Strominger made a proposal for supersymmetric compactification in the theory of the heterotic string. He proposed a system consisting of a pair (ω, h) – a hermitian metric ω on a Calabi-Yau threefold X and a hermitian metric h on a holomorphic vector bundle V over X. If (ω, h) is a solution of such a system, then ω is a (conformal) balanced metric, h is a hermitian-Yang-Mills metric with respect to ω, and together they satisfy a third equation (which is called the anomaly equation).

When the complex dimension of the manifold $n \geq 3$, as in the Kähler case, one can look for a canonical metric in each balanced class such that with respect to this metric, the norm of a non-vanishing holomorphic n-form is constant. In view of this point, one can derive an equation on real $(n-2, n-2)$-forms, which is called the form-type Calabi-Yau equation in [21].

In this article, we will survey some of results concerning the existence of balanced metrics on some non-Kähler manifolds and also the existence of solutions to the Strominger system. We mainly describe a joint result with Jun Li and Shing-Tung Yau [19] on the existence of balanced metrics on the connected sum of k copies of $S^3 \times S^3$ for $k \geq 2$, and also describe another joint work with Yau [22] on the existence of solutions to the Strominger system on a class of non-Kähler Calabi-Yau threefolds.

2. The Balanced Metrics

2.1. The definition and examples.

Definition 1. *Let X be an n-dimensional complex manifold with a hermitian metric g. Let ω be its hermitian form.*

(1) If $d\omega = 0$, then g (or ω) is called a Kähler metric;

(2) If $d(\omega^{n-1}) = 0$, then g (or ω) is called a balanced metric.

A complex manifold with a Kähler metric (resp. a balanced metric) is called a Kähler manifold (resp. a balanced manifold).

A. Gray and L. M. Hervella observed that on a compact complex manifold, if $d(\omega^k) = 0$ for some k with $2 \leq k \leq n-2$, then $d\omega = 0$. So it is reasonable to consider the balanced metric on non-Kähler complex manifolds.

There exists an obstruction to the existence of balanced metrics on a compact complex manifold [29]: In a compact complex manifold with a balanced metric, any compact complex hypersurface is not homologous to zero. For example, consider the Calabi-Eckmann complex structures on $S^{2p+1} \times S^{2q+1}$ [12]. These have the property that the product of the Hopf mappings:

$$\pi : S^{2p+1} \times S^{2q+1} \to \mathbb{P}^p \times \mathbb{P}^q$$

is holomorphic. Hence $\pi^{-1}(\mathbb{P}^{p-1} \times \mathbb{P}^q)$ is a codimension 1 complex submanifold in $S^{2p+1} \times S^{2q+1}$. This is of course homologous to 0 since the homology of $S^{2p+1} \times S^{2q+1}$ is 0 in real dimension $2(p+q)$. Therefore these manifolds are not balanced.

Now let us describe some examples and constructions of compact non-Kähler complex manifolds with balanced metrics.

The Calabi construction [11]. E. Calabi constructed his three dimensional complex manifolds as a complex tori bundle over a Riemann surface. He then proved that such manifolds cannot be Kähler. On the other hand, the natural metric (i.e. the product metric) is a balanced metric.

The twistor spaces over the self-dual Riemannian 4-manifolds. The natural metrics on these manifolds are balanced (c.f. [15]). However, N. J. Hitchin [24] showed that the only compact twistor spaces which are Kähler are those associated to S^4 and $\mathbb{P}^2$.

The torus bundle over a $K3$ surface or over a complex torus. Let S be a $K3$ surface or a complex torus. Let $\frac{\omega_1}{2\pi}, \frac{\omega_2}{2\pi} \in H^2(S, \mathbb{Z}) \cap H^{1,1}(S, \mathbb{C})$. Using these two forms one can construct a three dimensional complex manifold X such that X is the T^2-bundle over S. This construction can be viewed as the generalization of the above Calabi-Eckmann manifolds.

E. Goldstein and S. Prokushkin [23] proved that X is non-Kähler. They also observed that if ω_1 and ω_2 are anti-self-dual (1,1)-forms with respect to a Calabi-Yau metric ω_{CY} on S, then the natural metric on X is a balanced metric. Explicitly the natural metric is

$$\omega_0 = \omega_{CY} + \sqrt{-1}\ \theta \wedge \bar{\theta}. \tag{1}$$

Here θ is the connection form on the torus bundle (as the principal bundle) such that $d\theta = \omega_1 + \sqrt{-1}\ \omega_2$. Then, since $\omega_i \wedge \omega_{CY} = 0$ for $i = 1, 2$, it follows that $d(\omega_0^2) = 0$.

2.2. Some existence results of balanced metrics. M. L. Michelsohn found an intrinsic characterization of compact manifolds with balanced metrics by means of positive currents:

Theorem 2. [29] *A compact complex manifold X admits a balanced metric if and only if its every positive current of degree (1,1) which is the component of a boundary is zero.*

Using this characterization, L. Alessandrini and G. Bassanelli proved that the existence of balanced metrics is preserved under birational transformations:

Theorem 3. [2, 3] *Let X and X' be compact complex manifolds, and $f : X \to X'$ a modification. Then X has a balanced metric if and only if X' has a balanced metric.*

This theorem implies that compact complex manifolds bimeromorphic to Kähler manifolds are balanced. Note that the Kähler condition of a compact complex manifold is not preserved by modification. So the balanced metric condition is natural and important in complex geometry.

On the other hand, the balanced condition is not preserved under small deformation. The Iwasawa manifold gives such a counterexample [1]. Recall that the Kähler condition is preserved under small deformation. However, in case the complex manifold satisfies the $\partial\bar{\partial}$-lemma, the balanced condition is preserved under small deformation [33].

In 2004, Alessandrini and Bassanelli [4] proved that for a compact complex manifold X of dimension three, if X is Kähler outside a smooth (complex) curve, then X carries a balanced metric.

2.3. The construction of balanced metrics on $\#_k(S^3 \times S^3)$.

We begin with the Clemens-Friedman construction.

Let Y be a smooth Kähler Calabi-Yau threefold that contains a collection of mutually disjoint $(-1,-1)$-curves $E_1, \ldots, E_l \subset Y$; these are smooth, isomorphic to $\mathbb{P}^1$ and have normal bundles isomorphic to the direct sum of two copies of degree -1 line bundles over them. By contracting all E_i, we obtain a singular Calabi-Yau threefold X_0 with l ordinary double points $p_1, \ldots, p_l$:

$$\psi : Y \setminus \cup_{i=1}^{l} E_i \cong X_0 \setminus \{p_1, \cdots, p_l\}.$$

Friedman [17, 18] proved that there is an infinitesimal smoothing of X_0 if and only if the fundamental classes $[E_i]$ in $H^{2,2}(Y;\mathbb{Q})$ satisfy a relation $\Sigma_i n_i [E_i] = 0$ such that $n_i \neq 0$ for every i. Tian [32] and Kawamata [25] then used the different methods to proved that the infinitesimal smoothing can always be realized by a real smoothing, i.e., X_0 can be smoothed to a family of smooth complex manifolds X_t.

Friedman also proved that the canonical line bundle of X_t is trivial. But in general, X_t is not Kähler. Explicitly, Friedman observed that $\#_k(S^3 \times S^3)$ for any $k \geq 2$ can be given a complex structure in this way [18, 28]. Since the hodge number $h^{1,1}$ of these manifolds are zero, they can not be Kähler.

We can now state the main result jointly with J. Li and S.-T. Yau on the existence of balanced metrics on these non-Kähler manifolds:

Theorem 4. [19] *Let Y be a smooth Kähler Calabi-Yau threefold and let $Y \to X_0$ be a contraction of mutually disjoint $(-1,-1)$-curves. Suppose X_0 can be smoothed to a family of smooth complex manifolds X_t. Then for sufficiently small t, X_t admits a smooth balanced metric.*

Corollary 5. [19] *For any $k \geq 2$, $\#_k(S^3 \times S^3)$ admits a balanced metric.*

We outline the proof of our existence theorem here. Our first step is to modify a Kähler metric on Y near the $(-1,-1)$-curves E_i to get a balance metric ω_0 on the contraction X_0 that is smooth and balanced away from the singularities of X_0; near its singularities, ω_0 coincides with the Ricci-flat metric of Candelas-de la Ossa's (see [13]).

The second step is to deform ω_0 to a family of smooth balanced metrics on X_t. Since the Candelas-de la Ossa's metric on the cone singularity can be deformed to a family of smooth Ricci-flat metrics on the smoothing of the cone singularity, we can deform ω_0 to a family of smooth hermitian metrics ω_t that are Kähler near the singular points of X_0 and are almost balanced on X_t for small t. To get balanced metrics, we first perturb ω_t^2 by

$$\Omega_t = \omega_t^2 + \theta_t + \bar{\theta}_t, \qquad d\Omega_t = 0,$$

with $\theta_t = i\partial\mu_t$ for μ_t a $(1,2)$-form on X_t that solves the system

$$\partial_t\bar{\partial}_t\mu_t = \bar{\partial}_t\omega_t^2 \quad \text{and} \quad \mu_t \perp_{\omega_t} \ker \partial_t\bar{\partial}_t.$$

We then solve $\Omega_t = (\tilde{\omega}_t)^2$. For this to be possible, we need to prove that Ω_t is positive. We only need to prove that the C^0-norm $\| \theta_t \|_{\omega_t}$ approaches zero as t approaches zero.

To this end, we choose γ_t to be the solution to the Kodaira-Spencer equation $E_t(\gamma_t) = \bar{\partial}\omega_t^2$ subject to $\gamma_t \perp_{\omega_t} \ker E_t$. It then follows directly that the solution γ_t automatically satisfies $\partial_t\gamma_t = 0$ and $\mu_t = -i\bar{\partial}_t^*\partial_t^*\gamma_t$. Applying the elliptic estimates, the L^2-estimates and the vanishing of L^2-cohomology groups, we prove that $\lim_{t\to 0} t^\kappa \| \theta_t \|^2_{C^0(\omega_t)} = 0$ for $\kappa > -\frac{4}{3}$.

From the Clemens-Friedman construction and our main result, there exists a large class of non-Kähler Calabi-Yau threefolds admitting balanced metrics. Now the question is how to construct the canonical metrics on such manifolds. At first we consider the case of three dimensional since in this case we have the Strominger system.

3. The Strominger System

In heterotic string theory, the internal space X is a compact complex three-dimensional manifold with trivial canonical line bundle, i.e., with a

non-vanishing holomorphic three-form Ω. It also involves a holomorphic vector bundle V over X. Let ω be a hermitian metric on X and h a hermitian metric on V. In 1986, Strominger [31] proposed a system for (ω, h):

$$d(\| \Omega \|_\omega \omega^2) = 0;$$

$$F_h^{2,0} = F_h^{0,2} = 0, \quad F_h \wedge \omega^2 = 0;$$

$$\sqrt{-1}\partial\bar{\partial}\,\omega = \frac{\alpha'}{4}\left(\mathrm{tr}\big(R_\omega \wedge R_\omega\big) - \mathrm{tr}\big(F_h \wedge F_h\big)\right).$$

The first equation says that the metric ω is a conformal balanced metric. The second one is the hermitian-Yang-Mills equation. The existence of its solution is, by the Li-Yau theorem [26] which is the non-Kähler version of the Donaldson-Uhlenbeck-Yau theorem, equivalent to that V is stable with respective to the conformal balanced metric ω. The third equation is called the anomaly equation. Following Strominger, we take the curvature R in third equation to be defined by the hermitian connection. Thus the term $tr(R \wedge R)$ is always a $(2,2)$-form.

When V is the holomorphic tangent bundle $T'X$ and ω is Kähler, (ω, h) is a solution to the Strominger system if and only if $\omega = h$ is the Calabi-Yau metric. So this system should be viewed as a generalization of the Calabi conjecture for the case of non-Kähler Calabi-Yau threefolds with balanced metrics.

The existence of smooth solutions of the Strominger system has been studied since 2004. Using the perturbation method, J. Li and S.-T. Yau constructed irreducible smooth solutions to a class of Kähler Calabi-Yau threefolds on some $U(4)$ and $U(5)$ principle bundles. Shortly after, with Yau, we constructed solutions to this system on a class of non-Kähler Calabi-Yau threefolds. Our solutions were orbifolded by M. Becker, L.-S. Tseng and Yau to give many more solutions. With Tseng and Yau, we also presented explicit solutions on T^2-bundles over the Eguchi-Hanson space. We note further that nilmanifold solutions with different connections have been discussed recently in [16].

3.1. Non-Kähler solutions on some Kähler Calabi-Yau threefolds. We assume that X is a Kähler Calabi-Yau threefold and ω is a Calabi-Yau metric on it. Take $V = \mathbb{C}_X^{\oplus r} \oplus T'X$ and $h = h_1 \oplus \omega$. Here h_1 is a standard constant metric on $\mathbb{C}_X^{\oplus r}$. Then (X, ω, V, h) is a solution to the Strominger system, which is called a reducible solution. For any small deformations D_s'' of the holomorphic structure D_0'' of $\mathbb{C}_X^{\oplus r} \oplus T'X$, J. Li and S.-T. Yau derived a sufficient condition for the Strominger system to be solvable for (X, D_s''): it is that the Kodaira-Spencer class of the family D_s'' at $s = 0$ satisfies certain non-degeneracy condition. By showing this sufficient condition to hold on some projective Calabi-Yau threefolds, they provided the first example of regular irreducible solution to the Strominger system with gauge group $SU(4)$ and $SU(5)$.

Theorem 6. [27] *Let $X \subset \mathbb{P}^4$ be a smooth quintic threefold and $V = \mathbb{C}_X \oplus T'X$ or $X \subset \mathbb{P}^3 \times \mathbb{P}^3$ be a smooth Calabi-Yau threefold cut out by three homogeneous*

polynomials of bi-degree $(3,0)$, $(0,3)$ *and* $(1,1)$ *and* $V = \mathbb{C}_X^{\oplus 2} \oplus T'X$. *Let* ω *be a Calabi-Yau form (metric) on* X. *Then, there is a smooth deformation* D''_s *of* (V, D''_0) *so that for large* $c > 0$ *and small* s, *there are irreducible regular solutions* (h_s, ω_s) *to the Strominger system on the vector bundle* (V, D''_s) *so that* $\lim_{s\to 0} \omega_s = c\omega$ *and* $lim_{s\to 0} h_s$ *is a regular hermitian-Yang-Mills connection on* V.

3.2. Solutions on some non-Kähler Calabi-Yau threefolds.

With Yau [22], we gave the first existence result of solutions to Strominger system for a non-Kähler Calabi-Yau threefolds. Actually we constructed solutions on a class of torus bundles X over $K3$-surfaces twisted by two anti-self-dual $(1,1)$-forms ω_1 and ω_2, which have been mentioned in subsection 2.1. Based on physical arguments of superstring dualities, the existence of such solutions was suggested in [14, 6].

On such manifolds, we have showed that the natural metric ω_0 (see (1)) is the balanced metric. There also exists a non-vanishing holomorphic three form

$$\Omega = \Omega_{K3} \wedge \theta,$$

where Ω_{K3} is a non-vanishing holomorphic two form on S.

Moreover, one can define a hermitian metric on X:

$$\omega_u = e^u \omega_{K3} + \sqrt{-1}\ \theta \wedge \bar{\theta}.$$

Here u is any function of the $K3$ surface. This metric is not the balanced metric. The key point is that for any function u, the metric ω_u still satisfies the first equation of the Strominger system [23] (see also [22]).

Then let us consider the second equation. Take a stable vector bundle E over the $K3$ surface with respect to the metric ω_{CY}. By the Donaldson-Uhlenbeck-Yau theorem, there exists a hermitian-Yang-Mills metric h on E, i.e. its hermitian curvature F_h satisfies

$$F_h \wedge \omega_{CY} = 0.$$

So $\pi^* F_h \wedge \omega_u^2$ is also zero. This means that $\pi^* h$ is also the hermitian-Yang-Mills metric on $V = \pi^* E \to X$ with respect to any conformal balanced metric ω_u. So given a stable vector bundle E over the $K3$ surface, the second equation for the vector bundle $V = \pi^* E$ can always be solved for any metric ω_u.

Therefore we only need to consider the third equation. Certainly the term $\mathrm{tr} F_h \wedge F_h$ is a $(2,2)$-form defined on the $K3$ surface. For the metric ω_u, by explicit calculation, we found that the terms $\mathrm{tr}(R_{\omega_u} \wedge R_{\omega_u})$ and $\sqrt{-1}\partial\bar{\partial}\omega_u$ are also defined on the surface. Thus we reduced the third equation to the following Monge-Ampere equation defined on the $K3$ surface:

$$\triangle \left(e^u - \frac{\alpha'}{2} f e^{-u} \right) + 4\alpha' \frac{\det u_{i\bar{j}}}{\det g_{i\bar{j}}} + \mu = 0,$$

where f and μ are two functions on the $K3$ surface satisfying $f \geq 0$ and $\int_S \mu\omega_{K3}^2 = 0$. The last compatibility condition is equivalent to the condition

$$\alpha'(24 - c_2(E)) + Q(\omega_1/2\pi) + Q(\omega_2/2\pi) = 0. \tag{2}$$

Here 24 stands for the second Chern number of the $K3$ surface and $Q(\omega_i/2\pi)$, for $i = 1, 2$, denotes the intersection number of $\omega_i/2\pi$. We used the continuity method to solve the above equation. The estimate of the volume form is very complicated. Our main result is

Theorem 7. [22] *Let S be a $K3$ surface with a Calabi-Yau metric ω_{CY}. Let ω_1 and ω_2 be anti-self-dual $(1,1)$-forms on S such that $\omega_1/2\pi, \omega_2/2\pi \in H^2(S, \mathbb{Z})$. Let X be the T^2-bundle over S twisted by ω_1 and ω_2. Let E be a stable bundle over S with the gauge group $SU(r)$. Suppose ω_1, ω_2 and $c_2(E)$ satisfy the topological constraint (2). Then there exist a smooth function u on the $K3$ surface and a hermitian-Yang-Mills metric h on E such that (ω_u, h) is a solution of the Strominger system.*

3.3. Analysis and generalizations of the torus bundle over K3 solutions. The solution on the torus bundle over K3 was further generalized in [5]. With M. Becker, K. Becker, L.-S. Tseng, and Yau, we relaxed the conditions of Theorem 7 by allowing $\omega_1 + i\omega_2$ to contain a (2,0) component. A description of the allowable holomorphic vector bundles was also presented which led to a classification of all solutions in terms of the Chern classes of the torus and vector bundle.

M. Becker, L.-S. Tseng, and S.-T. Yau have analyzed further the non-Kähler torus bundle solutions. The linearized local moduli space of the solution within the Strominger system was given in [7]. They also found that when the solutions are analyzed within the context of string theory, both the Kähler $(K3 \times T^2)$ and non-Kähler solutions of the Strominger system can be connected inside the string theory moduli space [8]. Furthermore, many more solutions of the Strominger system can also be constructed by modding out the non-Kähler torus bundle solution by elements of its automorphism group. This was worked out explicitly in [9].

3.4. The explicit solution on the torus bundle over the Eguchi-Hanson space. With Tseng and Yau [20], we solved the Strominger system on this space. Actually, we can change the base space $K3$ surface in subsection 3.2 to be an ALE space. Simplest is the Eguchi-Hanson space: blow up of $\mathbb{C}^2/\mathbb{Z}_2$ at the origin of the $\mathbb{Z}_2$ action $\sigma(z_1, z_2) = (-z_1, -z_2)$. On this space, there is a Ricci-flat metric ω_{EH}. There is also a single anti-self dual (1,1)-form with respect to ω_{EH}. We can use this to twist the torus and as the curvature of a $U(1)$ vector bundle. Now the anomaly equation is reduced to an ODE on the Eguchi-Hanson space due to the dependence being only on the radial coordinate for all quantities on $\mathbb{C}^2/\mathbb{Z}_2$. By solving the ODE, we get

the explicit solution of the Strominger system on the torus bundle over the Eguchi-Hanson space.

3.5. The main problem. Now we return to the connected sum of k copies of $S^3 \times S^3$. By Corollary 6, there exist balanced metrics on such manifolds. So the first equation of the Strominger system is solvable. As to the second equation, since there are no non-trivial line bundles on $\#_k(S^3 \times S^3)$, its holomorphic tangent bundle is stable with respect to any Gauduchon metric [10]. Then by the Li-Yau theorem, there exists on the tangent bundle a hermitian-Yang-Mills metric with respect to the balanced metric. So the second equation is also solvable. Therefore we only need to consider the third equation. In view of the importance of $\#_k(S^3 \times S^3)$ in the study of the moduli space of Calabi-Yau threefolds and in superstring theory, we can ask

Question. Does there exist any solution to the Strominger system on $\#_k(S^3 \times S^3)$?

4. Form-type Calabi-Yau Equations

In this section, we assume X^n $(n \geq 3)$ is an n-dimensional Calabi-Yau manifold with balanced metrics. Let ω_0 be a balanced metric and Ω a non-vanishing holomorphic n-form. We want to look for a balanced metric ω such that

$$\omega^{n-1} = \omega_0^{n-1} + \frac{\sqrt{-1}}{2}\partial\bar{\partial}\varphi,$$

for some real $(n-2, n-2)$-form φ, and such that

$$\|\Omega\|_\omega = \text{constant}.$$

So we are looking for solutions in the balanced class of ω_0, which is the subset of the Bott-Chern cohomology class $[\omega_0^{n-1}] \in H^{n-1,n-1}_{BC}(X)$.

The relation between Ricci forms of the hermitian connection and the spin connection (i.e. the Bismut connection) with the metric ω is given by

$$Ric^s = Ric^h + dd^*\omega.$$

So ω is the balanced metric and $\| \Omega \|_\omega =$ const. if and only if $Ric^h = Ric^s = 0$.

As in the Kähler case, $\| \Omega \|_\omega =$ const. is equivalent to the equation

$$\frac{\det[\omega_0^{n-1} + (\sqrt{-1}/2)\partial\bar{\partial}\varphi]}{\det \omega_0^{n-1}} = e^{(n-1)f}\left(\frac{\int_X \omega^n}{\int_X \omega_0^n}\right)^{n-1},$$

for some function f. We call the above equation the *form-type Calabi-Yau equation*.

It seems very difficult to solve such form-type equations. To begin with, we consider the form-type Calabi–Yau equation on the complex n-torus T^n. Let ω_0 be a balanced metric on T^n. We can assume, without loss of generality, that ω_0 is a constant metric on T^n. With Z.-Z. Wang and D.-M. Wu, we have

Theorem 8. [21] *Let Ω be a non-vanishing holomorphic n-form on T^n, and ω_0 is a constant metric on T^n such that $\|\Omega\|_{\omega_0} = 1$. We denote by C_0 a positive constant.*

1. *If $C_0 \leq 1$, then for any metric ω on T^n such that $[\omega^{n-1}] = [\omega_0^{n-1}] \in H_{BC}^{n-1,n-1}(T^n)$ and that $\|\Omega\|_\omega = C_0$, we must have $C_0 = 1$ and*

$$\omega = \omega_0.$$

2. *For each $C_0 > 1$, there exists a non-Kähler balanced metric ω on T^n such that $[\omega^{n-1}] = [\omega_0^{n-1}]$ and that*

$$\|\Omega\|_\omega = C_0.$$

We also further generalized the uniqueness part of above theorem to an arbitrary Kähler Calabi-Yau manifold.

References

[1] L. Alessandrini and G. Bassanelli, *Small deformations of a class of compact non-Kähler manifolds*, Proc. AMS, **109**(1990), 1059–1062.

[2] L. Alessandrini and G. Bassanelli, *Metric properties of manifolds bimeromorphic to compact Kähler spaces*, J. Differential Geom. **37**(1993), 95–121.

[3] L. Alessandrini and G. Bassanelli, *Modifications of compact balanced manifolds*, C. R. Acad. Sci. Paris Sér. I Math. **320**(1995), 1517–1522.

[4] L. Alessandrini and G. Bassanelli, *A class of balanced manifolds*, Proc. Japan Acad. Ser. A Math. Sci. **80**(2004), 6–7.

[5] K. Becker, M. Becker, J.-X. Fu, L.-S. Tseng and S.-T. Yau, *Anomaly cancellation and smooth non-Kähler solutions in heterotic string theory*, Nucl. Physics B **751**(2006), 108–128.

[6] K. Becker and K. Dasgupta, *Heterotic strings with torsion*, JHEP **0211**(2002), 006.

[7] M. Becker, L.-S. Tseng and S.-T. Yau, *Moduli space of torsional manifolds*, Nuclear Phys. B **786**(2007), 119–134.

[8] M. Becker, L.-S. Tseng and S.-T. Yau, *Heterotic Kähler/non-Kähler transitions*, Adv. Theor. Math. Phys. **12**(2008), 1147–1162.

[9] M. Becker, L.-S. Tseng and S.-T. Yau, *New Heterotic Non-Kähler Geometries*, arXiv:0807.0827v2 [hep-th].

[10] Y. Bozhkov, *The geometry of certain three-folds*, Rend. Istit. Mat. Univ. Trieste **26**(1994), 79–93(1995).

[11] E. Calabi, *Construction and properties of some* 6*-dimensional almost complex manifolds*, Trans. Amer. Math. Soc. **87**(1958), 407–438.

[12] E. Calabi and B. Eckmann, *A class of compact, complex manifolds which are not algebraic*, Ann. Math. **(2) 58**(1953), 494–500.

[13] P. Candelas and Xenia C. de la Ossa, *Comments on conifolds.* Nuclear Phys. B **342**(1990), 246–268.

[14] K. Dasgupta, G. Rajesh and S.Sethi, *M theory, orientifolds and G-flux*, JHEP **9908**(1999), 023.

[15] P. de Bartolomeis and A. Nannicini, *Introduction to differential geometry of twistor spaces*, Geometric theory of singular phenomena in partial differential equations (Cortona, 1995), 91–160, Sympos. Math., XXXVIII, Cambridge Univ. Press, Cambridge, 1998.

[16] M. Fernandez, S. Ivanov, L. Ugarte and R. Villacampa, *Non-Kaehler heterotic string compactifications with non-zero fluxes and constant dilaton*, Comm. Math. Phys. **288**(2009), 677–697.

[17] R. Friedman, *Simultaneous resolution of threefold double points*, Math. Ann. **274**(1986), 671–689.

[18] R. Friedman, *On threefolds with trivial canonical bundle*, Complex geometry and Lie theory (Sundance, UT, 1989), 103–134, Proc. Sympos. Pure Math., **53**, Amer. Math. Soc., Providence.

[19] J.-X. Fu, J. Li and S.-T. Yau, *Constructing balanced metrics on some families of non-Kähler Calabi-Yau threefolds*, arXiv:0809.4748v1 [math.DG].

[20] J.-X. Fu, L.-S. Tseng and S.-T. Yau, *Local heterotic torsional models*, Comm. Math. Phys. **289**(2009), 1151–1169.

[21] J.-X. Fu, Z.-Z. Wang and D.-M. Wu, *Form-type Calabi-Yau equations*, arXiv:0908.0577v3 [math.DG].

[22] J.-X. Fu and S.-T. Yau, *The theory of superstring with flux on non-Kähler manifolds and the complex Monge-Ampère equation*, J. Differential Geom. **78**(2008), 369–428.

[23] E. Goldstein and S. Prokushkin, *Geometric model for complex non-Kähler manifolds with* $SU(3)$ *structure*, Comm. Math. Phys. **251**(2004), 65–78.

[24] N. J. Hitchin, *Kählerian twistor spaces*, Proc. London Math. Soc. **(3) 43**(1981), 133–150.

[25] Y. Kawamata, *Unobstructed deformations. A remark on a paper of Z. Ran: "Deformations of manifolds with torsion or negative canonical bundle"*, J. Algebraic Geom. **1**(1992), 183–190.

[26] J. Li and S.-T. Yau, *Hermitian Yang-Mills connections on non-Kähler manifolds*, Mathematical aspects of string theory (S.-T. Yau ed.), 560–573, World Scient. Publ. 1987.

[27] J. Li and S.-T. Yau, *The existence of supersymmetric string theorey with torsion*, J. Differential Geom. **70**(2005), 143–181.

[28] P. Lu and G. Tian, *Complex structures on connected sums of* $S^3 \times S^3$, Manifolds and geometry (Pisa, 1993), 284–293, Sympos. Math., **XXXVI**, Cambridge Univ. Press, Cambridge, 1996.

[29] M. L. Michelsohn, *On the existence of special metrics in complex geometry*, Acta Math. **149**(1982), 261–295.

[30] M. Reid, *The moduli space of* 3*-folds with* $K = 0$ *may nevertheless be irreducible*, Math. Ann. **278**(1987), 329–334.

[31] A. Strominger, *Superstings with Torsion*, Nuclear Phys. **B 274**(1986), 253–284.

[32] G. Tian, *Smoothing* 3*-folds with trivial canonical bundle and ordinary double points*, Essays on mirror manifolds, 458–479, Internat. Press, Hong Kong, 1992.

[33] C.-C. Wu, *On the geometry of superstrings with torsion*, thesis, Department of Mathematics, Harvard University, Cambridge MA 02138, April 2006.

[34] 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.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Locally Homogeneous Geometric Manifolds

William M. Goldman*

Abstract

Motivated by Felix Klein's notion that geometry is governed by its group of symmetry transformations, Charles Ehresmann initiated the study of geometric structures on topological spaces locally modeled on a homogeneous space of a Lie group. These locally homogeneous spaces later formed the context of Thurston's 3-dimensional geometrization program. The basic problem is for a given topology Σ and a geometry $X = G/H$, to classify all the possible ways of introducing the local geometry of X into Σ. For example, a sphere admits no local Euclidean geometry: there is no metrically accurate Euclidean atlas of the earth. One develops a space whose points are equivalence classes of geometric structures on Σ, which itself exhibits a rich geometry and symmetries arising from the topological symmetries of Σ.

We survey several examples of the classification of locally homogeneous geometric structures on manifolds in low dimension, and how it leads to a general study of surface group representations. In particular geometric structures are a useful tool in understanding local and global properties of deformation spaces of representations of fundamental groups.

Mathematics Subject Classification (2010). Primary 57M50; Secondary 57N16.

Keywords. Connection, curvature, fiber bundle, homogeneous space, Thurston geometrization of 3-manifolds, uniformization, crystallographic group, discrete group, proper action, Lie group, fundamental group, holonomy, completeness, development, geodesic, symplectic structure, Teichmüller space, Fricke space, hyperbolic structure, Riemannian metric, Riemann surface, affine structure, projective structure, conformal structure, spherical CR structure, complex hyperbolic structure, deformation space, mapping class group, ergodic action.

*Partially supported by the National Science Foundation.
Department of Mathematics, University of Maryland, College Park, MD 20742 USA. E-mail: wmg@math.umd.edu.

1. Historical Background

While geometry involves quantitative measurements and rigid metric relations, topology deals with the loose quantitative organization of points. Felix Klein proposed in his 1872 Erlangen Program that the classical geometries be considered as the properties of a space invariant under a transitive Lie group action. Therefore one may ask which topologies support a system of local coordinates modeled on a fixed homogeneous space $X = G/H$ such that on overlapping coordinate patches, the coordinate changes are locally restrictions of transformations from G.

In this generality this question was first asked by Charles Ehresmann [55] at the conference "Quelques questions de Geométrie et de Topologie," in Geneva in 1935. Forty years later, the subject of such *locally homogeneous geometric structures* experienced a resurgence when W. Thurston placed his 3-dimensional geometrization program [158] in the context of locally homogeneous (Riemannian) structures. The rich diversity of geometries on homogeneous spaces brings in a wide range of techniques, and the field has thrived through their interaction.

Before Ehresmann, the subject may be traced to several independent threads in the 19th century:

- The theory of monodromy of Schwarzian differential equations on Riemann surfaces, which arose from the integration of algebraic functions;
- Symmetries of crystals led to the enumeration (1891) by Fedorov, Schöenflies and Barlow of the 230 three-dimensional crystallographic *space groups* (the 17 two-dimensional *wallpaper groups* had been known much earlier). The general qualitative classification of crystallographic groups is due to Bieberbach.
- The theory of connections, curvature and parallel transport in Riemannian geometry, which arose from the classical theory of surfaces in $\mathbb{R}^3$.

The uniformization of Riemann surfaces linked complex analysis to Euclidean and non-Euclidean geometry. Klein, Poincaré and others saw that the moduli of Riemann surfaces, first conceived by Riemann, related (via uniformization) to the deformation theory of geometric structures. This in turn related to deforming discrete groups (or more accurately, representations of fundamental groups in Lie groups), the viewpoint of the text of Fricke-Klein [62].

2. The Classification Question

Here is the fundamental general problem: Suppose we are given a manifold Σ (a *topology*) and a homogeneous space $(G, X = G/H)$ (a *geometry*). Identify a space whose points correspond to equivalence classes of (G, X)-structures on

Σ. This space should inherit an action of the group of topological symmetries (*the mapping class group* $\mathsf{Mod}(\Sigma)$) of Σ. That is, how many inequivalent ways can one weave the geometry of X into the topology of Σ? Identify the natural $\mathsf{Mod}(\Sigma)$-invariant geometries on this deformation space.

3. Ehresmann Structures and Development

For $n > 1$, the sphere S^n admits no Euclidean structure. This is just the familiar fact there is no metrically accurate atlas of the world. Thus the deformation space of Euclidean structures on S^n is empty. On the other hand, the torus admits a rich class of Euclidean structures, and (after some simple normalizations) the space of Euclidean structures on T^2 identifies with the quotient of the upper half-plane H^2 by the modular group $\mathsf{PGL}(2, \mathbb{Z})$.

Globalizing the coordinate charts in terms of the *developing map* is useful here. Replace the coordinate atlas by a universal covering space $\tilde{M} \longrightarrow M$ with covering group $\pi_1(M)$. Replace the coordinate charts by a local diffeomorphism, the *developing map* $\tilde{M} \xrightarrow{\mathsf{dev}} X$, as follows. dev is equivariant with respect to the actions of $\pi_1(M)$ by deck transformations on $\tilde{M}$ and by a representation $\pi_1(M) \xrightarrow{h} G$, respectively. The coordinate changes are replaced by the *holonomy homomorphism* h. The resulting *developing pair* (dev, h) is unique up to composition/conjugation by elements in G. This determines the structure.

Here is the precise correspondence. Suppose that

$$\{(U_\alpha, \psi_\alpha) \mid U_\alpha \in \mathcal{U}\}$$

is a (G, X)-*coordinate atlas:* $\mathcal{U}$ is an open covering by coordinate patches U_α, with coordinate charts $U_\alpha \xrightarrow{\psi_\alpha} X$ for $U_\alpha \in \mathcal{U}$. For every nonempty connected open subset $U \subset U_\alpha \cap U_\beta$, there is a (necessarily unique)

$$g(U; U_\alpha, U_\beta) \in G$$

such that

$$\psi_\alpha|_U \;=\; g(U) \circ \psi_\beta|_U.$$

(Since a homogeneous space X carries a natural real-analytic structure invariant under G, every (G, X)-manifold carries an underlying real-analytic structure. For convenience, therefore, we fix a smooth structure on Σ, and work in the differentiable category, where tools such as transversality are available. Since we concentrate here in low dimensions (like 2), restricting to smooth manifolds and mappings sacrifices no generality. Therefore, when we speak of "a topological space Σ" we really mean a smooth manifold Σ rather than just a topological space.)

The coordinate changes $\{g(U; U_\alpha, U_\beta)\}$ define a *flat* (G, X)-*bundle* as follows. Start with the trivial (G, X)-bundle over the disjoint union $\coprod_{U_\alpha \in \mathcal{U}} U_\alpha$,

having components

$$E_\alpha := U_\alpha \times X \xrightarrow{\Pi_\alpha} U_\alpha.$$

Now identify, for

$$(u, u_\alpha, u_\beta) \in U \times U_\alpha \times U_\beta,$$

the two local total spaces $U \times X \subset E_\alpha$ with $U \times X \subset E_\beta$ by

$$\big(u, x\big)_\alpha \longleftrightarrow \big(u, g(U; U_\alpha, U_\beta)x\big)_\beta. \tag{1}$$

The fibrations Π_α over U_α piece together to form a fibration $E(M) \xrightarrow{\Pi} M$ over M with fiber X, and structure group G, whose total space $E = E(M)$ is the quotient space of the E_α by the identifications (1). The foliations $\mathcal{F}_\alpha$ of E_α defined locally by the projections $U_\alpha \times X \longrightarrow X$ piece together to define a foliation $\mathcal{F}(M)$ of $E(M)$ transverse to the fibration. In this atlas, the coordinate changes are locally constant maps $U_\alpha \cap U_\beta \longrightarrow G$. This *reduces the structure group* from G with its manifold topology to G *with the discrete topology.* We call the fiber bundle $\big(E(M), \mathcal{F}(M)\big)$ the *flat* (G, X)*-bundle tangent to* M.

Such a bundle pulls back to a trivial bundle over the universal covering $\tilde{M} \longrightarrow M$. Thus it may be reconstructed from the trivial bundle $\tilde{M} \times X \longrightarrow \tilde{M}$ as the quotient of a $\pi_1(M)$-action on $\tilde{M} \times X$ covering the action on $\tilde{M}$ by deck transformations. Such an action is determined by a homomorphism $\pi_1(M) \xrightarrow{h} G$, the *holonomy representation.* Isomorphism classes of flat bundles with structure group G correspond to G-orbits on $\mathsf{Hom}\big(\pi_1(M), G\big)$ by left-composition with inner automorphisms of G.

The coordinate charts $U_\alpha \xrightarrow{\psi_\alpha} X$ globalize to a section of the flat (G, X)-bundle $E \longrightarrow M$ as follows. The graph $\mathsf{graph}(\psi_\alpha)$ is a section transverse both to the fibration and the foliation $\mathcal{F}_\alpha$. Furthermore the identifications (1) imply that the restrictions of $\mathsf{graph}(\psi_\alpha)$ and $\mathsf{graph}(\psi_\beta)$ to $U \subset U_\alpha \cap U_\beta$ identify. Therefore all the ψ_α are the restrictions of a globally defined $\mathcal{F}$-transverse section $M \xrightarrow{\mathsf{Dev}} E$. We call this section the *developing section* since it exactly corresponds to a developing map.

Conversely, suppose that $(E, \mathcal{F})$ is a flat (G, X)-bundle over M and $M \xrightarrow{s} E$ is a section transverse to $\mathcal{F}$. For each $m \in M$, choose an open neighborhood U such that the foliation $\mathcal{F}$ on the local total space $\Pi^{-1}(U)$ is defined by a submersion $\Pi^{-1}(U) \xrightarrow{\Psi_U} X$. Then the compositions $\Psi_U \circ s$ define coordinate charts for a (G, X)-structure on M.

In terms of the universal covering space $\tilde{M} \longrightarrow M$ and holonomy representation h, a section $M \xrightarrow{s} E$ corresponds to a $\pi_1(M)$-equivariant mapping $\tilde{M} \xrightarrow{\tilde{s}} X$, where $\pi_1(M)$ acts on X via h. The section s is transverse to $\mathcal{F}$ if and only if the corresponding equivariant map $\tilde{s}$ is a local diffeomorphism.

4. Elementary Consequences

As the universal covering $\tilde{M}$ immerses in X, no (G, X)-structure exists when M is closed with finite fundamental group and X is noncompact. Furthermore if X is compact and simply connected, then every closed (G, X)-manifold with finite fundamental group would be a quotient of X. Thus by extremely elementary considerations, no counterexample to the Poincaré conjecture could be modeled on S^3.

When G acts properly on X (that is, when the isotropy group is compact), then G preserves a Riemannian metric on X which passes down to a metric on M. This metric lifts to a Riemannian metric on the the universal covering $\tilde{M}$, for which dev is a local isometry. Suppose that M is closed. The Riemannian metric on M makes M into a metric space, which is necessarily complete. By the Hopf-Rinow theorem, M is geodesically complete, and (after possibly replacing X with its universal covering space $\tilde{X}$, and G by an appropriate group $\tilde{G}$ of lifts), the local isometry dev is a covering space, and maps $\tilde{M}$ bijectively to $\tilde{X}$. In particular such structures correspond to discrete cocompact subgroups of $\tilde{G}$. In this way the subject of Ehresmann geometric structures extends the subject of discrete subgroups of Lie groups.

In general, even for closed manifolds, the developing map may fail to be surjective (for example, Hopf manifolds), and even may not be a covering space onto its image (Hejhal [103], Smillie [152], Sullivan-Thurston [155]).

5. The Hierarchy of Geometries

Often one geometry "contains" another geometry as follows. Suppose that G and G' act transitively on X and X' respectively, and $X \xrightarrow{f} X'$ is a local diffeomorphism equivariant respecting a homomorphism $G \xrightarrow{F} G'$. Then (by composition with f and F) every (G, X)-structure determines a (G', X')-structure. For example, when f is the identity, then G may be the subgroup of G' preserving some extra structure on $X = X'$. In this way, various flat pseudo-Riemannian geometries are refinements of affine geometry. The three constant curvature Riemannian geometries (Euclidean, spherical, and hyperbolic) have both realizations in conformal geometry of S^n (the Poincaré model) and in projective geometry (the Beltrami-Klein model) in $\mathbb{R}\mathsf{P}^n$. In more classical differential-geometric terms, this is just the fact that the constant curvature Riemannian geometries are *conformally flat* (respectively *projectively flat*). Identifying conformal classes of conformally flat Riemannian metrics as Ehresmann structures follows from Liouville's theorem on the classification of conformal maps of domains in $\mathbb{R}^n$ for $n \geq 3$.

An interesting and nontrivial example is the classification of closed similarity manifolds by Fried [63]. Here $X = \mathbb{R}^n$ and G is its group of similarity transformations. Fried showed that every closed (G, X)-manifold M is either

a Euclidean manifold (so G reduces to the group of *isometries*) or a *Hopf manifold,* a quotient of $\mathbb{R}^n \setminus \{0\}$ by a cyclic group of linear expansions. In the latter case M carries a $\big(\mathbb{R}^+ \cdot \mathsf{O}(n), \mathbb{R}^n \setminus \{0\}\big)$-structure. Such manifolds are finite quotients of $S^{n-1} \times S^1$.

6. Deforming Ehresmann Structures

One would like a space whose points are equivalence classes of (G, X)-structures on a fixed topology Σ. The prototype of such a *deformation space* is the *Teichmüller space* $\mathfrak{T}(\Sigma)$ of biholomorphism classes of complex structures on a fixed surface Σ. That is, we consider a Riemann surface M with a diffeomorphism $\Sigma \longrightarrow M$, which is commonly called a *marking.* Although complex structures are not Ehresmann structures, there is still a formal similarity. (This formal similarity can be made into an equivalence of categories via the uniformization theorem, but this is considerably deeper than the present discussion.) For example, every Riemann surface diffeomorphic to T^2 arises as $\mathbb{C}/\Lambda$, where $\Lambda \subset \mathbb{C}$ is a lattice. Two such lattices Λ, Λ' determine isomorphic Riemann surfaces if $\exists \zeta \in \mathbb{C}^*$ such that $\Lambda' = \zeta\Lambda$. The space of such equivalence classes identifies with the quotient $H^2/\mathsf{PSL}(2, \mathbb{Z})$. The quotient $H^2/\mathsf{PSL}(2, \mathbb{Z})$ has the natural structure of an *orbifold,*) and is not naturally a manifold.

In general deformation spaces will have very bad separation properties. (For example the space of complete affine structures on T^2 naturally identifies with the quotient of $\mathbb{R}^2$ by the usual linear action of $\mathsf{SL}(2, \mathbb{Z})$ (Baues, see [8].) This quotient admits no nonconstant continuous mappings into any Hausdorff space!) To deal with such pathologies, we form a *larger space* with a group action, whose orbit space parametrizes isomorphism classes of (G, X)-manifolds diffeomorphic to Σ. In general, passing to the orbit space alone loses too much information, and may result in an unwieldy topological space. For this reason, considering the *deformation groupoid,* consisting of structures (rather than equivalence classes) and isomorphisms between them, is a more meaningful and useful object to parametrize geometric structures.

Therefore we fix a smooth manifold Σ and define a *marked* (G, X)*-structure* on Σ as a pair (M, f) where M is a (G, X)-manifold and $\Sigma \xrightarrow{f} M$ a diffeomorphism. Suppose that Σ is compact (possibly $\partial\Sigma \neq \emptyset$). Fix a fiber bundle E over Σ with fiber X and structure group G. Give the set $\mathsf{Def}^{(G,X)}(\Sigma)$ of such marked (G, X)-structures on Σ the C^1-topology on pairs $(\mathcal{F}, \mathsf{Dev})$ of foliations $\mathcal{F}$ and smooth sections Dev. Clearly the diffeomorphism group $\mathsf{Diff}(\Sigma)$ acts on $\mathsf{Def}^{(G,X)}(\Sigma)$ by left-composition. Define marked (G, X)-structures (M, f) and (M', f') to be *isotopic* if they are related by an diffeomorphism of Σ isotopic to the identity.

Define the *deformation space* of isotopy classes of marked (G, X)-structures on Σ as the quotient space

$$\mathsf{Def}^{(G,X)}(\Sigma) \ := \ \mathsf{Def}^{(G,X)}(\Sigma)/\mathsf{Diff}_0(\Sigma).$$

Clearly the *diffeotopy group* $\pi_0\big(\mathsf{Diff}(\Sigma)\big)$ (which for compact surfaces Σ is the *mapping class group* $\mathsf{Mod}(\Sigma)$) acts on the deformation space.

7. Representations of the Fundamental Group

The set of isomorphism classes of flat G-bundles over Σ identifies with the set $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ of equivalence classes of representations $\pi_1(\Sigma) \longrightarrow G$, where two representations ρ, ρ' are equivalent if and only if $\exists g \in G$ such that $\rho' = \mathsf{Inn}(g) \circ \rho$, where $\mathsf{Inn}(g) : x \longmapsto gxg^{-1}$ is the inner automorphism associated to $g \in G$. Since $\pi_1(\Sigma)$ is finitely generated, $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)$ has the structure of a real-analytic subset in a Cartesian power G^N, and this structure is independent of the choice of generators. Give $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)$ the classical topology and note that it is stratified into smooth submanifolds. Give $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ the quotient topology.

The space $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ may enjoy several pathologies:

- The analytic variety $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)$ may have singularities, and not be a manifold;
- G may not act freely, even on the smooth points, so the quotient map may be nontrivially branched, and $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ may have orbifold singularities;
- G may not act properly, and the quotient space $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ may not be Hausdorff.

All three pathologies may occur.

The automorphism group $\mathsf{Aut}\big(\pi_1(\Sigma)\big)$ acts on $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)$ by right-composition. The action of its subgroup $\mathsf{Inn}\big(\pi_1(\Sigma)\big)$ is absorbed in the $\mathsf{Inn}(G)$-action, and therefore the quotient group

$$\mathsf{Out}\big(\pi_1(\Sigma)\big) \;:=\; \mathsf{Aut}\big(\pi_1(\Sigma)\big)/\mathsf{Inn}\big(\pi_1(\Sigma)\big)$$

acts on the quotient

$$\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G.$$

Associating to a marked (G, X)-structure the equivalence class of its holonomy representation defines a continuous map

$$\mathsf{Def}^{(G,X)}(\Sigma) \xrightarrow{\mathsf{hol}} \mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G \tag{2}$$

which is evidently $\pi_0\big(\mathsf{Diff}(\Sigma)\big)$-equivariant, with respect to the homomorphism

$$\pi_0\big(\mathsf{Diff}(\Sigma)\big) \longrightarrow \mathsf{Out}\big(\pi_1(\Sigma)\big).$$

Theorem (Thurston). *With respect to the above topologies, the holonomy map* hol *in* (2) *is a local homeomorphism.*

For hyperbolic structures on closed surfaces, which are special cases of (G, G)-structures (or discrete embeddings in Lie groups as above), this result is due to Weil [168, 169, 170]; see the very readable paper by Bergeron-Gelander [19]. This result is due to Hejhal [103] for $\mathbb{C}\mathsf{P}^1$-surfaces. The general theorem was first stated explicitly by Thurston [158], and perhaps the first careful proof may be found in Lok [125] and Canary-Epstein-Green [31]. Bergeron and Gelander refer to this result as the "Ehresmann-Thurston theorem" since many of the ideas are implicit in Ehresmann's viewpoint [56].

The following proof was worked out in [74] with Hirsch, and was also known to Haefliger. By the covering homotopy theorem and the local contractibility of $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)$, the isomorphism type of E as a G-bundle is constant. Thus one may assume that E is a fixed G-bundle, although the flat structure (given by the transverse foliation $\mathcal{F}$) varies, as the representation varies. However it varies continuously in the C^1 topology. Thus a given $\mathcal{F}$-transverse section Dev remains transverse as $\mathcal{F}$ varies, and defines a geometric structure. This proves local surjectivity of hol.

Conversely, if Dev' is a transverse section sufficiently close to Dev in the C^1-topology, then it stays within a neighborhood of $\mathsf{Dev}(\Sigma)$. For a sufficiently small neighborhood W of $\mathsf{Dev}(\Sigma)$, the foliation $\mathcal{F}|_W$ identifies with a product foliation of $W \approx \mathsf{Dev}(\Sigma) \times X$ defined by the projection to X. For each $m \in \Sigma$, the leaf of $\mathcal{F}|_W$ through $\mathsf{Dev}(m)$ meets $\mathsf{Dev}'(\Sigma)$ in a unique point $\mathsf{Dev}'(m')$ for $m' \in \Sigma$. The correspondence $m \longmapsto m'$ is the required isotopy, from which follows hol is locally injective.

8. Thurston's Geometrization of 3-manifolds

In 1976, Thurston proposed that every closed 3-manifold admits a canonical decomposition into pieces, by cutting along surfaces of nonnegative Euler characteristic. Each of these pieces has one of eight geometries, modeled on eight 3-dimensional Riemannian homogeneous spaces:

- **Elliptic geometry:** Here $X = S^3$ and $G = \mathsf{O}(3)$ its group of isometries. Manifolds with these geometries are the Riemannian 3-manifolds of constant positive curvature, that is, *spherical space forms,* and include lens spaces.

- $S^2 \times \mathbb{R}$: The only closed 3-manifolds with this geometry are $S^2 \times S^1$ and a few quotients.

- **Euclidean geometry:** Here $X = \mathbb{R}^3$ and G its group of isometries. These are the Riemannian manifolds of zero curvature, and are quotients by torsionfree Euclidean *crystallographic groups.* In 1912, Bieberbach proved every closed Euclidean manifold is a quotient of a flat torus by a finite group of isometries. Furthermore he proved there are only finitely many

topological types of these manifolds, and that any homotopy-equivalence is homotopic to an *affine isomorphism.*

- **Nilgeometry:** Here again $X = \mathbb{R}^3$, regarded as the Heisenberg group with a left-invariant metric and G its group of isometries. Manifolds with these geometry are covered by nontrivial oriented S^1-bundles over 2-tori.
- **Solvgeometry:** Once again $X = \mathbb{R}^3$, regarded as a 3-dimensional exponential solvable unimodular non-nilpotent Lie group and G the group of isometries of a left-invariant metric. Hyperbolic torus bundles (suspensions of Anosov diffeomorphisms of tori) have these structures.
- $H^2 \times \mathbb{R}$: Products of hyperbolic surfaces with S^1 have this geometry.
- **Unit tangent bundle of H^2:** An equivalent model is $\mathsf{PSL}(2, \mathbb{R})$ with a left-invariant metric. Nontrivial oriented S^1-bundles of hyperbolic surfaces (such as the unit tangent bundle) admit such structures.
- **Hyperbolic geometry:** Here $X = H^3$ and G its group of isometries.

For a description of the eight homogeneous Riemannian geometries and their relationship to 3-manifolds, see the excellent surveys by Scott [147] and Bonahon [21].

9. Complete Affine 3-manifolds

Manifolds modeled on *Euclidean geometry* are exactly the flat Riemannian manifolds. Compact Euclidean manifolds M^n are precisely the quotients $\mathbb{R}^n/\Gamma$, where Γ is a lattice of Euclidean isometries. By the work of Bieberbach (1912), such a Γ is a finite extension of a lattice Λ of translations. Thus M is finitely covered by the torus $\mathbb{R}^n/\Lambda$. Since all lattices $\Lambda \subset \mathbb{R}^n$ are *affinely* the homotopy type of M determines its affine equivalence class. When M is noncompact, but geodesically complete, then M is isometric to a flat orthogonal vector bundle over a compact Euclidean manifold.

These theorems give at least a qualitative classification of manifolds with Euclidean structures. The generalization to manifolds with *affine structures* is much more mysterious and difficult. We begin by restricting to ones which are *geodesically complete.* In that case the manifolds are quotients $\mathbb{R}^n/\Gamma$ but Γ is only assumed to consist of affine transformations. However, unlike Euclidean manifolds considered above, discreteness of $\Gamma \subset \mathsf{Aff}(\mathbb{R}^n)$ does not generally imply the properness of the action, and the quotient may not be Hausdorff. Characterizing which affine representations define proper actions is a fundamental and challenging problem.

In the early 1960's, L. Auslander announced that every compact complete affine manifold has virtually polycyclic fundamental group, but his proof was flawed. In this case, the manifold is finitely covered by an *affine solvmanifold*

$\Gamma\backslash G$ where G is a (necessarily solvable) Lie group with a left-invariant complete affine structure and $\Gamma \subset G$ is a lattice. Despite many partial results, ([64, 2, 3, 164, 87]) the *Auslander Conjecture* remains open.

Milnor [134] asked whether the virtual polycyclicity of Γ might hold even if the quotient $\mathbb{R}^n/\Gamma$ is *noncompact.* Using the Tits Alternative [162], he reduced this question to whether a rank two free group could act *properly* by affine transformations on $\mathbb{R}^n$. Margulis [128] showed, surprisingly, that such actions *do* exist when $n = 3$.

For $n = 3$, Fried and Goldman [64] showed that either Γ is virtually polycyclic (in which case all the structures are easily classified), or the linear holonomy homomorphism $\Gamma \xrightarrow{\mathsf{L}} \mathsf{GL}(3,\mathbb{R})$ maps Γ isomorphically onto a discrete subgroup of a conjugate of $\mathsf{O}(2,1) \subset \mathsf{GL}(3,\mathbb{R})$. Since $\mathsf{L}^{-1}\mathsf{O}(2,1)$ preserves a flat Lorentz metric on $\mathbb{R}^3$, the geometric structure on M refines to a flat Lorentz structure, modeled on $\mathbb{E}^3_1$, which is $\mathbb{R}^3$ with the corresponding flat Lorentz metric. In particular $M^3 = \mathbb{E}^3_1/\Gamma$ is a *complete flat Lorentz 3-manifold* and $\Sigma := H^2/\mathsf{L}(\Gamma)$ is a complete hyperbolic surface. This establishes the Auslander Conjecture in dimension 3: the cohomological dimension of $\Gamma \cong \pi_1(M^3)$ equals 3 since M is aspherical, but the cohomological dimension $\Gamma \cong \pi_1(\Sigma)$ is at most 2. In 1990, Mess [131] proved that the surface Σ is *noncompact,* and therefore Γ must be a free group. (Compare also Goldman-Margulis [90] and Labourie [119] for other proofs.)

Drumm [51, 52] (see also [39]) gave a geometric construction of these quotient manifolds using polyhedra in Minkowski space $\mathbb{R}^3_1$ now called *crooked planes.* Using crooked planes, he showed that every noncompact complete hyperbolic surface Σ arises from a complete flat Lorentz 3-manifold; that is, he showed that every non-cocompact Fuchsian group $\mathsf{L}(\Gamma) \subset \mathsf{O}(2,1)$ admits a *proper* affine deformation Γ.

The conjectural picture of these manifolds is as follows.

The space of equivalence classes of affine deformations of Γ is the vector space $H^1(\Gamma, \mathbb{R}^3_1)$, and the proper affine deformations define an open convex cone in this vector space. Goldman-Labourie-Margulis [89] have proved this when Γ is finitely generated and contains no parabolic elements. Furthermore a finite-index subgroup of Γ should have a fundamental domain which is bounded by crooked planes, and M^3 should be homeomorphic to a solid handlebody. Charette-Drumm-Goldman [37] have proved this when Σ is homeomorphic to a 3-holed sphere.

Translational conjugacy classes of affine deformations of a Fuchsian group $\Gamma_0 \subset \mathsf{O}(2,1)$ comprise the cohomology group $H^1(\Gamma_0; \mathbb{R}^3_1)$. As the $\mathsf{O}(2,1)$-module $\mathbb{R}^3_1$ identifies with the Lie algebra of $\mathsf{O}(2,1)$ with the adjoint representation, this cohomology group identifies with the *space of infinitesimal deformations of the hyperbolic surface* $\Sigma = H^2/\Gamma_0$. (Compare Goldman-Margulis [90] and [80].)

When Σ has no cusps, [89] provides a criterion for properness of an affine deformation corresponding to a deformation σ of the hyperbolic surface Σ. The affine deformation Γ_σ acts properly on $\mathbb{E}^3_1$ if and only if every probability

measure on $U\Sigma$ invariant under the geodesic flow *infinitesimally lengthens* (respectively *infinitesimally shortens* under σ. (We conjecture a similar statement in general.) Using ideas based on Thurston [161], one can reduce this to probability measures arising from measured geodesic laminations. When Σ is a three-holed sphere, [37] implies the proper affine deformations are precisely the ones for which the three components of $\partial\Sigma$ either all infinitesimally lengthen or all infinitesimally shorten.

Other examples of *conformally flat Lorentzian manifolds* have recently been studied by Frances [61], Zeghib [176], and Bonsante-Schlenker [22], also closely relating to hyperbolic geometry.

10. Affine Structures on Closed Manifolds

The question of which closed manifolds admit affine structures seems quite difficult. Even for complete structures, the pattern is mysterious. Milnor [134] asked whether every virtually polycyclic group arises as the fundamental group of a compact complete affine manifold. Benoist [9, 10] found 11-dimensional nilpotent counterexamples. However by replacing $\mathbb{R}^n$ by a simply connected nilpotent Lie group, one obtains more general structures. Dekimpe [48] showed that every virtually polycyclic group arises as the fundamental group of such a *NIL-affine* manifold.

For incomplete structures, the picture is even more unclear. The *Markus conjecture,* first stated by L. Markus as a homework exercise in unpublished lecture notes at the University of Minnesota in 1960 asserts that, for closed affine manifolds, geodesic completeness is equivalent to parallel volume (linear holonomy in $\mathsf{SL}(n, \mathbb{R})$. That this conjecture remains open testifies to our current ignorance.

An important partial result is Carrière's result [32] that a closed flat Lorentzian manifold is geodesically complete. This has been generalized in a different direction by Klingler [112] to all closed Lorentzian manifolds with *constant* curvature.

Using parallel volume forms, Smillie [153] showed that the holonomy of a compact affine manifold cannot factor through a free product of finite groups; his methods were extended by Goldman-Hirsch [85, 86] to prove nonexistence results for affine structures on closed manifolds with certain conditions on the holonomy. Using these results, Carrière, Dal'bo and Meigniez [33] showed that certain Seifert 3-manifolds with hyperbolic base admit no affine structures.

Perhaps the most famous conjecture about affine structures on closed manifolds is Chern's conjecture that a closed affine manifold must have Euler characteristic zero. For flat pseudo-Riemannian manifolds or *complex* affine manifolds, this follows from Chern-Gauss-Bonnet. Using an elegant argument, Kostant and Sullivan [114] proved this conjecture for complete affine manifolds. (This would follow immediately from the Auslander Conjecture.)

In a different direction, Smillie [151] found simple examples of closed manifolds with flat tangent bundles (these would have affine connections with zero curvature, but possibly nonzero torsion). Recent results in this direction have been obtained by Bucher-Gelander [26].

11. Hyperbolic Geometry on 2-manifolds

The prototype of geometric structures, and historically one of the basic examples, are hyperbolic structures on surfaces Σ with $\chi(\Sigma) < 0$. Here X is the hyperbolic plane and $G \cong \mathsf{PGL}(2, \mathbb{R})$. Fricke and Klein [62] studied the deformation space of hyperbolic structures on Σ as well as on 2-dimensional orbifolds. The deformation space $\mathfrak{F}(\Sigma)$ of marked hyperbolic structures on Σ (sometimes called *Fricke Space* ([20]) can also be described as the space of equivalence classes of discrete embeddings $\pi_1(\Sigma) \longrightarrow G$. The Poincaré-Klein-Koebe Uniformization Theorem relates hyperbolic structures and complex structures, so the Fricke space identifies with the *Teichmüller space* of Σ, which parametrizes Riemann surfaces homeomorphic to Σ. For this reason, although Teichmüller himself never studied hyperbolic geometry, the deformation theory of hyperbolic structures on surfaces is often referred to as *Teichmüller theory.*

Representations of surface groups in $G = \mathsf{PSL}(2, \mathbb{R})$ closely relate to geometric structures. A representation $\pi_1(\Sigma) \xrightarrow{\rho} G$ determines an oriented flat H^2-bundle over Σ. Oriented flat H^2-bundles are classified by their *Euler class*, which lives in $H^2(\Sigma; \mathbb{Z}) \cong \mathbb{Z}$ when Σ is closed and oriented. The Euler number of a flat oriented H^2-bundle satisfies

$$|\mathsf{Euler}(\rho)| \leq -\chi(\Sigma) \tag{3}$$

as proved by Wood[175], following earlier work of Milnor[134].

Theorem 1. *Equality holds in* (3) *if and only if ρ is a discrete embedding.*

This theorem was first proved in [69], using Ehresmann's viewpoint. Namely, the condition that $\mathsf{Euler}(\rho) = \pm\chi(\Sigma)$ means that the associated flat H^2-bundle E_ρ with holonomy homomorphism ρ is isomorphic (up to changing orientation) to the tangent bundle of Σ (as a topological disc bundle, or equivalently a microbundle over Σ). If ρ is the holonomy of a hyperbolic surface $M \approx \Sigma$, then $E(M) = E_\rho \approx T\Sigma$. Theorem 1 is a converse: if the flat bundle "is isomorphic to the tangent bundle (as a (G, X)-bundle)", then the flat (G, X)-bundle arises from a (G, X)-structure on Σ.

In the case the representation ρ has discrete torsionfree cocompact image, Theorem 1 reduces to a classical result of Kneser [113]. In 1930 Kneser proved that if $\Sigma \xrightarrow{f} \Sigma'$ is a continuous map of degree d, then

$$d|\chi(\Sigma')| \ \leq \ |\chi(\Sigma)|$$

with equality $\iff f$ is homotopic to a covering space. (In this case Σ' is the hyperbolic surface obtained as the quotient by the image of ρ, and $\mathsf{Euler}(\rho) = d\chi(\Sigma')$. Kneser's theorem is thus a *discrete* version of Theorem 1.

By now Theorem 1 has many proofs and extensions. One proof, using harmonic maps, begins by choosing a Riemann surface $M \approx \Sigma$. Then, by Corlette [47] and Donaldson [50], either the image of ρ is solvable (in which case $\mathsf{Euler}(\rho) = 0$) or the image is reductive, and there exists a ρ-equivariant harmonic map $\tilde{M} \xrightarrow{h} X$. By an adaptation of Eels-Wood [54], $\mathsf{Euler}(\rho)$ can be computed as the sum of local indices of the critical points of h. In particular, the assumption of *maximality:* $\mathsf{Euler}(\rho) = \pm\chi(M)$ implies that h must be holomorphic (or anti-holomorphic), and using the arguments of Schoen-Yau [142], h must be a diffeomorphism. In particular ρ must be a discrete embedding.

Shortly after [69], another proof was given by Matsumoto [129] (compare also Mess [131]), related to ideas of bounded cohomology. This led to the work of Ghys [67], who proved that the Euler class of an orientation-preserving action of $\pi_1(\Sigma)$ on S^1 is a *bounded* class, and its class in bounded cohomology determines the action up to topological semi-conjugacy. In particular maximality in the Milnor-Wood inequality (3) implies the topological action is conjugate to the projective action arising from (any) discrete embedding in $\mathsf{PSL}(2,\mathbb{R})$.

The Euler number classifies components of $\mathsf{Hom}\big(\pi_1(\Sigma), \mathsf{PSL}(2,\mathbb{R})\big)$. That is, if Σ is closed, oriented, of genus $g > 1$, the $4g-3$ connected components are the inverse images $\mathsf{Euler}^{-1}(j)$ where

$$j \;=\; 2-2g, 3-2g, \ldots, 2g-2$$

(Goldman [76]). Independently, Hitchin [104] gave a much different proof, using Higgs bundles. Moreover he identified the Euler class $2-2g+k$ component with a vector bundle over the k-th symmetric power of Σ (compare the expository article [84])

When G is a semisimple *compact* or *complex* Lie group, components of the representation space bijectively correspond to $\pi_1(G)$. In particular in these basic cases, the number of components is *independent* of the genus. (See Li [124] and Rapinchuk–Benyash-Krivetz–Chernousov[141].) Recently Florentino and Lawton [58] have determined the homotopy type of $\mathsf{Hom}(\Gamma, G)//G$ when Γ is free and G is a complex reductive group.

This simple picture becomes much more intricate and fascinating for higher dimensional noncompact real Lie groups; the most effective technique so far has been the interpretation in terms of Higgs bundles and the use of infinite-dimensional Morse theory; see Bradlow-Garcia-Prada-Gothen [23] for a survey of some recent results on the components when G is a simple *real* Lie group.

Theorem 1 leads to rigidity theorems for surface group representations as well. When G is the automorphism group of a Hermitian symmetric space X, integrating a G-invariant Kähler form on X over a smooth section of a flat (G, X)-bundle induces a characteristic class $\tau(\rho)$ first defined by Turaev [165] and Toledo [163]. This characteristic class satisfies an inequality similar to (3).

The *maximal representations,* (when equality is attained) have very special properties. When X is complex hyperbolic space, a representation $\pi_1(\Sigma) \xrightarrow{\rho} \mathsf{PU}(n,1)$ is maximal if and only if it stabilizes a totally geodesic holomorphic curve, and its restriction is Fuchsian (Toledo [163]).

In higher rank the situation is much more interesting and complicated. Burger-Iozzi-Wienhard [28] showed that maximal representations are discrete embeddings, with reductive Zariski closures. With Labourie, they proved [27] in the case of $\mathsf{Sp}(2n,\mathbb{R})$, that these representations quasi-isometrically embed $\pi_1(\Sigma)$ in G. Many of these properties follow from the fact that maximal representations are Anosov representations in the sense of Labourie [120]. Using Higgs bundle theory, Bradlow-Garcia-Prada-Gothen [23] have counted components of maximal representations. Guichard-Wienhard [101] have found components of maximal representations in $\mathsf{Sp}(2n,\mathbb{R})$, all of whose elements have Zariski dense image (in contrast to $\mathsf{PU}(n,1)$ discussed above). For a good survey of these results, see Burger-Iozzi-Wienhard [29].

12. Complex Projective 1-manifolds, Flat Conformal Structures and Spherical CR Structures

When X is enlarged to $\mathbb{C}\mathsf{P}^1$ and G to $\mathsf{PSL}(2,\mathbb{C})$, the resulting deformation theory of $\mathbb{C}\mathsf{P}^1$-structures is quite rich. A manifold modeled on this geometry is naturally a Riemann surface, and thus the deformation space fibers over the Teichmüller space of marked Riemann surfaces:

$$\mathsf{Def}^{(G,X)}(\Sigma) \longrightarrow \mathfrak{T}(\Sigma). \tag{4}$$

The classical theory of the Schwarzian derivative identifies this fibration with a holomorphic affine bundle, where the fiber over a point in $\mathfrak{T}(\Sigma)$ corresponding to a marked Riemann surface $\Sigma \xrightarrow{\approx} M$ is an affine space with underlying vector space $H^0(M;\kappa_M^2)$ consisting of holomorphic quadratic differentials on M.

In the late 1970's, Thurston (unpublished) showed that $\mathsf{Def}^{(G,X)}(\Sigma)$ admits an alternate description as $\mathfrak{F}(\Sigma)\times\mathcal{ML}(\Sigma)$ where $\mathcal{ML}(\Sigma)$ is the space of equivalence classes of measured geodesic laminations on Σ. (Compare Kamishima-Tan [108].) [73] gives the topological classification of $\mathbb{C}\mathsf{P}^1$-structures whose holonomy representation is a quasi-Fuchsian embedding. Gallo-Kapovich-Marden [65] showed that the image of the holonomy map hol consists of representations into $\mathsf{PSL}(2,\mathbb{C})$ which lift to an irreducible and unbounded representation into $\mathsf{SL}(2,\mathbb{C})$.

For an excellent survey of this subject, see Dumas [53].

These structures generalize to higher dimensions in several ways. For example $\mathsf{PSL}(2,\mathbb{C})$ is the group of orientation-preserving conformal automorphisms of $\mathbb{C}\mathsf{P}^1 \approx S^2$. A *flat conformal structure* is a geometric structure locally modeled

on S^n with its group of conformal automorphisms. This structure is equivalent to a conformal class of Riemannian metrics, which are *locally* conformally equivalent to Euclidean metrics. (Compare Matsumoto [130].) In the 1970's it seemed tempting to try to prove the Poincaré conjecture by showing that every closed 3-manifold admits such a structure. This was supported by the fact that these structures are closed under connected sums (Kulkarni [117]). This approach was further promoted by the fact that such structures arise as critical points of the Chern-Simons functional [42], and one could try to reach critical points by following the gradient flow of the Chern-Simons functional. However, closed 3-manifolds with nilgeometry or solvgeometry admit no flat conformal structures whatsoever [71]).

As $H^{n-1} \times \mathbb{R}$ embeds in S^n as the complement of a codimension-two subsphere, the conformal geometry of S^n contains $H^{n-1} \times \mathbb{R}$-geometry. Thus products of closed surfaces with S^1 do admit flat conformal structures, and Kapovich [109] and Gromov-Lawson-Thurston [97] showed that even some nontrivial S^1-bundles over closed surfaces admit flat conformal structures, although $T_1(H^2)$-geometry admits no conformal model in S^3.

Kulkarni-Pinkall [118] have extended Thurston's correspondence

$$\mathsf{Def}^{(G,X)}(\Sigma) \longleftrightarrow \mathfrak{F}(\Sigma) \times \mathcal{ML}(\Sigma)$$

to associate to a flat conformal structure on a manifold (satisfying a generic condition of "hyperbolic type") a hyperbolic metric with some extrinsic (bending) data. b

A similar class of structures are the *spherical CR-structures,* modeled on S^{2n-1} as the boundary of *complex hyperbolic n-space,* in the same way that S^{n-1} with its conformal structure bounds real hyperbolic n-space. Some of the first examples were given by Burns-Shnider [30]. 3-manifolds with nilgeometry naturally admits such structures, but by [71], closed 3-manifolds with Euclidean and solvgeometry do not admit such structures. Twisted S^1-bundles admit many such structures (see for example [88]), but recently Ananin, Grossi and Gusevskii [4, 5] have constructed surprising examples of spherical CR-structures on products of closed hyperbolic surfaces with S^1. Other interesting examples of spherical CR-structures on 3-manifolds have been constructed by Schwartz [144, 145, 146], Falbel [57], Gusevskii, Parker [137], Parker-Platis [138].

When $X = \mathbb{R}\mathsf{P}^n$ and $G = \mathsf{PGL}(n+1, \mathbb{R})$, then a (G, X)-structure is a *flat projective connection.*

In dimension 3, the only closed manifold known *not* to admit an $\mathbb{R}\mathsf{P}^3$-structure is the connected sum $\mathbb{R}\mathsf{P}^3 \# \mathbb{R}\mathsf{P}^3$ (Cooper-Goldman [46]). Many diverse examples of $\mathbb{R}\mathsf{P}^3$-structures on twisted S^1-bundles over closed hyperbolic surfaces arise from maximal representations of surface groups into $\mathsf{Sp}(4, \mathbb{R})$ by Guichard-Wienhard [101]. All eight of the Thurston geometries have models in $\mathbb{R}\mathsf{P}^3$ [136].

The 2-dimensional theory is relatively mature. The most important examples are the *convex* structures, namely those which arise as quotients Ω/Γ where

Ω is a convex domain in $\mathbb{R}\mathsf{P}^2$ and Γ is a group of collineations preserving Ω. Kuiper [115] showed that all convex structures on 2-tori are affine structures, and classified them. They are all quotients of the plane, a half-plane or a quadrant. In higher genus, he showed [116] that either $\partial\Omega$ is a conic (in which case the projective structure is a hyperbolic structure) or it fails to be C^2. Benzecri [18] showed that in the latter case, it is C^1 and is strictly convex. Using the analog of Fenchel-Nielsen coordinates, Goldman [77] showed that the deformation space $\mathfrak{C}(\Sigma)$ is a cell of dimension $-8\chi(\Sigma)$. (Kim [111] showed these coordinates are global Darboux coordinates for the symplectic structure, extending a result of Wolpert [174] for $\mathfrak{F}(\Sigma)$.) In his doctoral thesis, Choi showed that every structure on a closed surface *canonically* decomposes into convex structures with geodesic boundary, glued together along boundary components. Combining these two results, one identifies the deformation space *precisely* as a countable disjoint union of open $-8\chi(\Sigma)$-cells [44].

Using analytic techniques, Labourie [122] and Loftin [126], independently, described $\mathfrak{C}(\Sigma)$ as a cell in a quite different way. Associated to a convex $\mathbb{R}\mathsf{P}^2$-structure M is a natural Riemannian metric arising from representing M as a convex surface in $\mathbb{R}^3$, which is a *hyperbolic affine sphere.* The underlying conformal structure defines a point in $\mathfrak{T}(\Sigma)$ associated to the convex $\mathbb{R}\mathsf{P}^2$-manifold M. Its extrinsic geometry is described by a *holomorphic cubic differential* on the corresponding Riemann surface. In this way $\mathfrak{C}(\Sigma)$ identifies with the bundle over $\mathfrak{T}(\Sigma)$ whose fiber over a marked Riemann surface is the vector space of holomorphic cubic differentials on that Riemann surface. Loftin [127] relates the geometry of these structures to the asymptotics of this deformation space.

These results generalize in several directions. In a series of beautiful papers, Benoist [11, 12, 13, 14, 15, 16] studied convex projective structures Ω/Γ on compact manifolds. The natural *Hilbert metric* on Ω determines a (Finsler) metric on M, and if Ω is strictly convex, then this natural metric has negative curvature and Γ is a hyperbolic group. The corresponding geodesic flow is an Anosov flow, which if M admits a hyperbolic structure, is topologically conjugate to the geodesic flow of the hyperbolic metric. Furthermore, as in [43], the corresponding representations $\Gamma \longrightarrow \mathsf{PGL}(n+1,\mathbb{R})$ form a connected component of the space of representations. For compact quotients $\Omega/Gamma$, Benoist showed that the hyperbolicity of the group Γ is equivalent to the strict convexity of $\partial\Omega$. He constructed 3-dimensional examples of convex structures on 3-manifolds with incompressible tori and hyperbolic components, where $\partial\Omega$ is the closure of a disjoint countable union of triangles. In a different direction, Kapovich [110] constructed convex projective structures with $\partial\Omega$ strictly convex but Ω/Γ has no locally symmetric structure.

When G is a split real form of a complex semisimple Lie group, Hitchin [105] showed that $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ contains components homeomorphic to open cells. Specifically, these are the components containing Fuchsian representations into $\mathsf{SL}(2,\mathbb{R})$ composed with the Kostant principal representation $\mathsf{SL}(2,\mathbb{R}) \longrightarrow G$. When $G = \mathsf{SL}(3,\mathbb{R})$, then hol maps $\mathfrak{C}(\Sigma)$ diffeomorphically to Hitchin's com-

ponent (Choi-Goldman [43]). Guichard and Wienhard [100] have found interpretations of Hitchin components in $\mathsf{SL}(4,\mathbb{R})$ in terms of geometric structures. Recently [102] they have also shown that a very wide class of *Anosov representations* as defined by Labourie [120], correspond to geometric structures on *closed manifolds.* (A much different class of Anosov representations of surface groups has recently been studied by Barbot [6, 7].

The properness of the action of $\mathsf{Mod}(\Sigma)$ on $\mathfrak{F}(\Sigma)$ is generally attributed to Fricke. Many cases are known of components of deformation spaces when $\mathsf{Mod}(\Sigma)$ acts properly [94, 171, 27]. In many of these cases, these components consist of holonomy representations of uniformizable Ehresmann structures.

13. Surface Groups: Symplectic Geometry and Mapping Class Group

Clearly the classification of geometric structures in low dimensions closely interacts with the space of surface group representations. Many examples have already been discussed here. By the Ehresmann-Weil-Thurston holonomy theorem, the local geometry of $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ is the same local geometry of $\mathsf{Def}^{(G,X)}(\Sigma)$. When Σ is a compact surface, this space itself admits rich geometric structures.

Associated to an orientation on Σ and an $\mathsf{Ad}(G)$-invariant nondegenerate symmetric bilinear form $\mathbb{B}$ on the Lie algebra of G is a natural *symplectic structure* on the deformation space. (When $\partial\Sigma \neq \emptyset$, one obtains a *Poisson structure* whose symplectic leaves correspond to fixing the conjugacy classes of the holonomy along boundary components.) This extends the cup-product symplectic structure on $H^1(\Sigma,\mathbb{R})$ (when $G = \mathbb{R}$), the Kähler *form* on the Jacobian of a Riemann surface $M \approx \Sigma$, (when $G = \mathsf{U}(1)$), and the Weil-Petersson Kähler form on $\mathfrak{T}(\Sigma)$ (when $G = \mathsf{PSL}(2,\mathbb{R})$). Compare [72].

The symplectic geometry extends over the singularities of the deformation space as well. In joint work with Millson [93, 132], inspired by a letter of Deligne [49], it is shown that the germ at a reductive representation ρ, the analytic variety $\mathsf{Hom}(\pi_1(\Sigma), G)/G$ is locally equivalent to a cone defined by a system of homogeneous quadratic equations. Explicitly, this quadratic cone is defined by the cup-product

$$\mathsf{Z}^1(\Sigma, \mathfrak{g}_{\mathsf{Ad}\rho}) \times \mathsf{Z}^1(\Sigma, \mathfrak{g}_{\mathsf{Ad}\rho}) \xrightarrow{[,]_*\cup} \mathsf{H}^2\Sigma, \mathfrak{g}_{\mathsf{Ad}\rho})$$

using Weil's identification of the Zariski tangent space of $\mathsf{Hom}(\pi_1(\Sigma), G)/G$ at ρ with $\mathsf{Z}^1(\Sigma, \mathfrak{g}_{\mathsf{Ad}\rho})$. This quadratic singularity theorem extends to higher-dimensional Kähler manifolds [149] and relates to the stratified symplectic spaces considered by Sjamaar-Lerman [150].

The symplectic/Poisson geometry of the deformation spaces $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ and $\mathsf{Def}^{(G,X)}(\Sigma)$ associate vector fields to functions in the

following way (see [75]). A natural class of functions f_α on $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ arise from $\mathsf{Inn}(G)$-invariant functions $G \xrightarrow{f} \mathbb{R}$ and elements $\alpha \in \pi(\Sigma)$ by composition:

$$[\rho] \xrightarrow{f_\alpha} f\big(\rho(\alpha)\big).$$

For example, when ℓ is the geodesic length function on $\mathsf{PSL}(2,\mathbb{R})$, this construction yields the geodesic length functions ℓ_α on $\mathfrak{T}(\Sigma)$.

When α arises from a *simple closed curve* on Σ then the Hamiltonian flow associated to the vector field $\mathsf{Ham}(f_\alpha)$ admits a simple description as a *generalized twist flow.* Such a flow is "supported on α" in the sense that pulled back to the complement $\Sigma \setminus \alpha$ the flow is a trivial deformation. This extends the results of Wolpert [172, 173] for the Weil-Petersson symplectic form on $\mathfrak{T}(\Sigma)$, Fenchel-Nielsen twist flow (or *earthquake*) along α is $\mathsf{Ham}(\ell_\alpha)$. For the case of $G = \mathsf{SU}(2)$, Jeffrey and Weitsman [106] used these flows to define an "almost toric" structure on $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$ from which they deduced the Verlinde formulas.

The Poisson brackets of the functions f_α may be computed in terms of oriented intersections on Σ. For $G = \mathsf{GL}(n)$, and $f = \mathsf{tr}$, one obtains a topologically defined Lie algebra based on homotopy classes of curves on Σ with a representation in the Poisson algebra of functions on $\mathsf{Hom}\big(\pi_1(\Sigma), G\big)/G$. Turaev[167] showed this Lie algebra extends to a Lie bialgebra and found several quantizations. Recently Moira Chas [40] has discovered algebraic properties of this Lie algebra; in particular she proved that the ℓ^1 norm of a bracket $[\alpha, \beta]$ of two unoriented simple closed curves equals the geometric intersection number $i(\alpha, \beta)$.

These algebraic structures extend in higher dimensions to the *string topology* of Chas-Sullivan [41].

The symplectic geometry is $\mathsf{Mod}(\Sigma)$-invariant and in particular defines an invariant measure on the deformation space. Unlike the many cases in which $\mathsf{Mod}(\Sigma)$ acts properly discussed above, when G is compact, this measure-preserving action is ergodic on each connected component (Goldman [78], Pickrell-Xia [139], Goldman-Xia [96]). When G is noncompact, invariant open subsets of the deformation space exist where the action is proper (such as the subset of Anosov representations), but in general $\mathsf{Mod}(\Sigma)$ can act properly on open subsets containing non-discrete representations, even for $\mathsf{PSL}(2,\mathbb{R})$ ([81, 91, 156]).

Similar questions for the action of the outer automorphism group $\mathsf{Out}(\mathbb{F}_n)$ of a free group $\mathbb{F}_n$ on $\mathsf{Hom}(\mathbb{F}_n, G)/G$ have recently been studied [83]. In particular Gelander has proved that the action of $\mathsf{Out}(\mathbb{F}_n)$ is ergodic whenever G is a compact connected Lie group. For $G = \mathsf{SL}(2,\mathbb{C})$, Minsky [135] has recently found open subsets of $\mathsf{Hom}(\mathbb{F}_n, G)/G$ strictly containing the subset of Schottky embeddings for which the action is proper.

Acknowledgement

I would like to thank Virginie Charette, Son Lam Ho, Aaron Magid, Karin Melnick, and Anna Wienhard for helpful suggestions in the preparation of this manuscript.

References

[1] Abels, H., *Properly discontinuous groups of affine transformations: a survey,* Geom. Ded. **87** (2001), no. 1–3, 309–333.

[2] _____, Margulis, G., and Soifer, G., *The Auslander conjecture for groups leaving a form of signature* $(n-2,2)$ *invariant,* Probability in mathematics. Israel J. Math. **148** (2005), 11–21;

[3] _____, _____ and _____, *On the Zariski closure of the linear part of a properly discontinuous group of affine transformations,* J. Diff. Geo. **60** (2002), no. 2, 315–344.

[4] Ananin, A., Grossi, C., Gusevskii, N., *Complex Hyperbolic Structures on Disc Bundles over Surfaces I. General Settings. A Series of Examples,* arXiv:math/0511741.

[5] Ananin, A., Gusevskii, N., *Complex Hyperbolic Structures on Disc Bundles over Surfaces. II. Example of a Trivial Bundle,* arXiv:math/0512406

[6] Barbot, T., *Flag structures on Seifert manifolds,* Geom. Topol. **5** (2001), 227–266.

[7] _____, *Three-dimensional Anosov flag manifolds,* Geom. Topol. **14**, (2010), 153–191.

[8] Baues, O. and Goldman, W., *Is the deformation space of complete affine structures on the* 2*-torus smooth?,* in *Geometry and Dynamics,* J. Eels, E. Ghys, M. Lyubich, J. Palis and J. Seade (eds.), Contemp. Math. **389** (2006) 69–89 `math.DG/0401257.`

[9] Benoist, Y., *Nilvariétés projectives*, Comm. Math. Helv. 69 pp. 447–473 (1994)

[10] _____, *Une nilvariété non affine,* J. Diff. Geo. **41** (1995), no. 1, 21–52.

[11] _____, *Automorphismes des cônes convexes,* Inv. Math. **141** (2000), no. 1, 149–193;

[12] _____,*Convexes divisibles I,* in *Algebraic groups and arithmetic,* 339–374, Tata Inst. Fund. Res., Mumbai (2004).

[13] _____,*Convexes divisibles II,* Duke Math. J. **120** (2003), no. 1, 97–120.

[14] _____,*Convexes divisibles III,* Ann. Sci. Ècole Norm. Sup. (4) **38** (2005), no. 5, 793–832.

[15] _____,*Convexes divisibles IV, Structure du bord en dimension 3.* Inv. Math. **164** (2006), no. 2, 249–278.

[16] _____, *A survey on divisible convex sets,* in *Geometry, analysis and topology of discrete groups,* 1–18, Adv. Lect. Math. (ALM), **6**, Int. Press, Somerville, MA (2008).

[17] Benzécri, J. P., *Variétés localement affines,* Sem. Topologie et Géom. Diff., Ch. Ehresmann (1958–60), No. 7 (mai 1959)

[18] _____, *Sur les variétés localement affines et projectives,* Bull. Soc. Math. France **88** (1960), 229–332

[19] Bergeron, N. and Gelander, T., *A note on local rigidity.* Geom. Ded. **107** (2004), 111–131.

[20] Bers, L. and Gardiner, F., *Fricke spaces,* Adv. Math. **62** (1986), 249–284.

[21] Bonahon, F., *Geometric structures on 3-manifolds,* in *Handbook of geometric topology,* 93–164, North-Holland, Amsterdam, 2002.

[22] Bonsante, F. and Schlenker, J.-M., *AdS manifolds with particles and earthquakes on singular surfaces,* Geom. Func. Anal. **19** (2009), no. 1, 41–82.

[23] Bradlow, S., García-Prada, O. and Gothen, P., *Surface group representations and* $\mathsf{U}(p,q)$*-Higgs bundles,* J. Diff. Geo. **64** (2003), no. 1, 111–170;

[24] _____, _____, and _____, *Maximal surface group representations in isometry groups of classical Hermitian symmetric spaces,* Geom. Ded. **122** (2006), 185–213.

[25] _____, _____, and _____, *What is a Higgs bundle?* Notices Amer. Math. Soc. **54** no. 8, (2007), 980–981.

[26] Bucher, M. and Gelander, T., *Milnor-Wood inequalities for manifolds locally isometric to a product of hyperbolic planes,* C. R. Acad. Sci. Paris **346** (2008), no. 11–12, 661–666.

[27] M. Burger, A. Iozzi, Labourie, F., and Wienhard, A., *Maximal representations of surface groups: Symplectic Anosov structures,* Pure and Applied Mathematics Quarterly, Special Issue: In Memory of Armand Borel Part 2 of 3, **1** (2005), no. 3, 555–601.

[28] Burger, M., Iozzi, A. and Wienhard, A., *Surface group representations with maximal Toledo invariant*, C. R. Acad. Sci. Paris , Sér. I **336** (2003), 387–390; *Representations of surface groups with maximal Toledo invariant*, Preprint math.DG/0605656. Ann. Math. (to appear).

[29] _____, _____, and _____, *Higher Teichmüller spaces from* $\mathsf{SL}(2,\mathbb{R})$ *to other Lie groups, Handbook of Teichmüller theory, vol. III* (A. Papadopoulos, ed.), IRMA Lectures in Mathematics and Theoretical Physics European Mathematical Society (to appear).

[30] Burns, D., Jr. and Shnider, S., *Spherical hypersurfaces in complex manifolds,* Inv. Math. **33** (1976), no. 3, 223–246.

[31] Canary, R. D.; Epstein, D. B. A.; Green, P. *Notes on notes of Thurston,* in *Analytical and geometric aspects of hyperbolic space (Coventry/Durham, 1984),* 3–92, London Math. Soc. Lecture Note Ser., **111** Cambridge Univ. Press, Cambridge (1987).

[32] Carrière, Y., *Autour de la conjecture de L. Markus sur les variétés affines,* Inv. Math. **95** (1989), no. 3, 615–628.

[33] _____, Dal'bo, F. and Meigniez, G., *Inexistence de structures affines sur les fibrés de Seifert,* Math. Ann. **296** (1993), no. 4, 743–753.

[34] Charette, V., *Proper Actions of Discrete Groups on* $2+1$ *Spacetime,* Doctoral Dissertation, University of Maryland 2000.

[35] _____ and Drumm, T., *The Margulis invariant for parabolic transformations,* Proc. Amer. Math. Soc. **133** (2005), no. 8, 2439–2447.

[36] _____ and _____, *Strong marked isospectrality of affine Lorentzian groups,* J. Diff. Geo. **66** (2004), no. 3, 437–452.

[37] _____, _____ and Goldman, W. *Affine deformations of the three-holed sphere,* Geom. Topol. (to appear) arXiv:0907.0690

[38] _____,_____, _____, and Morrill, M., *Complete flat affine and Lorentzian manifolds,* Geom. Ded. **97** (2003), 187–198.

[39] Charette, V. and Goldman, W., *Affine Schottky groups and crooked tilings,* Contemp. Math. **262**, (2000), 69–98.

[40] Chas, M., *Minimal intersection of curves on surfaces,* Geom. Ded. **144** (2010), 25–60.

[41] _____ and Sullivan, D., *String Topology,* Ann. Math. (to appear)

[42] Chern, S. and Simons, J., *Characteristic forms and geometric invariants,* Ann. Math. **99** (2), (1974), 48–69.

[43] Choi, S., and Goldman, W., *Convex real projective structures on closed surfaces are closed*, Proc. Amer. Math. Soc. **118** (2) (1993), 657–661.

[44] _____, and _____, *The classification of real projective structures on compact surfaces*, Bull. Amer. Math. Soc. **34** (1997), 161–171.

[45] _____, and _____, *The deformation spaces of convex* RP^2*-structures on 2-orbifolds.* Amer. J. Math. **127** No. 5, (2005) 1019–1102

[46] Cooper, D., and Goldman, W., *A* 3*-manifold without a projective structure,* (in preparation).

[47] Corlette, K., *Flat G-bundles with canonical metrics*, J. Diff. Geo. **28** (1988), 361–382.

[48] Dekimpe, K., *Any virtually polycyclic group admits a NIL-affine crystallographic action,* Topology **42** (2003), no. 4, 821–832.

[49] Deligne, P., letter to W. Goldman and J. Millson (1986)

[50] Donaldson, S., *Twisted harmonic maps and the self-duality equations*, Proc. London Math. Soc. (3) **55** (1987), no. 1, 127–131.

[51] Drumm, T., *Fundamental polyhedra for Margulis spacetimes,* Doctoral Dissertation, University of Maryland 1990; Topology **31** (4) (1992), 677–683;

[52] _____, *Linear holonomy of Margulis space-times,* J. Diff. Geo. **38** (1993), 679–691.

[53] Dumas, D., *Complex Projective Structures,* Chapter 12, pp. 455–508, of "Handbook of Teichmüller theory, vol. II", (A. Papadopoulos, ed.), IRMA Lectures in Mathematics and Theoretical Physics volume 13, European Mathematical Society (2008).

[54] Eels, J., and Wood, J. C., *Restrictions on harmonic maps of surfaces,* Topology **15** (1976), no. 3, 263–266.

[55] Ehresmann, C., *Sur les espaces localement homogenes,* L'ens. Math. **35** (1936), 317–333

[56] _____, *Les connexions infinitésimales dans un espace fibré différentiable,* in "Colloque de topologie (espaces fibrés)," Bruxelles, (1950), 29–55.

[57] Falbel, E., *A spherical CR structure on the complement of the figure eight knot with discrete holonomy,* J. Diff. Geo. **79** (2008), no. 1, 69–110.

[58] Florentino, C. and Lawton, S., *The topology of moduli spaces of free group representations,* Math. Ann. **345** (2009), no. 2, 453–489.

[59] Fock, V. and Goncharov, A., *Moduli spaces of convex projective structures on surfaces*, Adv. Math. **208** (2007), no. 1, 249–273 math.DG/0405348, 2004;

[60] _____ and _____, *Moduli spaces of local systems and higher Teichmüller theory*, Publ. Math. d'I.H.E.S. **103** (2006), 1–211 math.AG/0311149, 2003. Math. Rev. 2233852

[61] Frances, C., *Lorentzian Kleinian groups,* Comm. Math. Helv. **80** (2005), no. 4, 883–910.

[62] Fricke, R. and Klein, F. , *Vorlesungen der Automorphen Funktionen,* Teubner, Leipzig, Vol. I (1897), Vol. II (1912)

[63] Fried, D., *Closed similarity manifolds,* Comm. Math. Helv. **55** (1980), 576–582

[64] _____ and Goldman, W. , *Three-dimensional affine crystallographic groups,* Adv. Math. **47** (1983), 1–49

[65] Gallo, D., Kapovich, M. and Marden, A. *The monodromy groups of Schwarzian equations on closed Riemann surfaces,* Ann. Math. (2) **151** (2000), no. 2, 625–704.

[66] Gelander, T., *On deformations of $\mathbb{F}_n$ in compact Lie groups,* Israel J. Math. **167** (2008), 15–26.

[67] Ghys, É., *Rigidité différentiable des groupes fuchsiens,* Publ. Math. d'I.H.E.S. **78** (1993), 163–185 (1994).

[68] Goldman, W., *Affine manifolds and projective geometry on surfaces,* Senior thesis, Princeton University (1977)

[69] _____, *Discontinuous groups and the Euler class,* Doctoral thesis, University of California, Berkeley (1980)

[70] _____, *Characteristic classes and representations of discrete subgroups of Lie groups,* Bull. Amer. Math. Soc. **6** (1982), 91–94.

[71] _____, *Conformally flat 3-manifolds with nilpotent holonomy and the uniformization problem for 3-manifolds,* Trans. Amer. Math. Soc. **278** (1983), 573–583.

[72] _____, *The symplectic nature of fundamental groups of surfaces,* Adv. Math. **54** (1984), 200–225.

[73] _____, *Projective structures with Fuchsian holonomy,* J. Diff. Geo. **25** (1987), 297–326

[74] _____, *Geometric structures and varieties of representations,* Contemp. Math. , Amer. Math. Soc. **74**, (1988), 169–198.

[75] _____, *Invariant functions on Lie groups and Hamiltonian flows of surface group representations,* Inv. Math. **85** (1986), 1–40.

[76] _____, *Topological components of spaces of representations*, Inv. Math. **93** (1988), no. 3, 557–607.

[77] _____, *Convex real projective structures on compact surfaces*, J. Diff. Geo. **31** (1990), no. 3, 791–845.

[78] _____, *Ergodic theory on moduli spaces,* Ann. Math. (2) **146** (1997), no. 3, 475–507.

[79] _____, *Complex hyperbolic geometry,* Oxford Mathematical Monographs. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1999. xx+316 pp. ISBN: 0-19-853793-X

[80] _____, *The Margulis Invariant of Isometric Actions on Minkowski (2+1)-Space,* in "Ergodic Theory, Geometric Rigidity and Number Theory", Springer-Verlag (2002), 149–164.

[81] _____, *Action of the modular group on real SL(2)-characters of a one-holed torus,* Geom. Topol. **7** (2003), 443–486.

[82] _____, *Mapping Class Group Dynamics on Surface Group Representations,* in "Problems on mapping class groups and related topics", (B. Farb, ed.) Proc. Sympos. Pure Math. **74**, Amer. Math. Soc. , Providence, RI, 2006, 189–214.. math.GT/0509114.

[83] _____, *An ergodic action of the outer automorphism group of a free group,* Geom. Func. Anal. **17–3** (2007), 793–805.

[84] _____, *Higgs bundles and geometric structures on surfaces,* to appear in "The Many Facets of Geometry: a Tribute to Nigel Hitchin", J.P. Bourgignon, O. Garcia-Prada & S. Salamon (eds.), Oxford University Press (to appear). `math.DG.0805.1793`

[85] _____ and Hirsch, M., *The radiance obstruction and parallel forms on affine manifolds,* Trans. Amer. Math. Soc. **286** (1984), 639–649

[86] _____, and _____, *Affine manifolds and orbits of algebraic groups,* Trans. Amer. Math. Soc. **295** (1986), 175–198

[87] _____ and Kamishima, Y., *The fundamental group of a compact flat Lorentz space form is virtually polycyclic,* J. Diff. Geo. **19** (1984), 233–240

[88] _____, Kapovich, M., and Leeb, B., *Complex hyperbolic manifolds homotopy equivalent to a Riemann surface,* Comm. Anal. Geom. **9** (2001), no. 1, 61–95.

[89] _____, Labourie, F., and Margulis, G., *Proper affine actions and geodesic flows on hyperbolic surfaces,* Ann. Math. **70** (2009), No. 3, 10511083. `math.DG/0406247`.

[90] _____ and Margulis, G., *Flat Lorentz 3-manifolds and cocompact Fuchsian groups,* Contemp. Math. **262**, (2000), 135–146. `math.DG/0005292.`

[91] _____, McShane, G., Stantchev, G., and Tan, S. P., *Dynamics of* $\mathsf{Out}(\mathbb{F}_2)$ *acting on isometric actions on the hyperbolic plane,* (in preparation).

[92] _____ and J. J. Millson, *Local rigidity of discrete groups acting on complex hyperbolic space,* Inv. Math. **88** (1987), 495–520.

[93] _____ and _____, The deformation theory of representations of fundamental groups of Kähler manifolds, Publ. Math. d'I.H.E.S. **67** (1988), 43–96.

[94] _____ and Wentworth, R., *Energy of twisted harmonic maps of Riemann surfaces,* in *In the tradition of Ahlfors-Bers. IV,* Proceedings of the Ahlfors-Bers Colloquium, University of Michigan,' H. Masur and R. Canary, eds., 45–61, Contemp. Math. **432** Amer. Math. Soc. , Providence, RI (2007), `math.DG/0506212`.

[95] _____ and Xia, E., *Rank one Higgs bundles and fundamental groups of Riemann surfaces,* Mem. Amer. Math. Soc. **904** (2008) `math.DG/0402429`.

[96] _____ and _____, *Ergodicity of mapping class group action on* $\mathsf{SU}(2)$*-character varieties,* (to appear in "Geometry, Rigidity, and Group Actions", Univ. of Chicago Press.)

[97] Gromov, M. , Lawson, H. B., Jr. and Thurston, W. , *Hyperbolic* 4*-manifolds and conformally flat* 3*-manifolds,* Publ. Math. d'I.H.E.S. **68** (1988), 27–45.

[98] Guichard, O., *Une dualité pour les courbes hyperconvexes*, Geom. Ded. **112** (2005), 141–164. Math. Rev. MR2163895

[99] _____, *Composantes de Hitchin et reprsentations hyperconvexes de groupes de surface*, J. Diff. Geo. **80** (3) (2008), 391–431.

[100] _____ and Wienhard, A., *Convex foliated projective structures and the Hitchin component for* $\mathrm{PSL}_4(R)$*,* Duke Math. J. **144** (3) (2008), 381–445.

[101] _____ and _____, *Topological Invariants for Anosov Representations,* J. Top. (to appear) `math/AT:0907.0273`.

[102] _____ and _____, *Domains of discontinuity for surface groups,* C. R. Acad. Sci. Paris **347** (17–18) (2009), 1057–1060.

[103] Hejhal, D., *Monodromy groups and linearly polymorphic functions,* Acta Math.**135** (1) (1975), 1–55.

[104] Hitchin, N., *The self-duality equations on Riemann surfaces,* Proc. Lond. Math. Soc. **55** (1987), 59–126.

[105] _____, *Lie groups and Teichmüller space*, Topology **31** (3) (1992), 449–473.

[106] Jeffrey, L. and Weitsman, J., *Bohr-Sommerfeld orbits in the moduli space of flat connections and the Verlinde dimension formula,* Comm. Math. Phys. **150** (3) (1992), 593–630.

[107] Kac, V., and Vinberg, E. B., *Quasi-homogeneous cones,* Math. Notes **1** (1967), 231–235,

[108] Kamishima, Y. and Tan, S. P., *Deformation spaces of geometric structures,* in *Aspects of low-dimensional manifolds,* 263–299, Adv. Stud. Pure Math. **20**, Kinokuniya, Tokyo (1992).

[109] Kapovich, M., *Flat conformal structures on three-dimensional manifolds: the existence problem. I.,* Siberian Math. J. **30** (1989), no. 5, 712–722 (1990), *Flat conformal structures on* 3*-manifolds. I. Uniformization of closed Seifert manifolds,* J. Diff. Geo. **38** (1) (1993), 191–215.

[110] _____, *Convex projective structures on Gromov-Thurston manifolds,* Geom. Topol. **11** (2007), 1777–1830.

[111] Kim, Hong Chan, *The symplectic structure on the moduli space of real projective structures* Doctoral Dissertation, University of Maryland (1999); *The symplectic global coordinates on the moduli space of real projective structures,* J. Diff. Geo. **53** (2) (1999), 359–401.

[112] Klingler, B., *Complétude des variétés lorentziennes à courbure constante,* Math. Ann. **306** (2) (1996), 353–370.

[113] Kneser, H., *Die kleinste Bedeckungszahl innerhalb einer Klasse von Flchenabbildungen,* Math. Ann. **103** (1) (1930), 347–358.

[114] Kostant, B. and Sullivan, D., *The Euler characteristic of a compact affine space form is zero,* Bull. Amer. Math. Soc. **81** (1975)

[115] Kuiper, N., *Sur les surfaces localement affines,* Colloque Int. Géom. Diff., Strasbourg, CNRS (1953), 79–87

[116] _____, *On convex locally projective spaces,* Convegno Int. Geometria Diff., Italy (1954), 200–213

[117] Kulkarni, R., *The principle of uniformization,* J. Diff. Geo. **13** (1978), 109–138

[118] _____ and Pinkall, U., *A canonical metric for Möbius structures and its applications,* Math. Z. **216** (1) (1994), 89–129.

[119] Labourie, F., *Fuchsian affine actions of surface groups,* J. Diff. Geo. **59** (1) (2001), 15–31.

[120] _____, *Anosov flows, surface group representations and curves in projective space,* Inv. Math. **165** (1) (2006), 51–114.

[121] _____, *Cross ratios, Anosov representations and the energy functional on Teichmüller space,* Annales Sci. de l'E. N. S. (4) **41** (3) (2008), 437–469.

[122] _____, *Flat Projective Structures on Surfaces and Cubic Holomorphic Differentials,* Pure Appl. Math. Q. **3** (2007), no. 4, part 1, 1057–1099.

[123] _____, *What is a ... cross ratio?* Notices Amer. Math. Soc. **55** (10) (2008), 1234–1235.

[124] Li, J., *Spaces of surface group representations,* Manuscripta Math. **78** (3), (1993), 223–243.

[125] Lok, W. L., *Deformations of locally homogeneous spaces and Kleinian groups,* Doctoral Thesis, Columbia University (1984).

[126] Loftin, J., *Affine spheres and convex* $\mathbb{R}\mathsf{P}^n$*-manifolds,* Amer. J. Math. **123** (2001), 255–274;

[127] _____, *The compactification of the moduli space of convex* $\mathbb{RP}^2$*-surfaces I,* J. Diff. Geo. **68** (2) (2004), 223–276.

[128] Margulis, G. A., *Free properly discontinuous groups of affine transformations,* Dokl. Akad. Nauk SSSR **272** (1983), 937–940; *Complete affine locally flat manifolds with a free fundamental group,* J. Soviet Math. **134** (1987), 129–134.

[129] Matsumoto, S., *Some remarks on foliated* S^1 *bundles,* Inv. Math. **90** (2) (1987), 343–358.

[130] _____, *Topological aspects of conformally flat manifolds,* Proc. Japan Acad. Ser. A Math. Sci. **65** (7) (1989), 231–234.

[131] Mess, G., *Lorentz spacetimes of constant curvature,* Geom. Ded. **126** (2007), 3–45.

[132] Millson, J., *Rational homotopy theory and deformation problems from algebraic geometry,* Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990), 549–558, Math. Soc. Japan, Tokyo, 1991.

[133] Milnor, J., *On the existence of a connection with curvature zero,* Comm. Math. Helv. **32** (1958), 215-223

[134] _____, *On fundamental groups of complete affinely flat manifolds,* Adv. Math. **25** (1977), 178–187

[135] Minsky, Y., *On dynamics of* $\mathsf{Out}(\mathbb{F}_n)$ *on* $\mathsf{PSL}(2,\mathbb{C})$*-characters,* (2009) `arXiv:0906.3491`.

[136] Molnar, E., *The projective interpretation of the eight 3-dimensional homogeneous geometries*, Beitrage zur Algebra und Geometrie **38** (1997), p. 262-288..

[137] Parker, J., *Complex hyperbolic lattices,* in *Discrete Groups and Geometric Structures,* Contemp. Math. **501** Amer. Math. Soc. , (2009) 1–42.

[138] _____ and Platis, I., *Complex hyperbolic Fenchel-Nielsen coordinates,* Topology **47** (2008), no. 2, 101–135.

[139] Pickrell, D. and Xia, E., *Ergodicity of mapping class group actions on representation varieties. I. Closed surfaces,* Comm. Math. Helv. **77** (2) (2002), 339–362.

[140] _____ and _____, *Ergodicity of mapping class group actions on representation varieties. II. Surfaces with boundary,* Transform. Groups **8** (4) (2003), 397–402.

[141] Rapinchuk, A., Benyash-Krivetz, V., Chernousov, V., *Representation varieties of the fundamental groups of compact orientable surfaces,* Israel J. Math. **93** (1996), 29–71.

[142] Schoen, R. and Yau, S. T., *Univalent harmonic maps between surfaces,* Inv. Math. **44** (1978), 265–278.

[143] _____ and _____, *Conformally flat manifolds, Kleinian groups and scalar curvature*, Inv. Math. **92** (1988), no. 1, 47–71

[144] Schwartz, R., *Complex hyperbolic triangle groups,* Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002), 339–349, Higher Ed. Press, Beijing, 2002.

[145] _____, *Spherical CR geometry and Dehn surgery,* Annals of Mathematics Studies, **165**. Princeton University Press, Princeton, NJ, 2007.

[146] _____, *A better proof of the Goldman-Parker conjecture,* Geom. Topol. **9** (2005), 1539–1601.

[147] Scott, G. P., *The geometries of 3-manifolds,* Bull. Lond. Math. Soc. **15** (1983), 401–487

[148] Sharpe, R., *Differential Geometry: Cartan's Generalization of Klein's Erlangen Program,* Graduate Texts in Mathematics **166**. Springer-Verlag, New York (1997).

[149] Simpson, C., *Higgs bundles and local systems,* Publ. Math. d'I.H.E.S. **75** (1992), 5–95.

[150] Sjamaar, R. and Lerman, E., *Stratified symplectic spaces and reduction,* Ann. Math. (2) **134** (2) (1991), 375–422.

[151] Smillie, J., *Flat manifolds with nonzero Euler characteristic,* Comm. Math. Helv. **52** (1977), 453–456

[152] _____, *Affinely flat manifolds,* Doctoral dissertation, University of Chicago (1977)

[153] _____, *An obstruction to the existence of affinely flat structures,* Inv. Math. **64** (1981), 411–415

[154] Steenrod, N. E., *The topology of fibre bundles,* Princeton University Press (1951)

[155] Sullivan, D. and Thurston, W., *Manifolds with canonical coordinates: Some examples,* L'ens. Math. **29** (1983), 15–25.

[156] Tan, S. P., Wong, Y. L., Zhang, Y., *Generalized Markoff maps and McShane's identity,* Adv. Math. **217** (2) (2008), 761–813.

[157] Thiel, B., *Einheitliche Beschreibung der acht Thurstonschen Geometrien*, Diplomarbeit, Universitat zu Gottingen, 1997.

[158] Thurston, W., *The geometry and topology of 3-manifolds,* Princeton University lecture notes (1979) (unpublished, available from: http://www.msri.org/communications/books/gt3m);

[159] _____, *Hyperbolic geometry, three-dimensional manifolds and Kleinian groups,* Bull. Amer. Math. Soc. (New Series) **6** (1982), 357–381

[160] _____, *Three-dimensional geometry and topology, Vol. 1,* Edited by Silvio Levy. Princeton Mathematical Series **35**. Princeton University Press, Princeton, NJ, (1997).

[161] _____, *Minimal stretch maps between hyperbolic surfaces,* `math.GT/9801039`.

[162] Tits, J., *Free subgroups in linear groups,* J. Algebra **20** (1972), 250–270.

[163] Toledo, D., *Representations of surface groups in complex hyperbolic space*, J. Diff. Geo. **29** (1) (1989), 125–133.

[164] Tomanov, G., *The virtual solvability of the fundamental group of a generalized Lorentz space form,* J. Diff. Geo. **32** (2) (1990), 539–547.

[165] Turaev, V. G., *A cocycle of the symplectic rst Chern class and Maslov indices,* Funktsional. Anal. i Prilozhen. **18** (1) (1984), 43–48;

[166] _____, *The first symplectic Chern class and Maslov indices,* Studies in topology, V. Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **143** (1985), 110D129, 178.

[167] _____, *Skein quantization of Poisson algebras of loops on surfaces,* Ann. Sci. École Norm. Sup. (4) **24** (6) (1991), 635–704.

[168] Weil, A. , *On discrete subgroups of Lie groups I,* Ann. Math. **72** (1960), 369–384.

[169] _____, *On discrete subgroups of Lie groups II,* **75** (1962), 578–602.

[170] _____, *Remarks on the cohomology of groups,* Ann. Math. (2) **80** (1964), 149–157.

[171] Wienhard, A., *The action of the mapping class group on maximal representations*, Geom. Ded. **120** (2006), 179–191.

[172] Wolpert, S. , *The Fenchel-Nielsen deformation,* Ann. Math. (2) **115** (3) (1982), 501–528.

[173] ———, *The symplectic geometry of deformations of a hyperbolic surface,* Ann. Math. **117** (1983), 207–234

[174] ———, *On the Weil Petersson geometry of the moduli space of curves,* Amer. J. Math. **107** (4) (1985), 969–997.

[175] Wood, J., *Bundles with totally disconnected structure group,* Comm. Math. Helv. **51** (1971), 183–199.

[176] Zeghib, A., *On closed anti-de Sitter spacetimes,* Math. Ann. **310** (4) (1998), 695–716.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Metaphors in Systolic Geometry

Larry Guth*

Abstract

We discuss the systolic inequality for n-dimensional tori, explaining different metaphors that help to organize the proof. The metaphors connect systolic geometry with minimal surface theory, topological dimension theory, and scalar curvature.

Mathematics Subject Classification (2010). Primary 53C23

Keywords. Systole, filling radius, isoperimetric inequality.

1. Introduction

This essay is an introduction to systolic geometry. Rather than surveying a lot of results, I'm going to focus on one central result, and I want to survey a lot of ways of thinking about it.

Systolic inequality for tori. *(Gromov, 1983 [10]) If (T^n, g) is an n-dimensional torus with a Riemannian metric, then there is a non-contractible curve $\gamma \subset T^n$ whose length obeys the inequality*

$$length(\gamma) \leq C_n \, Vol(T^n, g)^{1/n}.$$

This inequality is very general. It holds in every dimension n, and it holds for every metric g on T^n. (For example, there is no restriction on the curvature of g.) This result is difficult and significant because it applies to so many metrics.

In the early 80's, Gromov formulated several remarkable metaphors connecting the systolic inequality to important ideas in other areas of geometry, and these metaphors have guided most of the research in the subject. They connect the systolic problem with ideas about minimal surfaces, topological dimension, and scalar curvature. The main goal of this essay is to explain Gromov's metaphors.

*Mathematics department, University of Toronto, 40 St. George St., Toronto ON M5S 2E4, Canada. E-mail: lguth@math.toronto.edu.

The systole of (T^n, g) is defined to be the length of the shortest non-contractible curve in (T^n, g). We will denote it by $Sys(T^n, g)$. The systole of (T^n, g) and the volume of (T^n, g) are both ways of describing the *size* of (T^n, g). Size may sound like a basic issue in Riemannian geometry, but mathematicians have not spent much time exploring it. The proofs of the systolic inequality lead to some interesting perspectives about size in Riemannian geometry. At the end of the essay, I will discuss the issue of size and point out some open problems.

Acknowledgements. I would like to thank Hugo Parlier for the figure in Section 2, and Alex Nabutovsky for helpful comments on a draft of the essay.

2. Examples

To get a feeling for the systolic inequality, let's consider some examples.

First, suppose that (T^n, g) is a product of circles with lengths $L_1, \ldots, L_n$. The length of the shortest non-contractible curve in this metric is $\min_{i=1}^n L_i$, and the volume of the metric is $\prod_{i=1}^n L_i$. Hence we see that for product metrics, there is a non-contractible curve of length at most $Vol^{1/n}$.

Next let's consider some examples of two dimensional tori that we can visualize. The systolic inequality for two-dimensional tori was proven by Loewner in 1949 with a sharp constant.

Loewner's systolic inequality. *(1949) If (T^2, g) is a 2-dimensional torus with a Riemannian metric, then there is a non-contractible curve $\gamma \subset (T^2, g)$ whose length obeys the inequality*

$$length(\gamma) \leq C Area(T^2, g)^{1/2},$$

where $C = 2^{1/2} 3^{-1/4} \sim 1.1$.

The diagram below shows four different tori.
The first picture is supposed to show a torus of revolution, where we take the circle of radius 1 around the point $(2, 0)$ in the x-z plane and revolve it around the z-axis. It has systole 2π and area around 60, and so it obeys the systolic inequality. According to Loewner's theorem, there is nothing we can do to dramatically increase the systole while keeping the area the same. The second picture shows a long skinny torus. When we make the torus skinnier and longer, the systole goes down and the area stays about the same. The third picture shows a torus with a long thin spike coming out of it. When we add a long thin spike to the torus, the systole doesn't change and the spike adds to the area. The fourth picture shows a ridged torus with some thick parts and some thin parts. When we put ridges in the surface of the torus, the systole only depends on the thinnest part and the thick parts contribute heavily to the area.

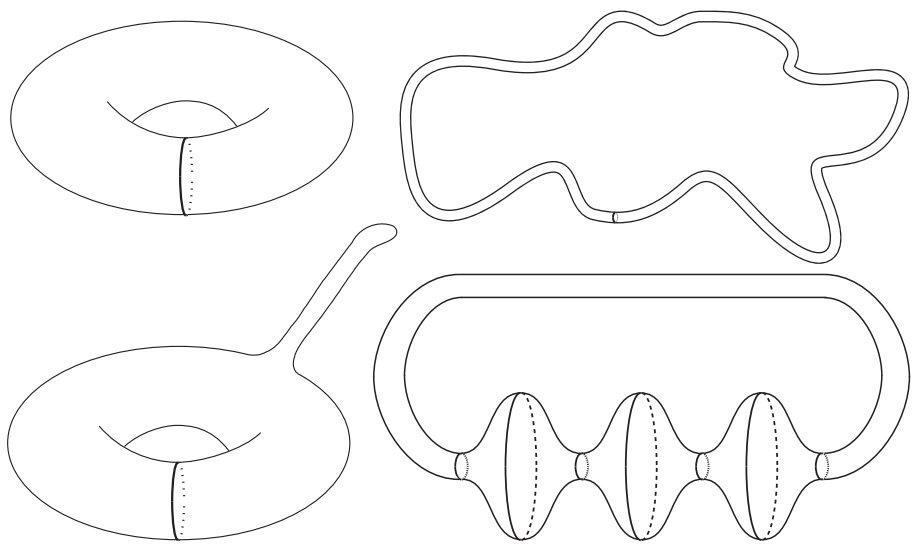


Figure 1. Pictures of tori

(Friendly challenge to the reader: can you think of a torus with geometry radically different from the pictures above?)

These pictures start to give a feel for the systolic inequality in two dimensions. In three dimensions it gets much harder to draw pictures. In fact, in three dimensions, there are examples of metrics much stranger than these. We touch on them more in the next section.

3. Why Is the Systolic Inequality Hard?

The systolic inequality has the same flavor as the isoperimetric inequality. To get a sense of the difficulty of the systolic inequality, let's recall the classical isoperimetric inequality and then compare them.

Isoperimetric inequality. *Suppose that $U \subset \mathbb{R}^n$ is a bounded open set. Then the volume of the boundary ∂U and the volume of U are related by the formula*

$$Vol_n(U) \leq C_n \, Vol_{n-1}(\partial U)^{\frac{n}{n-1}}.$$

From the Riemannian point of view, this domain U is a compact manifold with boundary equipped with a *flat* Riemannian metric (the Euclidean metric). The isoperimetric inequality can be considered as a theorem about flat Riemannian metrics. By contrast, the systolic inequality is a theorem about *all* Riemannian metrics on T^n. (To make the comparison tighter, the classical isoperimetric inequality holds for every *flat* metric on the n-ball. The systolic

inequality does not make sense on a ball, but we will meet below a covering inequality that holds for *every* metric on the n-ball.) Now the set of flat metrics is only a tiny sliver in the set of all metrics. Moreover, the flat metrics are probably the easiest metrics to understand. So we see that the systolic inequality is far more general than the classical isoperimetric inequality.

Loewner proved the systolic inequality for two-dimensional tori in 1949, but the three-dimensional case was open for more than thirty years after Loewner's proof. Why is three dimensions so much harder than two? The space of Riemannian metrics has many strange examples, disproving naive conjectures, and this is especially true in dimensions three and higher. For example, let us consider the following problem, raised by Berger and Gromov. Suppose that g is a metric on $S^n \times S^n$ with volume 1. Can we find a non-trivial copy of S^n with controlled n-dimensional volume? When $n = 1$, this is the systolic inequality for T^2. By analogy, it seems plausible that it should hold for all n, but it turns out that there are counterexamples for $n \geq 2$.

Gromov-Katz examples. *([28]) For each $n \geq 2$, and every number B, there is a metric on $S^n \times S^n$ with (2n-dimensional) volume 1, so that every non-contractible n-sphere in $S^n \times S^n$ has (n-dimensional) volume at least B.*

As we go from domains in Euclidean space to metrics on T^2 to metrics on T^3, the possible geometries become more complicated. To get a perspective on this, let me describe a naive conjecture about the sizes of level sets and trace how it plays out in the different settings.

Naive conjecture 1. *If $U \subset \mathbb{R}^n$ is a bounded open set, then there is a function $f : U \to \mathbb{R}$ so that the volume of every level set is controlled by the volume of U:*

$$\text{For every } y \in \mathbb{R},\ Vol_{n-1}[f^{-1}(y)] \leq C_n\, Vol_n(U)^{\frac{n-1}{n}}.$$

Naive conjecture 1 is true. I proved it in [18].

Naive conjecture 2. *If g is a metric on T^2, then there is a function $f : T^2 \to \mathbb{R}$ so that the length of every level set is controlled by the area of g:*

$$\text{For every } y \in \mathbb{R},\ Length[f^{-1}(y)] \leq C\, Area(T^2, g)^{1/2}.$$

Naive conjecture 2 is also true. This result is more surprising than the first one. The problem was open for a long time. It was proven by Balacheff and Sabourau in [5].

Naive conjecture 3. *If g is a metric on T^3, then there is a function $f : T^3 \to \mathbb{R}$ so that the area of every level set is controlled by the volume of (T^3, g):*

$$\text{For every } y \in \mathbb{R},\ Area[f^{-1}(y)] \leq C\, Vol(T^3, g)^{2/3}.$$

Naive conjecture 3 is wrong. (There are many counterexamples. I think that historically the first examples came from work of Brooks.)

This story is typical for naive conjectures in metric geometry. The space of all the metrics on T^3 is huge. There is a substantial zoo of strange examples, and there are probably many other strange metrics yet to be discovered. Universal statements about all metrics on T^3 are rare and significant.

4. The Role of Metaphors in Systolic Geometry

Reminiscing about his work in systolic geometry, Gromov wrote, "Since the setting was so plain and transparent, I expected rather straightforward proofs." (See the end of Chapter 4 in [11] for Gromov's recollections of working on the systolic problem.) But in spite of the plain and transparent setting, the result is difficult, and in particular, it's hard to see how to get started. In the early 1980's, he formulated several remarkable metaphors connecting the systolic inequality to important ideas in other areas of geometry. Guided by these metaphors, he proved the systolic inequality. We now have three independent proofs of the systolic inequality for the n-dimensional torus, each based on a different metaphor.

The goal of this essay is to explain Gromov's metaphors. In doing that, I hope to describe the flavor of this branch of geometry and put it into a broad context. The metaphors connect the systolic inequality to the following areas:

1. Minimal surface theory.
2. Topological dimension theory.
3. Scalar curvature.

Each metaphor gives a valuable perspective about the systolic problem and suggests an outline of the proof. It still takes substantial work to fill in the details of the proofs. Up to the present, every proof of the systolic inequality is based on one of these metaphors.

5. Minimal Surface Theory

In the early 1970's, Bombieri and Simon [6] proved the following sharp inequality about the geometry of minimal surfaces in Euclidean space.

Bombieri-Simon radius inequality. *Suppose that Z^n is a closed submanifold of $\mathbb{R}^N$, and that Y^{n+1} is a minimal surface with $\partial Y = Z$. Suppose that Z has the same volume as a round n-sphere of radius R. Then for each point $y \in Y$, the distance from y to Z is at most R.*

This inequality is sharp when Z is a round sphere of radius R and Y is the corresponding ball of radius R.

Using this inequality, Bombieri and Simon proved the Gehring link conjecture. If Z^n and W^{N-n-1} are disjoint closed surfaces in $\mathbb{R}^N$, then the linking number of Z with W is defined as follows. Let Y^{n+1} be a surface with $\partial Y = Z$. Put Y in general position, and consider $Y \cap W$, which will be a finite set of points. If we count these points with multiplicity we get the linking number of Z with W. This linking number doesn't depend on the choice of Y. If the number is non-zero, we say that Z and W are linked.

Gehring link conjecture. *Suppose that Z^n and W^{N-n-1} are linked submanifolds of $\mathbb{R}^N$. If Z has the same volume as a round n-sphere of radius R, then the distance from Z to W is at most R. In other words, there are points $z \in Z$ and $w \in W$ with $|z - w| \leq R$.*

Proof. By the solution of the Plateau problem, there is a minimal surface Y with $\partial Y = Z$. Since Z and W are linked, Y must intersect W in some point $y \in W$. But by the radius inequality, the distance from y to Z is at most R. $\square$

Gromov built an analogy between the Gehring link conjecture and the systolic problem. On the one hand, such an analogy sounds promising because both inequalities bound a 1-dimensional length (or distance) in terms of an n-dimensional volume.

$$Dist(Z^n, W^{N-n-1}) \leq C_n \, Vol(Z)^{1/n}. \qquad \text{(Gehring link inequality)}$$

$$Sys(T^n, g) \leq C_n \, Vol(T^n, g)^{1/n}. \qquad \text{(Systolic inequality)}$$

On the other hand, the analogy sounds far-fetched because the systolic problem is about an abstract Riemannian manifold, and the Gehring link conjecture is about a submanifold of Euclidean space $\mathbb{R}^N$.

Every closed Riemannian manifold admits a canonical embedding into a Banach space.

Kuratowski embedding. *Define the map $K : (M^n, g) \to L^\infty(M)$ by letting $K(p)$ be the distance function $dist_p$. The map K is an isometry in the strong sense that*

$$dist_{(M,g)}(p, q) = \|K(p) - K(q)\|_{L^\infty}.$$

The Kuratowski embedding is canonical and respects the geometry of (M, g). The target space $L^\infty(M)$ is infinite-dimensional, but we can approximate this embedding using a finite-dimensional Banach space. For each (M, g) there is a finite dimension N and an embedding $K_0 : (M, g) \to (\mathbb{R}^N, l^\infty)$ which is nearly isometric in the sense that

$$\frac{99}{100}\|K_0(p) - K_0(q)\|_{l^\infty} \leq dist_{(M,g)}(p, q) \leq \frac{100}{99}\|K_0(p) - K_0(q)\|_{l^\infty}.$$

The following striking observation relates the systole problem and the linking problem.

Linking observation. *([10]) Let (T^n, g) be any Riemannian metric on T^n. Let Z^n be the image $K_0(T^n) \subset (\mathbb{R}^N, l^\infty)$. Then Z is linked with a surface W^{N-n-1} with $dist(Z, W) \geq (1/8)Sys(T^n, g)$.*

We know that Z is linked with a faraway surface W, and we wish to conclude that Z has a large volume. This is a version of the Gehring link problem in $(\mathbb{R}^N, l^\infty)$.

Metaphor 1. *The systolic inequality is like the Gehring link problem in the Banach space $(\mathbb{R}^N, l^\infty)$.*

The method of Bombieri-Simon does not work in Banach spaces. In effect, their method uses the symmetry of Euclidean space. To get estimates for linked surfaces in $(\mathbb{R}^N, l^\infty)$, Gromov proved the following inequality.

Filling radius inequality. *([10]) If $Z^n \subset (\mathbb{R}^N, l^\infty)$ is a closed surface, then there exists a surface Y^{n+1} with $\partial Y = Z$ such that for each $y \in Y$,*

$$dist(y, Z) \leq C_n \, Vol_n(Z)^{1/n}.$$

The filling radius inequality implies a linking inequality in $(\mathbb{R}^N, l^\infty)$: if Z^n and W^{N-n-1} are linked in $(\mathbb{R}^N, l^\infty)$, then $dist(Z, W) \leq C_n \, Vol(Z)^{1/n}$. To prove the systolic inequality, we let $Z = K_0(T^n, g)$ and we let W be the surface mentioned in the linking observation above. Then we observe that

$$(1/8)Sys(T^n, g) \leq dist(Z, W) \leq C_n \, Vol(Z)^{1/n} \sim C_n \, Vol(T^n, g)^{1/n}.$$

There is an important story about the constant C_n in Gromov's filling radius inequality. It's comparatively easy to prove an inequality of the form $dist(y, Z) \leq C_N \, Vol_n(Z)^{1/n}$ with a constant C_N depending on the ambient dimension N. This inequality does not imply the systolic inequality. We can find a nearly isometric embedding from (T^n, g) into some $(\mathbb{R}^N, l^\infty)$, but the dimension N depends on the metric g. Roughly speaking, if g is complicated, then N will be large. To prove the systolic inequality for all g, we need a filling radius estimate for all N with a uniform constant. We discuss this issue more in Section 8 below.

(A note on vocabulary: I've been using the word surface a little bit loosely. For readers with background in geometric measure theory, surface means Lipschitz chain and closed surface means Lipschitz cycle. For readers with less background, surfaces (or Lipschitz chains) include smooth submanifolds and they are a little bit more general. A surface is a submanifold with mild singularities. For example, suppose that Z is a submanifold diffeomorphic to $\mathbb{CP}^2$. By the cobordism theory, $\mathbb{CP}^2$ is not the boundary of any 5-dimensional manifold. In this case, Y may be homeomorphic to a cone over $\mathbb{CP}^2$, which is a manifold except for one singularity at the cone point.)

6. Topological Dimension Theory

In the 1870's, Cantor discovered that $\mathbb{R}^q$ and $\mathbb{R}^n$ have the same cardinality even if $q < n$. This discovery surprised and disturbed him. He and Dedekind formulated the question whether $\mathbb{R}^q$ and $\mathbb{R}^n$ are homeomorphic for $q < n$. This question turned out to be quite difficult. It was settled by Brouwer in 1909.

Topological Invariance of Dimension. *(Brouwer 1909) If $q < n$, then there is no homeomorphism from $\mathbb{R}^n$ to $\mathbb{R}^q$.*

Cantor and Dedekind certainly knew that $\mathbb{R}^q$ and $\mathbb{R}^n$ were not *linearly* isomorphic. Linear algebra gives us two stronger statements:

Linear algebra lemma 1. *If $q < n$, then there is no surjective linear map from $\mathbb{R}^q$ to $\mathbb{R}^n$.*

Linear algebra lemma 2. *If $q < n$, then there is no injective linear map from $\mathbb{R}^n$ to $\mathbb{R}^q$.*

It seems reasonable to try to prove topological invariance of dimension by generalizing these lemmas. A priori, it's not clear which lemma is more promising. Cantor spent a long time trying to generalize Lemma 1 to continuous maps. (At one point, Cantor even believed he had succeeded [27].) In fact, Lemma 1 does not generalize to continuous maps.

Space-filling curve. *(Peano, 1890) For any $q < n$, there is a surjective continuous map from $\mathbb{R}^q$ to $\mathbb{R}^n$.*

In his important paper on topological invariance of dimension, Brouwer proved that Lemma 2 does generalize to continuous maps.

Brouwer non-embedding theorem. *If $n > q$, then there is no injective continuous map from $\mathbb{R}^n$ to $\mathbb{R}^q$.*

So it turns out that Lemma 2 is more robust than Lemma 1. A smaller-dimensional space may be stretched to cover a higher-dimensional space. But a higher-dimensional space may not be squeezed to fit into a lower-dimensional space. This fact is not obvious a priori - it is an important piece of acquired wisdom in topology. In this section, we're going to talk about the geometric consequences/cousins of this fundamental discovery of topology.

Shortly after Brouwer, Lebesgue introduced a nice approach to Brouwer's non-embedding theorem in terms of coverings. If U_i is an open cover of some set $X \subset \mathbb{R}^n$, we say that the multiplicity of the cover is at most μ if each point $x \in X$ is contained in at most μ open sets U_i. We say the diameter of a cover is at most D if each open set U_i has diameter at most D. For any $\epsilon > 0$, Lebesgue constructed an open cover of $\mathbb{R}^n$ with multiplicity $\leq n + 1$ and diameter at most ϵ. He then proposed the following lemma.

Lebesgue covering lemma. *If U_i are open sets that cover the unit n-cube, and each U_i has diameter less than 1, then some point of the n-cube lies in at least $n+1$ different U_i.*

(Brouwer gave the first proof of the Lebesgue covering lemma in 1913. See the interesting essay "The emergence of topological dimension theory" [27] for more information on the history.)

To see how the Lebesgue covering lemma implies the non-embedding theorem, suppose that we have a continuous map f from the unit n-cube to $\mathbb{R}^q$ for some $q < n$. Lebesgue constructed an open cover U_i of $\mathbb{R}^q$ with multiplicity $q+1$ and diameter $< \epsilon$. The preimages $f^{-1}(U_i)$ form an open cover of the unit n-cube with multiplicity $q+1$. Since $q+1 < n+1$, the Lebesgue covering lemma implies that some set $f^{-1}(U_i)$ must have diameter at least 1. On the other hand, the diameters of the sets U_i are as small as we like. By taking a limit as $\epsilon \to 0$, we can find a point $y \in \mathbb{R}^q$ such that the fiber $f^{-1}(y)$ has diameter at least 1. So the Lebesgue covering lemma implies the following large fiber lemma:

Large fiber lemma. *Suppose $q < n$. If f is a continuous map from the unit n-cube to $\mathbb{R}^q$, then one of the fibers of f has diameter at least 1. In other words, there exist points p, q in the unit n-cube with $|p-q| \geq 1$ and $f(p) = f(q)$.*

The large fiber lemma is a precise quantitative theorem saying that an n-dimensional cube cannot be squeezed into a lower-dimensional space.

What is it about the unit n-cube which makes it hard to cover with multiplicity n? Roughly speaking, the key point is that the unit n-cube is "fairly big in all directions". If every non-contractible curve in (T^n, g) has length at least 1, then in some sense, (T^n, g) is fairly big in all directions too. Gromov was able to make this precise and proved the following generalization of the Lebesgue covering lemma.

Generalized Lebesgue covering lemma. *([10]) Suppose that g is a Riemannian metric on the n-dimensional torus T^n with systole at least 1. In other words, every non-contractible loop in (M^n, g) has length at least 1.*

If U_i is an open cover of (M^n, g) with diameter at most 1/10, then some point of M lies in at least n+1 different sets U_i.

Topologists following Lebesgue used the covering lemma as a basis for defining the dimension of metric spaces [26]. They said that the Lebesgue covering dimension of a metric space X is at most n if X admits open covers with multiplicity at most $n+1$ and arbitrarily small diameters. Different notions of dimension were intensively studied in the first half of the twentieth century. The most well-known is the Hausdorff dimension of a metric space. The Hausdorff dimension and the Lebesgue covering dimension may be different. For example, the Cantor set has Lebesgue dimension zero and Hausdorff dimension strictly greater than zero. In 1937, Szpilrajn proved that $LebDim(X) \leq HausDim(X)$

for any compact metric space X. To do so, he constructed coverings of metric spaces with small diameters and bounded multiplicities.

Szpilrajn covering construction. *(1937) If X is a (compact) metric space with n-dimensional Hausdorff measure 0, and $\epsilon > 0$ is any number, then there is a covering of X with multiplicity at most n and diameter at most ϵ. Hence X has Lebesgue dimension $\leq n-1$.*

Gromov asked whether Szpilrajn's theorem is stable in the following sense: If X has very small n-dimensional Hausdorff measure, is there a covering of X with multiplicity at most n and small diameter? In 2008, I constructed such coverings for Riemannian manifolds.

Covering construction for Riemannian manifolds. *(Guth 2008, [19]) If (M^n, g) is an n-dimensional Riemannian manifold with volume V, then there is an open cover of (M^n, g) with multiplicity n and diameter at most $C_n V^{1/n}$.*

Combining this covering construction with the generalized Lebesgue covering lemma, we get a second proof of the systolic inequality. The second proof is summarized in the following metaphor.

Metaphor 2. *The systolic inequality is like topological dimension theory. In particular, it follows from robust versions of the Lebesgue covering lemma and the Szpilrajn covering construction.*

The inequality in my covering construction above and Gromov's filling radius inequality are actually quite similar to each other. The covering inequality implies the filling radius inequality, but the results are equally useful in practice. The methods of proof are quite different though. The proof of the covering construction uses ideas from topological dimension theory: we begin by choosing an open cover of (M, g) and mapping to the nerve of the cover. The main difficulty is that we need quantitative estimates that don't appear in topological dimension theory. We need to estimate the multiplicity the cover, the sizes of the open sets and their overlaps, etc. Taking classical ideas from topology and modifying them to get quantitative estimates is a developing area of research connecting geometry and topology. See Gromov's essay 'Quantitative topology' [15] for an introduction.

7. Scalar Curvature

The Geroch conjecture was one of the guiding problems in the history of scalar curvature.

Geroch conjecture. *The n-torus does not admit a metric of positive scalar curvature.*

In the late 1970's, there were two breakthroughs in the field of scalar curvature. Schoen and Yau invented the minimal hypersurface method, and used it to prove the Geroch conjecture for $n \leq 7$ (see [33] and [34]). We will discuss the minimal hypersurface method more below. Shortly afterwards, Gromov and Lawson used the Dirac operator method to prove the Geroch conjecture for all n.

Gromov's third metaphor connects the Geroch conjecture to the systolic inequality. The metaphor is based on the description of scalar curvature in terms of the volumes of small balls.

Scalar curvature and volumes of balls. *If (M^n, g) is a Riemannian manifold and p is a point in M, then the volumes of small balls in M obey the following asymptotic:*

$$Vol\, B(p,r) = \omega_n r^n - c_n Sc(p) r^{n+2} + O(r^{n+3}). \qquad (*)$$

In this equation, ω_n is the volume of the unit n-ball in Euclidean space, and $c_n > 0$ is a dimensional constant. So we see that if $Sc(p) > 0$, then the volumes of tiny balls $B(p,r)$ are a bit less than Euclidean, and if $Sc(p) < 0$ then the volumes of tiny balls are a bit more than Euclidean.

The scalar curvature measures the asymptotic behavior of volumes of tiny balls as the radius goes to zero. We will consider something analogous to scalar curvature but based on the volumes of balls with finite radius - we call it the "macroscopic scalar curvature at scale r". We define the macroscopic scalar curvature as follows. Let p be a point in (M^n, g). We let $V(p,r)$ be the volume of the ball of radius r around p. Then we let $\tilde{V}(p,r)$ be the volume of the ball of radius r around p in the universal cover of M. (We'll come back in a minute to discuss why it makes sense to use the universal cover here.) Now we compare the volume $\tilde{V}(p,r)$ with the volumes of balls of radius r in spaces of constant curvature. We let $\tilde{V}_S(r)$ denote the volume of the ball of radius r in a simply connected space with constant curvature and scalar curvature S. If we fix r, then $\tilde{V}_S(r)$ is a decreasing function of S; as $S \to +\infty$, $\tilde{V}_S(r)$ goes to zero, and as $S \to -\infty$, $\tilde{V}_S(r)$ goes to infinity. We define the "macroscopic scalar curvature at scale r at p" to be the number S so that $\tilde{V}(p,r) = \tilde{V}_S(r)$.

We denote the macroscopic scalar curvature at scale r at p by $Sc_r(p)$. In particular, if $\tilde{V}(p,r)$ is more than $\omega_n r^n$, then $Sc_r(p) < 0$, and if $\tilde{V}(p,r) < \omega_n r^n$, then $Sc_r(p) > 0$.

By formula $(*)$, it's straightforward to check that $\lim_{r\to 0} Sc_r(p) = Sc(p)$.

Let's work out a simple example. Suppose that g is a flat metric on the n-dimensional torus T^n. In this case, the universal cover of (T^n, g) is Euclidean space. Therefore, we have $\tilde{V}(p,r) = \omega_n r^n$ for each $p \in T^n$ and each $r > 0$. Hence $Sc_r(p) = 0$ for every r and p. If we had used volumes of balls in (T^n, g) instead of in the universal cover, then we would have $Sc_r(p) > 0$ for all r bigger than the diameter of (T^n, g). By using the universal cover, we arrange that flat metrics have $Sc_r = 0$ at every scale r.

Metaphor 3. *The macroscopic scalar curvature is like the scalar curvature.*

This metaphor leads to some deep, elementary, and wide open conjectures in Riemannian geometry.

Generalized Geroch conjecture. *(Gromov 1985) Fix $r > 0$. The n-dimensional torus does not admit a metric with $Sc_r > 0$. Equivalently, if g is any metric on T^n, then the universal cover (T^n, g) contains a ball of radius r and volume at least $\omega_n r^n$.*

The generalized Geroch conjecture is very powerful (if it's true). Since the scalar curvature is the limit of Sc_r as $r \to 0$, the generalized Geroch conjecture implies the original Geroch conjecture. The generalized Geroch conjecture also implies the systolic inequality, which we can see as follows. Suppose that (T^n, g) has systole at least 1. The generalized Geroch conjecture implies that the universal cover of (T^n, g) contains a ball of radius $(1/2)$ and volume $\geq \omega_n(1/2)^n$. Since the systole of (T^n, g) is at least 1, the covering projection $\tilde{T}^n \to T^n$ is injective on this ball. Therefore, (T^n, g) contains a ball of radius $(1/2)$ and volume at least $\omega_n(1/2)^n$. In particular, the total volume of (T^n, g) must be at least $\omega_n(1/2)^n$.

The generalized Geroch conjecture really appeals to me because it's so strong and so elementary to state, but I don't see any plausible tool for approaching the problem.

Now we return to the Schoen-Yau proof of the Geroch conjecture, and we discuss how to adapt it to systolic geometry. The key idea in the Schoen-Yau proof is an inequality for stable minimal hypersurfaces in a manifold of positive scalar curvature.

Stability inequality for scalar curvature. *If (M^n, g) is a Riemannian manifold with $Sc > 0$, and $\Sigma^{n-1} \subset M$ is a stable minimal hypersurface, then Σ has - on average - positive scalar curvature also.*

To see how to apply this observation, suppose that (M^3, g) has positive scalar curvature. Then a stable minimal hypersurface $\Sigma \subset M^3$ is 2-dimensional, and it has (on average) positive scalar curvature. In two dimensions, the scalar curvature is much better understood, and it's not so hard to get topological and geometric information about Σ. Now we know topological and geometric information about every minimal surface Σ in M, and we can use this to learn topological and geometric information about M itself. With this tool, Schoen and Yau proved the Geroch conjecture.

I proved an analogue of the Schoen-Yau stability inequality using volumes of balls instead of scalar curvature. Informally, the lemma says that if a Riemannian manifold has balls of small volume then an absolutely minimizing hypersurface also has balls of small volume.

Stability inequality for volumes of balls. *(Guth, 2009, [20]) Suppose that (M^n, g) is a Riemannian manifold where every ball of radius 1 has volume at*

most α, and suppose that (M, g) has systole at least 2. If $\Sigma^{n-1} \subset M$ is an embedded surface which is absolutely minimizing in its homology class, then every ball in Σ of radius $1/2$ has (n-1)-volume at most 2α.

Using this lemma, I proved a weak version of the generalized Geroch conjecture with a non-sharp constant.

Non-sharp generalized Geroch. *(Guth, 2009, [20]) For any metric g on T^n, the universal cover of T^n contains a ball of radius 1 and volume at least $c(n) > 0$. Therefore, if (T^n, g) has systole at least 2, then it contains a ball of radius 1 with volume at least $c(n) > 0$.*

It's unknown whether there is any systolic analogue of the Dirac operator method for positive scalar curvature.

The results of Schoen-Yau and Gromov-Lawson remain today the main theorems about scalar curvature. Now we turn to an open question in the field of scalar curvature, and we consider it from the viewpoint of systolic geometry.

Schoen conjecture. *Suppose that (M^n, hyp) is a closed hyperbolic manifold. Suppose that g is any metric on M obeying the scalar curvature estimate $Sc(g) \geq Sc(hyp)$. Then $Vol(M, g) \geq Vol(M, hyp)$.*

This elegant conjecture appears in connection with the Yamabe problem in conformal geometry [32], and it is also beautiful in its own right. In two dimensions, the conjecture follows from the Gauss-Bonnet formula. In three dimensions, it was proven by Perelman as a byproduct of the Ricci flow proof of geometrization. In four dimensions, the conjecture is open, but LeBrun proved a cousin of this conjecture for complex hyperbolic manifolds [31]. LeBrun's proof uses Seiberg-Witten theory. In dimensions $n \geq 5$, the problem is wide open. According to a deep theorem of Besson, Courtois, and Gallot, if $Ric(g) \geq Ric(hyp)$, then $Vol(M, g) \geq Vol(M, hyp)$ [4]. This theorem of Besson, Courtois, and Gallot is much weaker than the Schoen conjecture, but it is still a landmark result in comparison geometry. In dimensions $n \geq 5$ we don't have any lower bound at all for $Vol(M^n, g)$ with $Scal(g) \geq Scal(hyp)$.

The Schoen conjecture can be generalized to the macroscopic scalar curvature, producing an even more general and daunting conjecture.

Generalized Schoen conjecture. *Let $r > 0$ be any number. Suppose that (M^n, hyp) is a closed hyperbolic manifold. Suppose that g is any metric on M obeying the estimate $Sc_r(g) \geq Sc_r(hyp)$. Then $Vol(M, g) \geq Vol(M, hyp)$.*

Needless to say, this conjecture is far out of reach. But using methods from systolic geometry, I proved a weak version of this conjecture with a non-sharp constant.

Non-sharp generalized Schoen conjecture. *(Guth, [22]) Suppose that (M^n, hyp) is a hyperbolic manifold. Suppose that g is any metric on M obeying the estimate $Sc_1(g) \geq Sc_1(hyp)$. In other words, every unit ball in the universal*

cover of (M^n, g) *has volume at most the volume of a hyperbolic unit ball. Then* $Vol(M, g) \geq c(n)\, Vol(M, hyp)$.

The generalized Schoen conjecture implies the original Schoen conjecture by taking the limit as $r \to 0$, but my inequality is not sharp enough to give any information about scalar curvature.

The minimal hypersurface approach to scalar curvature is not enough to resolve the Schoen conjecture. Similarly, the minimal hypersurface approach to systolic geometry is not enough to prove the volume estimate above. The proof of this volume estimate uses the techniques coming from topological dimension theory.

8. The Federer-Fleming Averaging Argument

The three metaphors we have been discussing provide large-scale perspective on the systolic problem. They provide guidance about how the outline of the proof should go, but they usually don't provide guidance about how the details of the proof should go. One crucial idea that makes the details work is the Federer-Fleming averaging argument. It is the one ingredient which appears in some form in all three proofs of the systolic inequality.

Here is the first example of the Federer-Fleming averaging argument, coming from their paper [9] on the Plateau problem.

Deformation lemma. *Suppose that* z *is a* k*-dimensional surface in the unit* N*-ball* B^N*, and that* z *has a boundary* ∂z *lying in* ∂B^N*. If* $k < N$*, then there is a map* $\Phi : z \to \partial B^N$ *which fixes* ∂z *and obeys the volume estimate*

$$Vol_k[\Phi(z)] \leq C(k, N)\, Vol_k[z].$$

Informally, the proposition says that we can push z into the boundary of the ball without stretching it too much.

The simplest way one could think to map z into ∂B^N is to project z radially outward to the boundary. Let Φ_0 denote the radial projection outward from zero. In polar coordinates, $\Phi_0(r, \theta) = (1, \theta)$. This map Φ_0 is undefined at the point 0, but we can first put z into general position so that it avoids 0, and this operation has a negligible effect on the volume of z. But the radial projection Φ_0 may not obey the volume estimate. If a large fraction of z is concentrated near to 0, then the radial projection may badly stretch this portion of z leading to an image with a huge volume. Instead of projecting from 0, one can instead project outward from any point $p \in B^N$. We let $\Phi_p : B^N \setminus \{p\} \to \partial B^N$ denote the radial projection outward from the point p. Federer and Fleming discovered that for any fixed surface z, *most projections* Φ_p *obey the volume estimate.* To do that, they estimated the average volume of a projection, proving the inequality

$$\frac{1}{Vol\, B^N} \int_{B^N} Vol_k[\Phi_p z] dp \leq C(k, N)\, Vol_k z.$$

This inequality follows in a couple lines using Fubini's theorem.

This simple averaging method tells us something fundamental about surface areas. By using the averaging method many times, one can prove a surprising range of geometric estimates about surface areas. This approach to geometry problems originates with Federer and Fleming in 1959, but Gromov's proof of the systolic inequality really showed how powerful it is, starting a stream of results proven by using the averaging trick many times. Let's trace the history of this method.

1. (Isoperimetric inequalities) The method begins with Federer and Fleming who used the deformation lemma to prove a general isoperimetric inequality [9].

   Federer-Fleming isoperimetric inequality. *If Z is a k-dimensional closed surface in $\mathbb{R}^N$, then there is a (k+1)-dimensional surface Y with $\partial Y = Z$ obeying the volume estimate*

   $$Vol_{k+1}(Y) \leq C(k,N)\, Vol_k(Z)^{\frac{k+1}{k}}.$$

   Their proof also gives a filling radius estimate.

   Federer-Fleming filling radius inequality. *If Z is a k-dimensional closed surface in $\mathbb{R}^N$, then there is a (k+1)-dimensional surface Y with $\partial Y = Z$ so that every point $y \in Y$ obeys the distance estimate*

   $$dist(y, Z) \leq C(k,N)\, Vol_k(Z)^{\frac{1}{k}}.$$

2. (Isoperimetric inequalities in high dimensions) The constants in the Federer-Fleming estimates above are not sharp. They are particularly bad in large ambient dimensions N. As $N \to \infty$, the constant $c(k,N) \to \infty$. The sharp constants were found using geometric measure theory, and they occur when Z is a round sphere. (The sharp radius estimate is due to Bombieri-Simon [6] and the sharp isoperimetric inequality is due to Almgren [1].) In particular, the sharp constants do not depend on the ambient dimension N.

   Let us contrast the Federer-Fleming approach with the minimal surface approach. In the minimal surface approach to the filling radius inequality, one takes Y to be an absolutely minimizing chain with boundary Z. The existence of such a minimizer is a deep theorem (the solution of the Plateau problem). The variational method really doesn't tell us how to construct Y or even how to approximate Y. Next one proves that Y is smooth at most points. Finally, minimal surfaces enjoy special geometric properties such as the monotonicity formula, which then imply estimates about the radius or volume of Y. By contrast, Federer and Fleming construct the filling Y "by hand", using the deformation lemma repeatedly.

This construction is crude compared to the minimal surface filling, and hence it does not give sharp constants.

In the early 80's, one might have guessed that a direct construction of Y would be too crude to prove good isoperimetric estimates when the ambient dimension $N \to \infty$. Surprisingly, Gromov was able to adapt the Federer-Fleming method to prove isoperimetric and filling radius estimates with constants independent of the ambient dimension [11]. Moreover, the method was flexible enough to work in Banach spaces such as $(\mathbb{R}^N, l^\infty)$, where minimal surface techniques do not work. The main new idea in Gromov's proof was to use induction on k. The proof was further simplified and generalized by Wenger in [35]. His proof is only a couple pages long.

Isoperimetric inequality in Banach spaces. *Let B be a Banach space. Suppose that Z is a k-dimensional closed surface in B. Then there is a (k+1)-dimensional surface Y with $\partial Y = Z$ obeying the volume inequaliy*

$$Vol_{k+1}(Y) \leq C(k)\, Vol_k(Z)^{\frac{k+1}{k}}.$$

3. (Sweep out inequalities) In an appendix to [11], Gromov used the Federer-Fleming method to approach the Almgren sweepout inequality.

   Sweep out inequality. *(Almgren, 1962 [2]) Suppose that $\Phi : S^k \times S^{n-k} \to S^n$ is a map of non-zero degree. Equip the target S^n with the standard unit sphere metric. Then there exists some $\theta \in S^{n-k}$ so that $\Phi(S^k \times \{\theta\})$ has k-volume at least the volume of the unit k-sphere.*

   This is a deep result based on the variational theory of minimal surfaces. For a reader without a strong background in geometric measure theory, the proof is hundreds of pages long. Gromov proved a slightly weaker result by using the Federer-Fleming averaging lemma repeatedly. The lower bound on volume in Gromov's result is a non-sharp constant $c(k, n) > 0$, but the proof is only a few pages long.

4. (Isoperimetric inequalities on Lie groups) Gromov adapted the Federer-Fleming method to Lie groups such as the Heisenberg group. In [16] he proved an analogue of the filling radius inequality for surfaces in the Heisenberg group.

   Building on Gromov's work, Young proved an isoperimetric inequality in the Heisenberg group as follows.

   Isoperimetric inequality in the Heisenberg group. *(Young, 2008, [36]) Let (H^{2n+1}, g) be a left-invariant metric on the Heisenberg group H^{2n+1}. If Z is a k-dimensional closed surface in H^{2n+1} and $k < n$, then*

there is a (k+1)-dimensional surface Y with $\partial Y = Z$ obeying the volume estimate

$$Vol_{k+1}(Y) \leq C(k,n,g)\, Vol_k(Z)^{\frac{k+1}{k}}.$$

Young's main new idea was to use the averaging lemma at many scales.

5. (Area-expanding embeddings) I applied the Federer-Fleming method to the problem of area-expanding embeddings. If $U, V \subset \mathbb{R}^n$ are open sets, an embedding $\Psi : U \to V$ is called k-expanding if it increases the k-dimensional area of each k-dimensional surface. I studied when there is a k-expanding embedding from one n-dimensional rectangle into another, and I answered the question up to a constant factor [23]. This problem turns out to be fairly "rigid" in the sense that the optimal strategy for embedding one rectangle in another is simple. The difficult part of the problem is to prove that there are no k-expanding embeddings between certain rectangles.

   Area-expanding embeddings of rectangles. *If R is an n-dimensional rectangle with side lengths $R_1 \leq ... \leq R_n$, and R' is an n-dimensional rectangle with side lengths $R'_1 \leq ... \leq R'_n$, and if there is a k-expanding embedding from R into R', then the following inequalities hold*

   $$R_1 ... R_j (R_{j+1} ... R_l)^{\frac{k-j}{l-j}} \leq C(n) R'_1 ... R'_j (R'_{j+1} ... R'_l)^{\frac{k-j}{l-j}},$$

   for each $1 \leq j \leq k$ and $k \leq l \leq n$.

   Up to a constant factor, this list of inequalities is necessary and sufficient to find a k-expanding from R into R'.

6. (Point selection theorem in combinatorics) Gromov applied the Federer-Fleming method to give a new proof of the point selection theorem in combinatorics.

   Point selection. *(Barany [3]) If $p_1, ..., p_N$ are points in $\mathbb{R}^n$, consider the $\binom{N}{n+1}$ n-dimensional simplices with vertices among these points. Then there is a point $y \in \mathbb{R}^n$ which lies in at least $c(n)\binom{N}{n+1}$ of the $\binom{N}{n+1}$ n-simplices, for a universal constant $c(n) \geq (n+1)^{-(n+1)}$.*

   Gromov reproved this theorem and generalized it. Given N points in $\mathbb{R}^n$, we get a linear map L from the (N-1)-simplex Δ^{N-1} to $\mathbb{R}^n$, given by mapping the N vertices of the simplex to $p_1, ..., p_N$. The point selection theorem says that y lies in the image of at least $c(n)\binom{N}{n+1}$ of the n-faces of Δ^{N-1}. It turns out that this holds for all continuous maps, not only for linear maps.

   Topological simplex inequality. *(Gromov, 2009, [14]) Suppose that F is a continuous map from Δ^{N-1} to $\mathbb{R}^n$. Then there is a point $y \in \mathbb{R}^n$ which lies in the image of at least $c(n)\binom{N}{n+1}$ n-faces of Δ^{N-1}.*

Gromov's proof of this combinatorial theorem is closely based on his proof of the sweepout inequality, using a combinatorial analogue of the Federer-Fleming averaging argument.

In each of these theorems, using the Federer-Fleming averaging trick over and over is essentially the entire proof.

I want to end this section with a philosophical discussion of the Federer-Fleming averaging method.

The fundamental idea is that the average value of some function may be easier to understand than the function itself. This idea is certainly older than Federer and Fleming. As a dramatic example, Erdos used a similar averaging trick to prove that there are colorings of a graph with no cliques. Given appropriate bounds on the size of the graph and the size of the cliques, he proved that the average number of cliques in a coloring is less than 1. Hence colorings with no cliques exist, even though it is difficult to produce an explicit example. Federer and Fleming borrowed this idea and used it to prove inequalities in geometry. (It would be interesting to know more about the history of this averaging trick.)

The wonderful thing about the averaging trick is that it's so flexible. As we have seen, some of the results in the above list can also be approached by minimal surface theory, and the minimal surface techniques lead to the sharp constants. Using the averaging lemma repeatedly is not as precise but it's more flexible. It can be adapted to Banach spaces. It can be adapted to the Heisenberg group. It can be adapted to the geometry of surfaces inside a rectangle - measuring how the dimensions of the rectangle influence the isoperimetric inequalities. It can be adapted to the combinatorics of an N-dimensional simplex with $N \to \infty$.

In the small field of metric geometry, the Federer-Fleming averaging trick is the most common tool. When the averaging trick doesn't work, we often get stuck. Intuitively, we can only use the averaging trick to find a geometric object if the objects we are looking for are pretty common. Are there any geometric theorems about the existence of rare objects? What tools could we use to find those objects?

I think these issues may be related to the open problems at the end of this essay. Those problems have to do with notions of size in Riemannian geometry, and I need to lay a little groundwork before we get to them.

9. Notions of Size in Riemannian Geometry

Many of the arguments in systolic geometry have to do with various ways of measuring the 'size' of a Riemannian manifold.

Size invariants. *Let M be a smooth manifold. A size invariant for metrics on M is a function S which assigns a positive number to each metric on M, and which obeys the following axioms.*

1. *If g and g' are isometric, then $S(g) = S(g')$.*

2. *If $g \leq g'$, then $S(g) \leq S(g')$.*

(We say that $g \leq g'$ if for each point x and each tangent vector v in T_xM, $g(v,v) \leq g'(v,v)$.)

The volume and diameter are two fundamental size invariants. Many Riemannian invariants are not size invariants. For example, anything related to the curvature is not a size invariant. The injectivity radius is not a size invariant, and neither are the eigenvalues of the Laplacian or the lengths of closed geodesics. But the systole is a size invariant.

The most interesting size invariants I know came out of the proofs of the systolic inequality. We met these invariants implicitly in the discussion above, and now we turn our attention to them.

Filling radius. *If (M^n, g) is a closed Riemannian manifold, then we define its filling radius to be the smallest radius R so that the Kuratowski embedding of (M, g) into L^∞ bounds a chain inside its R-neighborhood.*

Uryson width. *If X is any metric space, such as a Riemannian manifold, and $q \geq 0$ is an integer, then we say that X has q-dimensional Uryson width at most W if there is an open cover of X with diameter $\leq W$ and multiplicity $\leq q+1$. We denote the q-dimensional Uryson width of X by $UW_q(X)$.*

Among the size invariants that I know, the Uryson width seems like the most useful one, so I will try to give a little intuition about it. In some sense, the definition goes back to topologists working on dimension theory, including Uryson. Gromov returned to the the definition and applied it to Riemannian geometry. He gives a long discussion of it in [17]. Recall that $\mathbb{R}^n$ has open covers of multiplicity $n+1$ with arbitrarily small diameters, so $UW_n(\mathbb{R}^n, g_{euclid}) = 0$. More generally, the Uryson n-width of any n-dimensional simplicial complex is equal to zero. Roughly speaking, X has a small Uryson q-width if it "looks q-dimensional". If X has an open cover with multiplicity $q+1$, then the nerve of the cover is a simplicial complex of dimension q. There is a continuous map Φ from X to the nerve so that each fiber of the map is contained in one of the open sets. Thus a metric space X with small q-dimensional Uryson width may be mapped into a q-dimensional complex and each fiber of the map will have small diameter. If the Uryson q-width of X is $< \epsilon$, then we can informally say, "when we look at X from far away and cannot distinguish points of distance $< \epsilon$, X appears to be q-dimensional".

So far in this essay, we have seen three universal inequalities about size functions.

1. The systolic inequality: $Sys(g) \leq C(n)\, Vol(g)^{1/n}$ for all metrics on T^n.

2. The filling radius inequality: $FillRad(g) \leq C(n)\, Vol(g)^{1/n}$ for all metrics on closed n-manifolds.

3. The Uryson width inequality: $UW_{n-1}(g) \leq C(n)\, Vol(g)^{1/n}$ for all metrics on n-manifolds.

These inequalities are closely related. The Uryson width inequality implies the filling radius inequality which implies the systolic inequality, but they all come from the same circle of ideas. Twenty-five years ago, Gromov proved 1 and 2 and conjectured 3. Since then, we have not found any really new universal inequality about sizes of Riemannian metrics. The inequalities we have proven since are either much easier than the filling radius inequality or else they are closely related to the filling radius inequality.

Are there other interesting universal inequalities about the sizes of Riemannian manifolds?

There may well be, but let me try to describe why it hasn't been easy to find any. It is easy to define size invariants of Riemannian manifolds. I know ten or twenty different kinds of size invariants for Riemannian manifolds. But it's often hard to evaluate these invariants, even roughly. For example, here is a simple size invariant for metrics on S^3.

Covering radius. *The covering radius of (S^3, g) is the smallest radius R so that we can find a degree 1 contracting map from the 3-sphere of radius R to (S^3, g).*

(A contracting map is a map that decreases distances.) The manifold S^3 is diffeomorphic to the Lie group $SU(2)$. The left-invariant metrics on $SU(2)$ are some of the simplest metrics on S^3. Gromov raised the problem of estimating the covering radius of left-invariant metrics on $SU(2)$. There is a huge gap between the best known upper and lower bounds, and the problem has been open for more than twenty five years.

There are lots of size invariants, and they are often hard to evaluate. I don't know any good perspective to organize the information. As we've seen, the space of Riemannian metrics is huge, so there are counterexamples for many naive conjectures about size invariants. And after defining ten or twenty size invariants it gets hard to see what's significant.

I want to end by putting forward two questions about sizes of Riemannian manifolds. I think that whether the answers are yes or no, some interesting new geometry will be involved.

The first question is about the geometry of high-genus surfaces. My main point is that we really don't have a good understanding of the geometry of high-genus Riemannian surfaces.

Question 1. *(Buser) If (Σ^2, g) is a closed Riemannian surface of arbitrary genus, is there a continuous map F from Σ to a graph Γ obeying the following inequality:*

$$\textit{for every } y \textit{ in } \Gamma,\ Length[F^{-1}(y)] \leq C Area(\Sigma, g)^{1/2}?$$

(This question is a small variation on Buser's question about the sharp value of the Bers constant — see [7].)

This question connects to topics we've seen above in a couple ways. First of all, the Uryson width inequality tells us that we can find a map F from (Σ, g) to a graph so that each fiber has *diameter* at most $C Area(\Sigma, g)^{1/2}$. This estimate does not imply the length estimate at all, because a fiber may be a very long curve which wiggles a lot and therefore has a small diameter. The most interesting examples of high genus Riemannian surfaces are probably the arithmetic hyperbolic surfaces studied by Buser and Sarnak in [8]. These surfaces have genus G, area around G, and diameter around $\log G$. Since the entire surface has diameter around $\log G$, any curve in it has diameter at most around $\log G$. When G is large, the diameters are much smaller than the square root of the area. So any map from an arithmetic hyperbolic surface to a graph has fibers of diameter at most $Area^{1/2}$, but it's not at all clear how small we can make the *lengths* of the fibers.

This question also fits in with the naive conjectures in Section 3 of this essay. In particular, if Σ is a small genus surface, then Balacheff and Sabourau proved that the answer to the question is yes. In a bit more generality, here is their result.

Balacheff-Sabourau inequality. *([5]) If (Σ^2, g) is a closed surface of genus G, then there is a function $f : \Sigma^2 \to \mathbb{R}$ so that for every $y \in \mathbb{R}$, the length of the level set $f^{-1}(y)$ obeys the inequality*

$$Length[f^{-1}(y)] \leq C\sqrt{G+1}Area(\Sigma^2, g)^{1/2}.$$

For large genus, the right-hand side grows like $\sqrt{G}$, and this behavior is sharp. But if we allow maps to a 1-dimensional complex Γ instead of maps to $\mathbb{R}$, we may get a better estimate for lengths. If the answer to Question 1 is yes, then we can look for similar inequalities in higher dimensions. Can every 3-manifold of volume 1 be mapped to a 2-dimensional complex with fibers of length $\leq C$? Can every 3-manifold of volume 1 be mapped to $\mathbb{R}^2$ with fibers of length $\leq C$?? Can every 3-manifold of volume 1 be mapped to a 1-dimensional complex with fibers of area $\leq C$?

The second problem is about Uryson widths. Recall the Uryson width inequality, $UW_{n-1}(M^n, g) \leq C(n)\, Vol(M^n, g)^{1/n}$, which says that an n-manifold of tiny n-dimensional volume looks (n-1)-dimensional. What conditions on g would force (M^n, g) to look (n-2)-dimensional?

This is an open-ended question that could go in many directions. For instance, Gromov has a conjecture that if the scalar curvature of g is at least 1, then $UW_{n-2}(M^n, g) \leq C(n)$.

Here is another direction suggested by the geometry of area-contracting maps. Suppose that M^n is just the standard unit n-ball, and we have the metric g_{ij} written in coordinates. What do we need to know pointwise about g_{ij} to control $UW_{n-2}(B^n, g)$?

Question 2. *Let B^n denote the standard (unit) n-ball in $\mathbb{R}^n$, and let g_0 denote the standard Euclidean metric. Suppose that g is another metric obeying $\Lambda^k g \leq \Lambda^k g_0$. This means that for every k-dimensional surface $\Sigma^k \subset B^n$, the g-volume of Σ is at most the Euclidean volume of Σ. Suppose that $n/k \geq d$. Is it true that $UW_{n-d}(B^n, g) \leq C(n)$?*

To get a sense of this question, let us first imagine that the metric $g_{ij}(x)$ is constant in x. In this case, (B^n, g_{ij}) is isometric to a Euclidean ellipsoid. If g is a constant metric and $\Lambda^k g \leq \Lambda^k g_0$, then linear algebra implies that $UW_{k-1}(B^n, g) \leq 1$. At this point, one might naively conjecture that all metrics g with $\Lambda^k g \leq \Lambda^k g_0$ obey $UW_{k-1}(B^n, g) \leq C(n)$. Moreover, the Uryson width inequality implies that if $\Lambda^n g \leq \Lambda^n g_0$, then $UW_{n-1}(B^n, g) \leq C(n)$. So the naive conjecture is true when $k = n$. But the naive conjecture is false for other values of k because of a counterexample coming from work of Zel'dovitch in astrophysics and Gehring in conformal geometry. Zel'dovitch's work has to do with the internal geometry of a neutron star. I think that this counterexample is the worst case, and the question asks whether this is true. See my paper [24] on area-contracting maps and topology for more context.

10. Reading Guide

For the reader who would like to learn more about this area of geometry, here are some resources.

Gromov wrote about systolic geometry in several places. The key research paper is "Filling Riemannian manifolds" [10]. His expository writing about systoles includes Chapter 4 of *Metric Structures* [11], and the essay "Systoles and isosystolic inqualities" [13].

Katz's expository work on systoles includes the book *Systolic Geometry and Topology* [29] and his website on systoles [30]. The website contains a lot of interesting stuff, including a list of open problems in the field.

I wrote a set of notes on the systolic inequality [21] which explains the original proof in detail in 14 pages. This talk is based on my essay [25], which includes several topics we didn't have time to discuss here: hyperbolic geometry, symmetry, calibrations, and Nabutovsky's work on the complexity of the space of metrics.

References

[1] Almgren, F. Optimal isoperimetric inequalities. Indiana Univ. Math. J. 35 (1986), no. 3, 451–547.

[2] Almgren, F., The theory of varifolds - a calculus of variations in the large for the k-dimensional area integrated, manuscript available in the Princeton math library.

[3] Barany, I., A generalization of Caratheodory's theorem, Discrete Math. 40 (1982), no. 2–3, 141–152.

[4] Besson, G.; Courtois, G.; Gallot, S., Volumes, entropies et rigidités des espaces localement symétriques de courbure strictement négative, Geom. Funct. Anal. 5 (1995), no. 5, 731–799.

[5] Balacheff, F.; Sabourau, S., Diastolic inequalities and isoperimetric inequalities on surfaces, preprint.

[6] Bombieri, E. ; Simon, L., On the Gehring link problem. Seminar on minimal submanifolds, 271–274, Ann. of Math. Stud., 103, Princeton Univ. Press, Princeton, NJ, 1983.

[7] Buser, P.; Seppl, M., Symmetric pants decompositions of Riemann surfaces. Duke Math. J. 67 (1992), no. 1, 39–55.

[8] Buser, P.; Sarnak, P. On the period matrix of a Riemann surface of large genus. Invent. Math. 117 (1994), no. 1, 27–56.

[9] Federer, H.; Fleming, W. Normal and integral currents. Ann. of Math. (2) 72 1960 458–520.

[10] Gromov, M., Filling Riemannian manifolds. J. Differential Geom. 18 (1983), no. 1, 1–147.

[11] Gromov, M., *Metric Structures on Riemannian and Non-Riemannian Space*, Based on the 1981 French original [MR0682063 (85e:53051)]. With appendices by M. Katz, P. Pansu and S. Semmes. Translated from the French by Sean Michael Bates. Progress in Mathematics, 152. Birkhuser Boston, Inc., Boston, MA, 1999.

[12] Gromov, M., Volume and bounded cohomology, Inst. Hautes tudes Sci. Publ. Math. No. 56 (1982), 5–99 (1983).

[13] Gromov, M., Systoles and intersystolic inequalities, Actes de la Table Ronde de Geometrie Differentielle (Luminy, 1992) 291-362, Semin. Cong., 1, Soc. Math. France, Paris, 1996.

[14] Gromov, M., Singularities, expanders, and topology of maps, part 2, preprint.

[15] Gromov, M., Quantitative homotopy theory. Prospects in mathematics (Princeton, NJ, 1996), 45–49, Amer. Math. Soc., Providence, RI, 1999.

[16] Gromov, M., Carnot-Caratheodory spaces seen from within. Sub-Riemannian geometry, 79–323, Progr. Math., 144, Birkhuser, Basel, 1996.

[17] Gromov, M. Width and related invariants of Riemannian manifolds. On the geometry of differentiable manifolds (Rome, 1986). Astrisque No. 163–164 (1988), 6, 93–109, 282 (1989).

[18] Guth, L., Width-volume inequality, Geom. Funct. Anal. 17 (2007), no. 4, 1139–1179.

[19] Guth, L. Uryson width and volume, preprint

[20] Guth, L., Systolic inequalities and minimal hypersurfaces, Geometric and Functional Analysis: Volume 19, Issue 6 (2010) , Page 1688.

[21] Guth, L., Notes on Gromov's systolic inequality, Geom. Dedicata 123 (2006), 113–129.

[22] Guth, L., Volumes of balls in large Riemannian manifolds, accepted for publication in Annals of Mathematics.

[23] Guth, L., Area-expanding embeddings of rectangles, preprint.

[24] Guth, L., Contraction of areas vs. topology of mappings, preprint.

[25] Guth, L., Metaphors in systolic geometry, preprint.

[26] Hurewicz, W., Wallman, H., *Dimension Theory*, Princeton University Press, Princeton, New Jersey, 1996.

[27] Crilly T. with Johnson, D., The emergence of topological dimension theory, in *History of Topology*, edited by I. M. James, North-Holland, Amsterdam, 1999.

[28] Katz, M., Counterexamples to isosystolic inequalities. Geom. Dedicata 57 (1995), no. 2, 195–206.

[29] Katz, M., *Systolic Geometry and Topology*, Mathematical surverys and monographs volume 137, American Mathematical Society, 2007.

[30] Website on systolic geometry and topology, maintained by Katz, M., http://u.cs.biu.ac.il/ katzmik/sgt.html

[31] LeBrun, C., Four-manifolds without Einstein metrics, Math. Res. Lett. 3 (1996) no. 2, 133–147.

[32] Schoen, R., Variational theory for the total scalar curvature functional for Riemannian metrics and related topics in *Topics in calculus of variations* (Montecatini Terme, 1987) 120–154, Lecture Notes in Math. 1365, Springer, Berlin, 1989.

[33] Schoen, R.; Yau, S. T. Incompressible minimal surfaces, three-dimensional manifolds with nonnegative scalar curvature, and the positive mass conjecture in general relativity. Proc. Nat. Acad. Sci. U.S.A. 75 (1978), no. 6, 2567.

[34] Schoen, R.; Yau, S.T., On the structure of manifolds with positive scalar curvature. Manuscripta Math. 28 (1979), no. 1–3, 159–183.

[35] Wenger, S., A short proof of Gromov's filling inequality. Proc. Amer. Math. Soc. 136 (2008), no. 8, 2937–2941.

[36] Young, R., Filling inequalities for nilpotent groups, preprint.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Volume Comparison via Boundary Distances

Sergei Ivanov*

Abstract

The main subject of this lecture is a connection between Gromov's filling volumes and a boundary rigidity problem of determining a Riemannian metric in a compact domain by its boundary distance function. A fruitful approach is to represent Riemannian metrics by minimal surfaces in a Banach space and to prove rigidity by studying the equality case in a filling volume inequality. I discuss recent results obtained with this approach and related problems in Finsler geometry.

Mathematics Subject Classification (2010). Primary 53C23; Secondary 53C60.

Keywords. Filling volume, minimal filling, boundary distance rigidity.

1. Introduction

1.1. A toy question. One of the goals of this lecture is to advertise a conjecture about filling volumes. It can be stated without preliminaries (although in an obscured way) as follows.

Question 1.1. Let N^{n+1} be a complete Riemannian manifold and $M^n \subset N$ a compact hypersurface with boundary. Suppose that M is convex in the following strong sense: for every two points $x, y \in M$, there is a unique shortest geodesic segment connecting x and y in N, and this segment lies in M. (In particular, M is totally geodesic.)

Is it true that every such M is an area minimizer? That is, does it have the least n-dimensional area among all compact (orientable) hypersurfaces in N with the same boundary?

*Supported by the Dynasty Foundation and RFBR grants 08-01-00079-a, 09-01-12130-ofi-m.

St. Petersburg Department of Steklov Institute of Mathematics, Fontanka 27, 191023, St. Petersburg, Russia. E-mail: svivanov@pdmi.ras.ru.

The wording of this question is deliberately chosen so as to make an affirmative answer sound more plausible. Actually the answer is not known, and an affirmative one would have strong implications.

The convexity assumptions in Question 1.1 imply that M is diffeomorphic to the n-disc, its boundary is convex, and all its geodesic are shortest paths. The latter is a crucial property while the former two could be relaxed: for example, non-convex regions in M are area minimizers if so is M.

Since the surface in question is totally geodesic, it is minimal in the variational sense: the mean curvature, and hence the first variation of area, is zero. Cutting off a neighborhood of the boundary yields a surface where geodesics have no conjugate points, and it is easy to see that in this case it is a stable minimal surface and hence minimizes the area locally (among all nearby surfaces). However the *global* area-minimality in Question 1.1 is a completely different issue.

1.2. Boundary rigidity. I postpone further discussion of Question 1.1 until subsection 1.3. This subsection is a brief introduction to boundary distance rigidity.

For a Riemannian manifold M, possibly with boundary, let d_M denote the induced length metric on M. This is a function on $M \times M$ measuring geodesic distances between points. The *boundary distance function* of M, denoted by bd_M, is the restriction of d_M to $\partial M \times \partial M$. It is natural to ask whether the metric in the interior can be determined if one knows the boundary distance function.

Inverse boundary problems of this type were originally motivated by geophysics: the inner structure of the Earth can be studied by measuring travel times of seismic waves between points at the surface. Assuming that the Earth is filled by isotropic media with variable speed of sound, the travel times represent the boundary distance function of a conformal metric on D^3, and the problem is to determine the conformal factor by these data. Under the assumption that the Earth is spherically symmetric, this inverse kinematic problem was solved by Herglotz [24] and Wiechert [39]. For a general simple conformal metric, the uniqueness of a solution was proved by Mukhometov and Romanov [33], see also [6], [32], [17].

If the metric is not supposed to be conformal, determining metric coefficients as functions of coordinates does not make sense: any Riemannian isometry that fixes the boundary obviously preserves the boundary distances. Two metrics related by such an isometry must be regarded as the same metric, hence the following definition.

Definition 1.2. A compact Riemannian manifold M with boundary is said to be *boundary rigid* if it is determined by its boundary distance function uniquely up to an isometry fixing the boundary.

In a more formal language this means the following: every compact Riemannian manifold M' such that $\partial M' = \partial M$ and $bd_{M'} = bd_M$ is isometric to M via an isometry $f : M \to M'$ such that $f|_{\partial M} = id_{\partial M}$.

It is easy to construct metrics that are *not* boundary rigid. For example, begin with an arbitrary metric and enlarge it near a point p so that no shortest path between boundary points goes through p. Then a perturbation of the metric near p does not affect the boundary distance function. Another example is the standard hemisphere: since the boundary distances are realized by boundary arcs, enlarging the metric in the interior does not change them.

Such examples should be excluded if one seeks boundary rigidity. A natural set of restrictions is contained in the following definition.

Definition 1.3. A compact Riemannian manifold M is said to be *simple* if

(1) The boundary ∂M is strictly convex, i.e. has positive definite second fundamental form.

(2) Every geodesic segment in M is minimal, i.e. realizes the distance between its endpoints.

(3) The geodesics in M have no conjugate points. (Or, equivalently, there is a larger manifold M^+ containing M in its interior and such that all geodesics in M^+ are minimal.)

For example, the standard hemisphere is not simple but cutting off an arbitrarily small neighborhood of the boundary makes it simple.

The first requirement of Definition 1.3 implies that all distances in M are realized by geodesics. Then one easily sees that the exponential map at every point is a diffeomorphism, and it follows that a simple manifold is diffeomorphic to a disc. Thus one may as well speak about *simple metrics* on D^n.

Note that simplicity of the metric can be observed via the boundary distance function. That is, if two metrics have the same boundary distance function, then either they are both simple or both are not. Indeed, the convexity of ∂M is equivalent to a sort of strict triangle inequality for bd_M, and the fact that geodesics are minimal and have no conjugate points is equivalent to smoothness of bd_M away from the diagonal.

Conjecture 1.4 (R. Michel [31]). *Every simple Riemannian manifold is boundary rigid.*

Pestov and Uhlmann [34] proved this conjecture in dimension 2. In higher dimensions the following types of spaces are known to be boundary rigid:

- regions in $\mathbb{R}^n$ and moreover all n-dimensional flat manifolds that admit an isometric immersion to $\mathbb{R}^n$ (Besikovitch [4]; Gromov [22]);
- regions in the standard open hemisphere S^n_+ (Michel [31]);

- regions is symmetric spaces of negative curvature (this follows from a volume entropy inequality proved by Besson, Courtois and Gallot [5]);
- regions in metric products of the form $M_0 \times \mathbb{R}$ where M_0 is a complete simply connected Riemannian manifold without conjugate points (Croke and Kleiner [21]);
- metrics sufficiently close in C^2 to the Euclidean metric of a region in $\mathbb{R}^n$ (Burago and Ivanov [12]);
- metrics sufficiently close in C^3 to the hyperbolic metric of a region in $\mathbb{H}^n$ (Burago and Ivanov [13]).

Proofs of the last two results are discussed in section 3.

Remark. More is known about the local variant of the conjecture, that is, when the metrics of M and M' in Definition 1.2 are assumed *a priori* close to each other. Local boundary rigidity is proved for a generic set of simple metrics including all analytic ones [37] and for all metrics with "not too much" positive curvature [19].

1.3. Filling volumes and minimal fillings. To simplify matters, all manifolds and surfaces in the sequel are assumed orientable. And for the most part one may assume that all Riemannian manifolds in question are just metrics on the disc D^n.

Definition 1.5. Let N be a closed $(n-1)$-dimensional manifold and $f : N \times N \to \mathbb{R}$ a nonnegative function. The *filling volume* of f, denoted by $\mathrm{FillVol}(N, f)$, is defined by

$$\mathrm{FillVol}(N, f) = \inf\{\mathrm{Vol}(M) : \partial M = N,\ bd_M \geq f\} \tag{1.1}$$

where the infimum is taken over all (orientable) compact n-dimensional Riemannian manifolds M such that $\partial M = N$ and $bd_M \geq f$. Such manifolds M are referred to as *fillings* of (N, f).

A compact Riemannian manifold M is said to be a *minimal filling* if it realizes the infimum in (1.1) for $S = \partial M$ and some function f (and hence for $f = bd_M$). In other words, M is a minimal filling if $\mathrm{Vol}(M) = \mathrm{FillVol}(\partial M, bd_M)$.

The notion of filling volume was introduced by Gromov [22], originally in the special case where f is a metric on N. The above definition assumes that there are no topological obstructions for N to be a boundary, cf. [22] for the general case.

Substituting intermediate definitions yields the following: M is a minimal filling if and only if, for every compact Riemannian manifold M' such that $\partial M' = \partial M$ and

$$d_{M'}(x, y) \geq d_M(x, y) \quad \text{for all } x, y \in \partial M, \tag{1.2}$$

one has

$$\mathrm{Vol}(M') \geq \mathrm{Vol}(M). \tag{1.3}$$

The following conjecture is the main topic of this lecture.

Conjecture 1.6. *Every simple manifold is a minimal filling.*

Note that a (C^0) limit of minimal fillings is also a minimal filling, and a limit of simple metrics can have a non-strictly convex boundary and non-strictly minimal geodesics. Thus the simplicity assumption in Conjecture 1.6 can be relaxed to allow for such cases. In particular, if the conjecture is true, then the standard hemisphere is a minimal filling.

Convexity of the boundary is a convenience assumption and it can be removed in some cases (see e.g. [29]). Observe that any subregion of a minimal filling is a minimal filling as well.

If a simple manifold M is found to be a minimal filling, one can try to analyze the equality case in (1.3) and hope that it is attained only if M' is isometric to M (via an isometry fixing the boundary). This hope is expressed in the following stronger variant of Conjecture 1.6.

Conjecture 1.6$^+$. *Every simple manifold is a unique minimal filling of its boundary distance function, up to an isometry fixing the boundary.*

It is easy to see that Conjecture 1.6$^+$ implies Michel's boundary rigidity conjecture 1.4. Almost all boundary rigid metrics listed above are also known to be minimal fillings (the exceptions are subsets of the hemisphere and product metrics). In dimension 2, all simple manifolds are minimal fillings within the class of manifolds homeomorphic to the disc [27], but the general filling minimality is not known even for the hemisphere.

Conjecture 1.6 is equivalent to the affirmative answer to Question 1.1. Indeed, let $M \subset N$ be as in Question 1.1 and suppose that there is a surface $M' \subset N$ with the same boundary but smaller area. Then M and M', regarded as Riemannian n-manifolds, satisfy (1.2) and hence provide a counterexample to Conjecture 1.6. Conversely, if manifolds M and M' satisfy (1.2) but do not satisfy (1.3), one can glue them together along the boundary and embed the resulting space into a suitable manifold N^{n+1} in order to produce a counterexample to Question 1.1. (One may need to change the metric of M' near the boundary to make a smooth gluing; this and other technical details are easy to handle.)

2. Some Implications

In this section I discuss some implications of the minimal filling conjectures.

2.1. Boundary rigidity. As I already mentioned, Conjecture 1.6^+ implies Conjecture 1.4. Moreover, this implication works for every individual manifold:

Proposition 2.1. *If a simple Riemannian manifold M is a unique minimal filling of its boundary distance function, then M is boundary rigid.*

The key to the proof is Santaló's integral geometric formula for the volume of a simple Riemannian manifold in terms of its boundary distance function and its first order derivatives (cf. [36], [22], [17]). This formula implies that two simple manifolds with the same boundary distance function have the same volume. Recall that if M is simple and M' has the same boundary distance function, then M' is simple as well, hence $\mathrm{Vol}(M') = \mathrm{Vol}(M)$ by Santaló's formula. Then the uniqueness assumption implies that M and M' are isometric.

This argument actually works not only for simple manifolds but for a large class of *strong geodesic minimizing* (SGM) manifolds, cf. [17].

2.2. Gromov's circle filling conjecture. What is the filling volume of the intrinsic metric of the circle? This was the first question asked by Gromov after the definition of filling volume in [22]. It is conjectured that this filling volume equals 2π, the value realised by the standard round hemisphere. In other words, the question is: is the hemisphere a minimal filling? Since the hemisphere is a limit of simple manifolds, Conjecture 1.6 would immediately imply the affirmative answer.

With definitions substituted, the circle filling conjecture boils down to the following. Let M be a compact orientable two-dimensional surface with a Riemannian metric such that ∂M is a circle of length 2π, and for every pair x, y of opposite points of this circle one has $d_M(x, y) = \pi$. Then (the conjecture asserts that) $\mathrm{area}(M) \geq 2\pi$.

This inequality is well-known if M is homeomorphic to D^2. In other words, the hemisphere is a minimal filling *within the class of surfaces homeomorphic to the disc*. Indeed, one can identify opposite points of the boundary circle and obtain a closed surface $M_1 \simeq \mathbb{RP}^2$ such that the length of a shortest non-contractible loop in M_1 equals π. Then Pu's isosystolic inequality [35] implies that $\mathrm{area}(M) = \mathrm{area}(M_1) \geq 2\pi$.

Pu's original proof uses uniformization and integral geometry, another proof can be found in [27]. The uniformization approach can be pushed further to cover the case when M has genus 1, cf. [2]. The case of a higher genus remains open.

The general case of the circle filling conjecture can be similarly reformulated in terms of a systolic inequality, and it has applications in higher-dimensional systolic geometry, see e.g. [30, §8.3].

2.3. E. Hopf's theorem. If M is an n-torus with a Riemannian metric without conjugate points, then M is flat (that is, locally isometric to $\mathbb{R}^n$). This

fact was proved for $n = 2$ by E. Hopf [26] and for all n by Burago and Ivanov [8]. Both proofs involve dynamical arguments. Croke and Kleiner [20] proposed a more geometric approach where E. Hopf's theorem is derived from asymptotic volume inequalities. Their approach led to a new proof of the theorem in the two-dimensional case. The following modification of their argument shows how the theorem (in all dimensions) follows from Conjecture 1.6 (with a relaxed boundary convexity assumption).

Let $\widetilde{M}$ denote the universal cover of M with the metric lifted from M. The *asymptotic volume* of M is defined by

$$\mathrm{AsVol}(M) = \liminf_{R\to\infty} \frac{\mathrm{Vol}(B_R)}{R^n}$$

where B_R is the metric ball in $\widetilde{M}$ centered at a fixed point $x_0 \in \widetilde{M}$. Let ω_n denote the Euclidean volume of a unit ball in $\mathbb{R}^n$. It can be shown that

$$\mathrm{AsVol}(M) \geq \omega_n, \tag{2.1}$$

with equality if and only if M is flat. This is proved in [18] for any closed Riemannian n-manifold without conjugate points and in [9] for a Riemannian n-torus (with or without conjugate points).

Actually the inequality (2.1) can be improved by inserting a factor depending on the affine type of the stable norm $\|\cdot\|$ of M, see [9] and [23, pp. 259–260]. Namely

$$\mathrm{AsVol}(M) \geq \frac{\mathrm{Vol}(B)}{\mathrm{Vol}(E)} \cdot \omega_n \tag{2.2}$$

where B is the unit ball of $\|\cdot\|$ and E is the ellipsoid of maximal volume contained in B. The equality in (2.2) is attained if and only if the metric is flat.

The universal cover $\widetilde{M}$ can be identified with $\mathbb{R}^n$ equipped with a $\mathbb{Z}^n$-periodic Riemannian metric. Then the distances in $\widetilde{M}$ differ from the distances in the normed space $(\mathbb{R}^n, \|\cdot\|)$ by a bounded function, cf. [7]. Let d_E denote the distance in the Euclidean metric associated with E, then

$$d_E(x,y) \geq d_{\widetilde{M}}(x,y) - const \tag{2.3}$$

for all $x, y \in \widetilde{M}$. If $\widetilde{M}$ has no conjugate points, Conjecture 1.6 (without the boundary convexity assumption) would imply that the ball $B_R \subset \widetilde{M}$ is a minimal filling. Apply the minimal filling inequality (1.3) to $M = B_R$ and $M' = (B_R, d_E)$, where d_E is modified near the boundary to get rid of the constant in (2.3). This yields the inequality opposite to (2.2), hence the metric of $\widetilde{M}$ is flat.

3. Minimality in a Banach Space

In this section I discuss one of the approaches to filling minimality and boundary rigidity and outline the proofs of the following two theorems.

Theorem 3.1 ([12]). *Let $D \subset \mathbb{R}^n$ be a compact region with a smooth boundary and g_0 the standard Euclidean metric on D. Then there is a neighborhood $\mathcal{U}$ of g_0 in the space of Riemannian metrics on D such that for every metric $g \in \mathcal{U}$ the space (D, g) is a minimal filling and boundary rigid.*

Theorem 3.2 ([13]). *Let $D \subset \mathbb{H}^n$ be a compact region with a smooth boundary and g_0 the standard hyperbolic metric on D. Then there is a neighborhood $\mathcal{U}$ of g_0 in the space of Riemannian metrics on D such that for every metric $g \in \mathcal{U}$ the space (D, g) is a minimal filling and boundary rigid.*

As explained above, it suffices to prove that the metric g in question is a unique minimal filling of its boundary distance function. The space of Riemannian metrics in these theorems is regarded with C^∞ topology. (In fact, one can lower it down to C^2 in Theorem 3.1 and to C^3 in Theorem 3.2.)

3.1. Isometric representations. It is well known that every metric space X can be isometrically embedded into an L^∞ type Banach space. A classic Kuratowski map embeds a bounded metric space X into $C^0(X)$ by sending every point $x \in X$ to the distance function $d_X(x, \cdot) \in C^0(X)$. For simple Riemannian metrics there are other natural constructions.

Let M be a simple Riemannian manifold and $S = \partial M$. The *boundary distance representation* is a map $\Phi : M \to C^0(S) \subset L^\infty(S)$ defined by

$$\Phi(x) = d_M(x, \cdot)|_S.$$

It is easy to see that this map is distance-preserving. Furthermore, it features additional nice properties: it is smooth away from the boundary and the gradients of its "coordinate functions" $d_M(\cdot, s)$, $s \in S$, at every point $x \in M \setminus \partial M$ define a diffeomorphism between S and the unit tangent bundle at x. This technical property plays an important role.

There is a similar construction for a complete simply connected manifold M of nonpositive curvature (or a compact region in such a manifold). Fix a point $o \in M$ and let $S = UT_oM$ be the unit tangent sphere at o. The *Busemann representation* $\Phi : M \to L^\infty(S)$ is defined by

$$\Phi(x)(v) = B_{\gamma_v}(x), \qquad x \in M,\ v \in S, \tag{3.1}$$

where γ_v is the geodesic ray from o defined by the initial data $\dot\gamma_v(0) = v$, and B_{γ_v} is its Busemann function. In the case $M = \mathbb{R}^n$ this map is linear:

$$\Phi(x) = \langle x, \cdot\rangle|_{S^{n-1}}, \qquad x \in \mathbb{R}^n,$$

where S^{n-1} is the standard unit sphere in $\mathbb{R}^n$. It is easy to see that the Busemann representation of a nonpositively curved metric is distance-preserving. If the metric has constant curvature outside a compact set, then the Busemann representation is smooth (in general, it may fail to be smooth even in the co-compact case).

The proofs of the above theorems are based on the following fact:

Theorem 3.3 ([28]). *Let M be a compact Riemannian manifold with boundary, S a σ-finite measure space and $\Phi : M \to L^\infty(S)$ a distance-preserving map. Then M is a minimal filling if and only if $\Phi(M)$ is an area minimizer, that is, it has the least area among all Lipschitz surfaces in $L^\infty(S)$ with the same boundary.*

Furthermore, if $\Phi(M)$ is a unique area minimizer spanning its boundary, then M is a unique minimal filling of its boundary distance function and hence is boundary rigid.

Here the surface area in $L^\infty(S)$ is defined as the Loewner area, see below.

The proof of Theorem 3.3 is similar to the argument in section 1.3 showing that Conjecture 1.6 is equivalent to Question 1.1. The "if" implication and the uniqueness assertion easily follow from the fact that any filling M' of $(\partial M, bd_M)$ admits a 1-Lipschitz map $\Phi' : M' \to L^\infty(S)$ such that $\Phi'|_{\partial M} = \Phi|_{\partial M}$. This part of the proof works for any definition of surface area satisfying the natural requirement that 1-Lipschitz maps do not increase areas.

The "only if" implication is not used in theorems 3.1 and 3.2 but it is important for motivation. This implication requires a careful choice of the surface area definition, see the next subsection.

Remark. Theorem 3.3 is a partial case of the following fact. Let N be a closed $(n-1)$-manifold, $d : N \times N \to \mathbb{R}$ is a metric on N and Ψ a distance-preserving map from (N, d) to $L^\infty(S)$. Then $\mathrm{FillVol}(N, d)$ equals the filling area of $\Psi(N)$ in $L^\infty(S)$, i.e. the infimum of the (Loewner) areas of Lipschitz n-surfaces in $L^\infty(S)$ whose boundaries are parametrized by Ψ.

In his founding paper [22] Gromov used the fact that filling volumes and filling areas in L^∞ are equal up to a factor bounded by a constant depending on n. This factor could not be removed because Gromov used another definition of area (namely Benson's area, cf. [38] and [3], denoted by $mass*$ in [22]). If one is interested in filling volumes up to a bounded factor, any definition of area works fine, and $mass*$ is technically easier than other definitions. However it is not suitable for finding precise filling volumes.

3.2. Defining the surface area in L^∞. There are two issues to sort out. First, we have to deal with surfaces of only Lipschitz regularity. For Lipschitz surfaces in $\mathbb{R}^n$ one uses Rademacher's theorem asserting that every Lipschitz map is differentiable almost everywhere. This gives one a Jacobian defined a.e. and then the surface area is defined by integration. This scheme does not work for surfaces in L^∞ due to the lack of Rademacher's theorem. This can be worked around by using *weak derivatives* (i.e., derivatives with respect to a weak topology on the target space). For a Lipschitz map from a smooth manifold M to L^∞, weak derivatives exist and have natural metric properties almost everywhere on M, cf. [1] or [28]. (This Rademacher-type theorem is the main reason why we prefer L^∞ over C^0 for the target space of our embeddings.)

Then, in order to define the surface area in L^∞, one uses weak derivatives in the same way as ordinary derivatives in $\mathbb{R}^n$.

The second issue is how to define the area integrand. Since the norm in L^∞ is not Euclidean, the induced metric of a surface (even of a smooth one) is not Riemannian in general. In fact, it can be an arbitrary Finsler metric. Contrary to the Riemannian case, there are many non-equivalent definitions of area and volume for Finsler metrics, see e.g. [38]. The most commonly used definitions are Busemann's [14] (the Hausdorff measure) and Holmes–Thompson's [25] (the projection of the Liouville measure from the unit tangent bundle).

In order to define an n-dimensional Finsler volume, one chooses a volume normalization factor in every (affine type of) n-dimensional Banach space. For example, Busemann's definition normalizes the volume of the norm's unit ball to be the same constant ω_n for all n-dimensional Banach spaces. The *Loewner volume* mentioned in Theorem 3.3 is defined as follows. Let $(V, \|\cdot\|)$ be an n-dimensional Banach space, B its unit ball and E the John–Loewner ellipsoid of B (i.e., the ellipsoid of maximal volume contained in B). Then the Loewner volume in $(V, \|\cdot\|)$ is normalized so that the volume of E equals ω_n. For a Finsler manifold $M = (M, \varphi)$, the Loewner volume equals the infimum of volumes of Riemannian metrics g on M satisfying $g(v, v) \geq \varphi^2(v)$ for all $v \in TM$. This definition extends to Lipschitz surfaces in L^∞ as explained above.

Remark. Theorem 3.3 is valid in a more general context of Finslerian minimal fillings. To define the notion of a Finslerian minimal filling, modify Definition 1.5 of filling volume so that the infimum in (1.1) is taken over Finsler manifolds M rather than Riemannian ones. Naturally one has to choose a definition of Finsler volume in (1.1), and the same definition should be used for the surface area in Theorem 3.3. Choosing Loewner's volume definition yields the Riemannian version of the theorem as a special case of the Finslerian one, cf. [28].

3.3. Sketch-proof of theorems 3.1 and 3.2. First I explain how the proof works in the (well-known) case when $g = g_0$, that is, M is a compact region $D \subset \mathbb{R}^n$ equipped with the Euclidean metric.

Let $S = S^{n-1}$ and $\Phi_0 : \mathbb{R}^n \to L^\infty(S)$ be the Busemann representation of the standard Euclidean metric. That is, Φ_0 is a linear map defined by

$$\Phi_0(x) = \langle x, \cdot \rangle|_S, \qquad x \in \mathbb{R}^n, \tag{3.2}$$

where S is identified with the unit sphere in $\mathbb{R}^n$. Denote $W = \Phi_0(\mathbb{R}^n)$ and $B = \Phi_0(D)$. By Theorem 3.3 it suffices to prove that B is a unique Loewner area minimizer in $L^\infty(S)$ among the Lipschitz surfaces with the same boundary. In fact, we can restrict ourselves to surfaces contained in a sufficiently large ball.

Equip $L^\infty(S)$ with a scalar product $\langle \cdot, \cdot \rangle_e$ defined by by

$$\langle u, v \rangle_e = n \int_S uv \, d\mu \tag{3.3}$$

where μ is the Haar probability measure on S. This defines a Euclidean norm on $L^\infty(S)$ that we denote by $\|\cdot\|_e$. One easily sees that $\|\cdot\|_e$ is Lipschitz w.r.t. the L^∞ norm and the two norms coincide on W. An easy application of Cauchy–Schwartz inequality shows that the Euclidean n-volume defined by the above scalar product is no greater than the Loewner n-volume defined by the L^∞ norm. Hence the Euclidean n-area of any Lipschitz surface in $L^\infty(S)$ is no greater that the Loewner n-area, and these areas are equal if the surface is contained in W. Thus it suffices to prove that $\Phi_0(D)$ minimize the Euclidean area among the surfaces with the same boundary. And this is trivial because the orthogonal projection onto W (with respect to our scalar product) does not increase areas.

Furthermore, one can compose the projection with a suitable shrinking in W to obtain a smooth retraction $P : L^\infty(S) \subset L^2(S) \to W$ such that

$$J_n P(u) \leq 1 - c \cdot \|u - P(u)\|_e^2 \tag{3.4}$$

for some $c > 0$ and all u from a large ball in $L^2(S)$. Here J_n denotes the n-dimensional Jacobian with respect to $\|\cdot\|_e$. This proves uniqueness and a sort of stability estimate.

Now consider the general case of Theorem 3.1 when the metric g of $M = (D, g)$ is close to Euclidean in C^r topology for a suitable r (in fact, $r = 3$ is sufficient for the argument presented here and a more delicate argument in [12] works for $r = 2$). The proof of Theorem 3.1 consists of three steps.

Step 1. Construct a smooth distance-preserving map $\Phi : M \to L^\infty(S)$ close to the above linear map Φ_0 (in a suitable topology). In order to do this, one can use a formula similar to (3.1) with Riemannian distances to hyperplanes rather than Busemann functions. By Theorem 3.3, it suffices to prove that $\Phi(M)$ is a unique Loewner area minimizer among the surfaces with the same boundary.

Step 2. Prove that the surface $\Phi(M)$ is minimal in a variational sense. This part of the proof is the most encouraging: it does not depend on the fact that the metric is close to Euclidean and works for any boundary distance representation of a simple metric, any smooth Busemann representation and, in fact, for any isometric embedding with a similar behavior of coordinate functions.

What is meant by being a minimal surface needs clarification. Unfortunately, the first variation of the Loewner area does not make sense since the Loewner area integrand is not differentiable (even in a finite-dimensional Banach space with a smooth norm). To work around this, we differentiate a smooth lower bound for the Loewner area. This lower bound is the n-area defined by a Riemannian metric $\mathcal{G}$ on $L^\infty(S)$ extending the metric of $\Phi(M)$.

The metric $\mathcal{G}$ is a smooth family of scalar products $\langle\cdot,\cdot\rangle_\varphi$, $\varphi \in L^\infty(S)$, on $L^\infty(S)$. Every scalar product $\langle\cdot,\cdot\rangle_\varphi$ is given by a formula similar to (3.3) where μ is replaced by a probability measure μ_φ depending on φ. The normalization of the measures μ_φ implies that the n-area defined by $\mathcal{G}$ is no greater than the Loewner n-area. In order to make $\mathcal{G}$ compatible with the metric of $\Phi(M)$,

one defines μ_φ explicitly for every $\varphi \in \Phi(M)$. Namely if $\varphi = \Phi(x)$ where $x \in M$, then the measure μ_φ is obtained from the normalized Haar measure on the unit sphere $UT_xM \subset T_xM$ via a natural diffeomorphism between UT_xM and S. (This diffeomorphism turns the derivative $d_x\Phi : T_xM \to L^\infty(S)$ into the standard linear map given by (3.2)).

The variational minimality of $\Phi(M)$ means that the first variation of the Riemannian n-area defined by $\mathcal{G}$ is zero for every (Lipschitz) variation, or, equivalently, the mean curvature w.r.t. any normal vector is zero. The proof is a direct computation of the mean curvature. It works for any Riemannian structure $\mathcal{G}$ defined as above, however the next step assumes that $\mathcal{G}$ is a small perturbation of the flat Riemannian structure defined by (3.3).

Step 3. Prove that $\Phi(M)$ is a unique area minimizer with respect to $\mathcal{G}$ provided that $\mathcal{G}$ is sufficiently close (in a suitable topology) to the constant scalar product (3.3). Since the n-area defined by $\mathcal{G}$ is a lower bound for the Loewner n-area and the two areas coincide on $\Phi(M)$, it follows that $\Phi(M)$ is a unique minimizer of the Loewner area and hence M is a minimal filling and boundary rigid.

The proof essentially establishes the fact that stable minimality that we had in the case $g = g_0$ is stable under small perturbations of the data. More precisely, one can construct a retraction from $L^\infty(S)$ to a (minimal) surface containing $\Phi(M)$ by perturbing the area-decreasing map P used in the flat case. The perturbation should preserve the property that pre-images of points are orthogonal to the surface. Since the surface is minimal, this implies that the n-dimensional Jacobian (with respect to $\mathcal{G}$) of the retraction has zero derivatives at the surface. And if its second derivatives are close to the original ones, the inequality (3.4) persists, implying the desired result.

Proof of Theorem 3.2. The proof goes along the same lines: first we prove the desired properties for the standard hyperbolic metric and then verify that they are stable under perturbations.

The only essential difference is the choice of an area non-increasing map in place of the linear orthogonal projection. We define a "projection" $P : L^\infty(S) \to \mathbb{H}^n$ as follows: for every $\varphi \in L^\infty(S^{n-1})$, $P(\varphi)$ is a (unique) point where the function $F_\varphi : \mathbb{H}^n \to \mathbb{R}$ defined by

$$F_\varphi(x) = \int_S e^{-n\varphi(s)} e^{B_{\gamma_s}(x)}\, ds$$

attains its minimum. Here $S = T_o\mathbb{H}^n$ where $o \in \mathbb{H}^n$ is a fixed origin, B_{γ_s} denotes the Busemann function of a geodesic ray starting from the origin in the direction s, and ds denotes the standard measure on S.

One can verify that P does not increase n-dimensional Loewner areas and that $\Phi_0 \circ P$ is a retraction of $L^\infty(S)$ onto $\Phi_0(\mathbb{H}^n)$ where Φ_0 is the Busemann representation of $\mathbb{H}^n$. This proves filling minimality and boundary rigidity for regions in $\mathbb{H}^n$. Then the proof of Theorem 3.2 is similar to that of Theorem 3.1.

4. Finslerian Case

As shown by Theorem 3.3, reducing filling minimality to area minimality is a natural approach (at least there is no loss of generality at this step). But some other tricks in the above proofs are too limited; it would be nice to replace them by a better technique. In particular, replacing the Loewner area by the area defined by an auxiliary Riemannian metric $\mathcal{G}$ is suspicious: this may not work for other minimal fillings, and there is no natural way to choose this auxiliary metric.

It would be more natural to utilize the Finslerian nature of surfaces in L^∞ and work with their natural Finsler areas, e.g. Busemann or Holmes–Thompson areas. Unfortunately very little is known about these surface areas in co-dimensions higher than 1. For example, the following basic question is not yet answered.

Question 4.1 (Busemann [15]). Let V be a finite-dimensional Banach space, D an n-disc in an n-dimensional affine subspace $W \subset V$ and F is an orientable surface in V such that $\partial F = \partial D$. Is it always true that $\mathrm{area}(F) \geq \mathrm{area}(D)$?

In other words, is the n-dimensional area integrand in a Banach space semi-elliptic (over $\mathbb{Z}$)? Actually this is a different question for every definition of area. In the cited paper [15] the question is asked for the Holmes–Thompson area, defined there in terms of the projection function of a convex body. For both Busemann and Holmes–Thompson areas, the answer is known to be affirmative in the case $\dim V = n+1$ but the question is open in higher co-dimensions (even in the special case when the restriction of the Banach norm to the subspace W is Euclidean). Contrary to this, Benson area and Loewner area are known to be semi-elliptic in all dimensions and co-dimensions, cf. [22] and [28].

An affirmative answer to Question 4.1 would have nice applications including a Finslerian generalization of the asymptotic volume estimate (2.1), cf. [10]. It would also imply that every region in an n-dimensional Banach space is a Finslerian minimal filling. This is especially interesting in the case of the Busemann volume because it is equal to the Hausdorff measure naturally defined for all metric spaces, not just Finslerian. Here is how one can formulate a filling question without referencing anything from differential geometry.

Question 4.2. Let d be a (continuous) metric on the standard unit ball $D^n \subset \mathbb{R}^n$ such that

$$d(x,y) \geq d_E(x,y) := |x-y|$$

for all $x, y \in \partial D^n = S^{n-1}$. Is it true that for all such metrics d one has

$$\mathcal{H}^n(D^n, d) \geq \mathcal{H}^n(D^n, d_E)$$

where $\mathcal{H}^n$ denotes the n-dimensional Hausdorff measure?

An affirmative answer to Question 4.1 would answer Question 4.2 for a Lipschitz metric d. I do not know whether the case of a general metric is different.

One may also seek a Finslerian generalization of the minimal filling conjecture 1.6. Although there is no boundary rigidity in the Finslerian case, simple Finsler metrics sharing the same boundary distance function have the same Holmes–Thompson volume. This leaves a possibility that the Finslerian generalization of Conjecture 1.6 might be true if the volume of a Finsler metric is defined as the Holmes–Thompson volume. This generalization is "almost proved" in dimension 2: every simple Finsler metric on D^2 is a minimal filling *among the Finsler fillings homeomorphic to* D^2, cf. [27] and [29].

This implies a partial answer to Question 4.1 for $n = 2$: an affine 2-disc in a Banach space minimizes the Holmes–Thompson area among the surfaces spanning the same boundary and homeomorphic to D^2. On the other hand, one can construct a Banach norm in $\mathbb{R}^4$ such that the resulting two-dimensional Holmes–Thompson area integrand is not convex (that is, it does not admit a convex extension to the exterior product $\Lambda^2\mathbb{R}^4$), cf. [16], [10]. And this implies that there is an affine 2-disc which does *not* minimize the Holmes–Thompson area among the Lipschitz (or polyhedral) chains with rational coefficients, cf. [11]. What is not known is whether an affine 2-disc minimizes area among the chains with integer coefficients, or, equivalently, among the orientable surfaces of arbitrary genus.

References

[1] L. Amrosio and B. Kirchheim, *Rectifiable sets in metric and Banach spaces*, Math. Ann. **318** (2000), 527–555.

[2] V. Bangert, C. Croke, S. Ivanov, M. Katz, *Filling area conjecture and ovalless real hyperelliptic surfaces*, Geom. Func. Anal. **15** (2005), no. 3, 577–597.

[3] R. V. Benson, *Euclidean geometry and convexity*, McGraw–Hill, New York, 1966.

[4] A. S. Besicovitch, *On two problems of Loewner*, J. London Math. Soc. **27** (1952), 141–144

[5] G. Besson, G. Courtois and S. Gallot, *Entropies et rigidités des espaces localement symétriques de courbure strictement négative*, Geom. Funct. Anal., **5** (1995), 731–799.

[6] G. Beylkin, *Stability and uniqueness of the solution of the inverse kinematic problem of seismology in higher dimensions*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. **84** (1979), 3–6 (Russian); J. Soviet Math. **21** (1983), 251–254 (English).

[7] D. Burago. Periodic Metrics. Advances in Soviet Math. **9** (1992), 205–210.

[8] D. Burago, S. Ivanov, *Riemannian tori without conjugate points are flat*, Geom. Funct. Anal. **4** (1994), no.3, 259–269.

[9] D. Burago, S. Ivanov, *On asymptotic volume of tori*, Geom. Funct. Anal. **5** (1995), no. 5, 800–808.

[10] D. Burago, S. Ivanov, *On asymptotic volume of Finsler tori, minimal surfaces in normed spaces, and symplectic filling volume.* Ann. of Math. (2) **156** (2002), no. 3, 891–914.

[11] D. Burago, S. Ivanov, *Gaussian images of surfaces and ellipticity of surface area functionals*, Geom. Funct. Anal. **14** (2004), no. 3, 469–490.

[12] D. Burago and S. Ivanov, *Boundary rigidity and filling volume minimality of metrics close to a flat one*, Ann. of Math. **171** (2010), no. 2, 1183–1211.

[13] D. Burago and S. Ivanov, *Area minimizers and boundary rigidity of almost hyperbolic metrics*, in preparation.

[14] H. Busemann, *Intrinsic area*, Ann. of Math. **48** (1947), 234–267.

[15] H. Busemann, *Convexity on Grassmann manifolds*, Enseignement Math. **7** (1961), 139–152.

[16] H. Busemann, G. Ewald, G. C. Shephard. *Convex bodies and convexity on Grassmann cones. I–IV*, Math. Ann. **151** (1963), 1–41.

[17] C. Croke, *Rigidity and the distance between boundary points*, J. Diff. Geom. **33** (1991), 445–464.

[18] C. Croke, *Volumes of balls in manifolds without conjugate points*, Internat. J. Math. **3** (1992), no. 4, 455–467.

[19] C. Croke, N. Dairbekov and V. Sharafutdinov, *Local boundary rigidity of a compact Riemannian manifold with curvature bounded above*, Trans. Amer. Math. Soc. **352** (2000), no. 9, 3937–3956.

[20] C. Croke and B. Kleiner, *On tori without conjugate points*, Invent. Math. **120** (1995), no. 2, 241–257.

[21] C. Croke and B. Kleiner, *A rigidity theorem for simply connected manifolds without conjugate points*, Ergodic Theory Dynam. Systems **18** (1998), no. 4, 807–812.

[22] M. Gromov, *Filling Riemannian manifolds*, J. Diff. Geom. **18** (1983), 1–147.

[23] M. Gromov, *Metric structures for Riemannian and non-Riemannian spaces*, Progr. in Mathematics, **152**, Birkhäuser, Boston, 1999.

[24] G. Herglotz, *Über das Benndorfsche Problem der Fortpflanzungsgeschwindigkeit der Erdbebenstrahlen*, Physikal. Zeitschr. **8** (1907), 145-147.

[25] R. D. Holmes, A. C. Thompson, *N-dimensional area and content in Minkowski spaces*, Pacific J. Math. **85** (1979), 77–110.

[26] E. Hopf, *Closed surfaces without conjugate points* Proc. Nat. Acad. of Sci. **34** (1948), 47–51.

[27] S. Ivanov, *On two-dimensional minimal fillings*, Algebra i Analiz **13** (2001), no. 1, 26–38 (Russian); St. Petersburg Math. J, **13** (2002), no. 1, 17–25.

[28] S. Ivanov, *Volumes and areas of Lipschitz metrics*, Algebra i Analiz **20** (2008), 74–111 (Russian); St. Petersburg Math. J. **20** (2009), no. 3, 381–405 (English).

[29] S. Ivanov, *Filling minimality of Finslerian 2-discs*, preprint, `arXiv:0910.2257`.

[30] Katz, M. *Systolic geometry and topology*, Mathematical Surveys and Monographs **137**, A.M.S., 2007.

[31] R. Michel, *Sur la rigidité imposeée par la longuer des géodésiques*, Invent. Math. **65** (1981), 71–83.

[32] R. G. Mukhometov, *On a problem of reconstructing Riemannian metrics*, Sibirsk. Mat. Zh. **22** (1981), no. 3, 119–135 (Russian); Siberian Math. J. **22** (1982), no. 3, 420–433 (English).

[33] R. G. Mukhometov and V. G. Romanov, *On the problem of finding an isotropic Riemannian metric in an n-dimensional space*, Dokl. Akad. Nauk SSSR **243** (1978), no. 1, 41–44 (Russian); Soviet Math. Dokl. **19** (1979), no. 6, 1330–1333 (English).

[34] L. Pestov and G. Uhlmann, *Two-dimensional compact simple Riemannian manifolds are boundary distance rigid*, Ann. of Math. (2) **161** (2005), 1093–1110.

[35] P. Pu, *Some inequalities in certain non-orientable Riemannian manifolds*, Pacific J. Math. **2** (1952), 55–71.

[36] L. A. Santaló, *Integral geometry and geometric probability*, Encyclopedia Math. Appl., Addison-Wesley, London, 1976.

[37] P. Stefanov and G. Uhlmann, *Boundary rigidity and stability for generic simple metrics*, J. Amer. Math. Soc. **18** (2005), 975–1003.

[38] A. C. Thompson, *Minkowski Geometry*, Encyclopedia of Math and Its Applications **63**, Cambridge Univ. Press., 1996.

[39] E. Wiechert, *Bestimmung des Weges von Erdbebenwellen im Erdinnern*, Theoretisches. Phys. Z. **11** (1910), 294–311.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Geometric Quantization on Kähler and Symplectic Manifolds

Xiaonan Ma*

Abstract

We explain various results on the asymptotic expansion of the Bergman kernel on Kähler manifolds and also on symplectic manifolds. We also review the "quantization commutes with reduction" phenomenon for a compact Lie group action, and its relation to the Bergman kernel.

Mathematics Subject Classification (2010). Primary 53D; Secondary 58J, 32A.

Keywords. Bergman kernel, Dirac operator, Geometric quantization, Index theorem.

0. Introduction

In the theory of quantization, one attempts to associate to a symplectic manifold (X, ω) a Hilbert space H and a mapping from the space of functions on X into the space of operators on H, and this in a canonical way. The mapping should give some reasonable relationship between the Poisson bracket on the function side and the commutator on the operator side. It is generally acknowledged that there is no canonical way to construct a quantization of X without making use of certain additional structures.

In the theory of the geometric quantization of Kostant and Souriau, (X, ω) is assumed to be prequantizable, that is, there exists a prequantum line bundle (L, h^L, ∇^L) on X (i.e., ω is the first Chern form of L associated with the Hermitian connection ∇^L). Given a compatible almost complex structure J and a Riemannian metric g^{TX}, we can define canonically a Dirac operator D^L acting on $\Omega^{0,\bullet}(X, L)$, the smooth $(0, \bullet)$-forms on X with coefficients in L.

*We wish to express our thanks to Jean-Michel Bismut for discussions on various subjects and for his enlightened support. We would like to thank our collaborators Xianzhe Dai, Kefeng Liu, George Marinescu and Weiping Zhang for many helpful discussions. Thanks are due also to Institut Universitaire de France for support.

Université Paris Diderot - Paris 7, UFR de Mathématiques, Case 7012, Site Chevaleret, 75205 Paris Cedex 13, France. E-mail: ma@math.jussieu.fr.

Assume that X is compact. Following an observation by Bott, we take, as a quantization of X, $\mathrm{Ind}(D^L_+) = \mathrm{Ker}(D^L_+) - \mathrm{Coker}(D^L_+)$ of $D^L_+ := D^L|_{\Omega^{0,\mathrm{even}}}$, which is a formal difference of finite dimensional Hilbert spaces. The virtual dimension of $\mathrm{Ind}(D^L_+)$, which can be computed by the Atiyah-Singer index theorem, does not depend on the choice of the connection and of the metric on L.

For $p \gg 1$, $\mathrm{Ind}(D^{L^p}_+) = \mathrm{Ker}(D^{L^p}_+)$ is an ordinary finite dimensional Hilbert space. The Bergman kernel is defined as the integral kernel $P_p(x, x')$ associated with the orthogonal projection P_p from $\Omega^{0,\bullet}(X, L^p)$ onto $\mathrm{Ker}(D^{L^p})$. We will show that when $p \to +\infty$, the Bergman kernel $P_p(x, x')$ has an asymptotic expansion whose coefficients contain interesting geometric informations about X and L. The kind of expansion obtained for the kernel $P_p(x, x')$ also characterizes the Berezin-Toeplitz operators. Their semi-classical limit provides a precise way to relate the classical and quantum observables.

Assume that a compact connected Lie group G acts on X, and that the action lifts to (L, h^L, ∇^L). Then the quantization of X is a G-virtual representation, and it is interesting to determine the multiplicity of the irreducible representations of G. The Guillemin-Sternberg conjecture "quantization commutes with reduction" gives a precise geometric answer to this problem by using the associated moment map. Here we explain the behavior of the G-invariant part of $P_p(x, x')$ as $p \to +\infty$, and we relate this behavior to the Guillemin-Sternberg conjecture.

New difficulties appear when the manifold X is no longer supposed to be compact, since in this case $\mathrm{Ind}(D^L_+)$ is not well defined. In her ICM 2006 plenary lecture, Michèle Vergne proposed to replace $\mathrm{Ind}(D^L_+)$ by a certain transversal index introduced by Atiyah, under the natural hypothesis that the moment map is proper, and that the zero-set of the vector field induced by the moment map is compact. She conjectured that "quantization commutes with reduction" still holds in this case.

If (X, ω, J) is a compact Kähler manifold and if L is holomorphic, then for $p \gg 1$, $\mathrm{Ker}(D^{L^p})$ is the space of holomorphic sections $H^0(X, L^p)$ of L^p on X. This leads to many applications of the asymptotic expansion of the Bergman kernel in Kähler geometry.

We refer the reader to our book with Marinescu [41] for a comprehensive study of the Bergman kernel and applications, and to the survey by Michèle Vergne [68] on the Guillemin-Sternberg conjecture. One can find more comments, references and motivations in [41] and [68].

This paper is organized as follows. The first two sections are based on our work with Dai, Liu and Marinescu, the last two sections are based on our work with Zhang. In Section 1, we review the definition of Bergman kernel and Berezin-Toeplitz quantization.

In Section 2, we discuss the asymptotic expansion of the Bergman kernel, and also Toeplitz operators.

In Section 3, we examine the corresponding results when a compact Lie group G acts on X and the action lifts to L.

In Section 4, we outline Ma-Zhang's solution of the Vergne conjecture.

1. Quantization on Symplectic Manifolds

In Section 1.1, we review the basic definitions, and the spectral gap property of the Dirac operator. Then we explain the model example $\mathbb{C}^n$ in Section 1.2.

1.1. Dirac operators and quantization. Let (X, ω) be a compact symplectic manifold of real dimension $2n$ with compatible almost complex structure J, i.e., $\omega(\cdot, J\cdot) > 0$, $\omega(J\cdot, J\cdot) = \omega(\cdot, \cdot)$. We endow X with a Riemannian metric g^{TX} compatible with J, i.e., $g^{TX}(J\cdot, J\cdot) = g^{TX}(\cdot, \cdot)$. Let (E, h^E) be a Hermitian vector bundle on X with Hermitian connection ∇^E and curvature $R^E = (\nabla^E)^2$.

The almost complex structure J induces a splitting of the complexification of the tangent bundle, $TX \otimes_{\mathbb{R}} \mathbb{C} = T^{(1,0)}X \oplus T^{(0,1)}X$, where $T^{(1,0)}X$ and $T^{(0,1)}X$ are the eigenbundles of J corresponding to the eigenvalues $\sqrt{-1}$ and $-\sqrt{-1}$ respectively. Let $T^{*\,(0,1)}X$ be the dual space of $T^{(0,1)}X$. For any $v \in T^{(1,0)}X$, let $\overline{v}^* \in T^{*\,(0,1)}X$ be the metric dual of v, then

$$\mathbf{c}(v) = \sqrt{2}\,\overline{v}^*\wedge, \quad \mathbf{c}(\overline{v}) = -\sqrt{2}\,i_{\overline{v}}, \tag{1.1}$$

define the Clifford actions of $v, \overline{v}$ on $\Lambda^{0,\bullet} := \Lambda^\bullet(T^{*\,(0,1)}X)$, where $\wedge$ and i denote the exterior and interior multiplications respectively.

Consider the Levi–Civita connection ∇^{TX} of (TX, g^{TX}) with associated curvature R^{TX}. Let $\nabla^{T^{(1,0)}X}$ be the connection on $T^{(1,0)}X$ induced by projecting ∇^{TX}; $\nabla^{T^{(1,0)}X}$ induces the connection $\nabla^{\det}$ on $\det(T^{(1,0)}X)$. The Clifford connection ∇^{Cl} on $\Lambda^{0,\bullet}$ is induced canonically by ∇^{TX} and $\nabla^{\det}$ (cf. [41, §1.3]). Finally, let $\nabla^{\Lambda^{0,\bullet}\otimes E}$ be the connection on $\Lambda^{0,\bullet} \otimes E$ induced by ∇^{Cl} and ∇^E.

Let dv_X be the Riemannian volume form of (TX, g^{TX}) and $\Omega^{0,\bullet}(X, E)$ be the space of smooth sections of $\Lambda^{0,\bullet} \otimes E$ endowed with the L^2-norm $\|\cdot\|_{L^2}$ induced by h^E, g^{TX}. Let $\{e_j\}_{j=1}^{2n}$ be an orthonormal frame of (TX, g^{TX}).

Definition 1.1. The *spin^c Dirac operator* D^E is defined by

$$D^E := \sum_j \mathbf{c}(e_j)\nabla_{e_j}^{\Lambda^{0,\bullet}\otimes E} : \Omega^{0,\bullet}(X, E) \longrightarrow \Omega^{0,\bullet}(X, E)\,, \quad D^E_{\pm} := D^E|_{\Omega^{0,\,\frac{\mathrm{even}}{\mathrm{odd}}}}\,. \tag{1.2}$$

The operator D^E is a formally self–adjoint, first order elliptic differential operator on $\Omega^{0,\bullet}(X, E)$, which interchanges $\Omega^{0,\mathrm{even}}(X, E)$ and $\Omega^{0,\mathrm{odd}}(X, E)$ (cf. [41, §1.3]).

Thus $\operatorname{Ker}\left(D_+^E\right)$, $\operatorname{Ker}\left(D_-^E\right)$ are finite dimensional Hilbert spaces and the *quantization space* of E is defined as their formal difference

$$Q(E) := \operatorname{Ind}(D_+^E) := \operatorname{Ker}\left(D_+^E\right) - \operatorname{Ker}\left(D_-^E\right). \tag{1.3}$$

The Atiyah-Singer index theorem [3, §4.1], [41, Th. 1.3.9] allows us to compute the virtual dimension of $Q(E)$ by using characteristic numbers:

$$\dim Q(E) = \int_X \operatorname{Td}(T^{(1,0)}X)\operatorname{ch}(E), \tag{1.4}$$

where $\operatorname{ch}(\cdot)$, $\operatorname{Td}(\cdot)$ are the Chern character and the Todd class of the corresponding complex vector bundles. In particular, the virtual dimension of $Q(E)$ does not depend on the choice of J, g^{TX} or the metric and connection on E. If $\operatorname{Ker}(D_-^E) = 0$, then the quantization space $Q(E)$ is an ordinary vector space.

We explain now the idea of the geometric quantization introduced by Kostant [33] and Souriau [62]. Let (L, h^L) be a Hermitian line bundle over X endowed with a Hermitian connection ∇^L with curvature $R^L = (\nabla^L)^2$. We assume that (L, h^L, ∇^L) satisfies the *prequantization condition*, that is

$$\omega = \tfrac{\sqrt{-1}}{2\pi} R^L\,. \tag{1.5}$$

For $p \in \mathbb{N}$, we denote by $D^{L^p\otimes E}$ the Dirac operator associated to $L^p \otimes E$ with $L^p := L^{\otimes p}$, and set

$$E_p := \Lambda^{0,\bullet} \otimes L^p \otimes E, \quad D_p := D^{L^p\otimes E}, \quad D_{\pm,p} := D_p|_{\Omega^{0,\frac{\text{even}}{\text{odd}}}}\,. \tag{1.6}$$

Let $L^2(X, E_p)$ be the L^2-completion of $(\Omega^{0,\bullet}(X, L^p \otimes E), \|\cdot\|_{L^2})$.

The following result is the starting point of the asymptotic expansion results for the Bergman kernel which we describe in the sequel. The proof is based on a direct application of the Lichnerowicz formula for D_p^2.

Theorem 1.2 (Ma-Marinescu [37, Th. 1.1, 2.5], [41, Th. 1.5.5]). *There exists $C > 0$ such that for any $p \in \mathbb{N}$, the spectrum of D_p^2 satisfies*

$$\operatorname{Spec}(D_p^2) \subset \{0\} \cup [2p\nu_0 - C, +\infty[\,, \tag{1.7a}$$

$$\operatorname{Ker}(D_{-,p}) = 0 \quad \textit{for } p \gg 1\,, \tag{1.7b}$$

where $\nu_0 = \inf\{R_x^L(u, \overline{u}) : u \in T_x^{(1,0)}X, |u|^2 = 1, x \in X\} > 0$.

Thus for $p \gg 1$, $Q(L^p \otimes E) = \operatorname{Ker}(D_p^2)$ is an ordinary vector space and its dimension is a polynomial in p of degree n given by (1.4). The analogue of Theorem 1.2 in the holomorphic setting was first obtained by Bismut and Vasserot [8, Th. 1.1] by using Demailly's version of the Bochner-Kodaira-Nakano formula (cf. [41, Th. 1.4.12]). Formula (1.7b) was first established by Borthwick-Uribe [10, Th. 2.3] and Braverman [14, Th. 2.6] by using Melin's inequality. Mathai-Zhang [46, Th. 1.3] obtained a version of (1.7b) for the proper cocompact group action case by applying the method in [37].

Definition 1.3. The orthogonal projection $P_p : L^2(X, E_p) \longrightarrow \mathrm{Ker}(D_p)$ is called the *Bergman projection*. The *Bergman kernel* of D_p is the smooth kernel $P_p(x, x') \in E_{p,x} \otimes E^*_{p,x'}$, $(x, x' \in X)$, of P_p with respect to $dv_X(x')$, i.e., for any $s \in L^2(X, E_p)$, we have

$$(P_p\, s)(x) = \int_X P_p(x, x') s(x')\, dv_X(x')\,. \tag{1.8}$$

For $f \in \mathscr{C}^\infty(X, \mathrm{End}(E))$, set

$$T_{f,\,p} : L^2(X, E_p) \longrightarrow L^2(X, E_p)\,, \quad T_{f,\,p} = P_p\, f\, P_p\,. \tag{1.9}$$

Here the action of f is the pointwise multiplication by f. The map which associates to $f \in \mathscr{C}^\infty(X, \mathrm{End}(E))$ the family of bounded operators $\{T_{f,\,p}\}_p$ on $L^2(X, E_p)$ is called the *Berezin-Toeplitz quantization*.

Definition 1.4. A *Toeplitz operator* is a sequence $\{T_p\}_{p\in\mathbb{N}}$ of linear operators $T_p : L^2(X, E_p) \longrightarrow L^2(X, E_p)$ satisfying $T_p = P_p\, T_p\, P_p$, such that there exists a sequence $g_l \in \mathscr{C}^\infty(X, \mathrm{End}(E))$ such that for all $k \geqslant 0$, there exists $C_k > 0$ with

$$\left\| T_p - \sum_{l=0}^{k} T_{g_l,p}\, p^{-l} \right\| \leqslant C_k\; p^{-k-1} \qquad \text{for any } p \in \mathbb{N}^*, \tag{1.10}$$

where $\|\cdot\|$ denotes the operator norm on the space of bounded operators. The section g_0 is called the *principal symbol* of $\{T_p\}$.

We express (1.10) symbolically by

$$T_p = \sum_{l=0}^{k} T_{g_l,p}\, p^{-l} + \mathcal{O}(p^{-k-1}). \tag{1.11}$$

If (1.10) holds for any $k \in \mathbb{N}$, then we write (1.11) with $k = +\infty$.

The Poisson bracket $\{\,\cdot\,,\,\cdot\,\}$ on (X, ω) is defined as follows. For $f, g \in \mathscr{C}^\infty(X)$, let $\xi_f \in \mathscr{C}^\infty(X, TX)$ be defined by $2\pi i_{\xi_f}\omega = df$. Then $\{f, g\} := \xi_f(dg)$.

In the spirit of the geometric quantization, (X, ω) represents the classical phase space and the Poisson algebra $(\mathscr{C}^\infty(X), \{\cdot\})$ represents the classical observables, while $\mathrm{Ker}(D_p)$ is the quantum space and the linear operators on $\mathrm{Ker}(D_p)$ are the quantum observables. The process $p \to +\infty$ is called the semiclassical limit, which is a way to relate the classical and quantum observables.

1.2. Bergman kernel on $\mathbb{C}^n$.

Let us consider the canonical real coordinates $(Z_1, \ldots, Z_{2n})$ on $\mathbb{R}^{2n}$ and the complex coordinates $(z_1, \ldots, z_n)$ on $\mathbb{C}^n$. The two sets of coordinates are linked by the relation $z_j = Z_{2j-1} + \sqrt{-1}Z_{2j}$, $j = 1, \ldots, n$. We consider the L^2-norm $\|\cdot\|_{L^2} = \big(\int_{\mathbb{R}^{2n}} |\cdot|^2\, dZ\big)^{1/2}$ on the obvious L^2-space on $\mathbb{R}^{2n}$, with $dZ = dZ_1 \cdots dZ_{2n}$ the Lebesgue measure. For $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, $z \in \mathbb{C}^n$, put $z^\alpha = z_1^{\alpha_1} \cdots z_n^{\alpha_n}$.

Let $L = \mathbb{C}$ be the trivial holomorphic line bundle on $\mathbb{C}^n$ with the canonical section $\mathbf{1} : \mathbb{C}^n \to L$, $z \mapsto (z, 1)$. Let h^L be the metric on L defined by

$$|\mathbf{1}|_{h^L}(z) := e^{-\frac{1}{4}\sum_{j=1}^n a_j |z_j|^2} = \rho(Z) \quad \text{for } z \in \mathbb{C}^n, \tag{1.12}$$

with $a_j > 0$ for $j \in \{1, \dots, n\}$. The space of L^2-integrable holomorphic sections of L with respect to h^L and dZ is the classical Segal-Bargmann space of L^2-integrable holomorphic functions with respect to the volume form $\rho\, dZ$. It is well-known that $\{z^\beta : \beta \in \mathbb{N}^n\}$ forms an orthogonal basis of this space.

To introduce the model operator $\mathscr{L}$ we set:

$$b_i = -2\frac{\partial}{\partial z_i} + \frac{1}{2} a_i \overline{z}_i\,, \quad b_i^+ = 2\frac{\partial}{\partial \overline{z}_i} + \frac{1}{2} a_i z_i\,, \quad \mathscr{L} = \sum_i b_i\, b_i^+\,. \tag{1.13}$$

We can interpret the operator $\mathscr{L}$ in terms of complex geometry. Let $\overline{\partial}^{L*}$ be the adjoint of the Dolbeault operator $\overline{\partial}^L$ on (L, h^L) over $(\mathbb{C}^n, \frac{\sqrt{-1}}{2}\sum_j dz_j \wedge d\overline{z}_j)$. We have the isometry $\Omega^{0,\bullet}(\mathbb{C}^n, \mathbb{C}) \to \Omega^{0,\bullet}(\mathbb{C}^n, L)$ given by $\alpha \mapsto \rho^{-1}\alpha$. If $\square^L = \overline{\partial}^{L*}\overline{\partial}^L + \overline{\partial}^L\overline{\partial}^{L*}$ denotes the Kodaira Laplacian acting on $\Omega^{0,\bullet}(\mathbb{C}^n, L)$, then $\rho\square^L\rho^{-1} : \Omega^{0,\bullet}(\mathbb{C}^n, \mathbb{C}) \to \Omega^{0,\bullet}(\mathbb{C}^n, \mathbb{C})$ is given by $\frac{1}{2}\mathscr{L} + \sum_j a_j d\overline{z}^j \wedge i\frac{\partial}{\partial \overline{z}_j}$, and its restriction on functions is $\frac{1}{2}\mathscr{L}$.

The operator $\mathscr{L}$ is the complex analogue of the harmonic oscillator, the operators b, b^+ are creation and annihilation operators respectively. Each eigenspace of $\mathscr{L}$ has infinite dimension, but we can still give an explicit description.

Theorem 1.5 (Ma-Marinescu [38, Th. 1.15], [41, Th. 4.1.20]). *The spectrum of $\mathscr{L}$ on $L^2(\mathbb{R}^{2n})$ is given by*

$$\mathrm{Spec}(\mathscr{L}) = \left\{ 2\sum_{i=1}^n \alpha_i a_i \,:\, \alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n \right\} \tag{1.14}$$

and an orthogonal basis of the eigenspace of $\lambda \in \mathrm{Spec}(\mathscr{L})$ is given by

$$B_\lambda = \left\{ b^\alpha \big(z^\beta \exp\big(-\tfrac{1}{4}\textstyle\sum_i a_i |z_i|^2\big)\big) \,:\, 2\textstyle\sum_i \alpha_i a_i = \lambda, \text{ with } \alpha, \beta \in \mathbb{N}^n \right\} \tag{1.15}$$

where $b^\alpha := b_1^{\alpha_1} \cdots b_n^{\alpha_n}$. Moreover, $\cup_\lambda \{B_\lambda : \lambda \in \mathrm{Spec}(\mathscr{L})\}$ forms a complete orthogonal basis of $L^2(\mathbb{R}^{2n})$.

Let $\mathscr{P}(Z, Z')$ be the smooth kernel of $\mathscr{P}$, which is the orthogonal projection from $(L^2(\mathbb{R}^{2n}), \|\cdot\|_{L^2})$ onto $\mathrm{Ker}(\mathscr{L})$, with respect to dZ'. Then $\mathscr{P}(Z, Z')$ is the classical *Bergman kernel* on $\mathbb{C}^n$ given by

$$\mathscr{P}(Z, Z') = \prod_{i=1}^n \frac{a_i}{2\pi} \ \exp\left(-\frac{1}{4}\sum_i a_i \big(|z_i|^2 + |z_i'|^2 - 2 z_i \overline{z}_i' \big) \right). \tag{1.16}$$

2. Asymptotic Expansion of Toeplitz Operators

The starting point for our work on the asymptotic expansion of the Bergman kernel has been the heat equation proof by Bismut [6] of Demailly's holomorphic Morse inequalities [21]. For a unified treatment of these two questions, we refer to the book [41]. Here, we give various results on expansions of Bergman kernels, and also on Toeplitz operators.

This Section is organized as follows. In Section 2.1, we give the asymptotic expansion of the Bergman kernel.

In Section 2.2, we describe a characterization of the Toeplitz operators in terms of their asymptotic expansion.

In Section 2.3, we specify the results to the Kähler case.

We will use the notation and assumptions of Section 1.1.

2.1. Asymptotic expansion of Bergman kernel. Let $d^X(x,x')$ be the Riemannian distance between x, $x' \in X$. Let a^X be the injectivity radius of (X, g^{TX}). We denote by $B^X(x,\varepsilon)$ and $B^{T_xX}(0,\varepsilon)$ the open balls in X and T_xX with centers x and 0 and radius ε, respectively. Then the exponential map $T_xX \ni Z \to \exp_x^X(Z) \in X$ is a diffeomorphism from $B^{T_xX}(0,\varepsilon)$ onto $B^X(x,\varepsilon)$ for $\varepsilon \leqslant a^X$. From now on, we identify $B^{T_xX}(0,\varepsilon)$ with $B^X(x,\varepsilon)$ via the exponential map for $\varepsilon \leqslant a^X$. When a function is calculated using normal coordinates based at x, we will add a subscript x.

We fix $x_0 \in X$. For $Z \in B^{T_{x_0}X}(0,\varepsilon)$, we identify $E_{p,Z}$ with E_{p,x_0} by parallel transport with respect to the connection $\nabla^{E_p} := \nabla^{\Lambda^{0,\bullet}\otimes L^p\otimes E}$ along the curve $\gamma_Z : [0,1] \ni u \to uZ$.

Let dv_{TX} be the Riemannian volume form on $(T_{x_0}X, g^{T_{x_0}X})$. There exists a smooth positive function κ_{x_0} on $B^{T_{x_0}X}(0,\varepsilon)$ defined by

$$dv_X(Z) = \kappa_{x_0}(Z) dv_{TX}(Z), \quad \kappa_{x_0}(0) = 1. \tag{2.1}$$

We will identify the 2-form R^L with the Hermitian matrix $\dot{R}^L \in \mathrm{End}(T^{(1,0)}X)$ such that for $W, Y \in T^{(1,0)}X$, $R^L(W,\overline{Y}) = \langle \dot{R}^L W, \overline{Y}\rangle$. We choose an orthonormal basis $\{w_i\}_{i=1}^n$ of $T_{x_0}^{(1,0)}X$ such that

$$\dot{R}^L(x_0) = \mathrm{diag}(a_1(x_0), \cdots, a_n(x_0)) \in \mathrm{End}(T_{x_0}^{(1,0)}X) \quad \text{with } a_j(x_0) > 0\,. \tag{2.2}$$

Then $e_{2j-1} = \frac{1}{\sqrt{2}}(w_j + \overline{w}_j)$ and $e_{2j} = \frac{\sqrt{-1}}{\sqrt{2}}(w_j - \overline{w}_j)\,, j = 1, \ldots, n\,,$ form an orthonormal basis of $T_{x_0}X$. We use the identification $(Z_1, \ldots, Z_{2n}) \in \mathbb{R}^{2n} \longrightarrow \sum_i Z_i e_i \in T_{x_0}X$. In what follows, we also use the corresponding complex coordinates $z = (z_1, \ldots, z_n)$ on $\mathbb{C}^n \simeq \mathbb{R}^{2n}$.

Let $\pi : TX \times_X TX \to X$ be the obvious projection. Let $\{\Theta_p\}_{p\in\mathbb{N}}$ be a sequence of linear operators $\Theta_p : L^2(X, E_p) \longrightarrow L^2(X, E_p)$ with smooth kernels $\Theta_p(x,y)$ with respect to $dv_X(y)$. In terms of our trivialization, $\Theta_p(x,y)$ induce smooth sections $\Theta_{p,\,x_0}(Z,Z')$ of $\pi^*(\mathrm{End}(\Lambda^{0,\bullet} \otimes E))$ over $TX \times_X TX$, with $Z, Z' \in T_{x_0}X$. Recall that $\mathscr{P}_{x_0} = \mathscr{P}$ was defined in (1.16).

Notation 2.1. Let $\{Q_{r,x_0}\}_{0\leqslant r\leqslant k, x_0\in X}$ be a family $Q_{r,x_0} \in \mathrm{End}(\Lambda^{0,\bullet}\otimes E)_{x_0}[Z,Z']$ of polynomials in Z, Z', smooth with respect to the parameter $x_0 \in X$. We will write

$$p^{-n}\Theta_{p,x_0}(Z,Z') \cong \sum_{r=0}^{k}(Q_{r,x_0}\mathscr{P}_{x_0})(\sqrt{p}Z,\sqrt{p}Z')p^{-\frac{r}{2}} + \mathcal{O}(p^{-\frac{k+1}{2}}), \tag{2.3}$$

if there exist $\varepsilon' \in]0, a^X[$, $C_0 > 0$ with the following property: for any $l \in \mathbb{N}$, there exist $C_{k,l} > 0$, $M > 0$ such that for any $x_0 \in X$, $Z, Z' \in T_{x_0}X$, $|Z|, |Z'| < \varepsilon'$ and $p \in \mathbb{N}^*$, the following estimate holds:

$$\begin{aligned}&\left|p^{-n}\Theta_{p,x_0}(Z,Z')\kappa_{x_0}^{\frac{1}{2}}(Z)\kappa_{x_0}^{\frac{1}{2}}(Z') - \sum_{r=0}^{k}(Q_{r,x_0}\mathscr{P}_{x_0})(\sqrt{p}Z,\sqrt{p}Z')p^{-\frac{r}{2}}\right|_{\mathscr{C}^l(X)}\\ &\leqslant C_{k,l}\,p^{-\frac{k+1}{2}}(1+\sqrt{p}\,|Z|+\sqrt{p}\,|Z'|)^M\exp(-\sqrt{C_0p}\,|Z-Z'|) + \mathscr{O}(p^{-\infty}).\end{aligned} \tag{2.4}$$

Here $|\cdot|_{\mathscr{C}^l(X)}$ is the $\mathscr{C}^l$ norm with respect to the parameter $x_0 \in X$.

If $K \subset X \times X$ is compact, we will write that as $p \to +\infty$, $P_p(x,x') = \mathscr{O}(p^{-\infty})$ for $x, x' \in K$ if for any $k, l \in \mathbb{N}$, the $\mathscr{C}^l$ norm of $P_p(x,x')$ for $x, x' \in K$ with respect to the connections ∇^L, ∇^E and the metrics h^L, h^E, g^{TX} is dominated by Cp^{-k}.

We denote by $I_{\mathbb{C}\otimes E}$ the projection from $\Lambda^{0,\bullet}\otimes E$ onto $\mathbb{C}\otimes E$ relative to the decomposition $\Lambda^{0,\bullet} = \mathbb{C}\oplus\Lambda^{0,>0}$.

We have the following full asymptotic expansion of the Bergman kernel.

Theorem 2.2 (Dai-Liu-Ma [20, Prop. 4.1 and Th. 4.18′], [41, Th. 8.1.4]). *For any $x_0 \in X$ and $r \in \mathbb{N}$, there exist polynomials $J_{r,x_0}(Z,Z') \in \mathrm{End}(\Lambda^{0,\bullet}\otimes E)_{x_0}$ in Z, Z' with the same parity as r and with $\deg J_{r,x_0} \leqslant 3r$, whose coefficients are functions of the curvatures and their derivatives, such that for any $k \in \mathbb{N}$, in the sense of Notation 2.1,*

$$p^{-n}P_{p,x_0}(Z,Z') \cong \sum_{r=0}^{k}(J_{r,x_0}\mathscr{P}_{x_0})(\sqrt{p}Z,\sqrt{p}Z')p^{-\frac{r}{2}} + \mathcal{O}(p^{-\frac{k+1}{2}}), \tag{2.5}$$

with $J_{0,x_0} = I_{\mathbb{C}\otimes E}$. Moreover, for any $\varepsilon > 0$, we have

$$P_p(x,x') = \mathscr{O}(p^{-\infty}) \quad \text{if } d^X(x,x') \geqslant \varepsilon. \tag{2.6}$$

Idea of the proof. Using the spectral gap property in Theorem 1.2, and finite propagation speed of solutions of hyperbolic equations, we get (2.6). Also we can localize the asymptotics of $P_p(x_0,x')$ in the neighborhood of x_0. The second step consists in working on $\mathbb{R}^{2n}$. To conclude the proof, we combine the spectral gap property, the rescaling of the coordinates and functional analytic techniques inspired by Bismut-Lebeau [7, §11]. □

By taking $\boldsymbol{b}_r(x_0) = (J_{2r,\,x_0}\mathscr{P}_{x_0})(0,0)$, we get from (2.5) that for any $k, l \in \mathbb{N}$, there exists $C_{k,l} > 0$ such that for any $p \in \mathbb{N}^*$,

$$\left| P_p(x,x) - \sum_{r=0}^{k} \boldsymbol{b}_r(x) p^{n-r} \right|_{\mathscr{C}^l(X)} \leqslant C_{k,l} p^{n-k-1}. \tag{2.7}$$

We will give an algorithm to compute the coefficients J_{r,x_0} in the expansion, by using a formal power series trick.

For $s \in \mathscr{C}^\infty(\mathbb{R}^{2n}, (\Lambda^{0,\bullet} \otimes E)_{x_0})$, $Z \in \mathbb{R}^{2n}$, $|Z| \leqslant \varepsilon$, and for $t = \frac{1}{\sqrt{p}}$, set

$$(S_t s)(Z) := s(Z/t), \quad \mathscr{L}_t := S_t^{-1} \kappa^{1/2}\, t^2 D_p^2 \kappa^{-1/2} S_t. \tag{2.8}$$

By [20, Th. 4.6] (cf. [41, Th. 4.1.7]), there exist second order differential operators $\mathcal{O}_r$ such that for any $m \in \mathbb{N}$, we have an asymptotic expansion when $t \to 0$,

$$\mathscr{L}_t = \mathscr{L}_0 + \sum_{r=1}^{m} t^r \mathcal{O}_r + \mathscr{O}(t^{m+1}), \quad \text{with } \mathscr{L}_0 = \mathscr{L} + 2\sum_j a_j \overline{w}^j \wedge i_{\overline{w}_j}. \tag{2.9}$$

Then $P^N = I_{\mathbb{C}\otimes E}\mathscr{P}$ is the orthogonal projection of $(L^2(\mathbb{R}^{2n}, (\Lambda^{0,\bullet} \otimes E)_{x_0}), \|\cdot\|_{L^2})$ onto $N = \operatorname{Ker}(\mathscr{L}_0)$. Set $P^{N^\perp} = \operatorname{Id} - P^N$. We define by recursion $f_r(\lambda) \in \operatorname{End}(L^2(\mathbb{R}^{2n}, (\Lambda^{0,\bullet} \otimes E)_{x_0}))$ by

$$f_0(\lambda) = (\lambda - \mathscr{L}_0)^{-1}, \quad f_r(\lambda) = (\lambda - \mathscr{L}_0)^{-1} \sum_{j=1}^{r} \mathcal{O}_j f_{r-j}(\lambda). \tag{2.10}$$

Let δ be the counterclockwise oriented circle in $\mathbb{C}$ of center 0 and radius $\nu_0/2$.

We denote by $\mathscr{F}_{r,x_0}$ the operator with smooth kernel

$$\mathscr{F}_{r,x_0}(Z,Z') = J_{r,x_0}(Z,Z')\mathscr{P}(Z,Z') \tag{2.11}$$

with respect to dZ'. Then by [38, (1.110)] (cf. also [41, (4.1.91)])

$$\mathscr{F}_{r,x_0} = \frac{1}{2\pi\sqrt{-1}} \int_\delta f_r(\lambda) d\lambda. \tag{2.12}$$

By Theorem 1.5, (2.10), (2.12) and by the residue formula, we can express $\mathscr{F}_{r,x_0}$ in terms of $\mathscr{L}_0^{-1}$, P^N, $P^{N^\perp}$, $\mathcal{O}_k$ (with $k \leqslant r$). This gives a direct method to compute $\mathscr{F}_{r,x_0}$. In [39, §2], we find an explicit computation for $\mathscr{F}_{2,x_0}$ when $\omega(\cdot,\cdot) = g^{TX}(J\cdot,\cdot)$ (i.e., $\dot{R}^L = 2\pi \operatorname{Id}$). We have in particular:

Theorem 2.3 (Ma-Marinescu [39, Th. 2.1]). *If $\omega(\cdot,\cdot) = g^{TX}(J\cdot,\cdot)$, we have*

$$\operatorname{Tr}|_{\Lambda(T^{*(0,1)}X)}[\boldsymbol{b}_1(x)] = \frac{1}{8\pi}\left[r^X + \frac{1}{4}|\nabla^X J|^2 + 4\sum_j R^E(w_j, \overline{w}_j) \right]. \tag{2.13}$$

Here $\nabla^X J$ is the covariant derivative of J with respect to ∇^{TX}, and r^X is the scalar curvature of (X, g^{TX}). In Donaldson [22], the term $r^X + \frac{1}{4}|\nabla^X J|^2$ in (2.13) is called the Hermitian scalar curvature. It is a natural substitute for the Riemannian scalar curvature in the almost-Kähler case. It was used by Donaldson to define the moment map on the space of compatible almost-complex structures.

Ma-Zhang [44] obtained a family version of Theorem 2.2.

2.2. Asymptotic expansion of Toeplitz operators. Here is a useful characterization of the Toeplitz operators in terms of their kernel.

Theorem 2.4. (Ma-Marinescu [40, Th. 4.9, Rem. 4.10], [41, Lemmas 7.2.2, 7.2.4, Th.7.3.1]) *Let $\{T_p : L^2(X, E_p) \longrightarrow L^2(X, E_p)\}$ be a family of bounded linear operators. Then $\{T_p\}$ is a Toeplitz operator if and only if it satisfies the following three conditions:*

(i) For any $p \in \mathbb{N}$, $P_p T_p P_p = T_p$.

(ii) For any $\varepsilon_0 > 0$, $T_p(x, x') = \mathscr{O}(p^{-\infty})$ if $d^X(x, x') \geqslant \varepsilon_0$.

(iii) There exists a family of polynomials $\{Q_{r,x_0} \in \mathrm{End}(\Lambda^{0,\bullet} \otimes E)_{x_0}[Z, Z']\}_{x_0 \in X}$ which has the same parity as r, such that for any $k \in \mathbb{N}$, we have in the sense of (2.3) and (2.4),

$$p^{-n} T_{p,x_0}(Z, Z') \cong \sum_{r=0}^{k} (Q_{r,x_0} \mathscr{P}_{x_0})(\sqrt{p}Z, \sqrt{p}Z') p^{-\frac{r}{2}} + \mathcal{O}(p^{-\frac{k+1}{2}}). \quad (2.14)$$

In this case, its principal symbol is $g_0(x_0) = Q_{0,x_0}(0,0)|_{\mathbb{C}\otimes E} \in \mathrm{End}(E_{x_0})$.

Remark 2.5. For $f \in \mathscr{C}^\infty(X, \mathrm{End}(E))$, conditions (i), (ii), (iii) of Theorem 2.4 for $\{T_{f,p}\}$ are consequences of Theorem 2.2 and of the Taylor expansion of f at x_0. The coefficients Q_{r,x_0} in (2.14) corresponding to the Toeplitz operator $\{T_{f,p}\}$ are denoted by $Q_{r,x_0}(f)$, and $Q_{0,x_0}(f) = f(x_0) I_{\mathbb{C}\otimes E}$.

By taking $\boldsymbol{b}_{r,f}(x_0) = (Q_{2r,x_0}(f)\mathscr{P}_{x_0})(0,0)$, we get from (2.14) that for any $k, l \in \mathbb{N}$, there exists $C_{k,l} > 0$ such that for any $p \in \mathbb{N}^*$, we have

$$\left| T_{f,p}(x,x) - \sum_{r=0}^{k} \boldsymbol{b}_{r,f}(x) p^{n-r} \right|_{\mathscr{C}^l(X)} \leqslant C_{k,l} p^{n-k-1}. \quad (2.15)$$

In [40, (4.15)] (cf. also [41, (7.2.16)]), we find a precise formula for $Q_{r,x_0}(f)$ by using the Taylor expansion of f at x_0, $J_{j,x_0} (j \leqslant r)$ and $\mathscr{P}_{x_0}$ in (2.5), from which the computation $\boldsymbol{b}_{r,f}(x_0)$ can be derived.

Theorem 2.6 (Ma-Marinescu [40, Th. 1.1], [41, Th. 7.4.1]). *The product of the Toeplitz operators $T_{f,p}$ and $T_{g,p}$, with $f, g \in \mathscr{C}^\infty(X, \mathrm{End}(E))$, is a Toeplitz*

operator, i.e., it admits the asymptotic expansion in the sense of (1.11):

$$T_{f,p}\,T_{g,p} = \sum_{r=0}^{\infty} p^{-r} T_{C_r(f,g),\,p} + \mathcal{O}(p^{-\infty}), \tag{2.16}$$

where C_r *are bidifferential operators,* $C_0(f,g) = fg$ *and* $C_r(f,g) \in \mathscr{C}^\infty(X, \mathrm{End}(E))$.
If $f, g \in (\mathscr{C}^\infty(X), \{\cdot,\cdot\})$ *with the Poisson bracket defined in Section 1.1, we have*

$$[T_{f,p}\,, T_{g,p}] = \frac{\sqrt{-1}}{p} T_{\{f,g\},\,p} + \mathcal{O}(p^{-2}). \tag{2.17}$$

Theorem 2.6 implies that the set of Toeplitz operators is closed under the composition of operators, and so it forms an associative algebra.

For $E = \mathbb{C}$, Theorem 2.6 shows that we can associate to $f, g \in \mathscr{C}^\infty(X)$ a formal power series $\sum_{l=0}^{\infty} \hbar^l C_l(f,g) \in \mathscr{C}^\infty(X)[[\hbar]]$, where C_l are bidifferential operators. Therefore, we have constructed in a canonical way an associative star-product $f * g = \sum_{l=0}^{\infty} \hbar^l C_l(f,g)$, called the *Berezin-Toeplitz star-product*. Note that the existence of formal star product on symplectic manifolds was established by De Wilde and Lecomte in 1983. We refer to Fedosov's book [24] for more information on the theory of deformation quantization. In Theorem 2.6, we gave a geometric realization of the associative star-product.

2.3. The Kähler case. In this subsection, we assume that (X, ω, J) is a compact Kähler manifold, (L, h^L) is a holomorphic Hermitian line bundle with Chern connection ∇^L verifying (1.5), and (E, h^E) is a holomorphic Hermitian vector bundle with Chern connection ∇^E. We assume also that $\omega = \frac{\sqrt{-1}}{2\pi} R^L$ is the Kähler form of (X, g^{TX}). Let $\overline{\partial}^{L^p \otimes E,*}$ be the adjoint of the Dolbeault operator $\overline{\partial}^{L^p \otimes E}$ on $\Omega^{0,\bullet}(X, L^p \otimes E)$. In this case, D_p in (1.6) is given by

$$D_p = \sqrt{2}(\overline{\partial}^{L^p \otimes E} + \overline{\partial}^{L^p \otimes E,*}). \tag{2.18}$$

Thus D_p^2 preserves the $\mathbb{Z}$-grading on $\Omega^{0,\bullet}(X, L^p \otimes E)$. By Hodge theory and the Kodaira vanishing theorem, we have

$$\mathrm{Ker}(D_p) = H^0(X, L^p \otimes E) \quad \text{for } p \gg 1. \tag{2.19}$$

The Bergman projection P_p reduces to a projection from $\mathscr{C}^\infty(X, L^p \otimes E)$ onto $H^0(X, L^p \otimes E)$, a Toeplitz operator $\{T_p\}$ is now a sequence of linear operators acting on $\mathscr{C}^\infty(X, L^p \otimes E)$. Thus we don't need to introduce differential forms, and we can work on $\mathscr{C}^\infty(X, L^p \otimes E)$. In this situation, J_{r,x_0}, $\boldsymbol{b}_r(x_0)$, $Q_{r,x_0}(f)$, $\boldsymbol{b}_{r,f}(x_0)$ introduced in (2.5), (2.7), Remark 2.5 and (2.15) take values in $\mathrm{End}(E)_{x_0}$.

Let $\mathbb{P}(H^0(X, L^p)^*)$ be the projective space associated to the dual of $H^0(X, L^p)$, and let ω_{FS} be the Fubini–Study $(1,1)$-form. The Kodaira map

$\phi_p : X \longrightarrow \mathbb{P}(H^0(X, L^p)^*)$ is defined by $\phi_p(x) = \{H^0(X, L^p) \ni s \to s(x) \in L^p_x\}$ for $x \in X$. The Kodaira embedding theorem asserts that for $p \gg 1$, ϕ_p is a holomorphic embedding and $\phi_p^* \mathscr{O}(1) = L^p$. Let $h^{\phi_p^* \mathscr{O}(1)}$ be the metric on $\phi_p^* \mathscr{O}(1)$ induced by the metric $h^{\mathscr{O}(1)}$ on $\mathscr{O}(1)$. Then for $E = \mathbb{C}$, we have (cf. [41, Th. 5.1.3])

$$h^{\phi_p^* \mathscr{O}(1)}(x) = P_p(x, x)^{-1} h^{L^p}(x). \tag{2.20}$$

The question of the convergence as $p \to +\infty$ of $\frac{1}{p}\phi_p^*(\omega_{FS})$ was raised by Yau [71, §6.1]. By (2.7) for $E = \mathbb{C}$, and (2.20), as $p \to +\infty$, $\frac{1}{p}\phi_p^*(\omega_{FS})$ converges to ω in the $\mathscr{C}^\infty$ topology: for any $l \geqslant 0$, there exists $C_l > 0$ such that

$$\left| \frac{1}{p}\ \phi_p^*(\omega_{FS}) - \omega \right|_{\mathscr{C}^l(X)} \leqslant C_l/p^2. \tag{2.21}$$

When $l = 2$, the estimate of the type (2.21) was obtained by Tian [64] with p^2 replaced by $\sqrt{p}$, by using the Bergman kernel on the diagonal, $P_p(x, x)$. Ruan [59] obtained (2.21) with p instead of p^2. Bouche [11] proved that $\lim_{p\to+\infty} p^{-n} P_p(x, x) = 1$ in the $\mathscr{C}^0$ topology. The expansion (2.7) was first established by Catlin [17] and Zelditch [72].

Lu [36] calculated more coefficients $\boldsymbol{b}_r$ via R^{TX}. Let $\mathrm{Ric} = \mathrm{Ric}_g(J\cdot, \cdot)$ be the $(1, 1)$-form associated to the Ricci curvature Ric_g of g^{TX}. Let Δ be the positive Laplacian acting on functions on X; set $|R^{TX}|^2 = \sum_{ijkl} |\langle R^{TX}(w_i, \overline{w}_j)w_k, \overline{w}_j\rangle|^2$.

Theorem 2.7 (Lu [36, Th. 1.1]). *When $E = \mathbb{C}$, we have*

$$\boldsymbol{b}_1 = \frac{r^X}{8\pi}, \quad \pi^2 \boldsymbol{b}_2 = -\frac{\Delta r^X}{48} + \frac{1}{96}|R^{TX}|^2 - \frac{1}{24}|\,\mathrm{Ric}\,|^2 + \frac{1}{128}(r^X)^2. \tag{2.22}$$

Wang [70] also computed $\boldsymbol{b}_1$ in (2.7) for general E. When $E = \mathbb{C}$, the existence of an asymptotic expansion similar to (2.5) for $|Z|, |Z'| \leqslant C/\sqrt{p}$ was also obtained in [61, Th. 1]. For other versions of the asymptotic expansion see [17], [31], [18], [4]. The main tool in [17], [72], [18], [31], and [61] is the Boutet de Monvel-Sjöstrand parametrix for the Szegö kernel [13], [25]. The coefficients were computed in [64], [36], [70] by constructing appropriate peak sections, using Hörmander's L^2 $\overline{\partial}$-method.

If $E = \mathbb{C}$, the existence of the expansion (2.16) was first established by Bordemann, Meinrenken and Schlichenmaier [9], Schlichenmaier [60], [31]. They used the theory of Toeplitz structures of Boutet de Monvel and Guillemin [12].

Lu's computation for $\boldsymbol{b}_1$ plays an important role in Donaldson's work [23] on Kähler metrics with constant scalar curvature. We refer to [5], [41] for further information. In [42], we computed the coefficients $\boldsymbol{b}_{1,f}, \boldsymbol{b}_{2,f}$, $C_1(f, g), C_2(f, g)$ from (2.15), (2.16). These computations are also relevant in Kähler geometry (cf. [26], [27], [35]).

Theorem 2.8 (Ma-Marinescu [42]). *If $E = \mathbb{C}$, for any $f \in \mathscr{C}^\infty(X)$, we have*:

$$\begin{aligned} &\boldsymbol{b}_{0,f} = f, \quad \boldsymbol{b}_{1,f} = \frac{r^X}{8\pi} f - \frac{1}{4\pi}\Delta f, \\ &\boldsymbol{b}_{2,f} = \boldsymbol{b}_2 f + \frac{1}{32\pi^2}\Delta^2 f - \frac{1}{32\pi^2} r^X \Delta f - \frac{\sqrt{-1}}{8\pi^2}\Big\langle \operatorname{Ric}, \partial\overline{\partial} f \Big\rangle. \end{aligned} \tag{2.23}$$

3. Quantization and Symplectic Reduction

We explain briefly the Guillemin-Sternberg conjecture in Section 3.1, then we review the asymptotic expansion of the G-invariant part of the Bergman kernel in Section 3.2, and we specialize the results in the Kähler case in Section 3.3. In particular, we show how to obtain the scalar curvature on the reduction from the G-invariant Bergman kernel on the total space, and we compare the metrics on the two sides of the "quantization commutes with reduction".

We use the same notation and assumptions as in Section 1.1.

3.1. Quantization commutes with reduction. Recall that (X, ω, J) is a compact symplectic manifold of real dimension $2n$ with compatible almost complex structure J, and (L, h^L, ∇^L) is a prequantum line bundle on X (cf. (1.5)).

Let G be a compact connected Lie group of dimension n_0 with Lie algebra $\mathfrak{g}$. We assume that G acts on the left on X and that this action lifts to L. Moreover, we assume that G preserves g^{TX}, J, h^L and ∇^L.

The G-action commutes with the Dirac operator D^L, and $\operatorname{Ker}\big(D^L_\pm\big)$ are finite dimensional G-representations. The quantization space $Q(L)$ of L (cf. (1.3)) is an element in the representation ring $R(G)$ of G.

For $K \in \mathfrak{g}$, let K^X be the vector field on X generated by K, and let L_K be the corresponding Lie derivative. Let $\Lambda^*_+ \subset \mathfrak{g}^*$ be the set of dominant weights, and let V^G_γ be the irreducible representation of G with highest weight $\gamma \in \Lambda^*_+$. Let $Q(L)^\gamma \in \mathbb{Z}$ be the multiplicity of V^G_γ in $Q(L)$. Then we have

$$Q(L) = \bigoplus_{\gamma \in \Lambda^*_+} Q(L)^\gamma \cdot V^G_\gamma \in R(G), \tag{3.1}$$

and there are only finitely many $\gamma \in \Lambda^*_+$ such that $Q(L)^\gamma \neq 0$.

It is not easy to read off $Q(L)^\gamma$ directly from the Atiyah-Bott-Segal-Singer equivariant index theorem for its character. Guillemin and Sternberg [29] suggested a geometric way to compute $Q(L)^\gamma$, by using the associated moment map.

Definition 3.1. The moment map $\mu : X \to \mathbf{g}^*$ is defined by the Kostant formula [33],

$$2\sqrt{-1}\pi\mu(K) = \nabla^L_{K^X} - L_K, \quad \text{for } K \in \mathfrak{g}. \tag{3.2}$$

Then μ is G-equivariant and one has $i_{K^X}\omega = d\,\mu(K)$.

For a regular value $\nu \in \mathfrak{g}^*$ of μ, the Marsden-Weinstein symplectic reduction $X_\nu := \mu^{-1}(G \cdot \nu)/G$ is a compact symplectic orbifold with the symplectic form ω_ν induced by ω. Moreover, L (resp. J) induces a prequantum line bundle L_ν (resp. an almost complex structure J_ν) over (X_ν, ω_ν). One can then construct the associated spin^c Dirac operator (twisted by L_ν), $D_+^{L_\nu}$ on X_ν, of which the index $Q(L_\nu) \in \mathbb{Z}$ (identified as the virtual dimension of $Q(L_\nu)$ in (1.3)).

If $\gamma \in \Lambda_+^*$ is not a regular value of μ, then by [49] (cf. [54, §7.4], [43, §3.5] for a standard perturbation definition), $Q(L_\gamma)$ is still well defined. Now we can state:

Guillemin-Sternberg conjecture: For any $\gamma \in \Lambda_+^*$,

$$Q(L)^\gamma = Q(L_\gamma). \tag{3.3}$$

By the classical shifting trick (i.e., by working on $X \times \mathcal{O}_\gamma$, where $\mathcal{O}_\gamma = G \cdot \gamma$ is the orbit of the co-adjoint action of G on $\mathfrak{g}^*$), we only need to prove (3.3) for $\gamma = 0$.

This conjecture was proved by Meinrenken [47] and Vergne [67] when G is abelian; by Meinrenken [48], Meinrenken-Sjamaar[49] for non-abelian groups G, by using the technique of symplectic cut of Lerman [34].

Tian and Zhang [65] gave an analytic proof of the Guillemin-Sternberg conjecture, using a deformation of the Dirac operator, which is associated with the function $|\mu|^2$. Their approach works for a general vector bundle E satisfying certain positivity conditions [65, (4.2)] (used afterwards by Paradan [54, p. 445] and Teleman [63, p. 6]), and also for manifolds with boundary [66]. Paradan [54] developed later a K-theoretic approach by making use of the theory of transversally elliptic operators. See [68] for a survey and complete references on this subject.

3.2. Berezin-Toeplitz quantization and reduction. We use the same notation and assumptions as in Sections 1.1 and 3.1. We assume also that the G-action lifts on E and preserves h^E and ∇^E.

Then G-action commutes with the Dirac operator D_p in (1.6). Let $\mathrm{Ker}(D_p)^G$ be the G-trivial component of $\mathrm{Ker}(D_p)$. Let P_p^G be the orthogonal projection from $\mathscr{C}^\infty(X, E_p)$ onto $\mathrm{Ker}(D_p)^G$. The *G-invariant Bergman kernel* is the $\mathscr{C}^\infty$ kernel $P_p^G(x, x')$, $(x, x' \in X)$ of P_p^G associated to $dv_X(x')$.

Assume for simplicity that G acts freely on $\mu^{-1}(0)$, and $g^{TX}(\cdot, \cdot) = \omega(\cdot, J\cdot)$. We will denote by $X_G = \mu^{-1}(0)/G$, and we add a subscript G to denote the objects on X_G induced by the corresponding objects on X.

By a result of Tian and Zhang [65, Th. 0.2], and (1.7b), we have

$$\dim \mathrm{Ker}(D_p)^G = \dim \mathrm{Ker}(D_{G,p}) \quad \text{for } p \gg 1. \tag{3.4}$$

We will describe how $P_p^G(x, x')$ "concentrates" on the Bergman kernel $P_{G,p}(x_0, x_0')$ on X_G, when $p \to +\infty$.

Theorem 3.2 (Ma-Zhang [43, Th. 0.1]). *For any open G-neighborhood U of $\mu^{-1}(0)$ and any $\varepsilon_0 > 0$, we have*

$$P_p^G(x,x') = \mathscr{O}(p^{-\infty}) \text{ if } (x,x') \notin U \times U \text{ or if } d^X(Gx,x') \geqslant \varepsilon_0. \tag{3.5}$$

Let U be an open G-neighborhood of $\mu^{-1}(0)$ such that G acts freely on U. For any G-equivariant vector bundle with connection (F, ∇^F) on U, we denote by (F_B, ∇^{F_B}) the bundle on $B := U/G$ induced by G-invariant sections of F on U.

For $x \in U$ denote by $\mathrm{vol}(Gx)$ the volume of the orbit Gx equipped with the metric induced by g^{TX}. Following [65, (3.10)], let $h(x)$ be the function on U defined by

$$h(x) = (\mathrm{vol}(Gx))^{1/2}. \tag{3.6}$$

Then h descends to a function on B.

Let pr_1 and pr_2 be the projections from $X \times X$ onto the first and the second factor X respectively. Then we can view $P_p^G(x,x')$ $(x,\, x' \in U)$ as a smooth section of $\mathrm{pr}_1^*(E_p)_B \otimes \mathrm{pr}_2^*(E_p^*)_B$ on $B \times B$.

We introduce the following coordinates: for any $x_0 \in X_G$, $Z \in T_{x_0}B$, we write $Z = Z^0 + Z^\perp$, with $Z^0 \in T_{x_0}X_G$, $Z^\perp \in N_{G,x_0}$, where N_G is the normal bundle of X_G in B. For $\varepsilon_0 > 0$ small enough, we identify $Z \in T_{x_0}B$, $|Z| < \varepsilon_0$ with $\exp^B_{\exp^{X_G}_{x_0}(Z^0)}(Z^\perp) \in B$, here we still denote by $Z^\perp \in N_{G,\exp^{X_G}_{x_0}(Z^0)}$, the parallel transport of $Z^\perp$ along the curve $u \to \exp^{X_G}_{x_0}(uZ^0)$ with respect to the connection on N_G induced by projecting the Levi-Civita connection on TB.

We identify $(E_p)_{B,Z}$ with $(E_p)_{B,x_0}$ by using parallel transport with respect to $\nabla^{(E_p)_B}$ (cf. §2.1) along the curve $[0,1] \ni u \to uZ$.

Let dv_B, dv_{X_G}, dv_{N_G} be the Riemannian volume forms on TB, TX_G, N_G induced by g^{TX}. Let $\varrho \in \mathscr{C}^\infty(TB|_{X_G}, \mathbb{R})$, with $\varrho = 1$ on X_G, be defined by

$$dv_B(x_0, Z) = \varrho(x_0, Z) dv_{X_G}(x_0) dv_{N_{G,x_0}} \quad \text{for } Z \in T_{x_0}B, x_0 \in X_G. \tag{3.7}$$

For $x_0 \in X_G$, $Z = (Z^0, Z^\perp), Z' = (Z'^0, Z'^\perp) \in T_{x_0}X_G \oplus N_{G,x_0} = T_{x_0}B$, set

$$\mathcal{P}(Z,Z') = 2^{\frac{n_0}{2}} \exp\left(-\frac{\pi}{2}\sum_i \left(|z_i^0|^2 + |z_i'^0|^2 - 2z_i^0 \overline{z}_i'^0\right)\right) \times \exp\Big(-\pi|Z^\perp|^2 - \pi|Z'^\perp|^2\Big), \tag{3.8}$$

with $n_0 = \dim G$. As in (1.16) and (2.9), $\mathcal{P}$ is the Bergman kernel of a limit operator, which itself is sum of two terms: one is defined on $T_{x_0}X_G$, and is equal $\mathscr{L}$ (cf. (1.13)); the other is defined on N_{G,x_0}, it is equal to a harmonic oscillator. This explains why we expect the G-invariant Bergman kernel $P_p^G(x,x')$ to exhibit the same sort of behavior, see (3.11).

Let $\{\Theta_p^G\}_{p\in\mathbb{N}}$ be a sequence of linear operators $\Theta_p^G : L^2(X, E_p) \longrightarrow L^2(X, E_p)$ with smooth kernel $\Theta_p^G(x,y)$ with respect to $dv_X(y)$. We assume

that $\Theta_p^G(x,y)$ is $G\times G$-invariant. Let $\pi_B : TB\times_{X_G} TB\to X_G$ be the obvious projection. Relative to our trivialization, $\Theta_p^G(x,y)$ induces a smooth section $\Theta^G_{p,\,x_0}(Z,Z')$ of $\pi_B^*(\mathrm{End}(\Lambda^{0,\bullet}\otimes E)_B)$ over $TB\times_{X_G} TB$ with $Z,Z'\in T_{x_0}B$. We introduce the following notation in analogy to Notation 2.1.

Notation 3.3. We write

$$p^{-n+\frac{n_0}{2}}\Theta^G_{p,x_0}(Z,Z')\overset{h}{\approx}\sum_{r=0}^{k}(Q^G_{r,\,x_0}\mathcal{P}_{x_0})(\sqrt{p}Z,\sqrt{p}Z')p^{-\frac{r}{2}}+\mathcal{O}(p^{-\frac{k+1}{2}})\,, \tag{3.9}$$

if there exists a family $\{Q^G_{r,\,x_0}\}_{0\leqslant r\leqslant k,\,x_0\in X_G}$ with $Q^G_{r,\,x_0}\in\mathrm{End}(\Lambda^{0,\bullet}\otimes E)_{B,x_0}[Z,Z']$ smooth with respect to the parameter $x_0\in X_G$, and there exist $\varepsilon'\in]0,a^X[$ and $C_0>0$ with the following property: for any $l,m\in\mathbb{N}$, there exist $C>0$, $M>0$ such that for any $x_0\in X_G$, $Z,Z'\in T_{x_0}B$, $|Z|,|Z'|<\varepsilon'$ and $p\in\mathbb{N}^*$, the following estimate holds:

$$\begin{aligned}(1+\sqrt{p}|Z^{\perp}|+\sqrt{p}|Z'^{\perp}|)^m\Big|&p^{-n+\frac{n_0}{2}}\Theta^G_{p,\,x_0}(Z,Z')(h\varrho^{\frac12})(Z)(h\varrho^{\frac12})(Z')\\&-\sum_{r=0}^{k}(Q^G_{r,x_0}\mathcal{P}_{x_0})(\sqrt{p}Z,\sqrt{p}Z')p^{-\frac{r}{2}}\Big|_{\mathscr{C}^l(X_G)}\\ \leqslant\; &C\,p^{-\frac{k+1}{2}}(1+\sqrt{p}\,|Z^0|+\sqrt{p}\,|Z'^0|)^M\exp(-\sqrt{C_0p}\,|Z-Z'|)+\mathscr{O}(p^{-\infty})\,.\end{aligned} \tag{3.10}$$

Theorem 3.4 (Ma-Zhang [43, Th. 0.2]). *There exists a family of polynomials $\{\mathcal{Q}_{r,\,x_0}\}_{r\in\mathbb{N},\,x_0\in X_G}\in\mathrm{End}(\Lambda^{0,\bullet}\otimes E)_{B,x_0}[Z,Z']$ on Z,Z' with the same parity as r, such that $\mathcal{Q}_{0,x_0}=I_{\mathbb{C}\otimes E,G}$, and for any $k\in\mathbb{N}$ the following expansion holds in the sense of Notation 3.3,*

$$p^{-n+\frac{n_0}{2}}P^G_{p,x_0}(Z,Z')\overset{h}{\approx}\sum_{r=0}^{k}(\mathcal{Q}_{r,x_0}\mathcal{P}_{x_0})(\sqrt{p}Z,\sqrt{p}Z')p^{-\frac{r}{2}}+\mathcal{O}(p^{-\frac{k+1}{2}}). \tag{3.11}$$

To read off the scalar curvature on the reduction from P_p^G, we define $\mathscr{I}_p(x_0)\in\mathrm{End}(\Lambda^{0,\bullet}\otimes E)_{G,x_0}$ for $x_0\in X_G$ by :

$$\mathscr{I}_p(x_0)=\int_{\substack{|Z|\leqslant\varepsilon_0,\\ Z\in N_G}}(\varrho h^2)(x_0,Z)P_p^G((x_0,Z),(x_0,Z))dv_{N_G}(Z). \tag{3.12}$$

By (3.4), (3.5), $\mathscr{I}_p(x_0)$ does not depend on ε_0 modulo $\mathscr{O}(p^{-\infty})$, and

$$\dim\mathrm{Ker}(D_{G,p})=\int_{X_G}\mathrm{Tr}[\mathscr{I}_p(x_0)]dv_{X_G}(x_0)+\mathscr{O}(p^{-\infty}). \tag{3.13}$$

From Theorem 3.4, we infer the existence of $\Phi_r\in\mathscr{C}^\infty(X_G,\mathrm{End}(\Lambda^{0,\bullet}\otimes E)_G)$, and $\Phi_0=I_{\mathbb{C}\otimes E,G}$, with the property that for all $k,m\in\mathbb{N}$, there exists $C_{k,m}>0$

such that for all $p \in \mathbb{N}^*$,

$$\left| p^{-n+n_0} \mathscr{I}_p(x_0) - \sum_{r=0}^{k} \Phi_r(x_0) p^{-r} \right|_{\mathscr{C}^m(X_G)} \leqslant C_{k,m} p^{-k-1}. \tag{3.14}$$

Using Theorems 3.2, 3.4, and the same argument as in Remark 2.5, we see that the analogue of Theorems 3.2, 3.4 still holds for the kernel $T^G_{f,p}(x,x')$ of the operator $T^G_{f,p} := P^G_p f P^G_p$, for $f \in \mathscr{C}^\infty(X, \mathrm{End}(E))$.

Theorem 3.5 (Ma-Zhang [43, p. 86-88]). *Let $f \in \mathscr{C}^\infty(X, \mathrm{End}(E))$. For any open G-neighborhood U of $\mu^{-1}(0)$, $\varepsilon_0 > 0$, we have*

$$T^G_{f,p}(x,x') = \mathscr{O}(p^{-\infty}) \text{ if } (x,x') \notin U \times U \text{ or if } d^X(Gx,x') \geqslant \varepsilon_0. \tag{3.15}$$

Moreover, there exists a family $\{\mathcal{Q}^G_{r,\,x_0}(f)\}_{r\in\mathbb{N},\,x_0\in X_G} \in \mathrm{End}(\Lambda^{0,\bullet} \otimes E)_{B,x_0}[Z,Z']$ of polynomials in Z, Z' with the same parity as r such that for any $k \in \mathbb{N}$, we have in the sense of Notation 3.3,

$$p^{-n+\frac{n_0}{2}} T^G_{f,\,p,\,x_0}(Z,Z') \overset{h}{\approx} \sum_{r=0}^{k} (\mathcal{Q}^G_{r,\,x_0}(f)\mathcal{P}_{x_0})(\sqrt{p}Z, \sqrt{p}Z') p^{-\frac{r}{2}} + \mathcal{O}(p^{-\frac{k+1}{2}}). \tag{3.16}$$

Moreover, $\mathcal{Q}^G_{0,x_0}(f) = f^G(x_0) I_{\mathbb{C}\otimes E,G}$, where f^G is the G-invariant component of f.

Since $\mathrm{Tr}\left[T^G_{f,p}\right] = \int_X \mathrm{Tr}\left[T^G_{f,p}(x,x)\right] dv_X(x)$, we deduce from Theorem 3.5 that there exists a sequence $B_{r,f}$ with $B_{0,f} = \int_{X_G} \mathrm{Tr}\left[f^G(x_0)\right] dv_{X_G}(x_0)$ and for any $k \in \mathbb{N}$,

$$p^{-n+n_0}\, \mathrm{Tr}\left[T^G_{f,p}\right] = \sum_{r=0}^{k} B_{r,f} p^{-r} + \mathscr{O}(p^{-k-1}). \tag{3.17}$$

Note that in [43, §4.1, §4.5] the case where 0 is a regular value of μ (so that X_G is an orbifold) is treated in detail. In [43, §4.2], it is shown by a shifting trick that Theorems 3.2 and 3.4 imply the expansion of the kernel of the orthogonal projection $P_p^{V_\gamma^G}$ from $\Omega^{0,\bullet}(X, L^p \otimes E)$ onto the V_γ^G-component of $\mathrm{Ker}(D_p)$ for any $\gamma \in \Lambda^*_+$.

3.3. The Kähler case. In this subsection, as in Section 2.3, we assume that (X, ω, J) is a compact Kähler manifold carrying a holomorphic Hermitian line bundle (L, h^L) and a holomorphic Hermitian vector bundle (E, h^E) and moreover $\omega = \frac{\sqrt{-1}}{2\pi} R^L$ is the Kähler form of (X, g^{TX}). We assume also that the G-action on X, L, E is holomorphic, and preserves the metrics.

By (2.19), we see as in Section 2.3 that the G-invariant Bergman projection P^G_p reduces to a projection from $\mathscr{C}^\infty(X, L^p \otimes E)$ onto $H^0(X, L^p \otimes E)^G$, and

the Toeplitz operator $\{T^G_{f,p}\}$ reduces to a sequence of linear operators acting on $\mathscr{C}^\infty(X, L^p \otimes E)$. In particular, $\mathcal{Q}_{r,x_0}$, $\mathscr{I}_p(x_0)$, $\Phi_r(x_0)$, $\mathcal{Q}^G_{r,x_0}(f)$ in (3.11), (3.14) and (3.16) take values in $\mathrm{End}(E_G)_{x_0}$.

Let $\widetilde{h}$ be the restriction of h on X_G. Let r^{X_G} be the scalar curvature on (X_G, ω_G, J_G), and Δ_{X_G} be the positive Laplacian on X_G. Let $\{w^0_j\}$ be an orthonormal frame of $T^{(1,0)}X_G$. The following result generalizes formula (2.13) for the coefficient $\boldsymbol{b}_1$ of the expansion (2.7).

Theorem 3.6 (Ma-Zhang [43, Th. 0.6]). *The coefficients* Φ_0 *and* Φ_1 *from* (3.14) *are given by,*

$$\Phi_0 = \mathrm{Id}_{E_G}, \quad \Phi_1(x_0) = \frac{1}{8\pi} r^{X_G}_{x_0} + \frac{3}{4\pi}\Delta_{X_G} \log \widetilde{h} + \frac{1}{2\pi}\sum_j R^{E_G}_{x_0}(w^0_j, \overline{w}^0_j). \tag{3.18}$$

We discuss now the metric aspect of quantization. Let $i : \mu^{-1}(0) \hookrightarrow X$ be the natural injection. Let $\pi_G : \mathscr{C}^\infty(\mu^{-1}(0), L^p \otimes E)^G \to \mathscr{C}^\infty(X_G, L^p_G \otimes E_G)$ be the natural identification. By a result of Zhang [73, Th. 1.1 and Prop. 1.2], for $p \gg 1$, the map $\pi_G \circ i^* : \mathscr{C}^\infty(X, L^p \otimes E)^G \to \mathscr{C}^\infty(X_G, L^p_G \otimes E_G)$ induces a natural isomorphism

$$\sigma_p = \pi_G \circ i^* : H^0(X, L^p \otimes E)^G \to H^0(X_G, L^p_G \otimes E_G). \tag{3.19}$$

(When $E = \mathbb{C}$, this result was first proved in [29, Th. 3.8] for $p \geqslant 1$). We denote by $\langle \cdot, \cdot \rangle$ the L^2-Hermitian products on these spaces. A corollary of Theorem 3.5 is as follows.

Theorem 3.7 (Ma-Zhang [43, Th. 4.8]). *Set* $\sigma^G_p = \sigma_p \circ P^G_p$ *and let* σ^{G*}_p *be the adjoint of* σ^G_p. *Then* $\mathcal{T}_{f,p} = p^{-\frac{n_0}{2}} \sigma^G_p f \sigma^{G*}_p \in \mathrm{End}(H^0(X_G, L^p_G \otimes E_G))$ *is a Toeplitz operator with principal symbol* $2^{\frac{n_0}{2}} f^G/\widetilde{h}^2$, *for any* $f \in \mathscr{C}^\infty(X, \mathrm{End}(E))$.

The natural Hermitian product $\langle \cdot, \cdot \rangle_{\widetilde{h}}$ on $\mathscr{C}^\infty(X_G, L^p_G \otimes E_G)$ is given by

$$\langle s_1, s_2 \rangle_{\widetilde{h}} = \int_{X_G} \langle s_1, s_2 \rangle(x_0)\, \widetilde{h}^2(x_0)\, dv_{X_G}(x_0). \tag{3.20}$$

Theorem 3.8 (Ma-Zhang [43, Th. 0.10]). *The isomorphism* $(2p)^{-\frac{n_0}{4}} \sigma_p$ *is an asymptotic isometry from* $(H^0(X, L^p \otimes E)^G, \langle \cdot, \cdot \rangle)$ *onto* $(H^0(X_G, L^p_G \otimes E_G), \langle \cdot, \cdot \rangle_{\widetilde{h}})$, *i.e., if* $\{s^p_i\}_{i=1}^{d_p}$ *is an orthonormal basis of* $(H^0(X, L^p \otimes E)^G, \langle \cdot, \cdot \rangle)$, *then*

$$(2p)^{-\frac{n_0}{2}} \left\langle \sigma_p s^p_i, \sigma_p s^p_j \right\rangle_{\widetilde{h}} = \delta_{ij} + \mathscr{O}\left(p^{-1}\right). \tag{3.21}$$

In [43, Remark 0.11], we find a natural symplectic extension of Theorem 3.8.

When $E = \mathbb{C}$ and G is a torus, Charles [19] first showed that $\mathcal{T}_{f,p}$ in Theorem 3.7 is a Toeplitz operator, and obtained (3.21).

Assume that $E = \mathbb{C}$. Then $P_p^G(x_0, x_0)$ becomes a positive function. By setting $Z = Z' = 0$ in (3.11), we get the following expansion on X_G for any k,

$$p^{-n+\frac{n_0}{2}} h^2(x_0) P_p^G(x_0, x_0) = \sum_{r=0}^{k} c_r(x_0) p^{-r} + \mathscr{O}(p^{-k-1}), \quad c_0(x_0) = 2^{n_0/2}. \tag{3.22}$$

Paoletti [50, Th. 1], [51, Th. 1] had obtained the expansion (3.22), but he claimed that $c_0(x_0) = 1$. After our preprint [43] was posted, Hall-Kirwin [30], Paoletti [52], [53] and Burns-Guillemin-Wang [16] have established related results.

4. Noncompact Case: Vergne's Conjecture

In this section, we use the same notation and assumptions as in Sections 1, 3.1, except that we assume now that X is noncompact. One asks naturally the following question: what is the quantization formula in this situation?

When (X, g^{TX}) is a complete Riemannian manifold, it is shown in [38, §3.5], [40, §5], [41, §6.1, §7.5], [43, §4.6] that under natural (positivity) conditions on R^L, R^E, the asymptotic expansion of the Bergman kernel holds. However, in this section, we do not assume (X, g^{TX}) to be complete.

In Section 4.1, the quantization formula is explained for the model example $\mathbb{C}^n$. In Section 4.2, we review briefly our solution with Zhang of Vergne's conjecture: "quantization commutes with reduction" in the noncompact setting.

4.1. Quantization formula on $\mathbb{C}^n$. We continue the discussion of Section 1.2. Let's assume now that $a_j = 2\pi$ for $j = 1, \cdots, n$. Then (L, h^L, ∇^L) is a prequantum line bundle on $(\mathbb{C}^n, \omega = \frac{\sqrt{-1}}{2} \sum_j dz_j \wedge d\bar{z}_j)$.

Let T^n be the n-dimensional torus with Lie algebra $\mathfrak{t}^n$. We define a holomorphic action of T^n on $\mathbb{C}^n$ by $e^{i\theta} \cdot z = (e^{i\theta_1} z_1, \cdots, e^{i\theta_n} z_n)$, with $\theta = (\theta_1, \cdots, \theta_n) \in \mathbb{R}^n$ and $e^{i\theta} = (e^{i\theta_1}, \cdots, e^{i\theta_n}) \in T^n$. For $\lambda = (\lambda_1, \cdots, \lambda_n) \in \mathbb{Z}^n$, we define a holomorphic T^n-action on L by $e^{i\theta} \cdot \mathbf{1} = e^{i\theta \cdot \lambda} \mathbf{1}$ with $\theta \cdot \lambda = \sum_j \theta_j \lambda_j$. Then the associated moment map $\mu : \mathbb{C}^n \to \mathbb{R}^{n*}$ (cf. (3.2)) is given by

$$\mu(z) = \frac{1}{2}(|z_1|^2, \cdots, |z_n|^2) + \lambda. \tag{4.1}$$

Given $\{u_i\}_{i=1}^n \subset \mathbb{Z}^m$, the *Delzant polytope* $\Delta \subset \mathbb{R}^{m*}$ [2, §VII. 1.c., 2.a.] is defined by

$$\Delta = \{x \in \mathbb{R}^{m*} : (u_i, x) \geqslant \lambda_i \text{ for } 1 \leqslant i \leqslant n\}, \tag{4.2}$$

if the vertices have integer coordinates and each vertex q has exactly m-edges, and the u_i such that $(u_i, q) = \lambda_i$ form a basis of $\mathbb{Z}^m$.

Let $\jmath : \mathbb{R}^n \to \mathbb{R}^m$ be the linear map defined by $\jmath(e_i) = u_i$ with $\{e_i\}$ the canonical basis of $\mathbb{R}^n$. Let $N = \mathrm{Ker}(\jmath)/(\mathrm{Ker}(\jmath) \cap (2\pi\mathbb{Z})^n) \subset \mathbb{R}^n/(2\pi\mathbb{Z})^n \simeq T^n$, so that N is a $(n-m)$-dimensional torus with Lie algebra $\mathfrak{n} \overset{\imath}{\hookrightarrow} \mathbb{R}^n \simeq \mathfrak{t}^n$. Thus we have the exact sequence: $0 \to \mathbb{R}^{m*} \xrightarrow{\jmath^*} \mathbb{R}^{n*} \xrightarrow{\imath^*} \mathfrak{n}^* \to 0$.

Now N acts naturally on $\mathbb{C}^n$ and L, the associated moment map is $\Phi = \imath^* \circ \mu : \mathbb{C}^n \to \mathfrak{n}^*$. Its symplectic reduction $X_\Delta = \Phi^{-1}(0)/N$ is a m-dimensional compact Kähler manifold, and L descends naturally to a positive holomorphic line bundle L_Δ on X_Δ. Then X_Δ is the toric variety associated to the Delzant polytope Δ.

Observe that if N acts trivially on a holomorphic section $z^\alpha \mathbf{1}$ of L for some $\alpha \in \mathbb{N}^n$, then $z^\alpha \mathbf{1}$ descends to a holomorphic section of L_Δ on X_Δ.

For $e^{i\theta} \in T^n$, we have $e^{i\theta} \cdot z^\alpha \mathbf{1} = e^{i\theta\cdot(\alpha+\lambda)} z^\alpha \mathbf{1}$. Thus N acts trivially on the holomorphic section $z^\alpha \mathbf{1}$ if and only if $\imath^*(\alpha+\lambda) = 0$, and this is equivalent to the existence of a $\nu \in \mathbb{R}^{m*}$ such that $\alpha_i + \lambda_i = (\nu, u_i)$, i.e., $\nu \in \Delta \cap \mathbb{Z}^m$ and $\alpha_i + \lambda_i = (\nu, u_i)$.

For $\nu \in \Delta \cap \mathbb{Z}^m$, we denote by s_ν the holomorphic section of L_Δ on X_Δ induced by $z^\alpha \mathbf{1}$, where $\alpha_i = (\nu, u_i) - \lambda_i$.

Theorem 4.1 ([28, §3.5]). *The cohomology of L_Δ on X_Δ is given by*

$$H^0(X_\Delta, L_\Delta) = \bigoplus_{\nu \in \Delta \cap \mathbb{Z}^m} \mathbb{C}\, s_\nu, \quad H^j(X_\Delta, L_\Delta) = 0 \ \ \textit{if } j > 0. \tag{4.3}$$

By Theorem 1.5, we see that the kernel of D^L on the noncompact space $\mathbb{C}^n$ is an infinite dimensional vector space. Moreover, by the discussion after (1.13) we deduce that all higher L^2 cohomology groups of $\mathbb{C}^n$ with values in L vanish. Theorem 4.1 implies that "quantization commutes with reduction" still holds. Note that the moment map $\Phi = \imath^* \circ \mu$ is proper here.

Example 4.2. Set $m = n-1$, $u_i = e_i$ for $i \leqslant m$, $u_n = -(1, \cdots, 1) = -\sum_{i=1}^m e_i$, $\lambda = (0, \cdots, 0, -1)$. Then $\mathrm{Ker}(\jmath) = \mathbb{R}(1, \cdots, 1)$, $\Phi(z) = \frac{1}{2}\sum_{i=1}^n |z_i|^2 - 1$. In this case, $(X_\Delta, L_\Delta) \simeq (\mathbb{CP}^{n-1}, \mathcal{O}(1))$ with $\mathcal{O}(1)$ the hyperplane line bundle on $\mathbb{CP}^{n-1}$.

4.2. Vergne's conjecture. Recall that (X, ω, J) is a noncompact symplectic manifold with the prequantum line bundle (L, h^L, ∇^L), and g^{TX} is a J-invariant Riemannian metric on X. Let $\tau : TX \to X$ be the natural projection. Following [1, p. 7] (cf. [54, §3]), set $T_G X = \big\{(x, v) \in T_x X : \langle v, K^X(x)\rangle = 0 \text{ for all } K \in \mathfrak{g}\big\}$.

Then the quantization space $Q(L) = \mathrm{Ind}(D^L)$ of L is not well defined, since usually D^L is not a Fredholm operator, and we need to make precise the self-adjoint extension of $D^L|_{\Omega_0^{0,\bullet}(X,L)}$, where $\Omega_0^{0,\bullet}(X, L)$ denotes the space of sections with compact support.

We suppose that the moment map $\mu : X \to \mathfrak{g}^*$ is proper. Then the right hand side of (3.3) is well defined.

We identify $\mathfrak{g}$ with $\mathfrak{g}^*$ by using an Ad_G-invariant metric on $\mathfrak{g}$. Let $\mu^X(x) := (\mu(x))^X(x)$ $(x \in X)$[1] be the vector field induced by $\mu : X \to \mathfrak{g}$.

We suppose for the moment that $\{x \in X : \mu^X(x) = 0\}$ is compact.

Recall that $\mathbf{c}(\cdot)$ is the Clifford action defined in (1.1). For $x \in X,\ \xi \in T_xX$, set[2]

$$\begin{aligned}\sigma^X_{L,\mu}(x,\xi) = \tau^*\left(\sqrt{-1}\mathbf{c}(\xi + \mu^X) \otimes \mathrm{Id}_L\right)\big|_{(x,\xi)} \\ : \tau^*(\Lambda^{\mathrm{even}}(T^{*(0,1)}X) \otimes L) \to \tau^*(\Lambda^{\mathrm{odd}}(T^{*(0,1)}X) \otimes L). \quad (4.4)\end{aligned}$$

Then $\sigma^X_{L,\mu}$ is a transversally elliptic symbol on T_GX in the sense of Atiyah [1, §1, §3] and Paradan [54, §3], [55, §3], which determines a transversal index $\mathrm{Ind}\left(\sigma^X_{L,\mu}\right)$ in the formal representation ring $R[G]$ of G,

$$\mathrm{Ind}\left(\sigma^X_{L,\mu}\right) = \bigoplus_{\gamma \in \Lambda^*_+} \mathrm{Ind}_\gamma\left(\sigma^X_{L,\mu}\right) \cdot V^G_\gamma \in R[G]. \quad (4.5)$$

The index $\mathrm{Ind}\left(\sigma^X_{L,\mu}\right)$ does not depend on g^{TX}, h^L, ∇^L, and it depends only on the homotopy classes of J, μ^X. The set $\{\gamma \in \Lambda^*_+ : \mathrm{Ind}_\gamma\left(\sigma^X_{L,\mu}\right) \neq 0\}$ can be infinite. Michèle Vergne suggested to use $\mathrm{Ind}_\gamma\left(\sigma^X_{L,\mu}\right)$ to replace the left hand side of (3.3).

Vergne's conjecture (ICM 2006 plenary lecture [69, §4.3]) : If $\mu : X \to \mathfrak{g}^*$ is proper and if $\{x \in X : \mu^X(x) = 0\}$ is compact, then for any $\gamma \in \Lambda^*_+$,

$$\mathrm{Ind}_\gamma\left(\sigma^X_{L,\mu}\right) = Q\left(L_\gamma\right). \quad (4.6)$$

Special cases of this conjecture, related to the discrete series of semi-simple Lie groups, have been proved by Paradan [55], [57].

For $a > 0$, set $X_a = \{x \in X : |\mu|^2(x) \leqslant a\}$. If a is a regular value of $|\mu|^2$, then X_a is a compact manifold with boundary ∂X_a, and μ^X is nowhere zero on ∂X_a. Thus $\sigma^{X_a}_{L,\mu}$ is a transversally elliptic symbol on X_a.

Theorem 4.3 (Quantization commutes with reduction, Ma-Zhang [45, Th. 0.2, 0.3]). *Suppose that $\mu : X \to \mathfrak{g}^*$ is proper. For any $\gamma \in \Lambda^*_+$, there exists $a_\gamma > 0$ such that the function $a \mapsto \mathrm{Ind}_\gamma\left(\sigma^{X_a}_{L,\mu}\right)$ is constant on $\{a > a_\gamma : a$ is regular value of $|\mu|^2\}$. Denote by $Q(L)^\gamma$ this constant. Then for any $\gamma \in \Lambda^*_+$, we have*

$$Q(L)^\gamma = Q(L_\gamma). \quad (4.7)$$

[1]The vector field μ^X is also called Kirwan vector field in view of [32].

[2]The symbol $\sigma^X_{L,\mu}$ is the (semi-classical) symbol of Tian-Zhang's [65] deformed Dirac operator (4.8) in their approach to the Guillemin-Sternberg geometric quantization conjecture.

If $\{x \in X : \mu^X(x) = 0\}$ is compact, then $Q(L)^\gamma = \mathrm{Ind}_\gamma\left(\sigma^X_{L,\mu}\right)$. Therefore Theorem 4.3 implies Vergne's conjecture. Note that Paradan [58] gives a new proof of Theorem 4.3 by using symplectic cuts and the wonderful compactifications of [56].

Idea of the proof. 1) Assume that $\{x \in X : \mu^X(x) = 0\}$ is compact. For $T > 0$, let D^L_T be the deformed Dirac operator introduced by Tian-Zhang [65, (1.20)]:

$$D^L_T = D^L + \sqrt{-1}T\mathbf{c}\left(\mu^X\right) : \Omega^{0,\bullet}(X, L) \to \Omega^{0,\bullet}(X, L). \tag{4.8}$$

A first step is to interpret the transversal index as the Atiyah-Patodi-Singer index of D^L_T for a manifold with boundary defined as in [66]. The proof uses Braverman's L^2-interpretation of the transversal index [15, §5]. The proof of (4.7) for $\gamma = 0$ is then easy.

2) A second key result is as follows. Let (N, ω^N, J^N) be a compact symplectic manifold with a prequantum line bundle (F, h^F, ∇^F) (see Section 1.1). We suppose that G acts on N and the action lifts to F as above with the associated moment map $\eta : N \to \mathfrak{g}^*$, etc. For $\gamma \in \Lambda^*_+$, set $Q(F)^{-\gamma} = \dim \mathrm{Hom}_G((V^G_\gamma)^*, Q(F))$, where Hom_G is the linear space of G-homomorphisms. Let $L \otimes F$ be the obvious prequantum line bundle over $X \times N$. Then we have

$$Q(L \otimes F)^{\gamma=0} = \sum_{\gamma \in \Lambda^*_+} Q(L)^\gamma \cdot Q(F)^{-\gamma}. \tag{4.9}$$

The proof of Theorem 4.3 is obtained in [45] by combining these two arguments. □

References

[1] M. F. Atiyah, *Elliptic operators and compact groups*, Lecture Notes in Mathematics, Vol. 401, Springer-Verlag, Berlin, 1974.

[2] M. Audin, *Torus actions on symplectic manifolds*. Second revised edition. Progress in Mathematics, 93. Birkhäuser Verlag, Basel, 2004. viii+325 pp.

[3] N. Berline, E. Getzler, and M. Vergne, *Heat kernels and Dirac operators*, Grundl. Math. Wiss. Band 298, Springer-Verlag, Berlin, 1992.

[4] R. Berman, B. Berndtsson, and J. Sjöstrand, *A direct approach to Bergman kernel asymptotics for positive line bundles*, Ark. Mat. **46** (2008), 197–217.

[5] O. Biquard, *Métriques kählériennes à courbure scalaire constante*, Séminaire Bourbaki. Vol. 2004/2005. Astérisque no. 307 (2006), Exp. No. 938, vii, 1–31.

[6] J.-M. Bismut, *Demailly's asymptotic inequalities: a heat equation proof*, J. Funct. Anal. **72** (1987), 263–278.

[7] J.-M. Bismut and G. Lebeau, *Complex immersions and Quillen metrics*, Inst. Hautes Études Sci. Publ. Math. (1991), no. 74, ii+298 pp.

[8] J.-M. Bismut and É. Vasserot, *The asymptotics of the Ray-Singer analytic torsion associated with high powers of a positive line bundle*, Comm. Math. Phys. **125** (1989), no. 2, 355–367.

[9] M. Bordemann, E. Meinrenken and M. Schlichenmaier, *Toeplitz quantization of Kähler manifolds and* $\mathrm{gl}(N)$*,* $N \to \infty$ *limits*, Comm. Math. Phys. **165** (1994), 281–296.

[10] D. Borthwick and A. Uribe, *Almost complex structures and geometric quantization*, Math. Res. Lett. **3** (1996), no. 6, 845–861. Erratum: **5** (1998), 211–212.

[11] Th. Bouche, *Convergence de la métrique de Fubini-Study d'un fibré linéare positif*, Ann. Inst. Fourier (Grenoble) **40** (1990), 117–130.

[12] L. Boutet de Monvel and V. Guillemin, *The spectral theory of Toeplitz operators*, Annals of Mathematics Studies, vol. 99, Princeton University Press, 1981.

[13] L. Boutet de Monvel and J. Sjöstrand, *Sur la singularité des noyaux de Bergman et de Szegő*, Soc. Math. France, Paris, 1976, pp. 123–164. Astérisque, No. 34–35.

[14] M. Braverman, *Vanishing theorems on covering manifolds*, Contemp. Math., **213** (1999), 1–23.

[15] M. Braverman, *Index theorem for equivariant Dirac operators on non-compact manifolds*, K-Theory **27** (2002), no. 1, 61–101.

[16] D. Burns, V. Guillemin, Z. Wang, *Stability Functions*, arXiv:0804.3225, Geom. Funct. Anal.19 (2010), 1258-1295.

[17] D. Catlin, *The Bergman kernel and a theorem of Tian*, Analysis and geometry in several complex variables (Katata, 1997), Trends Math., Birkhäuser Boston, 1999, 1–23.

[18] L. Charles, *Berezin-Toeplitz operators, a semi-classical approach*, Comm. Math. Phys. **239** (2003), 1–28.

[19] L. Charles, *Toeplitz operators and hamiltonian torus action*, J. Funct. Anal. **236** (2006), 299-350.

[20] X. Dai, K. Liu, and X. Ma, *On the asymptotic expansion of Bergman kernel*, J. Differential Geom. **72** (2006), 1-41; announced in C. R. Math. Acad. Sci. Paris **339** (2004), no. 3, 193–198.

[21] J.-P. Demailly, *Champs magnétiques et inegalités de Morse pour la* d''*–cohomologie*, Ann. Inst. Fourier (Grenoble) **35** (1985), 189–229.

[22] S. K. Donaldson, *Remarks on gauge theory, complex geometry and* 4*-manifold topology*, Fields Medallists' lectures, 384–403, World Sci. Ser. 20th Century Math., 5, World Sci. Publishing, River Edge, NJ, 1997.

[23] S. K. Donaldson, *Scalar curvature and projective embeddings. I*, J. Differential Geom. **59** (2001), no. 3, 479–522.

[24] B. Fedosov, *Deformation quantization and index theory.* Mathematical Topics. 9. Berlin: Akademie Verlag. 1996, 325 pp.

[25] C. Fefferman, *The Bergman kernel and biholomorphic mappings of pseudoconvex domains*, Invent. Math. **26** (1974), 1–65.

[26] J. Fine, *Calabi flow and projective embeddings*, with an appendix by K. Liu and X. Ma, *Asymptotic of the operators Q_k*, arXiv: 0811.0155. J. Differential Geom. to appear.

[27] J. Fine, to appear.

[28] W. Fulton, *Introduction to toric varieties*, Annals of Mathematics Studies, 131. Princeton University Press, Princeton, NJ, 1993. xii+157 pp.

[29] V. Guillemin and S. Sternberg, *Geometric quantization and multiplicities of group representations*, Invent. Math. **67** (1982), no. 3, 515–538.

[30] B. Hall, W. Kirwin, *Unitarity in "quantization commutes with reduction"*, arXiv:math/0610005, Comm. Math. Phys. 275 (2007), 401 – 442.

[31] A. V. Karabegov and M. Schlichenmaier, *Identification of Berezin-Toeplitz deformation quantization*, J. Reine Angew. Math. **540** (2001), 49–76.

[32] F. C. Kirwan, *Cohomology of quotients in symplectic and algebraic geometry.* Mathematical Notes, 31. Princeton University Press, Princeton, NJ, 1984. i+211 pp.

[33] B. Kostant, *Quantization and unitary representations.* Lect. Notes in Math. **170** (1970), 87-207.

[34] E. Lerman, *Symplectic cuts*, Math. Res. Lett. **2** (1995), 247–258.

[35] K. Liu and X. Ma, *A remark on 'some numerical results in complex differential geometry'*, Math. Res. Lett. **14** (2007), no. 2, 165–171.

[36] Z. Lu, *On the lower order terms of the asymptotic expansion of Tian-Yau-Zelditch*, Amer. J. Math. **122** (2000), no. 2, 235–273.

[37] X. Ma and G. Marinescu, *The* Spin^c *Dirac operator on high tensor powers of a line bundle*, Math. Z. **240** (2002), no. 3, 651–664.

[38] X. Ma and G. Marinescu, *Generalized Bergman kernels on symplectic manifolds*, Adv. Math. **217** (2008), no. 4, 1756–1815; announced in C. R. Math. Acad. Sci. Paris **339** (2004), no. 7, 493–498.

[39] X. Ma and G. Marinescu, *The first coefficients of the asymptotic expansion of the Bergman kernel of the spin^c Dirac operator*, Internat. J. Math. **17** (2006), 737–759.

[40] X. Ma and G. Marinescu, *Toeplitz operators on symplectic manifolds*, J. Geom. Anal. **18** (2008), 565–611.

[41] X. Ma and G. Marinescu, *Holomorphic Morse inequalities and Bergman kernels*, Progress in Mathematics, vol. 254, Birkhäuser Boston Inc., 2007, 422pp.

[42] X. Ma and G. Marinescu, *Berezin-Toeplitz quantization of Kähler manifolds*, Preprint 2010.

[43] X. Ma and W. Zhang, *Bergman kernels and symplectic reduction*, arXiv:math.DG/0607404, Astérisque **318** (2008), 154 pp; announced in C. R. Math. Acad. Sci. Paris **341** (2005), 297-302; announced also in *Toeplitz quantization and symplectic reduction*, Differential Geometry and Physics. Eds. M.-L. Ge and W. Zhang, Nankai Tracts in Mathematics Vol. **10**, World Scientific, (2006), 343-349.

[44] X. Ma and W. Zhang, *Superconnection and family Bergman kernels*, C. R. Math. Acad. Sci. Paris **344** (2007), 41–44.

[45] X. Ma and W. Zhang, *Geometric quantization for proper moment maps*, C. R. Math. Acad. Sci. Paris **347** (2009), 389-394. Full version: arXiv: 0812.3989.

[46] V. Mathai and W. Zhang, *Geometric quantization for proper actions*, with an appendix by U. Bunke, arXiv:0806.3138. Adv. Math. to appear.

[47] E. Meinrenken, *On Riemann-Roch formulas for multiplicities*, J. Amer. Math. Soc. **9** (1996), no. 2, 373–389.

[48] E. Meinrenken, *Symplectic surgery and the* Spin^c*-Dirac operator*, Adv. Math. **134** (1998), no. 2, 240–277.

[49] E. Meinrenken and R. Sjamaar, *Singular reduction and quantization*. Topology **38** (1999), 699-762.

[50] R. Paoletti, *Moment maps and equivariant Szegö kernels*, J. Symplectic Geom. **2** (2003), 133–175.

[51] R. Paoletti, *The Szegö kernel of a symplectic quotient*, Adv. Math. **197** (2005), 523-553.

[52] R. Paoletti, *Scaling limits for equivariant Szego kernels*, arXiv:math/0612547, J. Symplectic Geom. 6 (2008), 9–32.

[53] R. Paoletti, *Szego kernels, Toeplitz operators, and equivariant fixed point formulae*, arXiv:0707.1375, J. Anal. Math. 106 (2008), 209–236.

[54] P.-É. Paradan, *Localization of the Riemann-Roch character*. J. Funct. Anal. **187** (2001), no. 2, 442–509.

[55] P.-É. Paradan, Spin^c*-quantization and the* K*-multiplicities of the discrete series*. Ann. Sci. Ecole Norm. Sup. (4) **36** (2003), no. 5, 805–845.

[56] P.-É. Paradan, *Formal geometric quantization*, Ann. Inst. Fourier. **59** (2009), 199–238.

[57] P.-É. Paradan, *Multiplicities of the discrete series*. arXiv:0812.0059, 38 pp.

[58] P.-É. Paradan, *Formal geometric quantization II*, arXiv:0906.4436

[59] W.-D. Ruan, *Canonical coordinates and Bergman metrics*, Comm. Anal. Geom. **6** (1998), no. 3, 589–631.

[60] M. Schlichenmaier, *Deformation quantization of compact Kähler manifolds by Berezin-Toeplitz quantization*, Conférence Moshé Flato 1999, Vol. II (Dijon), Math. Phys. Stud., vol. 22, Kluwer Acad. Publ., Dordrecht, 2000, pp. 289–306.

[61] B. Shiffman and S. Zelditch, *Asymptotics of almost holomorphic sections of ample line bundles on symplectic manifolds*, J. Reine Angew. Math. **544** (2002), 181–222.

[62] J.-M. Souriau, Structure des systèmes dynamiques. Maîtrises de mathématiques Dunod, Paris 1970, xxxii+414 pp.

[63] C. Teleman, *The quantization conjecture revisited*, Ann. of Math. **152** (2000), 1–43.

[64] G. Tian, *On a set of polarized Kähler metrics on algebraic manifolds*, J. Differential Geom. **32** (1990), 99–130.

[65] Y. Tian and W. Zhang, *An analytic proof of the geometric quantization conjecture of Guillemin-Sternberg*, Invent. Math. **132** (1998), no. 2, 229–259.

[66] Y. Tian and W. Zhang, *Quantization formula for symplectic manifolds with boundary*, Geom. Funct. Anal. **9** (1999), no. 3, 596–640.

[67] M. Vergne, *Multiplicities formula for geometric quantization. I, II*, Duke Math. J. **82** (1996), no. 1, 143–179, 181–194.

[68] M. Vergne, *Quantification géométrique et réduction symplectique*. Séminaire Bourbaki, Vol. 2000/2001. Astérisque No. 282 (2002), Exp. No. 888, viii, 249–278.

[69] M. Vergne, *Applications of equivariant cohomology*, International Congress of Mathematicians. Vol. I, Eur. Math. Soc., Zürich, 2007, pp. 635–664.

[70] X. Wang, *Canonical metrics on stable vector bundles*, Comm. Anal. Geom. **13** (2005), 253–285.

[71] S. -T.Yau, *Nonlinear analysis in geometry*. Enseign. Math. (2) **33** (1987), 109–158.

[72] S. Zelditch, *Szegö kernels and a theorem of Tian*, IMRN (1998), 317–331.

[73] W. Zhang, *Holomorphic quantization formula in singular reduction*, Commun. Contemp. Math. **1** (1999), no. 3, 281–293.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Scalar Curvature, Conformal Geometry, and the Ricci Flow with Surgery

Fernando Codá Marques*

Abstract

In this note we will review recent results concerning two geometric problems associated to the scalar curvature. In the first part we will review the solution to Schoen's conjecture about the compactness of the set of solutions to the Yamabe problem. It has been discovered, in a series of three papers, that the conjecture is true if and only if the dimension is less than or equal to 24. In the second part we will discuss the connectedness of the moduli space of metrics with positive scalar curvature in dimension three. In two dimensions this was proved by Weyl in 1916. This is a geometric application of the Ricci flow with surgery and Perelman's work on Hamilton's Ricci flow.

Mathematics Subject Classification (2010). Primary 53C21; Secondary 83C05.

Keywords. Scalar curvature; Yamabe problem; Ricci flow with surgery.

1. The Compactness Conjecture

The celebrated Riemann's Uniformization Theorem states that any compact Riemannian surface can be conformally deformed to a surface of constant Gauss curvature. We will begin by describing the Yamabe Problem, which is a way of generalizing uniformization to higher-dimensional manifolds.

1.1. The Yamabe problem. Let (M^n, g) be a smooth compact Riemannian manifold of dimension $n \geq 3$. The *conformal class* of g is the set

$$[g] = \{\tilde{g} = \phi^2 g : \phi \in C^\infty(M), \phi > 0\}.$$

The *Yamabe Problem* consists of finding a metric $\tilde{g} \in [g]$ of constant scalar curvature.

*Instituto de Matemática Pura e Aplicada (IMPA), Estrada Dona Castorina 110, 22460-320, Rio de Janeiro - RJ, Brazil. E-mail: coda@impa.br.

If we write $\tilde{g} = u^{\frac{4}{n-2}} g$, $u > 0$, the transformation law for the scalar curvature is

$$R_{\tilde{g}} = -\frac{4(n-1)}{n-2} u^{-\frac{n+2}{n-2}} \Big(\Delta_g u - \frac{n-2}{4(n-1)} R_g u \Big).$$

Here R_g and $R_{\tilde{g}}$ denote the scalar curvatures of g and $\tilde{g}$, respectively, and Δ_g is the Laplace-Beltrami operator associated with g. The linear operator $L_g = \Delta_g - \frac{n-2}{4(n-1)} R_g$ is usually called the *conformal Laplacian* of g.

Therefore the Yamabe Problem is equivalent to finding a positive solution u to the partial differential equation

$$L_g(u) + c(n) K u^{\frac{n+2}{n-2}} = 0 \tag{1}$$

for some constant K, where $c(n) = \frac{n-2}{4(n-1)}$.

It turns out that the constant scalar curvature metrics $\tilde{g} \in [g]$ are the critical points of the functional

$$Q(\tilde{g}) = \frac{\int_M R_{\tilde{g}} \, dv_{\tilde{g}}}{\Big(\int_M dv_{\tilde{g}} \Big)^{\frac{n-2}{n}}} ,$$

known as the *normalized total scalar curvature functional*, when restricted to the conformal class $[g]$. This variational structure played a prominent role in the solution of the Yamabe Problem. In fact, after the initial paper of Yamabe [66], which contained a gap, the combined works of Trudinger [63], Aubin [2], and Schoen [52] established the existence of a minimizing solution in the conformal class $[g]$ of any given (M, g).

It is natural to ask if such solution is unique. In that respect, the conformal classes of compact Riemannian manifolds should be classified in three types, according to the sign of the *Yamabe quotient*:

$$Q(M, g) = \inf_{\tilde{g} \in [g]} Q(\tilde{g}).$$

If the Yamabe quotient is negative, it follows from the Maximum Principle, applied to equation (1), that the solution (of negative constant scalar curvature) is unique. If the Yamabe quotient is zero, the solution (of zero scalar curvature) is unique up to a constant factor. The structure of the set of solutions in the positive Yamabe quotient case can be very rich though.

1.2. The set of solutions and some conjectures. It is convenient, in the positive Yamabe quotient case, to normalize the scalar curvature of the solutions to be $n(n-1)$, for example. Hence let

$$\mathcal{M}_g = \{\tilde{g} \in [g] : R_{\tilde{g}} = n(n-1)\}$$

be the space of solutions with such normalization.

The simplest and most important example is given by the standard sphere (S^n, g_0). Here g_0 denotes the metric induced by the Euclidean metric on the unit sphere $S^n \subset \mathbb{R}^{n+1}$. This case is special because the standard sphere is the only compact manifold, up to conformal equivalence, which admits a noncompact group of conformal transformations $Conf(S^n)$. By looking at the transformation law for the Einstein tensor (traceless Ricci) under conformal changes, Obata (see [44]) proved that

$$\mathcal{M}_{g_0} = \{\psi^*(g_0) : \psi \in Conf(S^n)\}.$$

Since these metrics are all isometric to each other, every solution is minimizing in this particular example. Notice that $\mathcal{M}_{g_0}$ is noncompact.

The example of $S^1(L) \times S^{n-1}$ with the product metric (L denotes the length of the circle factor) was analyzed by R. Schoen in [54] . The set of solutions in this case can be described as a finite union of one-parameter families, parametrized by circles, and depending on L. If L is big, there exists a large number of high energy solutions with high Morse index. In fact, a theorem of Pollack ([49]) shows that every compact Riemannian manifold of positive scalar curvature can be perturbed, in the C^0 topology, to have as many solutions as desired. These solutions generally have high energy and index.

In order to obtain more refined information about $\mathcal{M}_g$, through the Morse theory of Palais and Smale (see [45]), one needs to prove a priori estimates (or compactness) for the set of solutions. The difficulty of that has its origin in the fact that these estimates fail in the case of the standard sphere (S^n, g_0).

In a topics course at Stanford in 1988 (see also [55] and [56]), motivated by the study of the locally conformally flat case, R. Schoen proposed the following conjecture, together with an outline of a strategy to prove it:

Compactness Conjecture

The set $\mathcal{M}_g$ of solutions to the Yamabe Problem, in the positive Yamabe quotient case, is compact (in any C^k topology) unless the manifold is conformally equivalent to the standard sphere.

The cases which were covered in the Stanford notes are the locally conformally flat case, published in [55], and the three dimensional case, the argument for which is in the paper of Schoen and Zhang [61] (used there to establish a single simple point of blow-up for the prescribed scalar curvature problem on S^3). In dimensions 4 and 5, the conjecture was proved by O. Druet (see [19]).

It follows from basic arguments in blow-up analysis that non-converging sequences of solutions to the Yamabe Problem have to concentrate and form bubbles (spheres) at some points of the manifold, referred to as *blow-up points*. This phenomenon can be explicitly illustrated in the case of the standard sphere (S^n, g_0). If $\pi : S^n - \{p\} \to \mathbb{R}^n$ denotes the stereographic projection, the metrics

$\tilde{g}_\varepsilon = \pi^*(4u_\varepsilon^{\frac{4}{n-2}}\delta)$, where

$$u_\varepsilon(x) = \Big(\frac{\varepsilon}{\varepsilon^2 + |x|^2}\Big)^{\frac{n-2}{2}},$$

lie in $\mathcal{M}_{g_0}$. Notice that u_ε blows-up at the origin as $\varepsilon \to 0$.

The main step in Schoen's program to establish compactness in dimensions greater than or equal to 6 consisted in proving a related statement, known as the *Weyl Vanishing Conjecture*, concerning the location of possible blow-up points:

If $\overline{x} \in M$ is a blow-up point of a sequence of solutions $\tilde{g}_\nu = u_\nu^{\frac{4}{n-2}} g$ to the Yamabe Problem, then the Weyl tensor of the metric g should satisfy

$$\nabla^k W_g(\overline{x}) = 0$$

for all $0 \leq k \leq [\frac{n-6}{2}]$.

Over the past several years many people have worked on these problems. It follows from the works of the author ([37]) and Y. Y. Li and L. Zhang ([33]) that compactness holds for $n \leq 7$ in general, and for arbitrary n under the assumption that the Weyl tensor vanishes nowhere to second order. In [34], Li and Zhang proved compactness for $n \leq 11$.

We should also point out that non-smooth blow-up examples were obtained by A. Ambrosetti and A. Malchiodi in [1], and by M. Berti and Malchiodi in [4]. In [20] O. Druet and E. Hebey have also obtained blow-up examples for Yamabe-type equations.

In the past few years the Compactness Conjecture was completely solved in a series of three articles ([11], [12], and [28]). The results of [11], [12], and [28] put together give the following answer to the Compactness Conjecture:

The Compactness Conjecture is true if and only if $n \leq 24$.

In the next sections we will give an overview of the results in these papers and of their proofs. We refer the reader to [9] and [38] for related accounts (see also [8]).

1.3. A compactness theorem. Throughout this section (M^n, g) will be a smooth compact Riemannian manifold of dimension $n \geq 3$ and of positive Yamabe quotient.

For any $p \in [1, \frac{n+2}{n-2}]$ we define

$$\Phi_p = \{u > 0, u \in C^\infty(M) : L_g u + K u^p = 0 \ \text{ on } \ M\}.$$

Although the geometric problem corresponds to the exponent $p = \frac{n+2}{n-2}$, critical with respect to the Sobolev embeddings, the consideration of the subcritical

solutions is useful for the purposes of applying Morse theory and computing the total Leray-Schauder degree of the problem.

In [28], M. Khuri, R. Schoen, and the author proved the following theorem:

Theorem 1.1. *Suppose $3 \leq n \leq 24$. If (M^n, g) is not conformally diffeomorphic to (S^n, g_0), then for any $\varepsilon > 0$ there exists a constant $C > 0$ depending only on g and ε such that*

$$C^{-1} \leq u \leq C \quad \text{and} \quad \| u \|_{C^{2,\alpha}} \leq C,$$

for all $u \in \cup_{1+\varepsilon \leq p \leq \frac{n+2}{n-2}} \Phi_p$, where $0 < \alpha < 1$.

The following compactness result is a corollary of the previous theorem and standard elliptic regularity theory:

Corollary. *Suppose $3 \leq n \leq 24$. If (M^n, g) is not conformally diffeomorphic to (S^n, g_0), then the set $\mathcal{M}_g$ is compact in any C^k topology.*

The proof of Theorem 1.1 is by contradiction and follows the strategy outlined in the notes of R. Schoen ([53]). In order to illustrate the ideas let us for simplicity restrict ourselves to the critical exponent $p = \frac{n+2}{n-2}$. Suppose then that there exists a sequence $u_\nu \in \Phi_{\frac{n+2}{n-2}}$ such that $\max_M u_\nu = u_\nu(x_\nu) \to \infty$ as $\nu \to \infty$. Suppose $\overline{x} = \lim x_\nu$.

The first step is to obtain sharp approximations of the blowing-up sequence of solutions in a neighborhood of the blow-up point. This is achieved by establishing optimal pointwise estimates which generalize the ones obtained by the author in [37]. These estimates assume the blow-up point is isolated simple (one bubble only). After the strategy is carried out with success for that particular case, the results can be used to handle the more general case of multiple blow-up by scaling arguments.

The important point of the estimates of [28] is that in high dimensions it is necessary to have an expansion of u_ν that goes beyond the rotationally symmetric first approximation (standard bubble). The approximate solutions used are the same ones introduced by S. Brendle in [10] to generalize the test function estimates of Aubin ([2]) and of Hebey and Vaugon ([26]), and prove convergence of the Yamabe flow in any dimension.

The basic idea then is to use the *Pohozaev Identity* as an obstruction tool in order to rule out the formation of bubbles at blow-up points. The following is a general version of it in geometric form:

Proposition 1.2 (Pohozaev Identity, [60]). *Let (Ω^n, g) be a Riemannian domain, $n \geq 3$. If X is a vector field on Ω, then*

$$\frac{n-2}{2n} \int_\Omega X(R_g)\, dv_g + \int_\Omega \langle \mathcal{D}_g X, T_g \rangle\, dv_g = \int_{\partial\Omega} T_g(X, \eta_g)\, d\sigma_g.$$

Here $T_g = Ric_g - \frac{R_g}{n} g$ is the traceless Ricci tensor, $(\mathcal{D}_g X)_{ij} = X_{i;j} + X_{j;i} - \frac{2}{n} div_g X\, g_{ij}$ is the conformal Killing operator, and η_g is the outward unit normal to $\partial\Omega$.

Since the scalar curvature of $g_\nu = u_\nu^{\frac{4}{n-2}} g$ is constant, the Pohozaev identity applied to the geodesic ball $B_\delta(x_\nu) = \{p \in M : r = d_g(x_\nu, p) \leq \delta\}$, endowed with the Riemannian metric g_ν and the radial vector field $X = r\frac{\partial}{\partial r}$, $r = d_g(x_\nu, \cdot)$, yields

$$\int_{B_\delta} \langle \mathcal{D}_{g_\nu} X, T_{g_\nu} \rangle \, dv_{g_\nu} = \int_{\partial B_\delta} T_{g_\nu}(X, \eta_{g_\nu}) \, d\sigma_{g_\nu}. \tag{2}$$

The idea then is to expand both integrals in powers of $\varepsilon_\nu = u_\nu(x_\nu)^{-\frac{2}{n-2}}$ and compare.

It turns out that the boundary integrals in the identity (2), when appropriately normalized, converge to a quantity which can be bounded above by $-m$, where m is the ADM mass of the asymptotically flat and scalar flat metric $\hat{g} = G_L^{\frac{4}{n-2}} g$. Here G_L denotes the Green's function of the conformal Laplacian with pole at $\overline{x}$. The vanishing of the Weyl tensor to order $[\frac{n-6}{2}]$ at $\overline{x}$ is necessary in order for the mass of $\hat{g}$ to be well-defined.

In order to analyze the interior integrals in (2), we let $(x_1, \ldots, x_n)$ be normal coordinates centered at x_ν. We write the components of the metric g in the form

$$g_{ij}(x) = \exp(h_{ij}(x)),$$

and look at the Taylor expansion of h_{ij} around the origin:

$$h_{ij}(x) = H_{ij}(x) + O(|x|^{d+1}),$$

where $d = [\frac{n-2}{2}]$. It is convenient to work in conformal normal coordinates (see [32]) to simplify the computations. In that case H_{ij} is a matrix whose entries are polynomials of degree less than or equal to d, and such that

1. $H_{ij}(x) = H_{ji}(x)$,
2. $\sum_k H_{kk}(x) = 0$,
3. $\sum_k x_k H_{ik}(x) = 0$,

for all $1 \leq i, j \leq n$ and $x \in \mathbb{R}^n$. Let us denote the vector space of such matrices by $\mathcal{V}_n$.

The optimal pointwise estimates established in [28] lead to an expansion of the interior integral of (2) in powers of ε_ν, much like in the work of Aubin [2]. It turns out that the relevant terms in this expansion are encoded in a canonical quadratic form $\mathcal{P}_n$ defined on $\mathcal{V}_n$. If n is odd, for instance, and $\sum_{i,j} \partial_i \partial_j H_{ij} = 0$, the quadratic form is given by

$$\mathcal{P}_n(H, H) = \sum_{i,j,l} \sum_{s,t=2}^{d} c_{s+t} \int_{S_1^{n-1}(0)} \left(-\frac{1}{2} \partial_j H_{ij}^{(s)} \partial_l H_{il}^{(t)} + \frac{1}{4} \partial_l H_{ij}^{(s)} \partial_l H_{ij}^{(t)} \right).$$

Here $H^{(s)}$ denotes the homegeneous component of H of degree s, and

$$c_k = \int_0^\infty \frac{(s^2-1)s^{k+n-3}}{(1+s^2)^{n-1}} ds$$

for $k < n-2$. We refer the reader to the appendix of [28] for a complete definition of $\mathcal{P}_n$.

The proof of Theorem 1.1 relies on an eigenvalue analysis done in [28]:

Proposition 1.3. *The quadratic form $\mathcal{P}_n$, defined on $\mathcal{V}_n$, is positive definite if $n \leq 24$. Moreover, it has negative eigenvalues if $n \geq 25$.*

Suppose $n \leq 24$. The vanishing of H_{ij} at $\overline{x}$, which is equivalent to the vanishing of the Weyl tensor to order $[\frac{n-6}{2}]$, follows from the positivity of $\mathcal{P}_n$ and estimates of the boundary term of (2). Hence the mass of $\hat{g}$ is well-defined. Since (M^n, g) is not conformally equivalent to the standard sphere, the metric $\hat{g}$ is not flat, and therefore $m > 0$ by the Positive Mass Theorem. It can be seen that this is in contradiction with the positivity of $\mathcal{P}_n$ by letting $\nu \to \infty$ in (2). Therefore we conclude that blowup cannot occur if $n \leq 24$.

Remark: The Positive Mass Theorem of General Relativity has been established by Schoen and Yau [57] in general for dimensions $n \leq 7$. In [65] E. Witten established it in any dimension for spin manifolds, while the locally conformally flat case was handled by a special argument in [59]. See [36] for work towards the general higher dimensional version.

As one of the consequences of compactness, we obtain the following statement about generic metrics (assuming $n \leq 24$):

Corollary. *Suppose that (M^n, g) satisfies the assumptions of Theorem 1.1, and assume that all critical points in $[g]$ are nondegenerate. Then there are a finite number of critical points $g_1, \ldots, g_k$ and we have*

$$1 = \sum_{j=1}^{k} (-1)^{I(g_j)},$$

where $I(g_j)$ denotes the Morse index of the variational problem with volume constraint.

1.4. Noncompactness results. One way to understand the noncompactness results is to look closely at the model case of the standard sphere (S^n, g_0). As was pointed out before, the set of solutions of scalar curvature $n(n-1)$ coincides in this particular case with the set of metrics coming from the action of the conformal group on g_0, and therefore it is noncompact. We might then be tempted to ask the following question:

Is there a way of perturbing the conformal structure of the standard sphere so that the noncompactness persists?

It follows from Theorem 1.1 that this is impossible if $n \leq 24$, but it turns out that the answer to this question is yes for all $n \geq 25$.

In a surprising paper ([11]), Simon Brendle constructed in 2008 the first examples of C^∞ metrics for which the compactness statement fails. These metrics were small perturbations of the standard sphere in dimensions greater than or equal to 52. In a subsequent article ([12]) Brendle and the author were able to extend these examples to the dimensions $25 \leq n \leq 51$.

The main theorems of [11] and [12] put together give:

Theorem 1.4. *Suppose $n \geq 25$. Given any $\varepsilon > 0$, there exists a smooth Riemannian metric g on S^n and a sequence of positive functions $v_\nu \in C^\infty(S^n)$ ($\nu \in \mathbb{N}$) with the following properties:*

(i) $\|g - g_0\|_{C^{[1/\varepsilon]}(S^n)} < \varepsilon$,

(ii) *g is not conformally flat,*

(iii) *v_ν is a solution of the Yamabe equation (1) for all $\nu \in \mathbb{N}$,*

(iv) $Q(v_\nu^{\frac{4}{n-2}} g) \nearrow Q(S^n, g_0)$ *as $\nu \to \infty$,*

(v) $\sup_{S^n} v_\nu \to \infty$ *as $\nu \to \infty$.*

The first step of the proof consists in reducing the construction to solving a finite dimensional variational problem. This follows from a procedure known as the Lyapunov-Schmidt reduction which we now briefly describe.

Since the standard sphere minus a point is conformally equivalent to the Euclidean space $(\mathbb{R}^n, \delta)$ through the stereographic projection, we can translate the problem to the Euclidean setting. In this setting the solutions of the Yamabe equation

$$\Delta u + n(n-2)u^{\frac{n+2}{n-2}} = 0 \tag{3}$$

on $\mathbb{R}^n$ are the functions

$$u_{(\xi,\varepsilon)}(x) = \left(\frac{\varepsilon}{\varepsilon^2 + |x-\xi|^2}\right)^{\frac{n-2}{2}},$$

where $(\xi, \varepsilon) \in \mathbb{R}^n \times (0, \infty)$. The solutions of the equation (3) can be also seen as the critical points (at the same energy level) of the functional

$$\mathcal{F}_\delta(u) = \int_{\mathbb{R}^n} \left(|\nabla u|^2 - (n-2)^2 |u|^{\frac{2n}{n-2}}\right) dx$$

restricted to the space

$$\mathcal{E} = \left\{ w \in L^{\frac{2n}{n-2}}(\mathbb{R}^n) \cap W^{1,2}_{loc}(\mathbb{R}^n) : \int_{\mathbb{R}^n} |\nabla w|^2 \, dx < \infty \right\}.$$

We consider Riemannian metrics g which are perturbations of the Euclidean metric with compact support. We write $g(x) = \exp(h(x))$, where $h(x)$ is a trace-free symmetric two-tensor on $\mathbb{R}^n$ satisfying $h(x) = 0$ for $|x| \geq 1$, and

$$|h(x)| + |\partial h(x)| + |\partial^2 h(x)| \leq \alpha$$

for some small $\alpha > 0$ and all $x \in \mathbb{R}^n$.

Although the linearization of the equation (3) has a kernel, it is possible to apply the Implicit Function Theorem if we restrict ourselves to the orthogonal subspace

$$\mathcal{E}_{(\xi,\varepsilon)} = \left\{ w \in \mathcal{E} : \int_{\mathbb{R}^n} \varphi_{(\xi,\varepsilon,k)}\, w\, dx = 0 \quad \text{for } k = 0, 1, \ldots, n \right\},$$

where

$$\varphi_{(\xi,\varepsilon,0)}(x) = \left(\frac{\varepsilon}{\varepsilon^2 + |x-\xi|^2} \right)^{\frac{n+2}{2}} \frac{\varepsilon^2 - |x-\xi|^2}{\varepsilon^2 + |x-\xi|^2}$$

and

$$\varphi_{(\xi,\varepsilon,k)}(x) = \left(\frac{\varepsilon}{\varepsilon^2 + |x-\xi|^2} \right)^{\frac{n+2}{2}} \frac{2\varepsilon\,(x_k - \xi_k)}{\varepsilon^2 + |x-\xi|^2}$$

for $k = 1, \ldots, n$.

As a consequence we can find an $(n+1)$-dimensional family of approximate solutions:

Proposition 1.5. *Let $\alpha > 0$ be sufficiently small, depending only on the dimension. Given $(\xi, \varepsilon) \in \mathbb{R}^n \times (0, \infty)$, there exists a function $v_{(\xi,\varepsilon)} \in \mathcal{E}$ such that $v_{(\xi,\varepsilon)} - u_{(\xi,\varepsilon)} \in \mathcal{E}_{(\xi,\varepsilon)}$ and*

$$\int_{\mathbb{R}^n} \left(\langle \nabla v_{(\xi,\varepsilon)}, \nabla \psi \rangle_g + \frac{n-2}{4(n-1)} R_g\, v_{(\xi,\varepsilon)}\, \psi - n(n-2)\, |v_{(\xi,\varepsilon)}|^{\frac{4}{n-2}}\, v_{(\xi,\varepsilon)}\, \psi \right) = 0$$

for all test functions $\psi \in \mathcal{E}_{(\xi,\varepsilon)}$.

The problem reduces to finding critical points of the finite-dimensional functional $\mathcal{F}_g : \mathbb{R}^n \times (0, \infty) \to \mathbb{R}$, given by

$$\mathcal{F}_g(\xi, \varepsilon) = \int_{\mathbb{R}^n} \left(|\nabla v_{(\xi,\varepsilon)}|_g^2 + \frac{n-2}{4(n-1)} R_g\, v_{(\xi,\varepsilon)}^2 - (n-2)^2\, |v_{(\xi,\varepsilon)}|^{\frac{2n}{n-2}} \right) dx.$$

We have that $(\xi, \varepsilon) \in \mathbb{R}^n \times (0, \infty)$ is a critical point of $\mathcal{F}_g$ if and only if $v_{(\xi,\varepsilon)}$ is a nonnegative weak solution (therefore smooth by a result of Trudinger [63]) of the Yamabe equation

$$\Delta_g v_{(\xi,\varepsilon)} - \frac{n-2}{4(n-1)} R_g v_{(\xi,\varepsilon)} + n(n-2) v_{(\xi,\varepsilon)}^{\frac{n+2}{n-2}} = 0.$$

The positivity of $v_{(\xi,\varepsilon)}$ then follows from the Maximum Principle and the fact that $v_{(\xi,\varepsilon)}$ and $u_{(\xi,\varepsilon)}$ are close to each other in the norm $\|w\|_{\mathcal{E}} = \int_{\mathbb{R}^n} |dw|^2 dx$.

The construction of the counterexample relies on a gluing procedure based on some local model metrics. The model metrics $g(x) = \exp(h(x))$ are such that

$$h_{ik}(x) = \mu\, \lambda^{2m}\, f(\lambda^{-2}\, |x|^2) \sum_{p,q} W_{ipkq} x_p x_q$$

for $|x| \leq \rho$, where μ, λ, ρ are positive constants satisfying $\mu \leq 1$ and $\lambda \leq \rho \leq 1$, f is a polynomial, and $W : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a nontrivial multi-linear form which satisfies all the algebraic properties of the Weyl tensor. It is also necessary that

$$2\deg(f) + 2 < \frac{n-2}{2}.$$

These choices make it possible to approximate the energy function $\mathcal{F}_g(\xi, \varepsilon)$ at appropriate scales by an auxiliary function $F(\xi, \varepsilon)$, $\xi \in \mathbb{R}^n$, $\varepsilon \in (0, \infty)$, and we are left with the algebraic problem of finding a polynomial f such that $F(\xi, \varepsilon)$ has a strict local minimum at $(0, 1)$.

Notice that

$$\mu\, \lambda^{2m}\, f(\lambda^{-2}\, |x|^2) \sum_{p,q} W_{ipkq} x_p x_q$$

belongs to $\mathcal{V}_n$ (as defined in the previous section), and it turns out that the algebraic problem of finding f can be solved when the quadratic form $\mathcal{P}_n$ has negative eigenvalues. It is proven in [11] that f can be chosen of degree 1 for all $n \geq 52$, and in [12] that it can be chosen of degree 3 for all $25 \leq n \leq 51$.

The counterexamples $g(x) = \exp(h(x))$ are obtained by gluing infinite copies of the local models supported in small disjoint balls placed along the x_1-axis. The N-th ball has radius $1/(2N^2)$ and is centered at $y_N = (\frac{1}{N}, 0, \ldots, 0) \in \mathbb{R}^n$, $N \in \mathbb{N}$. If $\eta : \mathbb{R} \to \mathbb{R}$ is a smooth cutoff function such that $\eta(s) = 1$ for $s \leq 1$ and $\eta(s) = 0$ for $s \geq 2$, the two-tensor $h(x)$ is given by

$$h_{ik}(x) = \sum_{N=N_0}^{\infty} \eta(4N^2\, |x - y_N|)\, 2^{-(m+\frac{1}{8})N}\, f(2^N\, |x - y_N|^2)\, H_{ik}(x - y_N),$$

where $y_N = (\frac{1}{N}, 0, \ldots, 0) \in \mathbb{R}^n$, $m = \deg(f)$, $H_{ik}(x) = \sum_{p,q} W_{ipkq} x_p x_q$, and N_0 is sufficiently large.

Since the metric g is in local model form in each of the infinitely many balls $B_{1/(2N^2)}(y_N)$, we can apply the Lyapunov-Schmidt reduction infinitely many times to obtain a sequence v_ν of solutions to the Yamabe equation as in Theorem 1.4.

We should notice that even though the Weyl tensor of these counterexamples vanishes to all orders at the blow-up point ($0 \in \mathbb{R}^n$), recent work of the author ([39]) shows that they can be perturbed to provide counterexamples to the Weyl Vanishing Conjecture as well.

2. The Connectedness Problem

In this section we will discuss a connectedness result for the space of positive scalar curvature metrics on an orientable compact 3-manifold. We refer the reader to [50] for a nice survey on related questions.

In 1916 H. Weyl proved the following connectedness result:

Theorem 2.1 ([64]). *Let g be a metric of positive scalar curvature on the two-sphere S^2. There exists a continuous path of metrics $\mu \in [0,1] \to g_\mu$ on S^2, of positive scalar curvature, such that $g_0 = g$ and g_1 has constant curvature.*

Weyl's proof is a nice application of the uniformization theorem. It is based on the existence of a constant curvature metric $\overline{g}$ in the conformal class of g. We can choose $\overline{g}$ so that $R_{\overline{g}} = 2$. If $\overline{g} = e^{2f}g$, we define $g_\mu = e^{2\mu f}g$. The transformation law for the scalar curvature in two dimensions gives

$$\begin{aligned} R_{g_\mu} &= e^{-2\mu f}\left(R_g - 2\mu\,\Delta_g f\right) \\ &= (1-\mu)R_g e^{-2\mu f} + 2\mu e^{2(1-\mu)f}. \end{aligned}$$

It is clear that $R_{g_\mu} > 0$ for all $\mu \in [0,1]$, if $R_g > 0$. The space of metrics of positive scalar curvature on S^2 is in fact contractible (see [51]).

There are several positive curvature conditions that are satisfied by the standard sphere in higher dimensions (positive scalar curvature, Ricci curvature, sectional curvature, etc). Each one of them leads to a different connectedness problem. Since there is no general uniformization theorem, other tools have to be developed.

In his famous 1982 paper R. Hamilton ([23]) introduced the equation

$$\frac{\partial g}{\partial t} = -2Ric_g,$$

known as the *Ricci flow*, and proved the existence of short time solutions with arbitrary compact Riemannian manifolds as initial data. In [23], Hamilton proved that the Ricci flow preserves positive scalar curvature in any dimension, and that positive Ricci curvature ($Ric > 0$) and positive sectional curvature ($sec > 0$) are both preserved in dimension three. He also proved a convergence result: if $g(t)$ denotes a solution to the Ricci flow on a compact 3-manifold M such that $g(0)$ has positive Ricci curvature, then the flow becomes extinct at finite time $T > 0$, and the volume one rescalings $\tilde{g}(t)$ of $g(t)$ converge to a constant curvature metric as $t \to T$. From that he could conclude that any compact 3-manifold of positive Ricci curvature is diffeomorphic to a spherical quotient S^3/Γ. It also follows from the method that Weyl's connectedness result extends to three dimensions under the conditions $Ric > 0$ or $sec > 0$.

We want to address the connectedness question in dimension three under the weaker condition of positive scalar curvature $R > 0$. In order to state the main result, let us introduce some notation. If M is a compact manifold, we will

denote by $\mathcal{R}_+(M)$ the set of Riemannian metrics g on M with positive scalar curvature R_g. The associated moduli space is the quotient $\mathcal{R}_+(M)/\mathrm{Diff}(M)$ of $\mathcal{R}_+(M)$ under the standard action of the group of diffeomorphisms $\mathrm{Diff}(M)$. Unless otherwise specified, the space of metrics on a given manifold will be endowed with the C^∞ topology.

In [40], we prove the following connectedness theorem:

Theorem 2.2. *Suppose that M^3 is a compact orientable 3-manifold such that $\mathcal{R}_+(M) \neq \emptyset$. Then the moduli space $\mathcal{R}_+(M)/\mathrm{Diff}(M)$ is path-connected.*

As a corollary we obtain:

Corollary. *Let g be a metric of positive scalar curvature on the three-sphere S^3. There exists a continuous path of metrics $\mu \in [0,1] \to g_\mu$ on S^3, of positive scalar curvature, such that $g_0 = g$ and g_1 has constant sectional curvature.*

Remark: Since the set $\mathrm{Diff}_+(S^3)$ of orientation-preserving diffeomorphisms of the 3-sphere is path-connected (J. Cerf, [15]), we have that the total space $\mathcal{R}_+(S^3)$ is path-connected.

We should point out that the results for scalar curvature in higher dimensions are quite different. This was first noticed by N. Hitchin ([27]) in 1974. He considered some index-theoretic invariants associated to the Dirac operator of spin geometry, and proved that the spaces $\mathcal{R}_+(S^{8k})$ and $\mathcal{R}_+(S^{8k+1})$ are disconnected for each $k \geq 1$. In 1988 R. Carr ([14]) proved that the space $\mathcal{R}_+(S^{4k-1})$ has infinitely many connected components for each $k \geq 2$, extending the 7-dimensional case ($k = 2$) established earlier by Gromov and Lawson in 1983 (see Theorem 4.47 of [22]). This result was improved by M. Kreck and S. Stolz ([31]) in 1993, where they show that even the moduli space $\mathcal{R}_+(S^{4k-1})/\mathrm{Diff}(S^{4k-1})$ has infinitely many connected components for $k \geq 2$. This means that on those spheres there are infinitely many nonequivalent metrics of positive scalar curvature which are *exotic* in the sense that they do not come from deformations of the standard metric. The same statement holds true for any nontrivial spherical quotient of dimension greater than or equal to five, as proved by B. Botvinnik and P. Gilkey in [7]. The surgery arguments used in these proofs break down in the three-dimensional case.

The evolution equation for the scalar curvature under the Ricci flow is

$$\frac{\partial R_g}{\partial t} = \Delta R_g + 2|Ric_g|^2.$$

It follows from the parabolic maximum principle that if $R_{g(0)} \geq R_0 > 0$, then

$$\min_M R_{g(t)} \geq \frac{1}{\frac{1}{R_0} - \frac{2}{n}t},$$

and the flow must end in finite time.

In two dimensions Hamilton ([24]) proved a convergence result: if g has positive scalar curvature (or Gauss curvature) on S^2, then the solution to the

normalized Ricci flow with initial condition (S^2, g) converges to a constant curvature metric. (See [17] for an extension to arbitrary g). It is interesting to note that his arguments were made independent of uniformization by Chen, Lu and Tian in [16]. This gives a Ricci flow proof of Weyl's theorem.

The great difficulty in studying the scalar curvature connectedness problem in dimensions greater than two is that the condition of $R_g > 0$ is too weak to imply convergence results. For instance, the condition of positive scalar curvature is stable under connected sums if $n \geq 3$. There is a construction of Gromov and Lawson ([21]) that starts with two compact Riemannian manifolds (M_1^n, g_1) and (M_2^n, g_2), with positive scalar curvature, and replaces the union of two small balls $B_\delta(p_1) \subset M_1$ and $B_\delta(p_2) \subset M_2$ with a small neck-like region N. The result is a metric $g_1 \# g_2$ of positive scalar curvature on the connected sum $M_1 \# M_2$ that coincides with the original metrics g_1 and g_2 outside N. Therefore, unlike in the case of positive Ricci curvature, neck-pinching singularities can occur under the Ricci flow.

In order to deal with this kind of situation Hamilton introduced in [25], in the context of four-manifolds with positive isotropic curvature, a discontinuous evolution process known as *Ricci flow with surgery.* The Ricci flow with surgery, with (M, g_0) as initial condition, can be thought of as a sequence of standard Ricci flows $(M_i, g_i(t))$, each defined for $t \in [t_i, t_{i+1})$ and becoming singular at $t = t_{i+1}$, where $0 = t_0 < t_1 < \cdots < t_i < t_{i+1} < \cdots < \infty$ is a discrete set, $M_0 = M$, and $g_0(0) = g_0$. The initial condition $(M_i, g_i(t_i))$ for each of these Ricci flows is a compact Riemannian manifold obtained from the preceding Ricci flow $(M_{i-1}, g_{i-1}(t))_{t \in [t_{i-1}, t_i)}$ by a specific process called surgery, which depends on some choice of parameters. Entire components with uniformly large curvature are discarded at each t_i. The flow becomes extinct in finite time $T > 0$ if $T = t_{j+1}$ for some $j \geq 0$ and $M_{j+1} = \emptyset$.

In three dimensions the existence of a Ricci flow with surgery and the study of its properties were accomplished by G. Perelman in a series of three papers [46], [47], [48]. It follows from Perelman's breakthroughs that the surgeries needed are of the simplest type, restricted to almost cylindrical regions. He is able to prove, through a backwards induction argument, that if the Ricci flow with surgery of an orientable compact Riemannian 3-manifold becomes extinct in finite time, then the manifold is diffeomorphic to a connected sum of spherical space forms and finitely many copies of $S^2 \times S^1$. Since this is the case if the fundamental group is trivial, a proof of the Poincaré Conjecture is obtained as an application. There is a different argument for the finite extinction time result that uses minimal surfaces, due to T. Colding and B. Minicozzi (see [18]).

Another application is the topological classification by Perelman of the orientable compact 3-manifolds which admit metrics of positive scalar curvature (see [58] and [22] for earlier results with different methods). Since the surgeries only increase scalar curvature, the associated Ricci flows with surgery have to become extinct in finite time. Therefore the assumption of Theorem 2.2 is

equivalent to asking that M is diffeomorphic to a connected sum of spherical space forms and finitely many copies of $S^2 \times S^1$.

In order to explain the strategy to prove Theorem 2.2 let us introduce the concept of a canonical metric. Let h be the metric on the unit sphere S^3 induced by the standard inclusion $S^3 \subset \mathbb{R}^4$. A *canonical metric* is any metric obtained from the 3-sphere (S^3, h) by attaching to it finitely many constant curvature spherical quotients (through the Gromov-Lawson procedure), and adding to it finitely many handles (Gromov-Lawson connected sums of S^3 to itself). The resulting manifold M is diffeomorphic to

$$S^3\#(S^3/\Gamma_1)\#\ldots\#(S^3/\Gamma_k)\#(S^2\times S^1)\#\ldots\#(S^2\times S^1),$$

where $\Gamma_1, \ldots, \Gamma_k$ are finite subgroups of $SO(4)$ acting freely on S^3. The resulting metric $\hat{g}$ is locally conformally flat and has positive scalar curvature. Two canonical metrics on M are in the same path-connected component of the moduli space $\mathcal{R}_+(M)/\mathrm{Diff}(M)$.

Given a metric g_0 in $\mathcal{R}_+(M)$, the strategy is to use the Ricci flow with surgery $(M_i^3, g_i(t))_{t\in[t_i,t_{i+1})}$ with initial condition $g_0(0) = g_0$ to construct a continuous path in $\mathcal{R}_+(M)$ that starts at g_0 and ends at a canonical metric. As in the proof of the Poincaré Conjecture we use backwards induction on the set of singular times t_i. We need a combination of the heat flow technique (Hamilton's convergence result [23]) and the conformal method (as in Weyl's proof) to deform the entire components that are discarded along the flow, including those at the extinction time. These components have known topology: S^3, $\mathbb{R}P^3$, $\mathbb{R}P^3\#\mathbb{R}P^3$, or $S^2\times S^1$. We also use the connected sum construction of Gromov and Lawson to undo the surgeries, making sure the final deformation is continuous despite the fact that the Ricci flow with surgery is a discontinuous process in its nature. A key observation for the induction is that a Gromov-Lawson connected sum of finitely many canonical metrics is in the same path-connected component (in the space of positive scalar curvature metrics) of a single canonical component. This follows from the conformal method.

The work of Perelman on the description of singularities (the existence of canonical neighborhoods, for example) under Hamilton's Ricci flow is fundamental ([46], [47], and [48]). We refer the reader to [13], [30], and [41] for some detailed presentations of the arguments due to Perelman. See also [5], [6] and [42].

2.1. Some applications to General Relativity. It turns out that we can use the previous results to study the topology of certain spaces of metrics which are relevant in General Relativity. In this section the spaces are always endowed with the topology associated to some natural weighted Hölder norm $C^{k,\alpha}_\beta$ (see [40] for more details).

We say that (g, h) is an *asymptotically flat initial data set* on $\mathbb{R}^3$ if g is a Riemannian metric and h is a symmetric $(0, 2)$-tensor on $\mathbb{R}^3$ such that

$$|g_{ij} - \delta_{ij}|(x) + |x||\partial g_{ij}|(x) + |x|^2|\partial^2 g_{ij}|(x) = O(1/|x|),$$
$$|h_{ij}|(x) = O(1/|x|^2),$$

as $x \to \infty$.

The full set of solutions to the *vacuum Einstein constraint equations* is the set $\mathcal{M}$ of all asymptotically flat initial data sets (g, h) defined on $\mathbb{R}^3$ such that

a) $R_g + (tr_g\, h)^2 - |h|^2 = 0$,

b) $\nabla_i h^i{}_j - \nabla_j (tr_g\, h) = 0$.

It goes back to the work of Choquet-Bruhat that the equations above are the precise conditions one needs in order to solve the Cauchy problem for Einstein equations: find a spacetime V (4-dimensional Lorentzian manifold) satisfying the vacuum Einstein equations $Ric_V = 0$ (zero Ricci curvature) and an embedded hypersurface $M^3 \subset V$ such that the induced metric on M is g and the second fundamental form of M is h. We refer the reader to [3] for a nice survey on the constraint equations.

Question: *Is the space $\mathcal{M}$ path-connected?*

These metrics no longer have nonnegative scalar curvature, so it would be interesting to find methods to study their deformations.

There is an important special case in which $R_g \geq 0$. Let $\mathcal{M}'$ be the set of all asymptotically flat initial data sets (g, h) on $\mathbb{R}^3$ such that

a) $tr_g\, h = 0$,

b) $R_g = |h|^2$,

c) $(div_g\, h)_j := \nabla_i h^i{}_j = 0$.

In [40] we prove

Theorem 2.3. *The set $\mathcal{M}'$ is path-connected.*

The idea is to first connect an initial data $(g_0, h_0) \in \mathcal{M}'$ into data of the form $(\hat{g}, 0)$ with $\hat{g}$ scalar-flat, through the conformal method (Lichnerowicz equation). We can then assume, by a perturbation argument, that $\hat{g}$ can be conformally compactified, i.e., $\hat{g}$ is a blow-up $G_x^4\, g$ of a positive scalar curvature metric g on S^3. Here G_x denotes the Green's function associated to the conformal Laplacian $L_g = \Delta_g - \frac{1}{8} R_g$ of g, with pole at $x \in S^3$. By deforming g, the connectedness of $\mathcal{R}_+(S^3)$ can be used to construct a continuous path of asymptotically flat and scalar-flat metrics on $\mathbb{R}^3$ connecting $G_x^4\, g$ to the flat metric. Along the way we also prove that the space of asymptotically flat metrics on $\mathbb{R}^3$ of nonnegative scalar curvature is path-connected. This space was studied previously by B. Smith and G. Weinstein in [62], where they established connectedness of the subspace of metrics that admit a quasi-convex global foliation.

References

[1] A. Ambrosetti and A. Malchiodi, *A multiplicity result for the Yamabe problem on S^n*, J. Funct. Anal., 168:529–561, 1999.

[2] T. Aubin, *Équations différentielles non linéaires et probléme de Yamabe concernant la courbure scalaire*, J. Math. Pures Appl., 55:269–296, 1976.

[3] R. Bartnik and J. Isenberg, *The constraint equations*, The Einstein equations and the large scale behavior of gravitational fields, 1–38, Birkhauser, Basel, 2004

[4] M. Berti and A. Malchiodi, *Non-compactness and multiplicity results for the Yamabe problem on S^n*, J. Funct. Anal., 180:210–241, 2001.

[5] L. Bessières, G. Besson, M. Boileau, S. Maillot and J. Porti, *Weak collapsing and geometrisation of aspherical 3-manifolds*, arXiv:0706.2065v2 [math.GT], 2007

[6] L. Bessières, G. Besson, M. Boileau, S. Maillot and J. Porti, *Geometrisation of 3-manifolds*, http://www-fourier.ujf-grenoble.fr/ besson/english_principal.pdf, 2009

[7] B. Botvinnik and P. Gilkey, *Metrics of positive scalar curvature on spherical space forms*, Canad. J. Math., 48(1):64–80, 1996.

[8] S. Brendle, *Elliptic and parabolic problems in conformal geometry*, International Congress of Mathematicians. Vol. II, 691–704, Eur. Math. Soc., Zrich, 2006.

[9] S. Brendle, *On the conformal scalar curvature equation and related problems*, Surveys in Differential Geometry XII, 2008.

[10] S. Brendle, *Convergence of the Yamabe flow in dimension 6 and higher*, Invent. Math., 170:541–576, 2007.

[11] S. Brendle, *Blow-up phenomena for the Yamabe equation*, J. Amer. Math. Soc., 21:951–979, 2008.

[12] S. Brendle and F. C. Marques, *Blow-up phenomena for the Yamabe equation II*, J. Differential Geom., 81:225–250, 2009.

[13] H-D. Cao and X-P. Zhu, *A complete proof of Poincaré and Geometrization conjectures - Application of the Hamilton-Perelman theory of the Ricci flow*, Asian J. Math., 10(2):169–492, 2006.

[14] R. Carr, *Construction of manifolds of positive scalar curvature*, Trans. Amer. Math. Soc., 307:63–74, 1988

[15] J. Cerf, *Sur les diffeomorphismes de la sphere de dimension trois* ($\Gamma_4 = 0$), Lecture Notes of Math., 53, Springer-Verlag, Berlin-New York, 1968

[16] X. Chen, P. Lu and G. Tian, *A note on uniformization of Riemann surfaces by Ricci flow*, Proc. Amer. Math. Soc., 134:3391–3393, 2006.

[17] B. Chow, *The Ricci flow on the 2-sphere*, J. Differential Geom., 33:325–334, 1991.

[18] T. Colding and B. Minicozzi, *Estimates for the extinction time for the Ricci flow on certain 3-manifolds and a question of Perelman*, J. Amer. Math. Soc., 18(3):561–569, 2005

[19] O. Druet, *Compactness for Yamabe metrics in low dimensions*, Int. Math. Res. Not., 23:1143–1191, 2004.

[20] O. Druet and E. Hebey, *Blow-up examples for second order elliptic PDEs of critical Sobolev growth*, Trans. Amer. Math. Soc., 357:1915–1929, 2004.

[21] M. Gromov and H. Blaine Lawson,Jr, *The classification of simply connected manifolds of positive scalar curvature*, Ann. of Math., 111:423–434, 1980

[22] M. Gromov and H. Blaine Lawson,Jr, *Positive scalar curvature and the Dirac operator on complete Riemannian manifolds*, Inst. Hautes tudes Sci. Publ. Math., 58:83–196, 1983

[23] R. Hamilton, *Three-manifolds with positive Ricci curvature*, J. Differential Geom., 17(2):255–306, 1982

[24] R. Hamilton, *The Ricci flow on surfaces*, In: Mathematics and General Relativity (Santa Cruz, CA, 1986), Contemp. Math., 71:237–262, Amer. Math. Soc., 1988

[25] R. Hamilton, *Four-manifolds with positive isotropic curvature*, Comm. Anal. Geom., 5(1):1–92, 1997

[26] E. Hebey and M. Vaugon, *Le problème de Yamabe équivariant*, Bull. Sci. Math., 117:241–286, 1993.

[27] N. Hitchin, *Harmonic spinors*, Adv. Math., 14:1–55, 1974

[28] M.A. Khuri, F.C. Marques, and R.M. Schoen, *A Compactness Theorem for the Yamabe Problem*, J. Differential Geom., 81:143-196, 2009.

[29] M.A. Khuri, F.C. Marques, and R.M. Schoen, *Details of Calculations from the Appendix*, http://math.stanford.edu/∼schoen/yamabe-paper/, 2007.

[30] B. Kleiner and J. Lott, *Notes on Perelman's papers*, Geom. Topol., 12:2587–2855, 2008

[31] M. Kreck and S. Stolz, *Nonconnected moduli spaces of positive sectional curvature metrics*, J. Amer. Math. Soc., 6(4):825–850, 1993

[32] J. Lee and T. Parker, *The Yamabe problem* Bull. Amer. Math. Soc., 17:37–91, 1987.

[33] Y. Li and L. Zhang, *Compactness of solutions to the Yamabe problem II*, Calc. Var. and PDEs, 25:185–237, 2005.

[34] Y. Li and L. Zhang, *Compactness of solutions to the Yamabe problem III*, J. Funct. Anal., 245(2):438–474, 2006.

[35] Y. Li and M. Zhu, *Yamabe type equations on three dimensional Riemannian manifolds*, Communications in Contemporary Math., 1:1–50, 1999.

[36] J. Lohkamp, *The higher dimensional positive mass theorem I*, Preprint, 2006.

[37] F.C. Marques, *A priori estimates for the Yamabe problem in the non-locally conformally flat case*, J. Differential Geom., 71:315–346, 2005.

[38] F.C. Marques, *Recent developments on the Yamabe problem*, Mat. Contemp., 35:115–130, 2008.

[39] F.C. Marques, *Blow-up examples for the Yamabe problem*, Calc. Var. and PDEs, 36 (3):377–397, 2009.

[40] F.C. Marques, *Deforming three-manifolds with positive scalar curvature*, arXiv:0907.2444v1 [math.DG], 2009.

[41] J. Morgan and G. Tian, *Ricci flow and the Poincaré conjecture*, Clay Mathematics Monographs, American Mathematical Society, 2007.

[42] J. Morgan and G. Tian, *Completion of the proof of the geometrization conjecture*, arXiv:0809.4040v1 [math.DG], 2008.

[43] L. Nirenberg, *Topics in Nonlinear Functional Analysis*, Courant Institute publication, 1973–74.

[44] M. Obata, *The conjectures on conformal transformations of Riemannian manifolds*, J. Diff. Geom., 6:247–258, 1972.

[45] R. Palais, *Critical point theory and the minimax principle*, Proc. Sympos. Pure Math., Amer. Math. Soc., Providence, 15:185–212, 1970.

[46] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications*, arXiv:math/0211159v1 [math.DG], 2002.

[47] G. Perelman, Ricci flow with surgery on three-manifolds, arXiv:math/0303109v1 [math.DG], 2003.

[48] G. Perelman, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*, arXiv:math/0307245v1 [math.DG], 2003.

[49] D. Pollack, *Nonuniqueness and high energy solutions for a conformally invariant scalar curvature equation*, Comm. Anal. and Geom., 1:347–414, 1993.

[50] J. Rosenberg, *Manifolds of positive scalar curvature: a progress report*, Surv. Differ. Geom., Int. Press, Somerville, MA, 11:259–294, 2007

[51] J. Rosenberg and S. Stolz, *Metrics of positive scalar curvature and connections with surgery. Surveys on surgery theory*, Ann. of Math. Stud., 149:353–386, Princeton Univ. Press, Princeton, NJ, 2001

[52] R. Schoen, *Conformal deformation of a Riemannian metric to constant scalar curvature*, J. Differential Geometry, 20:479–495, 1984.

[53] R. Schoen, *Courses at Stanford University*, 1989.

[54] R. Schoen, *Variational theory for the total scalar curvature functional for Riemannian metrics and related topics*, "Topics in Calculus of Variations", Lecture Notes in Mathematics, Springer-Verlag, New York, v. 1365, 1989.

[55] R. Schoen, *On the number of constant scalar curvature metrics in a conformal class*, Differential Geometry: A symposium in honor of Manfredo Do Carmo (H. B. Lawson and K. Tenenblat, eds.), Wiley, 311–320, 1991.

[56] R. Schoen, *A report on some recent progress on nonlinear problems in geometry*, Surveys in Differential Geometry, 1:201–241, 1991.

[57] R. Schoen and S.-T. Yau, *On the proof of the positive mass conjecture in General Relativity*, Comm. Math. Phys., 65:45–76, 1979.

[58] R. Schoen and S.T. Yau, *Existence of incompressible minimal surfaces and the topology of three-dimensional manifolds with nonnegative scalar curvature*, Ann. of Math., 110(1):127–142, 1979.

[59] R. Schoen and S.-T. Yau, *Conformally flat manifolds, Kleinian groups, and scalar curvature*, Invent. Math., 92:47–71, 1988.

[60] R. Schoen and S.-T. Yau, *Lectures on Differential Geometry*, Conference Proceedings and Lecture Notes in Geometry and Topology, International Press Inc., 1994.

[61] R. Schoen and D. Zhang, *Prescribed scalar curvature on the n-sphere*, Calc. Var. and PDEs, 4:1–25, 1996.

[62] B. Smith and G. Weinstein, *Quasiconvex foliations and asymptotically flat metrics of non-negative scalar curvature*, Comm. Anal. Geom., 12(3):511–551, 2004.

[63] N. Trudinger, *Remarks concerning the conformal deformation of Riemannian structures on compact manifolds*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 22(3):165–274, 1968.

[64] H. Weyl, *Uber die Bestimmung einer geschlossenen konvexen Flache durch ihr Linienelement*, Vierteljahrsschr. Naturforsch. Gesellschaft, 61:40–72, 1916

[65] E. Witten, *A new proof of the positive energy theorem*, Comm. Math. Phys., 80:381–402, 1981.

[66] H. Yamabe, *On a deformation of Riemannian structures on compact manifolds*, Osaka Math. J., 12:21–37, 1960.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Constant Mean Curvature Surfaces in 3-dimensional Thurston Geometries

Isabel Fernández* and Pablo Mira†

Abstract

This is a survey on the global theory of constant mean curvature surfaces in Riemannian homogeneous 3-manifolds. These ambient 3-manifolds include the eight canonical Thurston 3-dimensional geometries, i.e. $\mathbb{R}^3$, $\mathbb{H}^3$, $\mathbb{S}^3$, $\mathbb{H}^2 \times \mathbb{R}$, $\mathbb{S}^2 \times \mathbb{R}$, the Heisenberg space Nil_3, the universal cover of $\mathrm{PSL}_2(\mathbb{R})$ and the Lie group Sol_3. We will focus on the problems of classifying compact CMC surfaces and entire CMC graphs in these spaces. A collection of important open problems of the theory is also presented.

Mathematics Subject Classification (2010). 53A10, 53C42

Keywords. Constant mean curvature surfaces, homogeneous spaces, Thurston geometries, harmonic maps, minimal surfaces, entire graphs.

1. Introduction

Constant mean curvature (CMC) surfaces appear as critical points of a natural geometric variational problem: to minimize surface area with or without a volume constraint (the unconstrained case corresponds to zero mean curvature, i.e. to minimal surfaces). A fundamental problem of this discipline is the geometric study and classification of CMC surfaces under global hypotheses like compactness, completeness, properness or embeddedness. The study of this problem for CMC surfaces in the model spaces $\mathbb{R}^3$, $\mathbb{S}^3$ and $\mathbb{H}^3$ has produced a very rich theory, in which geometric arguments interact with complex analysis, harmonic maps, integrable systems, maximum principles, elliptic PDEs, geometric measure theory and so on.

*Isabel Fernández, Universidad de Sevilla (Spain). E-mail: isafer@us.es.
†Pablo Mira, Universidad Politécnica de Cartagena (Spain). E-mail: pablo.mira@upct.es.
The authors were partially supported by MEC-FEDER, Grant No. MTM2007-65249, Junta de Andalucía Grant No. FQM325 and the Programme in Support of Excellence Groups of Murcia, by Fundación Séneca, R.A.S.T 2007-2010, reference 04540/GERM/06 and Junta de Andalucía, reference P06-FQM-01642."

One of the most remarkable achievements of this field in the last decade has been the extension of this classical theory to the case of CMC surfaces in simply connected homogeneous 3-dimensional ambient spaces. Apart from $\mathbb{R}^3$, $\mathbb{S}^3$ and $\mathbb{H}^3$, these spaces are the remaining five Thurston 3-dimensional geometries (i.e. $\mathbb{H}^2 \times \mathbb{R}$, $\mathbb{S}^2 \times \mathbb{R}$, the Heisenberg group Nil_3, the universal covering of $\mathrm{PSL}_2(\mathbb{R})$ and the Lie group Sol_3), together with 3-dimensional Berger spheres and some other Lie groups with left-invariant metrics (see Section 2).

It must be said here that there is an important number of contributions regarding CMC surfaces in general Riemannian 3-manifolds (not even homogeneous), many of which deal for instance with isoperimetric questions or with geometric consequences derived from the *stability operator* associated to the second variation of the surface. The achievement in the case of homogeneous ambient 3-spaces has been the construction of a very rich global theory of CMC surfaces, analogous to the case of $\mathbb{R}^3$, $\mathbb{S}^3$ and $\mathbb{H}^3$, with an emphasis on the geometric classification (up to ambient isometries) of properly immersed or properly embedded CMC surfaces. The fact that the ambient space is homogeneous, i.e. it has the same local geometry at all points, makes this problem extremely natural.

Our aim here is to present a survey on some fundamental aspects of the global theory of CMC surfaces in homogeneous 3-manifolds. We do not plan, however, to give a systematic account of all important results of this already broad theory, but to discuss some specific problems at the core of it. Hence, there will be many important results omitted, and we apologize in advance for that.

In order to explain the problems we shall be dealing with, let us distinguish between compact and non-compact CMC surfaces in these spaces.

In the case of compact CMC surfaces, three fundamental problems are the Alexandrov problem (i.e. to classify compact embedded CMC surfaces), the Hopf problem (i.e. to classify CMC spheres), and the isoperimetric problem (recall that isoperimetric regions on a Riemannian 3-manifold are bounded by compact embedded CMC surfaces, but the converse is not always true). By classical results, round spheres constitute the solution to each of these three problems in the case of CMC surfaces in $\mathbb{R}^3$. One of our main objectives will be to explain what is known (and what is not known) for these problems in the broader context of CMC surfaces in homogeneous 3-manifolds.

In the case of non-compact CMC surfaces, one of the basic problems is to study the properly embedded CMC surfaces of finite topology. A classical result in that direction is given by *Bernstein's theorem:* planes are the only entire minimal graphs in $\mathbb{R}^3$. As in all Thurston 3-dimensional geometries there is a natural notion of entire graph, it is an important problem of the discipline to solve the *Bernstein problem* for CMC graphs, i.e. to classify all entire CMC graphs in these 3-dimensional ambient spaces. This will be our other main objective.

The theory of CMC surfaces in Thurston 3-dimensional geometries started to develop as a consistent unified theory after some pioneer works by Harold

Rosenberg, jointly with William H. Meeks [MeRo1, MeRo2, Ros] for the case of minimal surfaces in product spaces, and jointly with Uwe Abresch [AbRo1, AbRo2] for the case of CMC surfaces in homogeneous spaces with a 4-dimensional isometry group.

On one hand, Meeks and Rosenberg established many results on complete minimal surfaces in $M^2 \times \mathbb{R}$, what has guided a large number of subsequent works in the field. A recent major contribution in this sense is the Collin-Rosenberg theorem [CoRo] on the existence of harmonic diffeomorphims from $\mathbb{C}$ onto the hyperbolic plane $\mathbb{H}^2$, obtained by constructing an entire minimal graph of parabolic conformal type in $\mathbb{H}^2 \times \mathbb{R}$.

On the other hand, Abresch and Rosenberg discovered a holomorphic quadratic differential for CMC surfaces in these homogeneous spaces with 4-dimensional isometry group (the $\mathbb{E}^3(\kappa, \tau)$ spaces), and solved the Hopf problem for them. The general integrability theory of CMC surfaces in the homogeneous $\mathbb{E}^3(\kappa, \tau)$ spaces was then established by B. Daniel [Dan1]. The discovery by the authors of a harmonic Gauss map into $\mathbb{H}^2$ for $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$ turned into a series of papers by Daniel, Fernández, Hauswirth, Mira, Rosenberg, Spruck [FeMi1, Dan2, FeMi2, HRS, DaHa] in which the Bernstein problem for CMC graphs of *critical mean curvature* (including minimal graphs in Heisenberg space Nil_3, see Section 6) was solved. Very recently, the Hopf and Alexandrov problems for CMC surfaces have been solved by Daniel-Mira and Meeks [DaMi, Mee] in the remaining Thurston 3-dimensional geometry: the Lie group Sol_3, whose isometry group is only 3-dimensional.

We have organized this exposition as follows. In Section 2 we will introduce the 3-dimensional homogeneous ambient spaces. In Section 3 we will present the basic integrability equations by Daniel for CMC surfaces in the homogeneous spaces $\mathbb{E}^3(\kappa, \tau)$, together with the holomorphic Abresch-Rosenberg differential, and with some basic definitions on stability of CMC surfaces. In Section 4 we will discuss the Hopf, Alexandrov and isoperimetric problems in the homogeneous spaces $\mathbb{E}^3(\kappa, \tau)$. Section 5 will be devoted to solving the Hopf and Alexandrov problems in the eighth Thurston geometry, i.e. the Lie group Sol_3. In Section 6 we will present the solution to the Bernstein problem for entire graphs of critical CMC in the homogeneous $\mathbb{E}^3(\kappa, \tau)$ spaces. Finally, in Section 7 we shall expose the Collin-Rosenberg theorem on parabolic entire minimal graphs in $\mathbb{H}^2 \times \mathbb{R}$, together with some developments on the theory of complete minimal surfaces of finite total curvature in $\mathbb{H}^2 \times \mathbb{R}$. Most sections finish with a selection of important open problems. See [Mee, DHM] for more open problems in the theory.

A more detailed introduction to the global theory of CMC surfaces in homogeneous 3-spaces can be found in the Lecture Notes by Daniel, Hauswirth and Mira [DHM].

The authors are grateful to H. Rosenberg, B. Daniel and J.A. Gálvez for useful observations about this manuscript.

2. Homogeneous 3-spaces and Thurston Geometries

Homogeneous spaces are the natural generalization of space forms. By definition, a manifold is said to be homogeneous if the isometry group acts transitively on the manifold. Roughly speaking, the manifold looks the same at all the points, even though, standing at one point, the manifold can look different in different directions. In the simply connected case, the classification of the 3-dimensional homogeneous spaces is well-known. It turns out that any simply connected homogeneous 3-space must have isometry group of dimension 6, 4 or 3. The complete list of these spaces is the following (see subsections below for more details):

- The spaces with 6-dimensional isometry group are the space forms: the Euclidean space $\mathbb{R}^3$, the hyperbolic space $\mathbb{H}^3(\kappa)$, and the standard sphere $\mathbb{S}^3(\kappa)$. For simplicity we will assume that $\kappa = \pm 1$ and write $\mathbb{H}^3 = \mathbb{H}^3(-1)$ and $\mathbb{S}^3 = \mathbb{S}^3(1)$.
- The spaces with 4-dimensional isometry group are fibrations over the 2-dimensional space forms. They are the product spaces $\mathbb{H}^2 \times \mathbb{R}$ and $\mathbb{S}^2 \times \mathbb{R}$, the Berger spheres, the Heisenberg space Nil_3 and the universal covering of the Lie group $\mathrm{PSL}(2, \mathbb{R})$.
- The spaces with 3-dimensional isometry group are a certain class of Lie groups; among them we specially quote the space Sol_3.

These spaces are closely related with Thurston's Geometrization Conjecture. This recently proved conjecture states that any compact orientable 3-manifold can be cut by disjoint embedded 2-spheres or tori into pieces, each one of them, after gluing 2-balls or solid tori along its boundary components, admits a geometric structure. A 3-manifold without boundary is said to admit a geometric structure if it can be endowed with a complete locally homogeneous metric. In this case, by considering its universal covering we obtain a complete simply-connected locally homogeneous space and hence, by a result of Singer, homogeneous. Thus, a 3-manifold admitting a geometric structure can be realized as the quotient of a homogeneous simply connected 3-space under the action of a subgroup of a Lie group acting transitively by isometries. The list of the maximal geometric structures that give compact quotients consists of eight of the previously described spaces: the three space forms, the two product spaces, Nil_3, the universal covering of $\mathrm{PSL}(2, \mathbb{R})$ and Sol_3 (Berger spheres must be excluded from this list because they are not maximal, their isometry group are contained in the one of the standard sphere $\mathbb{S}^3$). We refer to [Sco, Bon] for more details.

2.1. Homogeneous spaces with 4-dimensional isometry group. Denote by $\mathcal{M}^2(\kappa)$ the 2-dimensional space form of constant curvature

κ (for example, $\mathcal{M}^2(\kappa) = \mathbb{R}^2, \mathbb{H}^2, \mathbb{S}^2$ for $\kappa = 0, -1, 1$ respectively). As commented above, any simply connected homogeneous 3-space with 4-dimensional isometry group admits a fibration over $\mathcal{M}^2(\kappa)$, for some $\kappa \in \mathbb{R}$. Moreover, these spaces can be parameterized in terms of the base curvature κ and the bundle curvature τ, that satisfy $\kappa - 4\tau^2 \neq 0$. We will use the notation $\mathbb{E}^3(\kappa, \tau)$ for these homogeneous spaces.

1. When $\tau = 0$, we have the product spaces $\mathcal{M}^2(\kappa) \times \mathbb{R}$, i.e. up to scaling, the spaces $\mathbb{S}^2 \times \mathbb{R}$ when $\kappa > 0$, and $\mathbb{H}^2 \times \mathbb{R}$ when $\kappa < 0$.

2. When $\tau \neq 0$ and $\kappa > 0$, the corresponding spaces are the Berger spheres, a family of 2-parameter (1-parameter after a homothetical change of coordinates) metrics on the sphere, obtained by deforming the standard metric in such a way that the Hopf fibration is still a Riemannian fibration. They can also be seen as the Lie group SU(2) endowed with a 1-parameter family of left-invariant metrics.

3. When $\tau \neq 0$ and $\kappa = 0$, $\mathbb{E}^3(\kappa, \tau)$ is the Heisenberg group Nil_3, the nilpotent Lie group
$$\left\{ \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \; ; \; a, b, c \in \mathbb{R} \right\},$$
endowed with a 1-parameter family of left-invariant metrics, all of them isometrically equivalent after a homothetical change of coordinates.

4. When $\tau \neq 0$ and $\kappa < 0$, we obtain the universal covering of the Lie group $\mathrm{PSL}(2, \mathbb{R})$, endowed with a 2-parameter (again 1-parameter after homotheties) family of left-invariant metrics.

There exists a common setting for all these spaces. Indeed, label $\mathbb{D}(\rho) = \{(x_1, x_2) \in \mathbb{R}^2 \, ; \, x_1^2 + x_2^2 < \rho^2\}$. Then, if $\kappa = 0$ (resp. $\kappa < 0$), the space $\mathbb{E}^3(\kappa, \tau)$ can be viewed as $\mathbb{R}^3$ (resp. $\mathbb{D}\left(2/\sqrt{-\kappa}\right) \times \mathbb{R}$) endowed with the metric

$$ds^2 = \lambda^2(dx_1^2 + dx_2^2) + \big(\tau\lambda(x_2 dx_1 - x_1 dx_2) + dx_3\big)^2, \qquad \lambda = \frac{1}{1 + \frac{\kappa}{4}(x_1^2 + x_2^2)}. \quad (1)$$

Also, for $\kappa > 0$, $(\mathbb{R}^3, ds^2)$ corresponds to the universal cover of $\mathbb{E}^3(\kappa, \tau)$ minus one fiber. In all cases, up to a homothetical change of coordinates we can suppose without loss of generality that $\kappa - 4\tau^2 = \pm 1$.

The corresponding Riemannian fibration $\pi : \mathbb{E}^3(\kappa, \tau) \to \mathcal{M}^2(\kappa)$ is given here by the projection on the first two coordinates. The unitary vector field

$$\xi = \frac{\partial}{\partial x_3}$$

is a Killing field tangent to the fibers of π, and will be referred to as the *vertical field* of the space $\mathbb{E}^3(\kappa, \tau)$. It satisfies the equation

$$\widehat{\nabla}_X \xi = \tau X \times \xi$$

for all vector fields X in $\mathbb{E}^3(\kappa,\tau)$. Here $\widehat{\nabla}$ is the Levi-Civita connection, $\times$ the cross product and τ the bundle curvature (this is basically the definition of τ).

A remarkable difference between the spaces $\mathbb{E}^3(\kappa,\tau)$ is that their isometry group has four connected components in the case $\tau = 0$, and only two when $\tau \neq 0$. This follows from the fact that any isometry in the product spaces can either preserve or reverse the orientation of the base and the fibers independently, while in the case $\tau \neq 0$ it can only either preserve or reverse both orientations. In particular, reflections only exist in product spaces.

Also, when $\tau \neq 0$ the spaces $\mathbb{E}^3(\kappa,\tau)$ are Lie groups, and if we set $\sigma := \frac{\kappa}{2\tau}$, an orthonormal frame of left-invariant vector fields (called the *canonical frame*) is given by

$$E_1 = \lambda^{-1}\left(\cos(\sigma x_3)\frac{\partial}{\partial x_1} + \sin(\sigma x_3)\frac{\partial}{\partial x_2}\right) + \tau(x_1\sin(\sigma x_3) - x_2\cos(\sigma x_3))\frac{\partial}{\partial x_3},$$

$$E_2 = \lambda^{-1}\left(-\sin(\sigma x_3)\frac{\partial}{\partial x_1} + \cos(\sigma x_3)\frac{\partial}{\partial x_2}\right) + \tau(x_1\cos(\sigma x_3) + x_2\sin(\sigma x_3))\frac{\partial}{\partial x_3},$$

$$E_3 = \xi = \frac{\partial}{\partial x_3}.$$

2.2. Homogeneous spaces with 3-dimensional isometry group. Of all homogeneous spaces with 3-dimensional isometry group, Sol_3 is specially important, since it is the only Thurston geometry among them. We will now describe some aspects of this space.

A useful representation of Sol_3 is the space $\mathbb{R}^3$ with the metric

$$ds^2 = e^{2x_3}dx_1^2 + e^{-2x_3}dx_2^2 + dx_3^2,$$

that is left-invariant for the structure of Lie group given by

$$(x_1,x_2,x_3)\cdot(y_1,y_2,y_3) = (x_1 + e^{-x_3}y_1, x_2 + e^{x_3}y_2, x_3 + y_3).$$

The following vector fields form an orthonormal left-invariant frame

$$E_1 = e^{-x_3}\frac{\partial}{\partial x_1}, \quad E_2 = e^{x_3}\frac{\partial}{\partial x_2}, \quad E_3 = \frac{\partial}{\partial x_3}.$$

The isometries in Sol_3 are generated by the three 1-parameter groups of translations

$$(x_1,x_2,x_3) \mapsto (x_1 + c, x_2, x_3), \qquad (x_1,x_2,x_3) \mapsto (x_1, x_2 + c, x_3),$$
$$(x_1,x_2,x_3) \mapsto (e^{-c}x_1, e^{c}x_2, x_3 + c),$$

and by the orientation reversing isometries fixing the origin

$$(x_1,x_2,x_3) \mapsto (-x_1, x_2, x_3), \qquad (x_1,x_2,x_3) \mapsto (x_2, -x_1, -x_3).$$

A remarkable fact is the existence of two canonical foliations, namely

$$\mathcal{F}_1 = \{x_1 = \text{constant}\}, \qquad \mathcal{F}_2 = \{x_2 = \text{constant}\},$$

whose leaves are totally geodesic surfaces isometric to the hyperbolic plane $\mathbb{H}^2$. Reflections across any of these leaves are orientation reversing isometries of Sol_3.

3. CMC Surfaces: Basic Equations

In this section we present three important tools for our study. One is the set of integrability equations of CMC surfaces in $\mathbb{E}^3(\kappa,\tau)$ by Daniel [Dan1]. Another one the *Abresch-Rosenberg differential*, a holomorphic quadratic differential geometrically defined on any CMC surface in $\mathbb{E}^3(\kappa,\tau)$. The third one is a local isometric correspondence for CMC surfaces in $\mathbb{E}^3(\kappa,\tau)$ via which one can pass from one homogeneous space into another when studying CMC surfaces [Dan1]. Some notions about the stability operator of CMC surfaces are also given.

3.1. Integrability equations in $\mathbb{E}^3(\kappa,\tau)$. It is well known that the Gauss-Codazzi equations are the integrability conditions of surface theory in $\mathbb{R}^3$, $\mathbb{S}^3$ and $\mathbb{H}^3$. In other homogeneous spaces, the situation is more complicated.

Let $\psi : \Sigma \to \mathbb{E}^3(\kappa,\tau)$ be an isometric immersion with unit normal map η, and consider on Σ the conformal structure given by its induced metric via ψ. Associated to a conformal parameter $z = s+it$ on Σ, we will consider the usual operators $\partial_z = (\partial_s - i\partial_t)/2$ and $\partial_{\bar z} = (\partial_s + i\partial_t)/2$. Also denote by ξ the vertical Killing field of $\mathbb{E}^3(\kappa,\tau)$.

Definition 3.1. *We call the* fundamental data *of ψ the 5-tuple $(\lambda|dz|^2, u, H, p\,dz^2, A\,dz)$ where H is the mean curvature and*

$$\lambda = 2\langle \psi_z, \psi_{\bar z}\rangle, \qquad u = \langle N, \xi\rangle, \qquad p = -\langle \psi_z, N_z\rangle, \qquad A = \langle \xi, \psi_z\rangle.$$

The function u is commonly called the *angle function* of the surface.

Once here, a set of necessary and sufficient conditions for the integrability of CMC surfaces in $\mathbb{E}^3(\kappa,\tau)$ can be written in terms of these fundamental data. This is a result by B. Daniel [Dan1], although the formulation that we expose here (i.e. in terms of a conformal parameter on the surface) comes from [FeMi2].

Theorem 3.2 ([Dan1, FeMi2]). *The fundamental data of an immersed surface $\psi : \Sigma \to \mathbb{E}^3(\kappa,\tau)$ satisfy the following integrability conditions:*

$$\left\{\begin{array}{llcl} (\mathbf{C.1}) & p_{\bar z} & = & \dfrac{\lambda}{2}(H_z + uA(\kappa - 4\tau^2)). \\ (\mathbf{C.2}) & A_{\bar z} & = & \dfrac{u\lambda}{2}(H + i\tau). \\ (\mathbf{C.3}) & u_z & = & -(H - i\tau)A - \dfrac{2p}{\lambda}\bar A. \\ (\mathbf{C.4}) & \dfrac{4|A|^2}{\lambda} & = & 1 - u^2. \end{array}\right. \tag{2}$$

Conversely, if Σ is simply connected, these equations are also sufficient for the existence of a surface $\psi : \Sigma \to \mathbb{E}^3(\kappa,\tau)$ with fundamental data $(\lambda|dz|^2, u, H, p\,dz^2, A\,dz)$. This surface is unique up to ambient isometries preserving the orientations of base and fiber of $\mathbb{E}^3(\kappa,\tau)$.

We see then that, in the spaces $\mathbb{E}^3(\kappa,\tau)$, more equations apart from the Gauss-Codazzi ones are needed, due to the loss of symmetries. As a matter of fact, **(C.1)** is the Codazzi equation, while the Gauss equation does not appear (it is deduced from the rest). These new equations evidence the special character of the vertical direction in the $\mathbb{E}^3(\kappa,\tau)$ spaces.

Definition 3.3. *The* Abresch-Rosenberg *differential of the immersion is defined as the quadratic differential on Σ given by*

$$Qdz^2 = \big(2(H+i\tau)p - (\kappa - 4\tau^2)A^2\big)\,dz^2.$$

It is then easy to see by means of (**C.2**) that the Codazzi equation **(C.1)** can be rephrased in terms of Q as

$$Q_{\bar z} = \lambda H_z + (\kappa - 4\tau^2)\frac{H_{\bar z}A^2}{(H+i\tau)^2}. \tag{3}$$

Consequently, one has the following theorem, which generalized the classical fact that the Hopf differential is holomorphic for CMC surfaces in $\mathbb{R}^3$, $\mathbb{S}^3$ and $\mathbb{H}^3$.

Theorem 3.4 ([AbRo1, AbRo2]). *Qdz^2 is a holomorphic quadratic differential on any CMC surface in $\mathbb{E}^3(\kappa,\tau)$.*

This is a crucial result of the theory, since it allows the use of holomorphic functions in the geometric classification of CMC surfaces in $\mathbb{E}^3(\kappa,\tau)$ (see Section 4 and Section 6, for instance).

An important tool in the description of CMC surfaces in $\mathbb{R}^3$, $\mathbb{S}^3$ and $\mathbb{H}^3$ is the classical Lawson correspondence. It establishes an isometric one-to-one local correspondence between CMC surfaces in different space forms that allows to pass, for instance, from minimal surfaces in $\mathbb{R}^3$ to $H=1$ surfaces in $\mathbb{H}^3$.

The Lawson correspondence was generalized by B. Daniel to the context of homogeneous spaces. Indeed, Daniel discovered in [Dan1] an isometric local correspondence for CMC surfaces in all the homogeneous spaces $\mathbb{E}^3(\kappa,\tau)$, which can be described as follows in terms of the *fundamental data* defined above.

Theorem 3.5 (Sister correspondence, [Dan1]). *Let $(\lambda|dz|^2, u, H_1, p_1\,dz^2, A_1\,dz)$ be the fundamental data of a simply connected H_1-CMC surface in $\mathbb{E}(\kappa_1,\tau_1)$, and consider $\kappa_2,\tau_2,H_2\in\mathbb{R}$ so that*

$$\kappa_2 - 4\tau_2^2 = \kappa_1 - 4\tau_1^2, \qquad H_2^2+\tau_2^2 = H_1^2+\tau_1^2.$$

Then if we set $\theta \in \mathbb{R}$ given by $H_2 - i\tau_2 = e^{i\theta}(H_1 - i\tau_1)$, the fundamental data given by

$$(\lambda|dz|^2, u, H_2, p_2\, dz^2 = e^{-i\theta} p_1\, dz^2, A_2\, dz = e^{-i\theta} A_1\, dz) \tag{4}$$

give rise to a (simply connected) H_2-CMC surface in $\mathbb{E}^3(\kappa_2, \tau_2)$, which is locally isometric to the original one.

Two surfaces related by the above correspondence are called *sister surfaces* with phase θ. In particular, the corresponding Abresch-Rosenberg differentials of sister surfaces are related by $Q_2 = e^{-2i\theta} Q_1$. As special cases of this correspondence we obtain the associate family of minimal surfaces in $\mathcal{M}^2(\kappa) \times \mathbb{R}$, and a correspondence between minimal surfaces in Nil_3 and CMC $\frac{1}{2}$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$. Generically, and up to ambient isometries and dilations, the family of sister surfaces for a given choice of (H, κ, τ) is a continuous 1-parameter family.

There is a natural notion of *graph* in these spaces. Since $\mathbb{E}^3(\kappa, \tau)$ has a canonical fibration over $\mathcal{M}^2(\kappa)$ (see Section 2), we will say that an immersed surface Σ in $\mathbb{E}^3(\kappa, \tau)$ is a (local) graph if the projection to the base is a (local) diffeomorphism. The CMC-equation for the graph of a function $u = u(x, y)$ is the PDE (see [Lee])

$$\frac{2H}{\delta^2} = \frac{\partial}{\partial x}\left(\frac{\alpha}{\omega}\right) + \frac{\partial}{\partial y}\left(\frac{\beta}{\omega}\right), \tag{5}$$

where

$$\delta = 1 + \frac{\kappa}{4}(x^2 + y^2), \quad \omega = \sqrt{1 + \delta^2(x^2 + y^2)},$$
$$\alpha = u_x + \tau\frac{y}{\delta}, \qquad \beta = u_y - \tau\frac{x}{\delta}.$$

For instance, a graph $u = u(x, y)$ in $\mathrm{Nil}_3 \equiv \mathbb{E}^3(0, \tau)$ is minimal if and only if it satisfies the elliptic PDE

$$(1 + \beta^2)u_{xx} - 2\alpha\beta\, u_{xy} + (1 + \alpha^2)u_{yy} = 0, \tag{6}$$

where $\alpha := u_x + y/2$ and $\beta := u_y - x/2$.

3.2. Stability and index of CMC surfaces. As it is well known, CMC surfaces in Riemannian 3-manifolds appear as the critical points for the area functional associated to variations of the surface with compact support and constant enclosed volume. Equivalently, an immersed surface S has constant mean curvature H if and only if it is a critical point for the functional Area $-$ $2H$ Vol. The second variation formula for this functional is given by

$$Q(f, f) = -\int_S f\mathcal{L}(f),$$

where $\mathcal{L}$ is the *Jacobi operator* (or *stability operator*) of the surface:

$$\mathcal{L} = \Delta + ||\mathcal{B}||^2 + \mathrm{Ric}(\eta).$$

Here Δ is the Laplacian for the induced metric on the surface, $\mathcal{B}$ is the second fundamental form, η is the unit normal vector field, and Ric is the Ricci curvature in the ambient manifold. As a particular case, the Jacobi operator for CMC surfaces in the spaces $\mathbb{E}^3(\kappa,\tau)$ can be rewritten (see [Dan1]) as

$$\mathcal{L} = \Delta - 2K + 4H^2 + 4\tau^2 + (\kappa - 4\tau^2)(1+u^2),$$

being K the Gaussian curvature of the surface and u the angle function (see Definition 3.1). A *Jacobi function* is a function f for which $\mathcal{L}(f) = 0$.

A CMC surface S is said to be *stable* (resp. *weakly stable*) if

$$Q(f,f) = -\int_S f\mathcal{L}(f) \geq 0$$

holds for any smooth function f on S with compact support (resp. with compact support and $\int_S f = 0$). For instance, CMC graphs in $\mathbb{E}^3(\kappa,\tau)$ are stable, and compact CMC surfaces bounding isoperimetric regions are weakly stable (but not necessarily stable, as round spheres in $\mathbb{R}^3$ show).

An important concept related to stability is the *index* of a CMC surface. The index of a compact CMC surface is defined as the number of negative eigenvalues of its Jacobi operator. Thus, stable CMC surfaces (in particular CMC graphs) have index zero. Round spheres in $\mathbb{R}^3$ have index one.

We refer to [MPR] for more details about stability of CMC surfaces.

4. Compact CMC Surfaces in $\mathbb{E}^3(\kappa,\tau)$

In this section we explain the most important results that are known regarding the existence and uniqueness of compact CMC surfaces in the homogeneous 3-spaces $\mathbb{E}^3(\kappa,\tau)$. The fundamental examples are the rotational CMC spheres, and we shall be interested in their uniqueness among compact embedded CMC surfaces, and among immersed CMC surfaces. These problems are called, respectively, the Alexandrov and Hopf problems.

4.1. Rotational compact CMC surfaces. Although round spheres in the model spaces $\mathbb{R}^3, \mathbb{S}^3, \mathbb{H}^3$ are CMC spheres, this does not hold for the rest of homogeneous spaces. However, in all the spaces $\mathbb{E}^3(\kappa,\tau)$ there exist rotations with respect to the vertical axis, and so there is a natural notion of *rotational surface*. It is hence natural to seek CMC spheres (and CMC tori) in $\mathbb{E}^3(\kappa,\tau)$ among the class of rotational surfaces. This can be done by ODE analysis, and the result of this can be summarized as follows:

Theorem 4.1. *(Structure of rotational CMC spheres in $\mathbb{E}^3(\kappa,\tau)$).*

1. *If $\kappa - 4\tau^2 > 0$, then for every $H \in \mathbb{R}$ there exists a unique rotational CMC H sphere (up to isometries) in $\mathbb{E}^3(\kappa,\tau)$. These spheres are embedded if*

$\tau = 0$, i.e. in $\mathbb{S}^2 \times \mathbb{R}$, and also for most Berger spheres. However, for some Berger spheres with small bundle curvature τ (with respect to a fixed κ) there is a certain region of variation of the parameters (H, τ) where the spheres are non-embedded. This region can be explicitly described, see [Tor].

2. *If $\kappa - 4\tau^2 < 0$, then*
   - *if $H^2 \leqslant -\frac{\kappa}{4}$, then there exists no rotational CMC H sphere in $\mathbb{E}^3(\kappa, \tau)$,*
   - *if $H^2 > -\frac{\kappa}{4}$, then there exists a unique rotational CMC H sphere (up to isometries) in $\mathbb{E}^3(\kappa, \tau)$. All these spheres are embedded.*

Let us remark that all these CMC spheres can be constructed explicitly. We shall call them *canonical rotational CMC spheres*. For example, the rotational CMC H spheres in $\mathbb{S}^2 \times \mathbb{R} \subset \mathbb{R}^4$ are given by the formula

$$\psi(u, v) = (-\cos k(u), \sin k(u) \cos v, \sin k(u) \sin v, h(u)),$$

where $-1 \leq u \leq 1$, $H \in \mathbb{R}$ and

$$k(u) := 2 \arctan\left(\frac{2H}{\sqrt{1-u^2}}\right), \qquad h(u) := \frac{4H}{\sqrt{4H^2+1}} \operatorname{arcsinh}\left(\frac{u}{\sqrt{1-u^2+4H^2}}\right).$$

Besides these rotational CMC spheres, there also exist rotational CMC tori in $\mathbb{E}^3(\kappa, \tau)$ when (and only when) $\kappa - 4\tau^2 > 0$ (excluding minimal surfaces in $\mathbb{S}^2 \times \mathbb{R}$). For $\mathbb{S}^2 \times \mathbb{R}$, they are all embedded (see Pedrosa [Ped]). For Berger spheres the situation is explained by Torralbo and Urbano in [Tor, ToUr]; one has for every H rotational embedded CMC tori given by the Hopf lift of a circle in $\mathbb{S}^2$, but there also exist some other non-flat rotational CMC tori. The embeddedness problem for such tori is open in general, but for the minimal case there are embedded rotational tori other than Clifford tori. This contrasts with the case of embedded minimal tori in $\mathbb{S}^3$.

A general study of CMC surfaces in $\mathbb{H}^2 \times \mathbb{R}$ and $\mathbb{S}^2 \times \mathbb{R}$ invariant by a continuous 1-parameter subgroup of ambient isometries can be found in [SaE, SaTo].

4.2. The Alexandrov problem in $\mathbb{E}^3(\kappa, \tau)$. One of the fundamental theorems of CMC surface theory is the so-called *Alexandrov theorem*.

Theorem 4.2 (Alexandrov). *Any compact embedded CMC surface in $\mathbb{R}^3$, $\mathbb{H}^3$ or a hemisphere of $\mathbb{S}^3$ is a round sphere.*

Proof. The proof relies on the so-called Alexandrov reflection principle, which we sketch for $\mathbb{R}^3$ although it works with great generality. Consider a plane P disjoint from the compact embedded CMC surface Σ, and start translating it in a parallel way towards Σ. After it first touches Σ, we start reflecting the piece

of Σ that has been left behind across this new translated plane. In this way we will eventually reach a first contact point with the unreflected part of Σ. By the maximum principle for elliptic PDEs, this means that Σ is symmetric with respect to such a plane. As the starting plane was arbitrary, the compact surface must be a round sphere. $\square$

It must be emphasized that there exist embedded CMC tori in $\mathbb{S}^3$, such as the product tori $\mathbb{S}^1(r) \times \mathbb{S}^1(\sqrt{1-r^2}) \subset \mathbb{S}^3$. Thus, the hemisphere hypothesis is necessary in the case of $\mathbb{S}^3$.

Motivated by this result, the problem of classifying all compact embedded CMC surfaces in a Riemannian 3-manifold $\bar{M}^3$ will be called the *Alexandrov problem* in $\bar{M}^3$.

In the case of CMC surfaces in the product spaces $\mathbb{H}^2 \times \mathbb{R}$ and $\mathbb{S}^2 \times \mathbb{R}$, the Alexandrov technique can be applied for *horizontal* directions, and so the following result holds.

Theorem 4.3 (Hsiang-Hsiang). *Any compact embedded CMC surface in* $\mathbb{H}^2 \times \mathbb{R}$ *or* $\mathbb{S}^2_+ \times \mathbb{R}$ *is a standard rotational CMC sphere.*

Again, the hemisphere hypothesis is necessary, since we know that there are embedded CMC tori in $\mathbb{S}^2 \times \mathbb{R}$.

As regards the homogeneous spaces $\mathbb{E}^3(\kappa, \tau)$ with $\tau \neq 0$, i.e. Heisenberg space, Berger spheres and the universal covering of $\mathrm{SL}_2(\mathbb{R})$, the Alexandrov problem is open. The main difficulty there is that these spaces do not admit reflections, and hence the reflection principle does not hold.

4.3. The Hopf problem in $\mathbb{E}^3(\kappa, \tau)$. Another fundamental result of CMC surface theory is the *Hopf theorem*:

Theorem 4.4 (Hopf). *Any immersed CMC sphere in* $\mathbb{R}^3$, $\mathbb{S}^3$ *or* $\mathbb{H}^3$ *is a round sphere.*

Proof. The Hopf differential (see Section 3) of any CMC surface in $\mathbb{R}^3$, $\mathbb{S}^3$ or $\mathbb{H}^3$ is holomorphic, and vanishes at the umbilical points of the surface. As any holomorphic quadratic differential must vanish on the Riemann sphere, we conclude that immersed CMC spheres are totally umbilical, and hence round spheres. $\square$

The *Hopf problem* in a Riemannian 3-manifold $\bar{M}^3$ will refer to the problem of classifying all immersed CMC spheres in $\bar{M}^3$.

As was proved in Theorem 3.4, CMC surfaces in the homogeneous spaces $\mathbb{E}^3(\kappa, \tau)$ have an associated holomorphic quadratic differential: the Abresch-Rosenberg differential Q_{AR}. This allows to solve the Hopf problem in $\mathbb{E}^3(\kappa, \tau)$, along the lines suggested by Hopf's classical theorem. We present here an alternative proof to the original one by Abresch and Rosenberg [AbRo1, AbRo2],

based on Daniel's integrability equations, and on some ideas in [FeMi2, GMM] (see [dCF, EsRo, DHM]).

Theorem 4.5 (Abresch-Rosenberg). *Any immersed CMC sphere in* $\mathbb{E}^3(\kappa,\tau)$ *is a standard rotational sphere.*

Proof. As the Abresch-Rosenberg Q_{AR} is holomorphic, it must vanish on any immersed CMC sphere. So, we need to prove that spheres with $Q_{AR}=0$ are rotational.

First, one can observe that on any CMC surface in $\mathbb{E}^3(\kappa,\tau)$, the equation $Q_{AR}=0$ together with the integrability conditions in Theorem 3.2 imply that the function $w := \operatorname{arctanh}(u)$ is a harmonic function on the surface (here u is the angle function of the surface). So, once we rule out the case $u=$ const. which does not produce CMC spheres (except for slices in $\mathbb{S}^2\times\mathbb{R}$), we can define ζ to be a local conformal parameter on the surface with $\operatorname{Re}\zeta = w$. Again from the integrability equations (**C.1**) to (**C.4**) we see that all fundamental data of the surface depend only on w (and not on $\operatorname{Im}\zeta$). This implies that the surface is a local piece of some CMC surface invariant by a continuous 1-parameter subgroup of ambient isometries of $\mathbb{E}^3(\kappa,\tau)$.

If the surface is compact, this isometry subgroup must be the group of rotations around the vertical axis, with the possible exception of Berger 3-spheres (the only space in which there are non-rotational compact continuous isometry subgroups). However, it is clear that any element of such a non-rotational isometry subgroup has no fixed points. Hence, by the invariance property, there is a globally defined non-zero vector field on the surface (that is tangent to the orbits). But this is impossible on a sphere. Hence, the isometry subgroup is always the group of rotations around the vertical axis, and thus the CMC sphere is rotational, as wished. □

4.4. The isoperimetric problem in $\mathbb{E}^3(\kappa,\tau)$. The Alexandrov problem is very relevant to the isoperimetric problem in a Riemannian 3-manifold $\bar{M}^3$; indeed, any solution to the isoperimetric problem in $\bar{M}^3$ is a region bounded by a compact embedded CMC surface. So, for instance, the only candidates to solve the isoperimetric problem for a given volume in $\mathbb{H}^2\times\mathbb{R}$ are rotational CMC spheres. Another geometric property satisfied by isoperimetric solutions is that they are weakly stable, see Section 2.

The class of isoperimetric solutions in $\mathbb{R}^3$, $\mathbb{S}^3$ and $\mathbb{H}^3$ is the class of round spheres. The isoperimetric problem in $\mathbb{S}^2\times\mathbb{R}$ and $\mathbb{H}^2\times\mathbb{R}$ has been also explicitly solved, as follows:

1. The isoperimetric regions in $\mathbb{H}^2\times\mathbb{R}$ are exactly the regions bounded by the canonical rotational CMC spheres. (Hsiang-Hsiang).

2. There is a value $H_1\approx 0.33$ such that the isoperimetric regions in $\mathbb{S}^2\times\mathbb{R}$ are exactly the regions bounded by the canonical rotational CMC H spheres with $H\geq H_1$. (Pedrosa).

So, regading complete simply connected Riemannian 3-manifolds, the isoperimetric problem is fully solved in $\mathbb{R}^3$, $\mathbb{S}^3$, $\mathbb{H}^3$, $\mathbb{S}^2 \times \mathbb{R}$ and $\mathbb{H}^2 \times \mathbb{R}$. A remarkable advance in this direction has been obtained very recently by F. Torralbo and F. Urbano [ToUr], who have added to this list a certain subfamily of Berger spheres:

Theorem 4.6 (Torralbo-Urbano). *The solutions to the isoperimetric problem in the Berger spheres $\mathbb{E}^3(\kappa,\tau)$ with $\frac{1}{3} \leq \frac{4\tau^2}{\kappa} < 1$ are the canonical rotational CMC spheres.*

The proof of this result relies on embedding the Berger spheres $\mathbb{E}^3(\kappa,\tau)$ into the 4-dimensional complex space $\mathbb{C}P^2$, and using a Willmore inequality in this space due to Montiel and Urbano [MoUr].

For the rest of the spaces, the isoperimetric problem is open. In any case, the general theory of the isoperimetric problem together with the Abresch-Rosenberg uniqueness theorem imply that, for small volumes, the isoperimetric solutions are canonical rotational CMC spheres with large H.

4.5. Open problems. One of the major unsolved problems in the theory is the Alexandrov problem when $\tau \neq 0$, i.e. in Nil_3, the universal cover of $\mathrm{PSL}(2,\mathbb{R})$ and Berger hemispheres. It is conjectured that canonical rotational spheres are the only compact embedded CMC surfaces in these spaces. A related open problem is the isoperimetric problem in Nil_3, the universal cover of $\mathrm{PSL}(2,\mathbb{R})$ and the Berger spheres not covered by Theorem 4.6. In the first two cases, it is conjectured that the isoperimetric solutions are exactly the canonical rotational spheres.

Besides, it is conjectured by Nelli and Rosenberg [NeRo2] that compact weakly stable CMC surfaces in $\mathbb{H}^2 \times \mathbb{R}$ are rotational CMC spheres.

Another important problem of the theory is the construction of higher genus compact (immersed) CMC surfaces, (e.g. CMC tori) in $\mathbb{E}^3(\kappa,\tau)$ with $\kappa \leq 0$.

5. CMC Spheres in Sol_3

In this section we will expose the recent solution to the Alexandrov problem (i.e. the classification of compact embedded CMC surfaces) and the Hopf problem (i.e. the classification of immersed CMC spheres) in the remaining Thurston 3-geometry: the homogeneous space Sol_3.

The first step in this direction is that we can solve the Alexandrov problem *from a topological point of view.*

Theorem 5.1 (Rosenberg). *Any compact embedded CMC surface in* Sol_3 *is, topologically, a sphere.*

Proof. By Alexandrov reflection principle using the two canonical foliations of Sol_3 (recall that reflections across their leafs are orientation-reversing isometries

of Sol_3), it turns out that any compact embedded CMC surface in Sol_3 is a bigraph with respect to two linearly independent directions in $\mathbb{R}^3$. Thus, the surface is, topologically, a sphere. $\square$

This result leaves us with the problem of classifying (embedded) CMC spheres. A substantial difficulty for this task is that Sol_3 has no rotations. Hence, there are no rotational CMC spheres to use in order to gain insight of the theory, and even the existence of CMC spheres for a given value of H needs to be settled.

The next theorem is the main result of the section, and solves the Hopf and Alexandrov problems in Sol_3.

Theorem 5.2 (Daniel-Mira, Meeks). *For every $H > 0$ there exists an embedded CMC H sphere S_H in Sol_3. This sphere is unique in the following sense:*

1. Hopf uniqueness: *every immersed CMC H sphere in Sol_3 is a left-translation of S_H.*
2. Alexandrov uniqueness: *every compact embedded CMC H surface in Sol_3 is a left-translation of S_H.*

Moreover, each sphere S_H has index one, it inherits all possible symmetries of the ambient space (its group of ambient isometries is the dihedral group D_4), its Lie group Gauss map is a diffeomorphism, and the family $\{S_H : H > 0\}$ is real analytic (up to left translations).

Remark 5.3. *Theorem 5.2 was obtained by Daniel and Mira [DaMi] for $H > 1/\sqrt{3}$. For the remaining values $H \in (0, 1/\sqrt{3}]$, Daniel and Mira also proved the uniqueness in the Hopf and Alexandrov sense for all values of H for which there exists an index one CMC H sphere. Finally, Meeks [Mee] obtained the existence of index one CMC H spheres for every $H > 0$ (and not just for $H > 1/\sqrt{3}$). This concluded the proof of Theorem 5.2.*

We shall split the sketch of the proof of Theorem 5.2 into two parts.

5.1. Proof of Theorem 5.2: uniqueness. The results of this part are contained in [DaMi]. The Lie group Gauss map $g : \Sigma \to \overline{\mathbb{C}}$ of a CMC surface $X : \Sigma \to \mathrm{Sol}_3$ satisfies the following elliptic PDE (here z is a conformal parameter on the surface):

$$g_{z\bar{z}} = A(g) g_z g_{\bar{z}} + B(g) g_z \bar{g}_{\bar{z}}, \tag{7}$$

where, by definition,

$$A(q) = \frac{R_q}{R} = \frac{2H(1+|q|^2)\bar{q} + 2q}{R(q)}, \quad B(q) = \frac{R_{\bar{q}}}{R} - \frac{\bar{R}_{\bar{q}}}{\bar{R}} = -\frac{4H(1+|q|^2)(\bar{q}+q^3)}{|R(q)|^2},$$
$$R(q) = H(1+|q|^2)^2 + q^2 - \bar{q}^2. \tag{8}$$

Moreover, the surface X is uniquely determined by the Gauss map g, and it can actually be recovered from g by means of an integral representation formula.

Once here, the first idea in order to prove a Hopf-type theorem is to look for a holomorphic quadratic differential for CMC surfaces in Sol_3. However, it seems that such a holomorphic object is not available in the theory; this constitutes another key difference from the theory of CMC surfaces in the other Thurston 3-geometries exposed in the previous section, where the Abresch-Rosenberg (or the Hopf differential) is holomorphic.

Still, it is not strictly necessary to obtain a holomorphic differential in order to prove a Hopf-uniqueness theorem: it suffices to find a geometrically defined quadratic differential with isolated zeros of negative index, so that it vanishes identically on spheres. This is done as follows.

Theorem 5.4 (Daniel-Mira). *Let $H > 0$, and assume that there exists an index one CMC H sphere S_H in Sol_3. Then there exists a quadratic differential Q_H, geometrically defined on any CMC H surface in Sol_3, with the following properties:*

1. *It has only isolated zeros of negative index (thus, it vanishes on spheres).*
2. *$Q_H = 0$ holds for a surface $X : \Sigma \to \mathrm{Sol}_3$ if and only if X is a left-translation of some piece of the sphere S_H.*

Moreover, the sphere S_H is embedded, and it is therefore unique in Sol_3 (up to left-translations) in the Hopf sense and in the Alexandrov sense.

The quadratic differential Q_H is constructed as follows. Let $G : S_H \equiv \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ denote the Gauss map of S_H. Then G is a diffeomorphism (otherwise one can construct a Jacobi function u on S_H with $u(p) = \nabla u(p) = 0$ at some $p \in S_H$, which contradicts the index one condition by Courant's nodal domain theorem).

Once here, the differential Q_H is defined for any CMC H surface $X : \Sigma \to \mathrm{Sol}_3$ with Gauss map $g : \Sigma \to \overline{\mathbb{C}}$ by

$$Q_H = (L(g)g_z^2 + M(g)g_z\bar{g}_z)\, dz^2, \tag{9}$$

where by definition

$$M(q) = \frac{1}{R(q)} = \frac{1}{H(1+|q|^2)^2 + q^2 - \bar{q}^2} \tag{10}$$

and $L : \overline{\mathbb{C}} \to \mathbb{C}$ is implicitly given in terms of the Gauss map G of S_H by

$$L(G(z)) = -\frac{M(G(z))\bar{G}_z(z)}{G_z(z)}. \tag{11}$$

It must also be emphasized that, by this uniqueness theorem, any index one CMC sphere S_H in Sol_3 is as symmetric as the ambient space allows: there is

a point $p \in \mathrm{Sol}_3$ such that S_H is invariant with respect to all the isometries of Sol_3 that leave p fixed.

5.2. Proof of Theorem 5.2: existence. Let us define

$$\mathcal{I} := \{H > 0 : \text{exists an index one CMC } H \text{ sphere } S_H \text{ in } \mathrm{Sol}_3\}.$$

We prove next the theorem by Meeks [Mee] that $\mathcal{I} = (0, \infty)$. (The fact that $(1/\sqrt{3}, \infty) \subset \mathcal{I}$ had been previously obtained in [DaMi]).

That $\mathcal{I} \neq \emptyset$ follows from the existence of isoperimetric spheres, which in Sol_3 must have index one. That $\mathcal{I}$ is open was proved in [DaMi], and follows from the implicit function theorem and from the continuity of the eigenvalues and eigenspaces in the deformation.

The proof that $\mathcal{I}$ is closed is the critical step. The key point is to prevent that the diameters of a sequence of CMC H_n spheres (S_{H_n}) with $H_n \to H_0 > 0$ tend to ∞. This was proved first by Daniel-Mira, but only for $H_0 > 1/\sqrt{3}$. The final proof for every $H_0 > 0$ was recently given by Meeks [Mee], using the following height estimate: *there exists a constant $K(H_0)$ such that for any CMC H_0 graph (possibly non-compact) with respect to one of the two canonical foliations of* Sol_3*, and with boundary on a leaf, the maximum height attained by the graph with respect to this leaf is $\leq K(H_0)$.*

Once this height estimate is ensured, Meeks concludes the proof by some elliptic theory and stability arguments.

5.3. Open problems. Are CMC spheres in Sol_3 weakly stable? Do they all bound isoperimetric regions in Sol_3? A positive answer is conjectured in [DaMi]. What happens in other homogeneous 3-spaces with 3-dimensional isometry group?

It seems very interesting to develop a global theory of minimal surfaces in Sol_3. Some natural problems would be proving half-space theorems, classifying entire minimal graphs, or finding properly embedded minimal surfaces of non-trivial topology.

6. Surfaces of Critical CMC

As we saw in Section 3, CMC H spheres in the homogeneous space $\mathbb{E}^3(\kappa, \tau)$ exist exactly for the values $H^2 > -\kappa/4$. Besides, one can easily see that there exist entire rotational CMC H graphs in $\mathbb{E}^3(\kappa, \tau)$, $\kappa \leq 0$, whenever $H^2 \leq -\kappa/4$. From these results and the maximum principle, we obtain

Theorem 6.1. *Any compact CMC H surface in $\mathbb{E}^3(\kappa, \tau)$ satisfies $H^2 > -\kappa/4$. Also, any entire CMC graph in $\mathbb{E}^3(\kappa, \tau)$, $\kappa \leq 0$, satisfies $H^2 \leq -\kappa/4$.*

There are several other properties that make the theory of CMC surfaces with $H^2 > -\kappa/4$ quite different from the theory of CMC surfaces with $H^2 \leq -\kappa/4$. For instance:

1. A properly embedded CMC surface in $\mathbb{H}^2 \times \mathbb{R}$ with $H > 1/2$ and finite topology cannot have exactly one end (Espinar, Gálvez, Rosenberg, [EGR]).

2. There exist horizontal and vertical height estimates for CMC surfaces with $H > 1/2$ in $\mathbb{H}^2 \times \mathbb{R}$ [NeRo2, AEG1, EGR].

3. There are no complete stable CMC surfaces in $\mathbb{H}^2 \times \mathbb{R}$ with $H > 1/\sqrt{3}$, and the result is expected for $H > 1/2$. (Nelli-Rosenberg, [NeRo2]). Besides, there are no complete stable CMC surfaces with $H > 1/2$ in $\mathbb{H}^2 \times \mathbb{R}$ of parabolic conformal structure (Manzano-Pérez-Rodríguez, [MaPR]).

It is hence natural to introduce the following definition.

Definition 6.2. *We say that a CMC surface in $\mathbb{E}^3(\kappa, \tau)$ with $\kappa \leq 0$ has* critical *CMC if its mean curvature H satisfies $H^2 = -\kappa/4$.*

The critical mean curvature is the largest value of $|H|$ for which compact CMC surfaces do not exist. Therefore we have $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$, minimal surfaces in Nil_3, and $H = \sqrt{-\kappa}/2$ surfaces in the universal covering of $\mathrm{PSL}(2, \mathbb{R})$. A remarkable property is that the sister correspondence preserves the property of having critical CMC, and that every simply connected surface of critical CMC is the sister surface of some minimal surface in Nil_3.

In this section we will study the global geometry of surfaces with critical CMC, focusing on the existence of harmonic Gauss maps and the classification of entire graphs.

6.1. Harmonic Gauss maps. A smooth map $G : M \to N$ between Riemannian manifolds is harmonic if it is a critical point for the total energy functional. When M is a surface, harmonicity is a conformal invariant, and it implies that the quadratic differential

$$Q_0 dz^2 = \langle G_z, G_z \rangle dz^2,$$

is holomorphic, where z is a conformal parameter on Σ, and $\langle , \rangle$ denotes the metric in N (see [FoWo]). We call $Q_0 dz^2$ the *Hopf differential* associated to G.

The Gauss map of CMC surfaces in $\mathbb{R}^3$ is harmonic into $\mathbb{S}^2$, and its Hopf differential agrees (up to a constant) with the Hopf differential of the surface. Moreover, the CMC surface can be recovered from the Gauss map by a representation formula. This Gauss map opens the door to the use of strong techniques from harmonic maps in the description of CMC surfaces.

The same holds for spacelike CMC surfaces in Minkowski 3-space $\mathbb{L}^3$, but this time the harmonic Gauss map takes values into $\mathbb{H}^2$. Let us briefly comment

this case, since it will play an important role in the development of the section. Let $f : \Sigma \to \mathbb{L}^3$ be a connected spacelike CMC surface, oriented so that its Gauss map G takes values in $\mathbb{H}^2$. Here $\mathbb{L}^3$ is $\mathbb{R}^3$ with the metric $dx^2 + dy^2 - dz^2$ and $\mathbb{H}^2$ is realized in $\mathbb{L}^3$ in the usual way. It turns out that G is harmonic into $\mathbb{H}^2$ and its associated Hopf differential agrees (up to a multiplicative constant) with the Hopf differential of the immersion f. Moreover, the metric of the CMC surface, $\langle df, df\rangle = \tau_0|dz|^2$, is related with G by

$$2\langle G_z, G_{\bar z}\rangle = \frac{\tau_0}{4} + \frac{4|Q_0|^2}{\tau_0}.$$

Definition 6.3. *We will say that a harmonic map G into $\mathbb{H}^2$ admits* Weierstrass data $\{Q_0, \tau_0\}$ *if the pullback metric induced by G can be written as*

$$\langle dG, dG\rangle = Q_0 dz^2 + \mu|dz|^2 + \bar{Q}_0 d\bar z^2, \qquad \mu = \frac{\tau_0}{4} + \frac{4|Q_0|^2}{\tau_0},$$

τ_0 being a positive smooth function.

6.1.1. $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$. We will regard $\mathbb{H}^2 \times \mathbb{R} = \mathbb{E}^3(-1, 0)$ in its Minkowski model, i.e.

$$\mathbb{H}^2 \times \mathbb{R} = \{(x_0, x_1, x_2, x_3) : x_0 > 0, -x_0^2 + x_1^2 + x_2^2 = -1\} \subset \mathbb{L}^3 \times \mathbb{R} = \mathbb{L}^4.$$

Using this model, the unit normal vector η of an immersed surface $\psi = (N, h) : \Sigma \to \mathbb{H}^2 \times \mathbb{R}$ takes values in the de Sitter 3-space, and $\{\eta, N\}$ is an orthonormal frame for the Lorentzian normal bundle of ψ in $\mathbb{L}^4$. Moreover, if u is the angle function of the surface (that is, the last coordinate of η) and we assume that $u \neq 0$ (that is, that ψ is nowhere vertical, or equivalently, that it is a multigraph), then we can write

$$\frac{1}{u}(\eta + N) = (G, 1), \tag{12}$$

for a certain map $G : \Sigma \to \mathbb{H}^2$.

Definition 6.4 ([FeMi1]). *The map G given by* (12) *will be called the* hyperbolic Gauss map *of an immersed (nowhere vertical) surface in $\mathbb{H}^2 \times \mathbb{R}$.*

The main property of the hyperbolic Gauss map is the following [FeMi1]:

Theorem 6.5 (Fernández-Mira). *The hyperbolic Gauss map of a CMC surface with $H = 1/2$ in $\mathbb{H}^2 \times \mathbb{R}$ is a harmonic map into $\mathbb{H}^2$, and admits Weierstrass data $\{-Q, \lambda u^2\}$, where Qdz^2, $\lambda|dz|^2$ and u are, respectively, the Abresch-Rosenberg differential, the metric, and the angle function of the surface.*

Conversely, if Σ is simply connected, any harmonic map $G : \Sigma \to \mathbb{H}^2$ admitting Weierstrass data is the hyperbolic Gauss map of some $H = 1/2$ surface in $\mathbb{H}^2 \times \mathbb{R}$.

Moreover, the space of $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$ with the same hyperbolic Gauss map G is generically two-dimensional, and it can be recovered from G by a representation formula.

The proof of the direct part of the above result follows from equations (2) and the very definition of G. The converse part is an integrability argument. This result is of great importance for the rest of this section, since it allows the use of harmonic maps in the description of surfaces of critical CMC.

6.1.2. Minimal surfaces in Nil_3. The existence of this harmonic Gauss map for $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$ was extended by B. Daniel [Dan2] to the case of minimal surfaces in $\mathrm{Nil}_3 = \mathbb{E}^3(0, \frac{1}{2})$.

This time, the harmonic Gauss map is given by the Lie group Gauss map of the surface. Indeed, if we identify the Lie algebra of Nil_3 with the tangent space at a point by left multiplication, we can stereographically project the unit normal vector field to obtain a map taking values in the extended complex plane. More specifically, we will consider the model of Nil_3 given in Section 2 and its canonical frame of left-invariant fields $\{E_1, E_2, E_3\}$. If $N = \sum N_i E_i$ is the unit normal of $X : \Sigma \to \mathrm{Nil}_3$, then the Gauss map of X is given by

$$g = \frac{N_1 + iN_2}{1 + N_3} : \Sigma \to \overline{\mathbb{C}}.$$

Now, if the surface is nowhere vertical we can orient it so that $u = \langle N, E_3 \rangle$ is positive, and so g takes values in the unit disc $\mathbb{D}$. By identifying $\mathbb{H}^2$ with $(\mathbb{D}, ds_P^2)$, where ds_P^2 is the Poincaré metric, Daniel obtained in [Dan2]:

Theorem 6.6 (Daniel). *The Gauss map of a nowhere vertical minimal surface is harmonic into $\mathbb{H}^2$.*

Conversely, let $g : \Sigma \to \mathbb{H}^2$ be a harmonic map defined on a simply connected oriented Riemann surface into $\mathbb{H}^2$, and assume that g is nowhere antiholomorphic (i.e., g_z does not vanish at any point). Take $z_0 \in \Sigma$ and $X_0 \in \mathrm{Nil}_3$.

Then there exists a unique conformal nowhere vertical minimal immersion $X : \Sigma \to \mathrm{Nil}_3$ with $X(z_0) = X_0$ having g as its Gauss map. Moreover, X can be uniquely recovered from g through an adequate representation formula.

Furthermore, it can be checked that the Weierstrass data of g as above are $\{-Q, \lambda u^2\}$, where Qdz^2, $\lambda|dz|^2$ and u are, respectively, the Abresch-Rosenberg differential, the metric, and the angle function of the surface defined in Section 2.

As we saw in Section 3, minimal surfaces in Nil_3 and $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$ are related by the sister correspondence, and sister surfaces have the same metric and angle function (in particular, the condition of being nowhere vertical is preserved). As in this case the sister surfaces have opposite Abresch-Rosenberg differentials, it turns out that their respective harmonic Gauss maps are conjugate to each other.

The relation between minimal surfaces in Nil_3 and $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$ can be made more explicit by means of the theory of spacelike CMC surfaces in $\mathbb{L}^3$, as follows.

Theorem 6.7 ([FeMi3]). *Let $X = (F,t) : \Sigma \to \mathrm{Nil}_3$ be a simply connected nowhere vertical minimal surface with metric $\lambda|dz|^2$ and angle function u, and $\psi = (N,h) : \Sigma \to \mathbb{H}^2 \times \mathbb{R}$ its sister surface.*

Then $f := (F,h) : \Sigma \to \mathbb{L}^3$ is a spacelike $H = 1/2$ surface in the Minkowski 3-space with metric $\lambda u^2|dz|^2$ and Hopf differential $-Qdz^2$, where Qdz^2 is the Abresch-Rosenberg differential of X.

6.1.3. CMC $\sqrt{-\kappa}/2$ surfaces in $\widetilde{\mathrm{PSL}(2,\mathbb{R})}$. In a forthcoming paper [DFM], the authors and B. Daniel will prove that there exists also a harmonic Gauss map for critical CMC surfaces in the remaining case, i.e. the universal covering of the group $\mathrm{PSL}(2,\mathbb{R})$, and will derive a representation formula for them.

6.2. Half-space theorems.

One of the most important results in the global study of minimal surfaces in $\mathbb{R}^3$ is the classical half-space theorem by Hoffman and Meeks [HoMe]. This theorem says that any properly immersed minimal surface in $\mathbb{R}^3$ lying in a half-space must be a plane parallel to the one determining the half-space. The main tools used here are the maximum principle and the existence of catenoids, a 1-parameter family of minimal surfaces converging to a doubly-covered punctured plane P, and intersecting the planes parallel to P in compact curves.

The analogous version for CMC one half surfaces in $\mathbb{H}^2 \times \mathbb{R}$ was proved in [HRS]. In this setting, horocylinders play the role of the planes in $\mathbb{R}^3$.

Theorem 6.8 (Hauswirth-Rosenberg-Spruck). *The only properly immersed CMC one half surfaces in $\mathbb{H}^2 \times \mathbb{R}$ that are contained in the mean convex side of a horocylinder C are the horocylinders parallel to C.*

Also, the only properly embedded CMC one half surfaces in $\mathbb{H}^2 \times \mathbb{R}$ containing a horocylinder in its mean convex side are the horocylinders.

Proof. The main point here is to construct a family of CMC one half surfaces in $\mathbb{H}^2 \times \mathbb{R}$ to be used in the same way as catenoids in the proof of the half-space theorem in $\mathbb{R}^3$. This is achieved by means of compact annuli with boundaries, contained between two horocylinders. □

For the case of Nil_3, we must distinguish between horizontal and vertical half-spaces. The equivalent to the half-space theorem for surfaces lying in a horizontal half-space is proved by using the family of rotational annuli [AbRo2]. The corresponding vertical version has been obtained in [DaHa], by constructing first a family of *horizontal catenoids*, i.e. properly embedded minimal annuli (non-rotational) with a geometric behaviour good enough to apply the Hoffman-Meeks technique.

Theorem 6.9 (Daniel-Hauswirth). *The only properly immersed minimal surfaces in Nil_3 that are contained in a vertical half space are the vertical planes parallel to the one determining the half-space.*

Proof. Using the representation formula for minimal surfaces in Nil_3 (see Theorem 6.6), it is possible to construct *horizontal catenoids* in Nil_3. These surfaces are a 1-parameter family of properly embedded minimal annuli, intersecting vertical planes $\{x_2 = c\}$ in a non-empty closed convex curve. Moreover, the family converges to a double covering of $\{x_2 = 0\}$ minus a point. They are obtained by integrating a family of harmonic maps that belong to a more general family used in the construction of Riemann type minimal surfaces in $\mathbb{H}^2 \times \mathbb{R}$ [Ha]. Once we have these *catenoids*, we finish by using the maximum principle similarly to the Euclidean case. □

6.3. The classification of entire graphs. In this section we will describe the space of entire graphs of critical CMC in $\mathbb{E}^3(\kappa, \tau)$. Such a description follows from the works of Fernández-Mira [FeMi1, FeMi3], Hauswirth-Rosenberg-Spruck [HRS] and Daniel-Hauswirth [DaHa], and is contained in Theorems 6.10 and 6.11. We expose here a unified perspective to this subject. First, we have

Theorem 6.10 ([DaHa, FeMi3, HRS]). *The following conditions are equivalent for a surface of critical CMC in $\mathbb{E}^3(\kappa, \tau)$:*

(1) *It is an entire graph.*

(2) *It is a complete multigraph.*

(3) *$u^2\, ds^2$ is a complete Riemannian metric (where u is the angle function and ds^2 the metric of the surface).*

In particular, the sister correspondence preserves entire graphs of critical CMC.

Let us make some comments on this theorem. First, Hauswirth, Rosenberg and Spruck proved (2) ⇒ (1) for $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$. Second, the authors proved in [FeMi3] that (3) ⇒ (1) (for any surface in $\mathbb{E}^3(\kappa, \tau)$, not necessarily CMC), and that (1) ⇒ (3) holds for minimal surfaces in Nil_3. Finally, Daniel and Hauswirth showed that (2) ⇒ (1) holds for minimal surfaces in Nil_3. The rest of the cases can be easily obtained from these results and the sister correspondence (this was first observed in [DHM]).

Proof. It is immediate that (1) ⇒ (2). Also, by an eigenvalue estimate, the authors proved in [FeMi3] that for arbitrary surfaces in $\mathbb{E}^3(\kappa, \tau)$ it holds $u^2\, ds^2 \leq g_F$, where $F = \pi \circ \psi$ is the projection onto $\mathcal{M}^2(\kappa)$ of ψ. Thus, if $u^2\, ds^2$ is complete, F is a local diffeomorphism with complete pullback metric, and by standard topological arguments, F is a diffeomorphism, i.e. (3) ⇒ (1) holds.

That (1) ⇒ (3) holds for minimal surfaces in Nil_3 was also proved in [FeMi3]: let $X = (F, t) : \Sigma \to \mathrm{Nil}_3$ be an entire minimal graph. By Theorem 6.7, there is an entire spacelike CMC graph $f = (F, h) : \Sigma \to \mathbb{L}^3$, whose induced metric is $ds_f^2 = u^2\, ds^2$. Now we can apply a theorem by Cheng and Yau [ChYa] which

says that spacelike entire CMC graphs in $\mathbb{L}^3$ have complete induced metric. Hence $u^2\,ds^2$ is complete, as wished.

We will now prove that (2) $\Rightarrow$ (1) holds for minimal surfaces in Nil_3. Let us observe that once this is done, we can also prove the theorem for surfaces of critical CMC in all the spaces $\mathbb{E}^3(\kappa,\tau)$. Indeed, as any simply connected surface of critical CMC is the sister surface of some minimal surface in Nil_3, and as the correspondence preserves the metric and the angle function (therefore it preserves conditions (2) and (3) by passing to the universal covering), we can easily translate the theorem for the case of minimal surfaces in Nil_3 to the rest of the spaces. It is important here that we proved (3) $\Rightarrow$ (1) in all spaces.

So, we only need to prove (2) $\Rightarrow$ (1) for minimal surfaces in Nil_3. This was done by Daniel and Hauswirth [DaHa]. For that, they used their half-space theorem in Nil_3 (Theorem 6.9) and an adaptation to Nil_3 of the previous proof of (2) $\Rightarrow$ (1) for the case of $H = 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$ given by Hauswirth-Rosenberg-Spruck [HRS].

In order to prove (2) $\Rightarrow$ (1) for minimal surfaces in Nil_3, we argue by contradiction. Assume that there exists a complete multigraph Σ that is not entire. Then there exists an open set $\Sigma_0 \subset \Sigma$ that is a graph over a disc $D \subsetneq \mathbb{R}^2$ of a function f, and a point $q \in \partial D$ such that f does not extend to q.

Step 1: For any sequence of points $\{q_n\}$ in D converging to q, the sequence of normal vectors at the points $p_n = (q_n, f(q_n)) \in \Sigma_0$ converges to the horizontal vector orthogonal to ∂D at q.

Indeed, as the surface is a multigraph, its angle function $u = \langle N, E_3\rangle$, where N denotes the unit normal vector, is a Jacobi function that does not vanish. As a result of this, Σ is (strongly) stable, and has bounded geometry. This means that locally around any p_n we can write the surface as the graph (in exponential coordinates) over a disc of radius δ of its tangent plane, where δ is a universal constant depending only on Σ. This neighborhood of p_n will be denoted by $\mathcal{G}(p_n)$. The limit of the normal vectors $\{N(p_n)\}$ must be a horizontal vector since otherwise, the piece $\mathcal{G}(p_n)$ of bounded geometry could be extended as a graph beyond q, which is impossible. Moreover, the limit vector must be normal to ∂D at q since Σ_0 is a graph over D.

Step 2: The function f defining the graph Σ_0 diverges at q. Moreover, as we approach q, and after translating the surface to the origin, the surfaces converge to a piece of the (translated) vertical plane P passing through q and tangent to ∂D.

That f diverges at q is a consequence of the completeness of Σ, and the last part can be proved by following the ideas of Collin and Rosenberg in [CoRo]. We will assume that P is the plane $\{x_1 = c\}$.

Step 3: Σ contains a graph $\mathcal{G}$ over a domain of the form $U_\epsilon = (c-\epsilon, c) \times \mathbb{R} \subset \mathbb{R}^2$. Moreover, this graph is disjoint from P and asymptotic to it as one approaches q.

The graph $\mathcal{G}$ is obtained by analytical continuation of the surfaces $\mathcal{G}(p_n)$ used in the first step, and after a careful study of the behavior of the intersection curves of these graphs and the planes parallel to P.

Finally, the contradiction follows from the half-space theorem (Theorem 6.9). Recall that, although $\mathcal{G}$ has boundary and the theorem is formulated for surfaces without boundary, its proof applies to this case, and so we are done. $\square$

Once here, we investigate the *Bernstein problem* for entire graphs of critical CMC in $\mathbb{E}^3(\kappa,\tau)$, i.e. the classification of such entire graphs (recall here that CMC graphs in $\mathbb{E}^3(\kappa,\tau)$ satisfy the elliptic PDE (5)). The terminology comes from the classical Bernstein theorem: *entire minimal graphs in* $\mathbb{R}^3$ *are planes.* Equivalently, any solution to the minimal graph equation

$$(1+f_y^2)f_{xx} - 2f_xf_yf_{xy} + (1+f_x^2)f_{yy} = 0 \tag{13}$$

defined on the whole plane is linear.

It is interesting to compare this result with the Bernstein problem in Nil_3, i.e. the classification of entire minimal graphs in Nil_3. This corresponds to classifying all global solutions to the PDE (6). Observe that taking $\tau = 0$ in (6) we obtain the classical equation (13), i.e. the classical case considered by Bernstein appears as a limit of the Heisenberg case.

There exists, however, a great difference between both situations. The following result classifies the entire graphs of critical CMC in $\mathbb{E}^3(\kappa,\tau)$, by parametrizing the moduli space of such entire graphs in terms of holomorphic quadratic differentials. It was obtained first for minimal graphs in Nil_3 by the authors [FeMi3], and shortly thereafter by Daniel and Hauswirth [DaHa] for $H = 1/2$ graphs in $\mathbb{H}^2\times\mathbb{R}$. The general case follows easily from the Heisenberg case and Theorem 6.10, using the sister correspondence (this was observed first in [DHM]).

Theorem 6.11 (Fernández-Mira, Daniel-Hauswirth). *Let Qdz^2 denote a holomorphic quadratic differential on $\Sigma\equiv\mathbb{C}$ or $\mathbb{D}$, such that $Q\not\equiv 0$ if $\Sigma\equiv\mathbb{C}$, and let $H^2 = -\kappa/4$.*

There exists a 2-parameter family of entire CMC H graphs in $\mathbb{E}^3(\kappa,\tau)$ whose Abresch-Rosenberg differential agrees with Qdz^2. These graphs are generically non-congruent.

And conversely, these are all the entire graphs of critical CMC in $\mathbb{E}^3(\kappa,\tau)$.

At this point, the proof for the case of minimal surfaces in Nil_3 is a consequence of Theorem 6.7 and the following result by Wan and Wan-Au [Wan, WaAu] on spacelike entire CMC graphs in $\mathbb{L}^3$: *for any holomorphic quadratic differential as above, there exists a unique (up to isometries) spacelike entire CMC 1/2 graph in $\mathbb{L}^3$ with Hopf differential Qdz^2.* The 2-parameter family of non-congruent graphs in $\mathbb{E}^3(\kappa,\tau)$ comes from the loss of ambient isometries (from 6 dimensions to 4 dimensions) when passing from $\mathbb{L}^3$ to Nil_3.

The remaining cases of critical CMC graphs follow since by Theorem 6.10 the sister correspondence preserves entire graphs.

6.4. Open Problems. As explained in Section 3, entire graphs are stable. It is conjectured that entire graphs and vertical cylinders are the only stable critical CMC surfaces (this has been proved for parabolic conformal type in [MaPR]). Related to this is the question of non-existence of complete stable $H > 1/2$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$ (proved for $H > 1/\sqrt{3}$ by Nelli-Rosenberg, [NeRo2]).

Also, not much is known about properly embedded surfaces of critical CMC and non-trivial topology. Can one obtain them by conjugate Plateau constructions, or by integrable systems techniques? Another remarkable problem is to establish the strong half-space theorem in Nil_3: are two disjoint properly embedded minimal surfaces in Nil_3 necessarily two parallel vertical planes, or two parallel entire minimal graphs?

7. Minimal Surfaces in $\mathbb{H}^2 \times \mathbb{R}$ and $\mathbb{S}^2 \times \mathbb{R}$

Minimal surfaces in product spaces admit a special treatment, due to several reasons. One of them is the following: if $\psi = (N, h) : \Sigma \to M^2 \times \mathbb{R}$ is a minimal surface immersed in the product space $M^2 \times \mathbb{R}$, where (M^2, g) is a Riemannian surface, then the horizontal projection $N : \Sigma \to M^2$ is a harmonic map and the height function $h : \Sigma \to \mathbb{R}$ is a harmonic function. This implies, for instance, that compact minimal surfaces in $M^2 \times \mathbb{R}$ only exist if M^2 is compact (in particular, if $M^2 = \mathbb{S}^2$), and the only ones are the slices $M^2 \times \{t_0\}$.

Another important fact about minimal surfaces in $M^2 \times \mathbb{R}$ is that there is a natural notion of *minimal graph* over a domain $\Omega \subset M^2$, and that this graph satisfies a simple elliptic PDE in divergence form. This fact together with general existence results for solutions to the Plateau problem in Riemannian 3-manifolds allows a good control on the geometry of the surface. Some of the most interesting results of the theory of minimal surfaces in product spaces come from the interplay between the information provided by harmonic maps and by Plateau constructions and the minimal graph equation.

Starting with the pioneer work of H. Rosenberg [Ros], and W.H Meeks and H. Rosenberg [MeRo1, MeRo2], the theory of minimal surfaces in $M^2 \times \mathbb{R}$ has developed substantially in the last decade. We will only talk here about a few results of special relevance to the theory, and not mention many other important results.

7.1. The Collin-Rosenberg theorem. The classical Bernstein theorem in $\mathbb{R}^3$ states that planes are the only entire minimal graphs in $\mathbb{R}^3$. This theorem can be extended to the case of product spaces: *any entire minimal graph in $M^2 \times \mathbb{R}$, where (M^2, g) is a complete surface of non-negative curvature, is totally geodesic.*

In contrast, in the product space $\mathbb{H}^2 \times \mathbb{R}$ there is a wide variety of entire minimal graphs. For instance, in [NeRo1] Nelli and Rosenberg solved the Dirichlet problem at infinity for the minimal graph equation in $\mathbb{H}^2 \times \mathbb{R}$. They proved that any Jordan curve at the ideal boundary $\mathbb{S}^1 \times \mathbb{R} \equiv \partial_\infty \mathbb{H}^2 \times \mathbb{R}$ of $\mathbb{H}^2 \times \mathbb{R}$ which is a graph over $\mathbb{S}^1 \equiv \partial_\infty \mathbb{H}^2$ is the asymptotic boundary of a unique entire minimal graph in $\mathbb{H}^2 \times \mathbb{R}$ (see [GaRo] for a proof of this in the more general case of entire minimal graphs in $M^2 \times \mathbb{R}$, where (M^2, g) is complete, simply connected and with $K_M \leq c < 0$).

All these entire minimal graphs are hyperbolic, that is, they have the conformal type of the unit disk. The problem of existence of entire minimal graphs of parabolic type (i.e. with the conformal type of $\mathbb{C}$) is much harder, and was solved recently by Collin and Rosenberg [CoRo].

Theorem 7.1 (Collin-Rosenberg). *There exist entire minimal graphs in $\mathbb{H}^2 \times \mathbb{R}$ of parabolic conformal type.*

As the projection onto $\mathbb{H}^2$ of a minimal graph is a harmonic diffeomorphism, the above theorem has the following consequence, which solves a major problem in the theory of harmonic maps and disproves a conjecture by R. Schoen and S.T. Yau.

Corollary 7.2 (Collin-Rosenberg). *There exist harmonic diffeomorphisms from $\mathbb{C}$ onto $\mathbb{H}^2$.*

The proof by Collin and Rosenberg is a good example of the interaction between the harmonicity properties of the minimal immersion and the use of Plateau constructions and the minimal graph equation.

The main idea in the proof is to construct first (non-entire) minimal graphs in $\mathbb{H}^2 \times \mathbb{R}$ of Scherk type over ideal geodesic polygons, having alternating asymptotic values $+\infty$ and $-\infty$ on the sides of the polygon. This generalizes a classical construction by Jenkins and Serrin in the case of minimal graphs over bounded domains in $\mathbb{R}^3$. This construction is done as follows:

Let Γ be an ideal polygon of $\mathbb{H}^2$, so that all the vertices of Γ are at the ideal boundary of $\mathbb{H}^2$ and Γ has an even number of sides $A_1, B_1, A_2, B_2 ..., A_k, B_k$, ordered clockwise. At each vertex a_i, we consider a small enough horocycle H_i with $H_i \cap H_j = \emptyset$. Each A_i (resp. B_i) meets exactly two horocycles. Denote by $\widetilde{A}_i$ (resp. $\widetilde{B}_i$), the compact arc of A_i(resp B_i) which is the part of A_i outside the two horodisks. We denote by $|A_i|$ the length of $|\widetilde{A}_i|$. Define $\widetilde{B}_i$ and $|B_i|$ in the same way.

Now we can consider $a(\Gamma) = \sum_{i=1}^k |A_i|$ and $b(\Gamma) = \sum_{i=1}^k |B_i|$. We observe that $a(\Gamma) - b(\Gamma)$ does not depend on the choice of the horocycle H_i at a_i, since horocycles with the same point at infinity are equidistant. Keeping in mind

these data, we can state the following theorem by Collin-Rosenberg [CoRo] (see also Nelli-Rosenberg [NeRo1]):

Theorem 7.3. *([NeRo1], [CoRo]) There is a (unique up to additive constants) solution to the minimal surface equation in the polygonal domain P, equal to $+\infty$ on A_i and $-\infty$ on B_i, if and only if the following conditions are satisfied:*

1. $a(\Gamma) = b(\Gamma)$,
2. *For each inscribed polygon $\mathcal{P}$ in Γ, $\mathcal{P} \neq \Gamma$, and for some choice of horocycles at the vertices, one has*

$$2a(\mathcal{P}) < |\mathcal{P}| \text{ and } 2b(\mathcal{P}) < |\mathcal{P}|.$$

All these examples have the conformal type of $\mathbb{C}$. Once there, Collin and Rosenberg designed a way of enlarging a given Scherk-type graph over the interior of some $\Gamma \subset \mathbb{H}^2$ into another one with more sides, and so that: (1) the extended surface is C^2-close to the original one over an arbitrary compact set in the interior of Γ, and (2) there is a control on the conformal radius on adequate compact annuli on the surface.

By passing to the limit in this sequence of minimal graphs over larger and larger domains, they obtained an entire minimal graph in $\mathbb{H}^2 \times \mathbb{R}$ which, by the control on the conformal radii of these annuli, has the conformal type of $\mathbb{C}$.

Remark 7.4. *The Collin-Rosenberg theorem has been extended by J.A. Gálvez and H. Rosenberg [GaRo] to more general product spaces $M^2 \times \mathbb{R}$:* there exist entire minimal graphs of parabolic conformal type on $M^2 \times \mathbb{R}$, where (M^2, g) is any complete simply connected Riemannian surface with Gaussian curvature $K_M \leq c < 0$ (K_M not constant).

7.2. Minimal surfaces of finite total curvature in $\mathbb{H}^2 \times \mathbb{R}$.

One of the most studied families among minimal surfaces in $\mathbb{R}^3$ are the complete minimal surfaces of finite total curvature (FTC for short). A minimal surface Σ is said to have FTC if its Gaussian curvature K satisfies

$$\left| \int_\Sigma K \, dA \right| < \infty.$$

By classical theorems of Huber and Osserman, complete FTC minimal surfaces in $\mathbb{R}^3$ are conformally equivalent to a compact Riemann surface minus a finite number of points. Moreover, the Gauss map extends meromorphically to the punctures, and the total curvature of the surface is a multiple of -4π. A key point here is that the Gauss map of a minimal surface in $\mathbb{R}^3$ is conformal.

In $\mathbb{H}^2 \times \mathbb{R}$ there is no conformal Gauss map for minimal surfaces. Nonetheless, using the global theory of harmonic maps into $\mathbb{H}^2$, L. Hauswirth and H. Rosenberg [HaRo] were able to prove that a similar situation holds in $\mathbb{H}^2 \times \mathbb{R}$.

Theorem 7.5 (Hauswirth-Rosenberg). *Let X be a complete minimal immersion of Σ in $\mathbb{H}^2 \times \mathbb{R}$ with finite total curvature. Then*

1. *Σ is conformally equivalent to a Riemann surface punctured at a finite number of points, $\Sigma \equiv \overline{M}_g - \{p_1 ...., p_k\}$.*

2. *$Qdz^2 := h_z^2\, dz^2$ is holomorphic on M and extends meromorphically to each puncture. If we parameterize each puncture p_i by the exterior of a disk of radius r, and if $Q(z)dz^2 = z^{2m_i}(dz)^2$ at p_i, then $m_i \geqslant -1$.*

3. *The third coordinate u of the unit normal tends to zero uniformly at each puncture.*

4. *The total curvature is a multiple of 2π:*

$$\int_\Sigma K dA = 2\pi \left(2 - 2g - 2k - \sum_{i=1}^{k} m_i \right).$$

As a consequence, every end of a finite total curvature surface is uniformly asymptotic to a Scherk type graph described in Theorem 7.3.

Proof. The first step is to prove that locally around an end, Qdz^2 only has at most a finite number of zeroes. Then a Huber theorem and an argument of Osserman give that the ends are conformally a punctured disk, and Qdz^2 extends meromorphically to the puncture. The final part of the behavior of Qdz^2 follows from the fact that $Qdz^2 = h_z^2 dz^2$, where h is the height function of the surface.

To prove that u goes to 0 at the ends, take an annular neighborhood of an end where Qdz^2 does not vanish. Then reparameterize this annulus by $w = \int \sqrt{Q} dz$. The metric conformal factor in these coordinates satisfies a sinh-Gordon equation, and the Gaussian curvature monotonically decreases to zero. Then, estimates on the growth of solutions of the sinh-Gordon equation allows one to conclude that, at a finite total curvature end, the tangent plane becomes vertical and the metric becomes flat.

Finally, the expression for the total curvature follows from Gauss-Bonnet formula and the estimates for the sinh-Gordon equation obtained before. □

In [HaRo], the following question was also raised: *are there complete non simply connected minimal surfaces with FTC in $\mathbb{H}^2 \times \mathbb{R}$?* Notice that rotational catenoids have infinite total curvature. Actually, at that time, the only known complete FTC minimal surfaces were the Scherk type graphs.

This question was positively answered by J. Pyo [Pyo] and also, independently, by Rodríguez and Morabito [RoMo]. Pyo constructed a 1-parameter family of genus zero properly embedded minimal surfaces in $\mathbb{H}^2 \times \mathbb{R}$ with k

ends for $k \geq 2$, similar to the k-noids in $\mathbb{R}^3$ (although the first ones are embedded and the k-noids in $\mathbb{R}^3$ are not). They have total curvature $4\pi(1-k)$, and are asymptotic to vertical planes at infinity. These surfaces are obtained as the conjugate surfaces of minimal graphs over infinite geodesic triangles in $\mathbb{H}^2$ that are asymptotic to vertical planes at infinity.

Very shortly thereafter, M. Rodríguez and F. Morabito discovered independently a larger family of FTC minimal surfaces, containing the previous ones. It is a $(2k-2)$-parameter of properly embedded FTC minimal surfaces of genus zero with k ends, obtained as the limits of simply periodic minimal surfaces called *saddle towers*, that are invariant by a vertical translation of vector $(0,0,2l)$. Taking limits when $l \to \infty$, they obtain genus zero minimal surfaces with k ends and total curvature $4\pi(1-k)$ that are symmetric with respect to the reflection over the slice $\mathbb{H}^2 \times \{0\}$. The surfaces found by Pyo appear when the ends are placed in symmetric positions.

7.3. Open problems. In [Ha], L. Hauswirth constructed a family of Riemann type minimal surfaces in $\mathbb{H}^2 \times \mathbb{R}$ and $\mathbb{S}^2 \times \mathbb{R}$, characterized by the property of being foliated by curves of constant curvature. It is a conjecture by W. Meeks and H. Rosenberg that in $\mathbb{S}^2 \times \mathbb{R}$ they are the only properly embedded minimal annuli. An approach for solving this conjecture using integrable systems techniques has been recently developed by L. Hauswirth and M. Schmidt. Another natural problem is to obtain classification results for properly embedded minimal surfaces of finite total curvature and a given simple topology in $\mathbb{R}^3$.

Schoen and Yau proved there is no harmonic diffeomorphism from the disk to a complete surface of non-negative curvature. Can there be such a harmonic diffeomorphism onto a complete parabolic surface? This is a question by J.A. Gálvez.

References

[AbRo1] U. Abresch, H. Rosenberg, A Hopf differential for constant mean curvature surfaces in $\mathbb{S}^2 \times \mathbb{R}$ and $\mathbb{H}^2 \times \mathbb{R}$, *Acta Math.* **193** (2004), 141–174.

[AbRo2] U. Abresch, H. Rosenberg, Generalized Hopf differentials, *Mat. Contemp.* **28** (2005), 1–28.

[AEG1] J.A. Aledo, J.M.Espinar, J.A. Gálvez, Height estimates for surfaces with positive mean curvature in $\mathbb{M} \times \mathbb{R}$. *Illinois Journal of Math.,* **52** (2008), 203–211.

[Bon] F. Bonahon, Geometric structures on 3-manifolds. In Handbook of Geometric Topology, pages 93–164. North-Holland, Amsterdam, 2002.

[ChYa] S.Y. Cheng, S.T. Yau, Maximal spacelike hypersurfaces in the Lorentz-Minkowski spaces, *Ann. of Math.* **104** (1976), 407–419.

[CoRo] P. Collin, H. Rosenberg, Construction of harmonic diffeomorphisms and minimal graphs, *Ann. of Math.*, to appear (2007).

[Dan1] B. Daniel, Isometric immersions into 3-dimensional homogeneous manifolds, *Comment. Math. Helv.* **82** (2007), 87–131.

[Dan2] B. Daniel, The Gauss map of minimal surfaces in the Heisenberg group, preprint, 2006, arXiv:math/0606299.

[DFM] B. Daniel, I. Fernández, P. Mira, Surfaces of critical constant mean curvature. Work in progress.

[DaHa] B. Daniel, L. Hauswirth, Half-space theorem, embedded minimal annuli and minimal graphs in the Heisenberg group. *Proc. Lond. Math. Soc. (3)*, **98** no.2 (2009), 445–470.

[DHM] B. Daniel, L. Hauswirth, P. Mira, Constant mean curvature surfaces in homogeneous manifolds, preprint, 2009. Published preliminarly by the Korea Institute for Advanced Study.

[DaMi] B. Daniel, P. Mira, Existence and uniqueness of constant mean curvature spheres in Sol_3. Preprint, 2008, arXiv:0812.3059

[dCF] M.P. do Carmo, I. Fernández, Rotationally invariant CMC disks in product space, *Forum Math.* **21** (2009), 951–963.

[EGR] J.M. Espinar, J.A. Gálvez, H. Rosenberg, Complete surfaces with positive extrinsic curvature in product spaces, *Comment. Math. Helv.*, **84** (2009), 351–386.

[EsRo] J.M. Espinar, H. Rosenberg, Complete constant mean curvature surfaces in homogeneous spaces, *Comment. Math. Helv.*, to appear (2009).

[FeMi1] I. Fernández, P. Mira, Harmonic maps and constant mean curvature surfaces in $\mathbb{H}^2 \times \mathbb{R}$, *Amer. J. Math.* **129** (2007), 1145–1181.

[FeMi2] I. Fernández, P. Mira, A characterization of constant mean curvature surfaces in homogeneous 3-manifolds, *Diff. Geom. Appl.*, **25** (2007), 281–289.

[FeMi3] I. Fernández, P. Mira, Holomorphic quadratic differentials and the Bernstein problem in Heisenberg space. *Trans. Amer. Math. Soc.*, **361**, no 11, (2009), 5737–5752.

[FoWo] A.P. Fordy, J.C. Wood. Harmonic maps and integrable systems. Aspects of Mathematics, vol. E23, by Vieweg, Braunschweig/Wiesbaden, 1994.

[GMM] J.A. Gálvez, A. Martínez, P. Mira, The Bonnet problem for surfaces in homogeneous 3-manifolds, *Comm. Anal. Geom.* **16** (2008), 907–935.

[GaRo] J.A. Gálvez, H. Rosenberg, Minimal surfaces and harmonic diffeomorphisms from the complex plane onto a Hadamard surface. Preprint, 2008, arXiv:0807.0997.

[Ha] L. Hauswirth, Minimal surfaces of Riemann type in three dimensional product manifolds. *Pacific J. Math.*, **224**, no.1 (2006), 91–117.

[HaRo] L. Hauswrith, H. Rosenberg. Minimal surfaces of finite total curvature in $\mathbb{H} \times \mathbb{R}$. *Mat. Contemp.* **31** (2006), 65–80.

[HRS] L. Hauswirth, H. Rosenberg, J. Spruck. On complete mean curvature $\frac{1}{2}$ surfaces in $\mathbb{H}^2 \times \mathbb{R}$. *Comm. Anal. Geom.*, **16**, no.5 (2008), 989–1005.

[HoMe] D. Hoffman, W. H. Meeks III. The strong halfspace theorem for minimal surfaces. *Invent. Math.* **101**, no.2 (1990), 373–377.

[HsHs] W.Y. Hsiang, W.T. Hsiang, On the uniqueness of isoperimetric solutions and imbedded soap bubbles in noncompact symmetric spaces I, *Invent. Math.* **98** (1989), 39–58.

[Lee] H. Lee. Extension of the duality between minimal surfaces and maximal surfaces. Preprint, 2009.

[MaPR] M. Manzano, J. Pérez, M. Rodríguez. Parabolic stable surfaces with constant mean curvature. Preprint, 2009, arXiv:0910.5373.

[MPR] W. H. Meeks III, J. Pérez, A. Ros. Stable constant mean curvature surfaces, *Handbook of Geometric Analysis* no.1 (2008).

[Mee] W.H. Meeks. Constant mean curvature surfaces in homogeneous 3-manifolds. Preprint (2009).

[MeRo1] W.H. Meeks, H. Rosenberg, The theory of minimal surfaces in $M \times \mathbb{R}$, *Comment. Math. Helv.* **80** (2005), 811–858.

[MeRo2] W.H. Meeks, H. Rosenberg, Stable minimal surfaces in $M \times \mathbb{R}$, *J. Differential Geom.* **68** (2004), 515–534.

[MoUr] S. Montiel, F. Urbano, A Willmore functional for compact surfaces in the complex projective plane, *J. Reine Angew. Math.* **546** (2002), 139–154.

[NeRo1] B. Nelli, H. Rosenberg. Minimal surfaces in $\mathbb{H}^2 \times \mathbb{R}$. *Bull. Braz. Math. Soc.*, **33**, no.2 (2002), 263–292.

[NeRo2] B. Nelli, H. Rosenberg. Global properties of constant mean curvature surfaces in $\mathbb{H}^2 \times \mathbb{R}$. *Pacific J. Math.* **226**, no-1 (2006), 137–152.

[Oss] R. Osserman. A survey on minimal surfaces. Dover Publications Inc., New York, second edition, 1986.

[Ped] R. Pedrosa. The isoperimetric problem in spherical cylinders, *Ann. Global Anal. Geom.* **26** (2004), 333–354.

[Pyo] J. Pyo. New examples of minimal surfaces in $\mathbb{H}^2 \times \mathbb{R}$. Preprint, 2009, arXiv:0911.5577.

[RoMo] M. Rodríguez, F. Morabito. Saddle towers in $\mathbb{H}^2 \times \mathbb{R}$. Preprint, 2009, arXiv:0910.5676.

[Ros] H. Rosenberg, Minimal surfaces in $M^2 \times \mathbb{R}$, *Illinois J. Math.* **46** (2002), 1177–1195.

[SaE] R. Sa Earp, Parabolic and hyperbolic screw motion surfaces in $\mathbb{H}^2 \times \mathbb{R}$, *J. Austr. Math. Soc.*, **85** (2008), 113–143.

[SaTo] R. Sa Earp, E. Toubiana, Screw motion surfaces in $\mathbb{H}^2 \times \mathbb{R}$ and $\mathbb{S}^2 \times \mathbb{R}$, *Illinois J. Math.* **49** (2005), 1323–1362.

[Sco] P. Scott. The geometries of 3-manifolds, *Bull. London Math. Soc.* **15** (1983), 401–487.

[Tor] F. Torralbo. Rotationally invariant constant mean curvature surfaces in homogeneous 3-manifolds. *Diff. Geom. Appl.*, to appear (2009), arXiv:0911.5128.

[ToUr] F. Torralbo, F. Urbano. Compact stable constant mean curvature surfaces in the Berger 3-spheres. Preprint, 2009, arXiv:0906.1439.

[Wan] T.Y. Wan, Constant mean curvature surface harmonic map and universal Teichmuller space, *J. Differential Geom.* **35** (1992), 643–657.

[WaAu] T.Y. Wan, T.K. Au, Parabolic constant mean curvature spacelike surfaces, *Proc. Amer. Math. Soc.* **120** (1994), 559–564.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Morse Landscapes of Riemannian Functionals and Related Problems

Alexander Nabutovsky*

Abstract

The subject of this talk is Morse landscapes of natural functionals on infinite-dimensional moduli spaces appearing in Riemannian geometry.

First, we explain how recursion theory can be used to demonstrate that for many natural functionals on spaces of Riemannian structures, spaces of submanifolds, etc., their Morse landscapes are always more complicated than what follows from purely topological reasons. These Morse landscapes exhibit non-trivial "deep" local minima, cycles in sublevel sets that become nullhomologous only in sublevel sets corresponding to a much higher value of functional, etc.

Our second topic is Morse landscapes of the length functional on loop spaces. Here the main conclusion (obtained jointly with Regina Rotman) is that these Morse landscapes can be much more complicated than what follows from topological considerations only if the length functional has "many" "deep" local minima, and the values of the length at these local minima are not "very large".

Mathematics Subject Classification (2010). Primary 53C23, 58E11, 53C20; Secondary 03D80, 68Q30, 53C40, 58E05.

Keywords. Non-computability, geometric calculus of variations, best Riemannian metrics, algorithmic unsolvability, quantitative topology, Riemannian functionals, the length functional, thick knots, curvature-pinching, loop spaces.

1. Introduction

In this talk we will discuss Morse landscapes of functionals on infinite-dimensional moduli spaces naturally arising in Differential Geometry.

Our first message is that in many cases these Morse landscapes are much more complicated than what would follow just from the Morse theory. The

*The author gratefully acknowledges a partial support from his NSERC Discovery Grant.
Department of Mathematics, 40 St. George st., University of Toronto, Toronto, ON, M5S2E4, Canada. E-mail: alex@math.toronto.edu.

examples include some Riemannian functionals (i.e. functionals on the space of isometry classes of Riemannian metrics), functionals on spaces of submanifolds, and so on. Our approach initiated in [N1], [N2], [N3] and further developed in collaboration with Shmuel Weinberger ([NW1], [NW2], [NW3]) is based on recursion theory. In many cases we are able to prove disconnectedness of sublevel sets of a functional of interest, and, moreover, the exponential growth of the number of connected components that merge only inside a much larger sublevel set. Sometimes this technique implies the existence of an infinite set of distinct local minima of a functional of interest, where only the existence of the global minimum was previously known. In other cases this recursion-theoretic approach is the only known method that can be used to establish the existence of critical points of a functional of interest.

In particular, methods using ideas from mathematical logic led to only known general results on the following problem posed by R. Thom "What is the best (or the nicest) metric on a given smooth manifold?" for compact manifolds of dimension ≥ 5 (joint work with Shmuel Weinberger). They also constitute the only known tool to demonstrate that the theory of (high-dimensional) "thick" knots is drastically different from the "usual" knot theory.

In a different direction I will discuss Morse landscapes of the length functional on loop spaces $\Omega_p M^n$ and spaces $\Omega_{pq} M^n$ of paths between points p, q on a closed simply-connected Riemannian manifold M^n. In [N5] I proved that if the length functional has a "very deep" non-trivial local minimum on $\Omega_p M^n$, then it has "many" "deep" local minima. [0] The proof used the idea of "effective universal coverings". A stronger form of this result can be proven using direct geometric methods recently invented by Regina Rotman. These methods also can be used to demonstrate that if the length functional has a critical point of a positive index of a "large" but finite depth, then it must have "many" "deep" local minima ([NR]).

2. "Thick" Knots

Knots are sometimes defined as submanifolds of R^3 (or S^3) diffeomorphic to S^1. More generally, one can consider higher-dimensional knots that are submanifolds of R^{n+k} (or S^{n+k}) diffeomorphic to S^n, where k is usually equal to two. Two knots have the same knot type, if they are isotopic. It makes sense to consider also "physical" or "thick" knots on a rope of small but non-zero thickness; two thick knots have the same type if they can be connected by an isotopy that preserves the thickness of the rope and does not increase its length. In the multidimensional case the isotopy should not increase the volume. Note

[0] The depth of a local minimum μ of a functional f can be defined as $\inf_\gamma \sup_t f(\gamma(t)) - f(\mu)$, where the infimum is taken over the set of all paths γ starting at μ and ending at a point $\gamma(1)$ such that $f(\gamma(1)) < f(\mu)$. If μ is a global minimum, then, by definition, it has infinite depth.

that the thickness of the rope cannot exceed the injectivity radius of the normal exponential map. Therefore, the study of "thick" knot types is equivalent to the study of connected components of sublevel sets of the crumpledness functional $\kappa_v = \frac{vol^{\frac{1}{n}}}{r}$, where vol denotes the volume, and r denotes the injectivity radius of the normal exponential map. (In other words, r is equal to the supremum of x such that every two normals to the knot of length $\leq x$ starting at its different points do not intersect. Informally speaking, $r(\Sigma)$ is the largest radius of a nonself-intersecting tube arounf Σ.)

The paper [N1] was one of the first papers on "thick" knots. [1] The most basic question (in every dimension and codimension) is whether or not there exist "thick" knots that are trivial as usual knots but not trivial as "thick" knots. This question still remains open for the "classical" dimension one and codimension two, despite the fact that it is easy to sketch plausible candidates. The question becomes especially interesting for $n = 2$ and codimension one, where I cannot even guess what answer to expect. In [N1] I answered this question in affirmative for the dimension $n \geq 5$ and codimension one. Observe, that Smale's h-cobordism theorem implies that every embedded n-sphere of codimension one in R^{n+1}, $(n > 3)$, is isotopic to the standard sphere. In other words, there exists only the trivial knot type. Yet in [N1] (see also [N2]) I proved that:

Theorem 2.1. *For every $n \geq 5$ and for each sufficiently large x the set of n-dimensional hypersurfaces $\Sigma^n \subset R^{n+1}$ diffeomorphic to S^n and such that $\kappa_v(\Sigma^n) \leq x$ is not connected.*

For every knot $\Sigma^n \subset R^{n+1}$ denote the infimum of y such that there exists an isotopy that passes through knots with $\kappa_v(\Sigma^n) \leq y$ and connects Σ^n with the round sphere of radius one by $C(\Sigma^n)$. An easy compactness argument implies that for every positive x there exists the supremum of $C(\Sigma^n)$ over the set of all knots Σ^n such that $\kappa_v(\Sigma^n) \leq x$. Denote this supremum by $C_n(x)$. Note that the previous theorem follows from the assertion that for every $n \geq 5$ for all sufficiently large x $C_n(x) > x$. We deduced this assertion (and, thus, the previous theorem) from the following much stronger assertion: [2]

[1] I am sure that the very natural idea to study knots of non-zero thickness occured independently to many other mathematicians, yet I found only one paper on "thick" knots preceding [N1], namely, [KV]. Although [N1] dealt mainly with high-dimensional "thick" knots of codimension one, some of its results, such as a $C^{1,1}$-compactness theorem remain valid for an arbitrary dimension/codimension. It also contained several basic problems about 1-dimensional "thick" knots in R^3 (Section 4, B,C,D in [N1]) that still remain unsolved. In recent years "thick" 1-dimensional knots in R^3 became the subject of a constantly growing number of publications -cf. [LSDR], [Dur] or [CFKSW].

[2] To be more precise in [N1] we proved that this inequality holds only for an infinite unbounded sequence of values of x. To prove that it holds for all sufficiently large values of x one needs either to apply a trick from [N2] involving the busy beaver function or to use time-bounded Kolmogorov complexity as in [N3] and subsequent papers [NW1], [NW2], [NW3].

Theorem 2.2. *Let ϕ be any computable* [3] *function. Then for every $n \geq 5$ and for all sufficiently large x $C_n(x) > \phi(\lfloor x \rfloor))$.*

We will explain the proof of this theorem in the next section. The preceding discussion does not depend on a particular choice of the smoothness of considered knots as long as the considered knots are at least $C^{1,1}$-smooth, which is the minimal smoothness required for r to be defined and positive. Now consider the space of all $C^{1,1}$-smooth n-dimensional knots in R^{n+1} with C^1 topology. Consider the following equivalence relation on this space: two knots are equivalent if they can be transformed one into the other by a similarity transformation of the ambient Euclidean space. Clearly each equivalence class is connected, and the crumpledness functional is constant on every equivalence class. Denote the space of the equivalence classes by $Knots_{n,1}$. In [N1] we proved that sublevel sets of κ_v are compact subsets of $Knots_{n,1}$. Therefore κ_v attains its local minimum on every connected component of each of its sublevel sets. These local minima will be automatically local minima of κ_v on the whole space $Knots_{n,1}$. The disconnectedness of $\kappa_v^{-1}((0,x])$ for arbitrarily large values of x implies that the set of local minima of κ_v is unbounded, and the set of values of κ_v at its local minima is infinite. Combining this observation with the previous theorem we see that there exists an infinite set of local minima of κ_v, where the depth is much higher than the value of κ_v.

Theorem 2.3. *For every $n \geq 5$ and every computable function $\phi : N \longrightarrow N$ there exists an infinite sequence Σ_i^n of local minima of κ_v on $Knots_{n,1}$ such that the set of values of κ_v at these local minima is unbounded, and the depth of each of these local minima Σ_i^n is greater than $\phi(\lfloor \kappa_v(\Sigma_i^n) \rfloor)$.*

This theorem holds for other versions of the crumpledness functional , e.g. $\kappa_d = \frac{diam}{r}$ as well as for many other functionals (see [N1]). The theorem can be also generalized to spaces of trivial knots of arbitrary codimension (of dimension $n > 4$), as well as for the spaces of trivial knots of dimension 3 or 4 and codimension 2. (The last fact follows from the results of [NW0].) The theorem and its generalizations for other crumpledness functionals obviously hold for the space of trivial $C^{1,1}$-smooth n-dimensional knots in R^{n+k}, where one does not take the quotient with respect to the action of the group of similarities of the ambient Euclidean space. In this form the theorem can be generalized for the cases, when 1) The submanifold can be diffeomorphic to an arbitrary closed manifold M^n , $(n > 4)$, instead of S^n; and 2) The ambient manifold can be not R^n but an arbitrary closed Riemannian manifold, as well as a complete non-compact Riemannian manifold from a wide class. (Of course, the considered

[3]Formally speaking, here "computable" means "Turing computable" or, equivalently, "recursive". Equivalently, a reader can take any computer programming language, strip it of all restrictions on the size of data (if there are any), and strip it of all data types but the integer numbers. A function is computable if and only if it can be described by a computer program in this language.

space of submanifolds needs to be non-empty; if it is not connected, the theorem holds for each of its connected components.)

3. Methods I: Algorithmic Unsolvability of the Diffeomorphism Problem and its Applications

The following theorem was first proven by Sergei Novikov (see its proof in the Appendix of [N4]):

Theorem 3.1. *For every $n \geq 5$ there is no algorithm deciding whether or not a given manifold M^n is diffeomorphic to the n-sphere.*

To make this theorem precise one needs to explain how M^n is presented in a finite form. In [N4] we observed that this theorem is true even in the case when M^n is a non-singular real algebraic hypersurface $\{x \in R^{n+1} | p(x) = 0\}$, where p is a polynomial with rational coefficients. In this case M^n can be presented by the vector of coefficients of p. The other ways to present M^n in a finite form include: 1) C^∞-semialgebraic atlases (also known as Nash atlases; see [BHP]); 2) Smooth triangulations; 3) Smooth real algebraic subvarieties of Euclidean spaces of a higher codimension defined over the field of algebraic numbers.

Here is a very brief sketch of the proof of this theorem. According to the classical theorem independently proven by S. Adyan and M. Rabin there exists an infinite sequence of finite presentations of groups $G_i, i = 1, 2, \ldots$ such that there is no algorithm deciding for every given i whether or not G_i is isomorphic to the trivial group. (In other words, the set I of all i such that G_i is trivial is non-recursive.) The standard proof of this theorem (cf. [Mil]) produces G_i that are perfect, that is $H_1(G_i) = G_i/[G_i, G_i]$ is trivial. S. Novikov observed that one can alter finite presentations of these groups in a certain explicit way to obtain a new sequence of finite presentations of *superperfect* groups $\bar{G}_i$ so that $\bar{G}_i$ is trivial if and only if G_i is trivial. (Thus, there is no algorithm deciding for a given value of i whether or not $\bar{G}_i$ is trivial. Superperfectness of a group means the vanishing of the first two homology groups of a group. Also, note that groups $\bar{G}_i$ are universal central extensions of groups G_i.) According to [Ke] the superperfectness of a finitely presented group G is the necessary and sufficient condition of the realizibility of G as the fundamental group of a smooth homology n-sphere, that is a smooth closed manifold with the same homology groups as S^n, for every $n \geq 5$. Thus, we can effectively realize groups $\bar{G}_i$ as fundamental groups of homology spheres Σ_i^n. Moreover, the proof of the quoted result from [Ke] implies that this construction can be carried in R^{n+1} so that Σ_i^n will be a smooth hypersurface in R^{n+1}. Smale's h-cobrdism theorem implies that a homology sphere Σ_i^n embedded as a hypersurface in R^{n+1} is diffeomorphic to S^n if and only if it is simply-connected, and, therefore, if and only if $\bar{G}_i$ is trivial. This completes the proof of the theorem.

This theorem (or rather its proof outlined above) has the following immediate corollary:

Corollary 3.2. *For every closed smooth manifold M_0^n of dimension $n > 4$ there is no algorithm that decides whether or not a given manifold M^n is diffeomorphic to M_0^n.*

Indeed, we can just construct a sequence M_i^n by forming connected sums of a copy of M_0^n with smooth homology spheres Σ_i^n from the outline of a proof of the previous theorem. The manifold M_i^n is diffeomorphic to M_0^n if and only if the fundamental group of Σ_i^n is trivial.

Note that it is not known whether or not this theorem remains true in dimension four. However, A. Markov proved that this theorem is true for manifolds M_0^4 diffeomorphic to the connected sum of a sufficient number N_0 of copies of $S^2 \times S^2$ with an arbitrary closed 4-manifoldi (cf. [BHP], [Sh]). Here one can take $N_0 = 14$([Sh]). This theorem enables us to extend some of our techniques that we are going to describe below to such four-dimensional manifolds.

Now the general idea behind the proof of Theorem 2.2 as well as of some results stated in the next sections can be described as follows. Consider a class C of diffeomorphism types of compact smooth n-dimensional manifold, where $n > 4$. The class C can be the class of all n-manifolds, or, for example, the class of all manifolds embeddable in R^{n+1}. We require that the class C is large enough to ensure that S. Novikov's theorem will be true in this class: For every manifold M^n from C there is no algorithm deciding whether or not a given manifold from C is diffeomorphic to M^n.

We consider situations, when for every manifold in $M^n \in C$ there is a natural "moduli space" $Moduli(M^n)$. associated with this manifold. (In the situation of Theorem 2.2 $Moduli(M^n) = Knots_{n,1}$. To prove theorems stated in section 5 below we will be choosing $Moduli(M^n)$ as certain subsets of the space of Riemannian structures on M^n.) Let ϕ be a non-negative functional on a moduli space $Moduli_n$ defined as the disjoint union of connected spaces $Moduli(M^n)$ associated with all manifolds $M^n \in C$. We are assuming that $Moduli_n$ is endowed with a metric, ρ. First, we are going to make the following assumptions about ϕ, ρ and the class C:

0) There exists a countable dense set $D \in Moduli_n$. Elements of D are representable in a finite form. For any $M^n \in C$ there is no algorithm deciding whether or not a given element $\mu \in D$ is in $Moduli(M^n)$ (that is, represents M^n).

1) There exists an algorithm computing the distance ρ between every pair of elements of D within to any prescribed (rational) accuracy.

2) The function ϕ can be effectively majorized: There exists an algorithm that for a given element $\mu \in D$ computes an upper bound for $\phi(\mu)$.

3) For every x the sublevel set $\phi^{-1}([0, x]) \subset Moduli_n$ is precompact. Moreover, there exists an algorithm that for every given positive rational x

and ϵ constructs a finite ϵ-net in $\phi^{-1}([0,x])$. All elements of this ϵ-net are in the countable set D.

4) There exists a computable decreasing positive function $\delta_n(x)$ such that every two $\delta_n(x)$-close points from $\phi^{-1}([0,x])$ are points from $Moduli(M^n)$ for the same manifold M^n.

Now we are going to demonstrate that for every $M^n \in C$ there exists an unbounded increasing sequence of values of x such that sublevel sets $S_x = \phi^{-1}([0,x]) \bigcap Moduli(M^n)$ are disconnected. Moreover, for these values of x S_x is a union of two non-empty subsets S_{1x}, S_{2x} such that the distance between each pair of points $\mu_1 \in S_{1x}$, $\mu_2 \in S_{2x}$ is at least $\delta_n(x)$.

Indeed, assume the opposite. Then we will construct an algorithm deciding whether or not a given manifold $N^n \in C$ is diffeomorphic to M^n, thus obtaining a contradiction with our assumptions. We start from calculating an upper bound y for the value of ϕ at the given manifold (which is presented as an element from D). We can always make it large enough to ensure that $\phi^{-1}([0,y]) \bigcap Moduli(M^n)$ is connected. Then we construct $\delta_n(y)/10$-net in $\phi^{-1}([0,y]) \subset Moduli_n$. The next step is to construct a graph such that the points of the constructed net will be its vertices, and two vertices are connected by an edge, if the corresponding points are approximately $\delta_n(y)/2$-close. Here we allow ourselves an error in these calculations that does not exceed $\delta_n(y)/4$. Now our connectedness assumption implies that exactly one component of the constructed graph contains elements of $Moduli(M^n)$. We can assume that our algorithm knows one vertex v_0 from this connected component. (This vertex will be in this connected component for all sufficiently large values of y). Now our algorithm needs to determine a vertex w of the net which is $\delta_n(y)/10$-close to the given element of $Moduli_n$, and to determine whether or not w and v_0 are in the same component of the constructed graph. The given element of $Moduli_n$ represents a manifold diffeomorphic to M^n if and only if w and v_0 are in the same component.

The obtained contradiction demonstrates that the sets S_x must be disconnected for some arbitrarily large values of x. An argument from [N2] (that involves Rado's busy beaver function) can be used to prove this assertion for all sufficiently large values of x. Yet there is another method using the notion of Kolmogorov complexity that can be used to prove not only the disconnectedness of sets S_x but lower bounds for the number of their connected components. This idea will be described in more details in the next section.

If sublevel sets of ϕ are not only precompact, but compact, then we can find distinct local minima of ϕ at the bottom of different connected components of its sublevel sets. In some applications of this method sublevel sets of ϕ are compact, when the manifold belongs to a class $C' \subset C$ but not necessarily in the general case. Then we establish the compactness of *some* of the connected components of sublevel sets of ϕ by using the fact that there is no algorithm that distinguishes a manifold $M^n \in C$ of interest for us from manifolds known

to be in the subclass C'. (Of course, this fact should be true for this idea to work.)

To prove the theorems stated in the previous section one uses this idea in the situation when $Moduli_n$ is the space of equivalence classes of codimension one closed submanifolds of R^{n+1}. (Two submanifolds are equivalent if they can be transformed one into the other by a similarity of R^{n+1}.) Further, C is the class of closed n-manifolds embeddable into R^{n+1}, $M^n = S^n$, and $\phi = \kappa_v$ or κ_d.

However, note that in most of the situations, when we would like to apply this method, the assumptions 3) or 4) either do not hold, or are difficult to establish. Nevertheless, the method sometimes can be salvaged using new ideas some of which will be explained in sections 6 and 7.

4. Methods II: Kolmogorov Complexity and Time-bounded Kolmogorov Complexity

In this section we will explain how to modify the method sketched in the previous section so as to obtain not merely disconnectedness of sublevel sets $\phi^{-1}([0,x])$ of a functional of interest on $Moduli(M^n)$, but a lower bound for the number of connected components of these sets.

A *decision problem* consists of a countable set A and its subset B. Elements of A are presentable in a finite form, and there is a computable complexity function $A \longrightarrow Z^+$. For each L there exist only finitely many elements of A of complexity $\leq L$. One is interested in existence/non-existence of an algorithm deciding whether or not a given element of A is an element of B. Assume that there is no such algorithm. Then one can ask for such an algorithm that uses arbitrary *oracle information*. The amount of oracle information is allowed to grow with the complexity of instances of the problem. We are assuming that the information is presented as a sequence of 0s and 1s. The "amount of information" is just the length of this sequence. Of course, one can ask for the list of all answers for all instances of the problem of complexity $\leq L$. Yet one is interested in the *minimal* amount of oracle information sufficient to solve the problem. The minimal number of bits of oracle information sufficient to solve the problem for all instances of complexity $\leq L$ is called *Kolmogorov complexity* of the decision problem. Of course, one can "hide" a constant number of bits of oracle information in the algorithm, so the Kolmogorov complexity is a function of L defined only up to adding a constant summand. For example, let G be a finitely presented group with unsolvable word problem, A the set of all words in the considered finite presentation, B the set of all words representing trivial elements, and assume that we define the complexity of words as their length. The resulting decision problem is the word problem for G; it can be solved in a computable time using the following oracle information: For each L we request (the binary representation of) the number $w(L)$ of all trivial words

with $\leq L$ letters. To use this information we start generating trivial words of length $\leq L$ using longer and longer products of conjugates of relations, and stop when the length of the list reaches $w(L)$. One can be sure that all the remaining words correspond to non-trivial elements of G. So, the Kolmogorov complexity of the word problem for words of length $\leq L$ grows not faster than a linear function of L. However, it is not difficult to note that the time of work of this "algorithm" grows faster than any computable function. Assume that we impose an additional constraint: the time of work of the algorithm that uses the oracle information should not exceed a given computable function λ. The resulting notion is called time-bounded Kolomogorov complexity of the considered decision problem (cf. [LV] for an introduction to its properties). A theorem of Barzdin ([B]) can be used to show that, in general, one cannot now do much better than to ask for the list of all answers for the word problem: There exists a finitely presented group G such that for every computable λ the time-bounded Kolmogorov complexity of the word problem is not less than $\frac{Const^L}{c(\lambda)} - const$ for some $Const > 1$, $c(\lambda) > 0$. In [N3] we prove that for every closed smooth manifold M_0^n of dimension $n > 4$ and computable time λ the time-bounded Kolomogorov complexity of the decision problem "Is a given smooth manifold diffeomorphic to M_0^n?" is also not less than $\frac{Const(n)^L}{c(\lambda)} - const$ for some universal $Const(n) > 1$. Here the complexity L can be, for example, the number of simplices in a smooth triangulation of the given manifold. To relate this result to geometry of sublevel sets of ϕ note that the mentioned diffeomorphism problem can be solved using a set of representatives from every connected component of $\phi^{-1}([0,x]) \bigcap Moduli(M_0^n)$ as the oracle information (see the previous section for the notations). Indeed, the diffeomorphism problem can be restated as the decision problem of recognizing whether or not a given element $\mu \in Moduli_n$ is in $Moduli(M_0^n)$; the oracle information enables one to solve the diffeomorphism problem for all $\mu \in Moduli_n$ such that $\phi(\mu) \leq x$. For this purpose one just needs to check whether or not an approximation to μ can be connected with one of the elements provided by the oracle by a finite sequence of sufficiently short "jumps" in $Moduli_n$. Now our lower bound for the time-bounded Kolmogorov complexity can be used to produce a lower bound for the number of the connected components of $\phi^{-1}([0,x]) \bigcap Moduli(M^n)$. In many interesting cases this lower bound is at least exponential in x.

5. Disconnectedness of Sublevel Sets of Riemannian Functionals

In this section M^n denotes a closed Riemannian manifold of dimension $n \geq 5$. Consider the space of Riemannian structures $Riem(M^n)$ (=isometry classes of Riemannain metrics) on M^n endowed with the Gromov-Hausdorff metric. In this section we will consider geometry of sublevel sets of various Riemannian

functionals on this space. Our goals are to prove that their sublevel sets are disconnected with a growing number of connected components, and, when possible, to prove the existence of infinitely many locally minimal values.

The first result of this kind was proven in [N2]: Let $I_{M^n}(\epsilon)$ denote the space of Riemannian structures on M^n of volume equal to one and injectivity radius $\geq \epsilon$.

Theorem 5.1. *If $n \geq 5$, then for all sufficiently small ϵ $I_{M^n}(\epsilon)$ is not connected. Moreover, there exist two non-empty subsets of $I_{M^n}(\epsilon)$ such that the Gromov-Hausdroff distance between each point of one of these sets and each point of the other is at least $\epsilon/9$.*

In fact, one can use the notion of time-bounded Kolmogorov complexity as described in the previous section to show that there exist $\sim \frac{1}{\epsilon^n}$ non-empty subsets of $I_{M^n}(\epsilon)$ such that the distance between each pair of points in different subsets is at least $\epsilon/9$. Moreover, assume that one would like to connect a point in one of these subsets to a point in the other by a path in $I_{M^n}(\delta)$ for some positive $\delta < \epsilon$. It is not difficult to prove using a precompactness argument that some such $\delta = \delta_{M^n}(\epsilon)$ must exist. Yet $\frac{1}{\delta_{M^n}(\epsilon)}$ grows faster than any computable function of $\lfloor \frac{1}{\epsilon} \rfloor$.

Studies of variational problems for Riemannian functionals are motivated by the following problem posed by R. Thom (cf. [Be] , p. 499): "What is the best Riemannian structure on a given compact manifold?". (This question also appears in a well-known list of unsolved problems in Differential Geometry composed by S.T. Yau ([Y]).) A possible idea here is to choose a natural Riemannian functional and to look for its minima (or local minima) on the set of Riemannian structures on a given closed manifold M^n. However, for $n \geq 5$ (and probably $n = 4$) there is no really good notion of the "best" Riemannian structures on all n-dimensional manifolds ([N4]): Assume that for every M^n there exists a non-empty subset $Best(M^n) \subset Riem(M^n)$. Also assume that there exists an algorithm recognizing when a given Riemannian metric is very close to one of the best Riemannian metrics. (This assumption is required to eliminate the following "solution" of the problem: Use the axiom of choice to choose one Riemannian structure on every M^n.) Then for every M^n $Best(M^n)$ is an infinite set.

Indeed, assume the opposite. Then there exists the following algorithm deciding whether or not a given n-dimensional manifold is diffeomorphic to M^n yielding a contradiction with S.P. Novikov theorem: Start from any Riemannian metric on a given manifold. Do a trial and error search until we find a Riemannian metric close to one of the best Riemannian metrics on the considered manifold. As we assumed that the set of the best Riemannian metrics on M^n is finite, we can assume that the algorithm "knows" them all (or, more precisely, it knows a sufficiently close approximation to each of them). Now we can check if the found approximation to a best metric is sufficiently close

to one of the known best Riemannian metrics on M^n. The given manifold is diffeomorphic to M^n if and only if the answer is positive.

This argument strongly suggests that if a Riemannian functional ϕ has local minima on $Riem(M^n)$ for every M^n, and the set of its local minima is locally compact, then for every M^n the set of local minima of ϕ must have an infinite set of connected components. Thus, the following result obtained by the author and Shmuel Weinberger seems to provide a reasonably good solution of the problem posed by R. Thom. The naive idea is that one can try to define the best Riemannian metrics by fixing a scale (i.e. the diameter or volume) and looking for (local) minima of a curvature functional, for example, $\sup|K|$, where K denotes the sectional curvatuture. Equivalently, one can consider Riemannian metrics with $\sup|K| \leq 1$ and to look for local minima of the diameter. More formally, let $Al(M^n)$ denote the Gromov-Hausdorff closure of the subset of $Riem(M^n)$ formed by all Riemannian structures satisfying $\sup|K| \leq 1$ in the space of all metric spaces homeomorphic to M^n. The elements of this space are Alexandrov structures on M^n with curvature bounded above and below. They have virtually the same nice analytic and geometric properties as smooth Riemannian manifolds with sectional curvature between -1 and 1 (see [BN]). In particular, they are $C^{1,\alpha}$-smooth Riemannian manifolds for each $\alpha < 1$. For each element of $Al(M^n)$ its sectional curvature is defined at almost all points, and the absolute value of the sectional curvature does not exceed 1. It is well-known that sublevel sets of the diameter d regarded as a functional on $\bigcup_{M^n} Al(M^n)$ are precompact. However, there exist manifolds M^n such that $Al(M^n)$ is complete, and, therefore, sublevel sets of d on $Al(M^n)$ are compact, as well as manifolds M^n such that sublevel sets of d on $Al(M^n)$ are not compact. For example, tori T^n admit flat metrics with arbitrarily small diameter, so that $\inf_{Al(T^n)} d = 0$ for every n. Therefore, even the existence part in the following theorem proven by the author and Shmuel Weinberger is non-obvious:

Theorem 5.2. *([NW1]) For every closed manifold M^n of dimension $n > 4$ the set of locally minimal values of d on $Al(M^n)$ is an unbounded set.*

In particular, this theorem implies that the set of locally minimal values of d on $Al(M^n)$ is infinite. However, it is not difficult to see that the set of its locally minimal values is countable. We also proved many additional results about distribution of local minima of d on $Al(M^n)$ and geometry of connected components of sublevel sets $d^{-1}((0,x])$ of $d : Al(M^n) \longrightarrow (0,\infty)$. For example, we proved that the assertion of Theorem 5.2 will remain true for the values of d at its "very deep" local minima. Here one can define "very deep" local minima by first choosing a (preferably rapidly growing) strictly increasing *computable* function $\phi : N \longrightarrow N$ and postulating that a local minimum μ of d is "very deep" if there is no path $\gamma : [0,1] \longrightarrow Al(M^n)$ starting at μ such that $d(\gamma(1)) < d(\gamma(0)) = d(\mu)$, and $d(\gamma(t)) \leq \phi(\lfloor d(\mu) \rfloor)$ for each $t \in [0,1]$. Moreover, the result remains true if one considers only those "very deep" local minima of d, where the value of the volume is not less than 1 (or any other fixed value). Furthermore, we

proved that the number of these "very deep" local minima of d on $Al(M^n)$, such that the value of d does exceed x, grows at least exponentially with x^n. Later Shmuel Weinberger observed that this distribution function for the number of very deep local minima has even a doubly exponential lower bound ([We]). (To explain the last observation note that the volume of manifolds with $|K| \leq 1$ and $diam \leq x$ can be as large as $\exp(c(n)x)$. Thus, one can "fit" an exponential number of nonintersecting metric balls of radius ~ 1 and volume ~ 1 inside such a manifold. Therefore, one can reduce the halting problem for a universal Turing machine with inputs of lengths up to $\exp(const(n)x)$ to a certain version of the diffeomorphism problem relevant here and explaned in the next section. This version of diffeomorphism problem involves only Riemannian manifolds with $|K| \leq 1$ and $diam \leq x$. The time-bounded Kolmogorov complexity of the halting problem grows exponentially with the length of the inputs, and the number of the local minima grows at least as the time-bounded Kolmogorov complexity, as it was explained in the previous section.)

We refer the reader to our paper [NW2] for further results about depths of the local minima of d on $Al(M^n)$ and the distribution of local minima of different depths.

6. Methods III: Simplicial Norm, Homology Surgery, Arithmetic Groups

Our proof of Theorem 5.2 follows the scheme outlined in section 3 but contains several new ideas. We start from recalling a classical result of Gromov ([G1], [Gr]) that if a closed Riemannian manifold has a positive simplicial volume, then a lower bound for the Ricci curvature implies a positive lower bound for the volume. More precisely, if $Ric \geq -(n-1)$, then $vol(M^n) \geq c(n)\|M^n\|$, where $\|M^n\|$ denotes the simplicial volume, and $c(n)$ is an explicit constant depending only on the dimension. Simplicial volume is a homotopy invariant of manifolds (see [G1] for its definition and basic properties.) It depends only on the fundamental group of the manifold M^n and the image $\phi_*([M^n])$ of the fundamental homology class of M^n under the homomorphism induced by the classifying map $\phi : M^n \longrightarrow K(\pi_1(M^n), 1)$. As the isomorphism problem for groups is algorithmically unsolvable, the following theorem proven in [NW1] is not especially surprising:

Theorem 6.1. *Let M^n be a closed manifold of dimension $n > 4$. There is no algorithm that decides whether or not a given manifold N^n is diffeomorphic to M^n even if it is a priori known that, if N^n is not diffeomorphic to M^n, then it has a simplicial volume greater than* 1.

Here one can replace 1 by any constant, if desired. Thus, having a large simplicial volume is not helpful, when one tries to distinguish between manifolds by means of an algorithm. This theorem immediately follows from its

particular case, when $M^n = S^n$. To prove this theorem we need a large stock of n-dimensional smooth homology spheres of non-zero simplicial volume. (Homology groups are computable, and there exists an easy algorithm that is able to distinguish between S^n and a manifold which is not a homology n-sphere.) Further, it turned out that given *one* homology n-sphere with a non-zero simplicial volume, one is able to construct a collection of different homology spheres with simplicial volume > 1 which is sufficiently rich to prove Theorem 6.1. Thus, proving the following theorem turned out to be by far the most difficult part of the proof of Theorem 6.1:

Theorem 6.2. *([NW1]) For every $n \geq 5$ there exists an n-dimensional smooth homology sphere of a non-zero simplicial volume.*

Prior to our work such homology spheres were known only for $n = 3$. Very informally speaking, such manifolds enjoy simultaneously certain hyperbolicity properties (namely, non-zero simplicial volume) as well as ellipticity properties (homology of a sphere). Their construction starts from an application of work of J.P. Hausmann and P. Vogel ([H], [V]) based on the theory of homology surgery by S. Cappell and J. Shaneson ([CS]). This work enables us to reduce the topological problem to an algebraic problem of constructing finitely presented groups with certain homological properties. The resulting algebraic problem can be essentially resolved by using certain discrete cocompact subgroups of $SU(2n-1,1)$ investigated by L. Clozel ([Cl]), who proved that these groups have very few non-trivial real homology classes below dimension n. We obtain the desired groups from the groups investigated by Clozel by taking certain amalgamated free products and passing to the universal central extension to kill the remaining real homology classes below dimension n.

Once Theorem 6.1 was established, we followed a line of reasoning similar to the outline described in section 3. In particular, we needed to design an algorithm that constructed sufficiently dense nets in the spaces of Riemannian structures on all closed n-dimensional manifolds satisfying $|K| \leq 1$, $vol \geq const > 0$ and $diam \leq x$ for a variable x. For this purpose we used the Ricci flow to smooth out the Riemannian metric and to obtain a control over derivatives of the curvature tensor, and a subsequent algebraic approximation to reduce the infinite-dimensional situation to a finite dimensional one.

7. Disconnectedness of Sublevel Sets of Riemannian Functionals: Current Work and Some Open Questions

The smoothing out of Riemannian metrics by means of the Ricci flow is not available if one replaces the two-sided bound for the sectional curvature by the lower bound (or by the two-sided bound for the Ricci curvature). Therefore, we do not know how to prove the existence of an algorithm constructing a

sufficiently dense net in the space of Riemannian structures on all n-dimensional manifolds satisfying $K \geq -1$ (or $|Ric| \leq 1$) and $diam \leq x$ despite the fact that these spaces are well-known to be precompact. The difficulty can be captured in the following problem:

Problem: Does there exist an algorithm that given a positive ϵ and a finite metric space X decides whether or not X is ϵ-close (in the Gromov-Hausdorff metric) to an n-dimensional Riemannian manifold with $K \geq -1$? We allow here a certain room for an error: a positive answer must imply only the 1.01ϵ-closeness, whenever a negative answer needs to imply only that X is not 0.99ϵ-close to any such manifold. (Here we assume that ϵ and all distances between points of X are algebraic numbers.)

The problem remains open if one would consider the class of n-dimensional Alexandrov spaces with $K \geq -1$ instead of Riemannian manifolds with $K \geq -1$. (It is also open, if one would replace the condition $K \geq -1$ by $Ric \geq -(n-1)$ or $|Ric| \leq n-1$.)

The main purpose of our paper [NW3] is to bypass this difficulty, and to prove the analogues of Theorem 5.2 and all results about geometry of sublevel sets of diameter on $Al(M^n)$ mentioned in section 5 in the situation, when the two-sided bound for the sectional curvature is replaced by the lower bound. In other words, we replace $Al(M^n)$ by a (larger) space $al(M^n)$ of Alexandrov structures with curvature ≥ -1 on M^n. More formally, $al(M^n)$ is the Gromov-Hausdorff closure of the set of all Riemannian structures on M^n satisfying $K \geq -1$ in the space of isometry classes of metric spaces homeomorphic to M^n.

Our basic idea is to "approximate" the space of Riemannian structures of sectional curvature bounded below on closed n-manifolds by a space of isometry classes of simplicial length spaces that share some important metric and topological properties with manifolds with curvature bounded from below. One chooses these properties so that they can be verified by means of an algorithm (in order to be able to construct the desired nets). For example, one needs to have a lower bound for the volume in terms of the simplicial volume for these length spaces. To ensure this property one can use Theorem 5.38 in [Gr]. This theorem yelds a desired generalization of the mentioned result from [G 1] providing a lower bound for the volume in terms of the simplicial volume in the case, when the Ricci curvature is bounded from below. According to Theorem 5.38 in [Gr] an analogous lower bound will be valid for a length space if a packing function for its universal covering admits an upper bound which is similar to Bishop-Gromov upper bounds for manifolds with Ricci curvature bounded below. However, note that universal coverings cannot be constructed by means of an algorithm (as, for example, there is no algorithm deciding whether or not the fundamental group is trivial). Nevertheless, one can modify this constraint so that it becomes verifiable by means of an algorithm: It is sifficient to require the desired upper bound for the packing function for an *effective universal covering* (that will be explained below in section 9) instead of the usual universal covering.

We believe that this approach can also be used to generalize Theorem 5.2 to the situation, when the bound $|K| \leq 1$ is replaced by $|Ric| \leq n-1$ (or by $Ric \geq -(n-1)$).

Furthermore, we conjecture that an analogue of Theorem 5.2 will hold in the situation, when one replaces *diam* by *vol*. In particular, we would like to establish disconnectedness of sublevel sets of *vol* on $Al(M^n)$ (and $al(M^n)$). This problem is interesting, because sets $vol^{-1}([v,V]) \subset Al(M^n)$ are not precompact, yet the failure of the precompactness is not too "severe".

Finally note, that it is possible that the technique used to prove Theorem 5.2 is applicable to the Einstein-Hilbert action, and can even lead to a proof of the existence of infinitely many isometry classes of singular Einstein metrics of scalar curvature equal to -1 on every compact manifold of dimension > 4.

8. Higher-dimensional Cycles in Sublevel Sets

In the previous sections we discussed deep local minima (or, more generally, deep basins on graphs) of some functionals. In principle, one can regard a non-trivial deep local minimum of a functional F on a simply-connected space X as a homologically non-trivial 0-dimensional cycle in a sublevel set of F that becomes trivial in an ambient sublevel set that corresponds to a much higher value of F. (This 0-cycle is the linear combination of the deep non-trivial local minimum and the global minimum taken with opposite signs.)

One can provide the following intuitive explanation of the appearance of the non-trivial deep basins in situations that we have considered: As the manifold of interest is algorithmically indistinguishable from other manifolds of the same dimension but with different fundamental groups, there will be deep basins where the manifold "looks" like it has a certain fundamental group which is different from what it actually is.

Similar phenomena for higher-dimensional cycles were explored in [NW2]. We found various sources of higher-dimensional cycles in sublevel sets of Riemannian functionals, say *diam* on $Al(M^n)$, that become null-homologous only in a much larger sublevel set corresponding to a much higher value of the functional.

To explain this phenomenon note that $Al(M^n)$ is weakly homotopy equivalent to the space $Riem(M^n)$ of Riemannian structures on M^n. This last space is the quotient of the contractible space of Riemannian metrics on M^n by the pullback action of the diffeomorphism group. Therefore, the topology of $Riem(M^n)$ (and $Al(M^n)$) is closely related to the topology of $BDiff(M^n)$. (For example, if all compact groups non-trivially acting on M^n are finite, then $Riem(M^n)$ is rationally homotopy equivalent to $BDiff(M^n)$.) On the other hand, $BDiff(M^n)$ has a very rich topology - especially in the case, when M^n has a non-trivial fundamental group. For example, in many interesting cases one can identify a subgroup of a homology group of $BDiff(M^n)$ isomorphic to a lattice in a homology group of the fundamental group of M^n with real

coefficients. (For this purpose one can use $H\rho$-invariant introduced by Shmuel Weinberger in [We0].) Now we can use the logical method, and to argue that as M^n is algorithmically undistinguishable from manifolds N^n with arbitrary large fundamental groups, the homology classes of $\pi_1(N^n)$ corresponding to non-trivial homology classes of $BDiff(N^n)$ and $Al(N^n)$ will correspond to "virtual" homology classes of $Al(M^n)$, that is, to cycles in sublevel sets of $diam$ on $Al(M^n)$ that will become null-homologous only in much large sublevel sets. This approach works under some restrictions on topology of M^n, and produces "virtual" k-cycles for $k << n$. Another approach to constructing "virtual" k-cycles in $Al(M^n)$ is based on connections between $Diff(N^n)$ and $Out(\pi_1(N^n))$ and works for all closed manifolds M^n of dimension > 4 and all k. For example, one can always choose N^n (algorithmically undistinguishable from M^n), so that $Diff(N^n)$ admits a split surjection on Z^m for arbitrarily large values of m. As the result, for every k one obtains "virtual" k-cycles in $Al(M^n)$.

On the other hand, Shmuel Weinberger noticed that if M^n admits a non-trivial smooth compact group action, then one can similarly exploit a part of topology of $Riem(M^n)$ based on singularities that does not come from the topology of $BDiff(M^n)$. In particular, in [NW2] we used the non-existence of an algorithm deciding whether or not the fixed point set of an S^1-action on S^n is diffeomorphic to S^{n-2} to prove the existence of 5-dimensional (or, more generally, $(4i+1)$-dimensional) "virtual" rational cycles in $Al(S^n)$ that are close to the round metric in the path metric on $Al(S^n)$ (see Theorem 17.1 in [NW2] and Theorem 5 in section 4.1 of [We] for precise statements).

9. Morse Landscapes of the Length Functional

Assume that M^n is a simply-connected Riemannian manifold, $p \in M^n$. Consider the length functional l on the space $\Omega_p M^n$ of loops on M^n based at p. Note that, in principle, l can have no local minima other than the trivial loop. If there exists another local minimum α, we can define its depth as the minimal possible difference between the length of the longest loop in a path homotopy connecting α with a loop of a smaller length and the length of α. One can generalize this definition for the situation, when the length is regarded as a functional on the space $\Omega_{p,q} M^n$ of paths connecting a pair of points $p, q \in M^n$. (Of course, $\Omega_{p,p} M^n = \Omega_p M^n$.)

It is clear that one can give a similar definition of depth in the case, when α is a critical point of the length functional of a higher index i. (One needs to look at the minimal $x \geq l(\alpha)$ such that an appropriately defined i-cycle in $l^{-1}([0, l(\alpha)])$ that "hangs" at α becomes a boundary in $l^{-1}([0, x])$; the depth is then defined as $x - l(\alpha)$. If no such x exists, then we say that α has infinite depth.)

In [N5] we proved a theorem with the following informal meaning (see Theorem 2.1 in [N5] for an exact statement):

Theorem 9.1. *Let M^n be a simply-connected Riemannian manifold. Assume that the length functional has a "very deep" non-trivial local minimum on $\Omega_p M^n$. Then it has "many" "deep" local minima.*

In other words, this theorem asserts that if there exists a loop γ based at p that cannot be contracted to a point via loops of length $\leq L+$ length(γ), then there exist at least $k(L)$ geodesic loops providing "deep" local minima for the length functional on $\Omega_p M^n$, where $k(L) \longrightarrow \infty$, as $L \longrightarrow \infty$.

Note that a counterexample to Theorem 9.1 must "look" like a Riemannian manifold with a "small" finite fundamental group. Otherwise we will be able to construct "many" deep local minima of the length functional by taking powers and products of powers of the already constructed geodesic loops based at p and shortening them to geodesic loops providing new local minima. Therefore, informally speaking, Theorem 9.1 implies that a closed simply-connected Riemannian manifold cannot "look" like it has a finite fundamental group. (Of course, we saw in the previous sections that a closed simply-connected Riemannain manifold can "look" like it has an infinite fundamental group, and this fact was one of the cornerstones of all applications of recursion theory to geometry discussed in this paper.)

The proof of this theorem given in [N5] is based on the idea of the "effective univeral covering". (Recall that this concept can also be used for proving analogues of Theorem 5.2 for weaker curvature constraints - see section 7 above.) This idea can be explained as follows: [4]

The universal covering space of a topological space X is usually constructed as the quotient of the space of paths on X starting at a base point $x \in X$ by the following equivalence relation: Two paths are equivalent if they end at the same point, and together form a contractible loop (based at x). Let $X = M^n$ be a closed Riemannian manifold. One can try to make the following natural modification of this construction: Assume that one takes into consideration the length of paths, and allows only a controlled increase of length during a homotopy contracting the loop formed by two paths. More specifically, one can choose parameters U and $V > 2U$, consider the set $P(U)$ of all paths of length $\leq U$ based at x, and then try to introduce the equivalence relation $\sim_V$ on this set by identifying paths forming loops contractible via loops of length $\leq V$. However, in general, this relation will not be an equivalence relation. Nevertheless, we observed that there exists a "large" set of values of V such that $\sim_V$ is an equivalence relation. In particular, one can choose a "controllably" large $V = V(U, M^n)$, and to obtain an effectively constructible connected space $P(U, V)$ of $\sim_V$-equivalence classes of elements of $P(U)$ so that the map $P(U, V) \longrightarrow M^n$ sending each equivalence class of paths into their common endpoint is a covering "away from the boundary" in the sense of Definition 1.1

[4] Note similarities between this idea and the notion of fundamental pseudogroups introduced by Gromov in [G2].

in [N5]. One can regard sets $P(U,V)$ as constructive analogs of metric balls of radius U in the universal covering of M^n.

Now one can demonstrate Theorem 9.1 by contradiction. Assume that there exists a counterexample. It must "look" like a manifold with a finite fundamental group formed by "few" "deep" local minima of the length functional on $\Omega_p M^n$. Observe that when one constructs the usual universal covering of a closed Riemannian manifold with a finite fundamental group, one does not need to consider arbitrarily long paths. Paths of length $\leq 2d|\pi_1 M^n|$ are sufficient. (Longer paths are equivalent to some of the shorter paths.) A similar phenomenon occurs, when we construct the "effective universal covering" $P(U,V)$ of M^n for approprately chosen U and V using p as the base point: Longer paths become equivalent to shorter paths, $P(U,V)$ becomes a closed manifold and the covering "away from the boundary" becomes the covering of M^n in the usual sense. Our assumption about the existence of at least one "very deep" non-trivial local minimum of the length functional implies that the cardinality of the fiber of this covering is at least two, and so it cannot be a homeomorphism. However, all coverings of M^n are trivial, as M^n was assumed to be simply-connected, and we obtain a desired contradiction.

Note that this proof implies that if the depth of a non-trivial local minimum is λd, then there exist at least $k(\lambda) \sim \sqrt{\lambda}$ local minima. Moreover, according to Theorem 2.1 in [N5] the lengths of the geodesic loops γ_i providing these local minima do not exceed $4id, i = 1, \ldots, k$.

Both these estimates were recently improved in [NR] using a different approach that was based on geometric constructions invented largely by Regina Rotman. In particular, we demonstrated that if the the length functional on $\Omega_p M^n$ has a non-trivial local minimum of depth $> \lambda d + S$, for some $S \geq 2d$ and λ, then there exist $k \geq [\frac{\lambda}{6} + \frac{1}{2}]$ non-trivial local minima of depth $> S$. In addition, one can ensure that the length of γ_i is in the interval $(2(i-1)d, 2id]$. (This is a direct corollary of Theorem 7.3 in [NR] for $m = 1$.) The same technique also implies that the existence of a "very deep" critical point of any index $m \geq 0$ of a finite depth of the length functional on $\Omega_p M^n$ also implies the existence of "many" "deep" local minima of the length functional; explicit bounds for the number, lengths and depths of these minima are available. For example, if the depth of a critical point of index m is finite but greater than $\lambda d + (2m-1)S$, for some $S \geq 2d$ and λ, then one is guaranteed $k \geq [\frac{\lambda}{4m+2} - \frac{2m-5}{4m+2}]$ local minima of depth $> S$ with lengths in the intervals $(2(i-1)d, 2id]$, $i = 1, \ldots, k$. Furthermore, Theorems 7.3, 7.4 in [NR] immediately imply similar results for the length functional on spaces $\Omega_{p,q}(M^n)$. Thus, in particular, the results of [NR] imply that:

Theorem 9.2. *(Imprecise version) If the length functional l on $\Omega_{p,q}(M^n)$ has a critical point of an arbitrary index of a "large" finite depth, then l has "many" "deep" local minima.*

For the lack of time I will not attempt to give a more detailed presentation of these and related results and methods, and most notably applications of

these methods to quantitative geometric calculus of variations. Instead I refer the readers to [NR], [NR0], [R1], [R2], [R3] for some of the highlights of this emerging theory that has its origins in some of Gromov's ideas from [G3].

References

[B] Barzdin Ja. M., *Complexity of programs which recognize whether natural numbers not greater than n belong to a recursively enumerable set*, Soviet Math. Dokl. 9(1968), 1251–1254.

[BN] V. N. Berestovskij, I.G. Nikolaev, *Multidimensional generalized Riemannian spaces*, in *Geometry IV*, ed. by Yu. G. Reshetnyak, Springer, 1993.

[Be] M. Berger, *A panoramic view of Riemannian geometry*, Springer, Berlin, 2003.

[BHP] W. Boone, W. Haken, V. Poenaru, *On recursively unsolvable problems in topology and their classification*, in *Contributions to Mathematical Logic*, ed. H. Arnold Schmidt et al., North-Holland, 1968.

[CFKSW] J. Cantarella, J. Fu, R. Kusner, J. Sullivan, N. Wrinkle, *Criticality for the Gehring link problem*, Geom. Top. 10(2006), 2055–2116.

[CS] S. Cappell, J. Shaneson, *The codimension two placement problem and homology equivalent manifolds*, Ann. Math. 99(1974), 277–348.

[Cl] L. Clozel, *On the cohomology of Kottwitz's arithmetic varieties*, Duke Math. J. 72(1993), 757–795.

[Dur] O. Durumeric, *Local structure of ideal shapes of knots*, Topology Appl. 154(2007), 3070–3089.

[Gr] M. Gromov, *Metric structures for Riemannian and non-Riemannian spaces*, Birkhäuser, 1999.

[G1] M. Gromov, *Volume and bounded cohomology*, Inst. Hautes Études Sci. Publ. Math., 56(1982), 5–100.

[G2] M. Gromov, *Almost flat manifolds*, J. Diff. Geometry, 13(1978), 231–241.

[G3] M. Gromov, *Filling Riemannian manifolds*, J. Diff. Geometry, 18(1983), 1–147.

[H] J.-C. Hausmann, *Manifolds with a given homology and fundamental group*, Comment. Math. Helv. 53(1978), 113–134.

[Ke] M. Kervaire, *Smooth homology spheres and their fundamental groups*, Trans. Amer. Math. Soc. 144(1969), 67–72.

[KV] O. Krötenheerdt, S. Veit, *Zur Theorie massiver Knoten*, Wiss. Beitr. Martin-Luther-Univ. Halle-Wittenberg Reihe M. Math. 7(1976), 61–74.

[LV] M. Li, P.M.B. Vitanyi, *Kolmogorov complexity and its applications*, in *Handbook of Theoretical Computer Science*, ed. by J. van Leeuwen, Elsevier, 1990, 187–254.

[LSDR] R. A. Litherland, J. Simon, O. Durumeric, E. Rawdon, *Thickness of knots*, Topology and its Appl. 91(1999), 233–244.

[Mil] C. F. Miller, *Decision problems for groups - survey and reflections*, in *Combinatorial group theory*, (ed. by G. Baumslag, C. F. Miller), Springer, 1989, 1–59.

[N1] A. Nabutovsky, *Non-recursive functions, knots "with thick ropes", and self-clenching "thick" hyperspheres*, Comm. Pure Appl. Math. 48(1995), 381–428.

[N2] A. Nabutovsky, *Disconnectedness of sublevel sets of some Riemannian functionals*, Geom. Funct. Anal. 6(1996), 703–725.

[N3] A. Nabutovsky, *Geometry of the space of triangulations of a compact manifold*, Comm. Math. Phys. 181(1996), 303–330.

[N4] A. Nabutovsky, *Einstein metrics: existence versus uniqueness*, Geom. Funct. Anal. 5(1995), 76–91.

[N5] A. Nabutovsky, *Effective universal coverings and local minima of the length functional on the loop spaces*, to appear in Geom. Funct. Anal.

[NR0] A. Nabutovsky, R. Rotman, *Curvature-free upper bounds for the smallest area of a minimal surface*, Geom. Funct. Anal. 16(2006), 453–475.

[NR] A. Nabutovsky, R. Rotman, *Length of geodesics and quantitative Morse theory on loop spaces*, preprint, 2009.

[NW0] A. Nabutovsky, S. Weinberger, *Algorithmic unsolvability of the triviality problem for multidimensional knots*, Comment. Math. Helv. 71(1996), 426–434.

[NW1] A. Nabutovsky, S. Weinberger, *Variational problems for Riemannian functionals and arithmetic groups*, Inst. Hautes Études Sci. Publ. Math. 92(2000), 5–62(2001).

[NW2] A. Nabutovsky, S. Weinberger, *The fractal nature of Riem/Diff I*, Geom. Dedicata 101(2003), 1–54.

[NW3] A. Nabutovsky, S. Weinberger, *Local minima of diameter on spaces of Riemannian structures with curvature bounded below*, in preparation.

[R1] R. Rotman, *The length of a shortest geodesic net on a closed Riemannian manifold*, Topology 46(2007), 343–356.

[R2] R. Rotman, *The length of a shortest geodesic loop at a point*, J. Differential Geom. 78(2008), 497–519.

[R3] R. Rotman, *Flowers on Riemannian manifolds*, preprint.

[Sh] M. A. Shtanko, *A theorem of A.A. Markov and algorithmically nonrecognizable combinatorial manifolds*, Izv. Math. 68(2004), 205–221.

[V] P. Vogel, *Un theoreme d' Hurewicz homologique*, Comment. Math. Helv. 52(1977), 393–413.

[We0] S. Weinberger, *Higher ρ-invariants*, Contemp. Math. 231, 315–320.

[We] S. Weinberger, *Computers, Rigidity and Moduli*, Princeton University Press, 2005.

[Y] S. T. Yau, *Open problems in geometry*, in *Differential Geometry: Riemannian Geometry*, ed. by R. Greene and S. T. Yau, Proceedings of AMS Symposia in Pure Mathematics, 54:1(1993), 1–28.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Constant Scalar Curvature and Extremal Kähler Metrics on Blow ups

Frank Pacard*

Abstract

Extremal Kähler metrics were introduced by E. Calabi as best representatives of a given Kähler class of a complex compact manifold, these metrics are critical points of the L^2 norm of the scalar curvature function. In this paper, we report some joint works with C. Arezzo and M. Singer concerning the construction of extremal Kähler metrics on blow ups at finitely many points of Kähler manifolds which already carry an extremal Kähler metric. In particular, we give sufficient conditions on the number and locations of the blown up manifold points for the blow up to carry an extremal Kähler metric.

Mathematics Subject Classification (2010). Primary 32J27; Secondary 53C21.

Keywords. Extremal metrics, Kähler geometry, perturbation methods.

1. Introduction

In [6], [7] E. Calabi proposed, as best representatives of a given Kähler class $[\omega]$ of a complex compact manifold (M, J), a special type of metric baptized *extremal*. If ω is a positive Kähler form on M and if $\mathbf{s}(\omega)$ denotes the scalar curvature of (the riemannian metric associated to) ω, E. Calabi proposed the study of the functional

$$\tilde{\omega} \in [\omega]^+ \longmapsto \int_M \mathbf{s}(\tilde{\omega})^2 \mathrm{dvol}_{\tilde{\omega}} \,,$$

where $[\omega]^+$ denotes the set of positive Kähler forms in the Kähler class $[\omega]$. The corresponding Euler-Lagrange equation reduces to the fact that the vector field

$$\Xi_{\mathbf{s}} := X_{\mathbf{s}} - i\, J\, X_{\mathbf{s}} \,, \quad \text{with} \quad X_{\mathbf{s}} := J\, \nabla \mathbf{s} \,,$$

is a holomorphic vector field on M. Obviously, if $\mathbf{s}(\omega)$ is constant, then $\Xi_{\mathbf{s}} \equiv 0$

*Université Paris-Est Créteil and Institut Universitaire de France, 61 Avenue du Général de Gaulle, 94010, Créteil. E-mail: pacard@univ-paris12.fr.

and so the set of extremal metrics contains the set of constant scalar curvature Kähler ones (and in particular the set of Kähler-Einstein metrics). Conversely, if (M, J) admits no non-trivial holomorphic vector field, then every extremal Kähler metric has constant scalar curvature. However, E. Calabi also proved that some extremal metrics with non constant scalar curvature do exist [6].

C. LeBrun and S. Simanca [20] have proved that the set of Kähler classes having an extremal representative is an open subset of $H^{1,1}(M, \mathbb{C}) \cap H^2(M, \mathbb{R})$. In the presence of holomorphic vector fields, if a Kähler class $[\omega]$ contains a metric of constant scalar curvature, then every nearby Kähler class will contain an extremal metric but these will not in general have constant scalar curvature. Another illustration of this phenomenon will be given in Proposition 4 below. Finally, recall that X.X. Chen and G. Tian [9] have proved the uniqueness of extremal metrics in a given Kähler class up to the action of automorphisms.

E. Calabi's intuition for looking at extremal metrics as canonical representatives of a given Kähler class has found a number of important confirmations and also (unfortunately) nontrivial constraints [22], [11]. E. Calabi himself proved that an extremal Kähler metric must have the maximal possible symmetry allowed by the complex manifold M, and, as observed by C. LeBrun and S. Simanca [19], this symmetry group can be fixed in advance. More precisely, the identity component of the isometry group of any extremal metric g must be a maximal compact subgroup of $\mathrm{Aut}_0(M, J)$, the identity component of $\mathrm{Aut}(M, J)$, the group of biholomorphic maps of M to itself. This group thus contains the complexification of the isometry group, but may be strictly larger (the blow-up of $\mathbb{P}^2$ at a point is the simplest example of such a situation).

Also, the important relationship between the existence of extremal metrics and various stability notions of the corresponding polarized manifolds (algebraic if the class is rational, analytic otherwise) has been deeply investigated for example by S. Donaldson [11], [14], T. Mabuchi [25], G. Tian [33], [34] and G. Szekelyhidi [31]. Yet, a complete understanding of the existence theory for extremal metrics is still missing. Given this last fact, one interesting problem is to give sufficient conditions for the existence of an extremal Kähler metric on the blow up at finitely many points of a manifold which already carries an extremal Kähler metric. Also, we would like to characterize the Kähler classes on the blown up manifold for which we are able to find such an extremal metric. The aim of the present paper is to present various results which were obtained along these lines in [1], [2], [3] and [4].

The author would like to thank C. Arezzo and M. Singer for valuable comments and help during the writing of this survey.

2. Statement of the Result

Let (M, J, g, ω) be a Kähler manifold with complex structure J and Kähler form ω and let g denote the Riemannian metric associated to the Kähler form ω, so that

$$\omega(X, Y) = g(J\,X, Y)\,.$$

Further assume that g is an extremal Kähler metric. Since the automorphism group of any blow up of M can be identified with a subgroup of $\mathrm{Aut}(M, J)$, and in light of the above mentioned result of Calabi-LeBrun-Simanca about the isometry group of any extremal Kähler metric, our strategy is to fix *a priori* a compact subgroup K of $\mathrm{Isom}(M, g)$ and work K-equivariantly. The identity component of K will be denoted by K_0.

Such a K will then be contained in the isometry group of the extremal Kähler metric we are seeking on the blow up of M at any set of points $p_1, \ldots, p_n \in M$ in $\mathrm{Fix}\,(K_0)$, the fixed locus of K_0. We also require that $\{p_1, \ldots, p_n\}$ is globally invariant under the action of K. We will denote by $\mathfrak{k}$ the Lie algebra associated to K_0. Observe that elements of $\mathfrak{k}$ vanish at the points $p_1, \ldots, p_n$ to be blown up and hence these vector fields can be lifted to the blown up manifold.

In order to produce extremal Kähler metrics on the blown up manifold, we have to identify, among all smooth functions on the blown up manifold, those who generate real-holomorphic vector fields, since these can arise as scalar curvatures of extremal Kähler metrics. To this aim, we define $\mathfrak{h}$ to be the vector space of K-invariant hamiltonian real-holomorphic vector fields on M or equivalently, the Lie algebra of the group H of holomorphic exact symplectomorphisms commuting with K. The correspondence between real-holomorphic vector fields and the scalar functions on M can be encoded in a compact way in a *moment map* for the action of H

$$\xi_\omega \colon M \longrightarrow \mathfrak{h}^* \,,$$

uniquely determined by requiring the Hamiltonian functions to have mean-value zero. More explicitly, the function $f := \langle \xi_\omega, X \rangle$ associated to the vector field $X \in \mathfrak{h}$ is defined to be the unique solution of

$$-\mathrm{d}f = \omega(X, \cdot)\,,$$

whose mean value over M is 0.

In general, K is not necessarily connected and $\mathfrak{h}$ is not necessarily included in $\mathfrak{k}$. There is a natural orthogonal decomposition

$$\mathfrak{h} = \mathfrak{h}' \oplus \mathfrak{h}''\,,$$

where

$$\mathfrak{h}' := \mathfrak{h} \cap \mathfrak{k}\,,$$

is the subspace of K-invariant real-holomorphic vector fields in $\mathfrak{k}$ and where the scalar product is taken to be

$$(X, \tilde{X})_{\mathfrak{h}} := \int_M \langle \xi_\omega, X \rangle \, \langle \xi_\omega, \tilde{X} \rangle \, \mathrm{dvol}_\omega \,.$$

Under the above assumptions, in the following general result, we give *sufficient conditions* for the existence of an extremal Kähler metric on the blow up of M at finitely many points:

Theorem 1. *[4] Assume that $\mathfrak{k}$ contains the vector field $X_{\mathbf{s}}$ and assume that we are given $p_1, \ldots, p_n \in Fix(K_0)$ such that $\{p_1, \ldots, p_n\}$ is invariant under the action of K and:*

(i) The projections of $\xi_\omega(p_1), \ldots, \xi_\omega(p_n)$ onto $\mathfrak{h}''^*$ span $\mathfrak{h}''^*$.

(ii) There exist $a_1, \ldots, a_n > 0$ satisfying

$$\sum_{j=1}^{n} a_j^{m-1}\, \xi_\omega(p_j) \in \mathfrak{h}'^*\,,$$

and $a_{j_1} = a_{j_2}$ if p_{j_1} and p_{j_2} belong to the same K-orbit.

(iii) There is no nontrivial element of $\mathfrak{h}''$ which vanishes at $p_1, \ldots, p_n$.

Then, there exists $\varepsilon_0 > 0$ and, for all $\varepsilon \in (0, \varepsilon_0)$, there exists a K-invariant extremal Kähler metric g_ε on $\tilde{M}$, the blow up of M at $p_1, \ldots, p_n$, whose associated Kähler form ω_ε lies in the class

$$\pi^*[\omega] - \varepsilon^2\, \left(a_1\, PD[E_1] + \ldots + a_n\, PD\,[E_n]\right)\,,$$

where $\pi\colon \tilde{M} \longrightarrow M$ is the standard projection map, the $PD[E_j]$ are the Poincaré duals of the $(2m-2)$-homology classes of the exceptional divisors of the blow up at p_j.

Finally, the sequence of metrics $(g_\varepsilon)_\varepsilon$ converges to g (in the smooth topology) on compacts, away from the exceptional divisors.

It is important to stress that our analytical construction does not give *one* extremal metric but *a family* converging to the starting metric on the base manifold. Observe that we assume that $X_{\mathbf{s}} \in \mathfrak{k}$ and hence $X_{\mathbf{s}} \in \mathfrak{h}' := \mathfrak{k} \cap \mathfrak{h}$. Since $p_1, \ldots, p_n$ are fixed by K_0, we conclude that the vector $X_{\mathbf{s}}$ vanishes at each $p_1, \ldots, p_n$. In particular, it can be lifted to $\tilde{M}$. This is an important property since our result is perturbative away from the exceptional divisors of $\tilde{M}$ and hence it is natural to be able to lift $X_{\mathbf{s}}$ to $\tilde{M}$ to guarantee that our construction will be successful.

Conditions (i) and (ii) which appear in the statement of Theorem 1 are known to be related to T. Mabuchi's T-stability [25] and to G. Szekelyhidi's relative K-stability [31] in the same way the analogous conditions for constant scalar curvature metrics are related to the asymptotic Chow semi-stability along the line of ideas described by R. Thomas in [32] (pages 27 and 28) [30], [10].

Remark 1. When $\mathfrak{h}' = \{0\}$, and in particular g is a constant scalar curvature metric (since $X_{\mathbf{s}} \in \mathfrak{h}'$), then Theorem 1 yields constant scalar curvature metrics [2].

Condition (iii) can be removed at the expense of leaving some freedom on the weights of the exceptional divisor on the blown up manifold. More precisely, Theorem 1 still holds without assuming (iii) but in this case, the only

information we have about $[\omega_\varepsilon]$ reads

$$[\omega_\varepsilon] = \pi^*[\omega] - \varepsilon^2 \left(\tilde{a}_1\, PD[E_1] + \ldots + \tilde{a}_n\, PD\,[E_n]\right) ,$$

where $\tilde{a}_1, \ldots, \tilde{a}_n > 0$ depend on ε and satisfy

$$|\tilde{a}_j - a_j| \leq c\, \varepsilon^\kappa ,$$

for some constant $\kappa > 0$ only depending on m. In other words, by removing (iii), we slightly lose control on the Kähler classes.

There are special important situations to which Theorem 1 applies. Let us describe some of them.

2.1. The non-obstructed case. Assume that g is a constant scalar curvature Kähler metric and that there exists K a discrete subgroup of $\mathrm{Isom}\,(M, g)$ such that $\mathfrak{h} = \{0\}$. Then, conditions (i), (ii) and (iii) become vacuous and the result applies to any finite of points $p_1, \ldots, p_n \in M$ which are globally invariant under the action of K to produce constant scalar curvature Kähler metrics on the blow up of M at these points [1], [2].

If the scalar curvature of g is not zero then the scalar curvatures of g_ε and of g have the same signs. Also, if the scalar curvature of g is zero and the first Chern class of M is non zero, then one can arrange so that the scalar curvature of g_ε is also equal to 0. This last result complements in any dimension previous constructions which have been obtained in complex dimension $m = 2$ and for zero scalar curvature metrics by J. Kim, C. LeBrun and M. Pontecorvo [17], C. LeBrun and M. Singer [21] and Y. Rollin and M. Singer [28].

As an application, let us consider (M, J, g, ω) to be the projective space $\mathbb{P}^m$ endowed with the Kähler form ω_{FS} associated to a Fubini-Study metric, we let $(z_1, \ldots, z_{m+1})$ be complex coordinates in $\mathbb{C}^{m+1}$ and we agree that ω_{FS} is normalized so that $[\omega_{FS}] = PD[\mathbb{P}^{m-1}]$, where $\mathbb{P}^{m-1} \subset \mathbb{P}^m$ is a linear subspace. We consider the group K generated by

$$[z_1 : \ldots : z_{m+1}] \longmapsto [\pm z_1 : \ldots : \pm z_{m+1}] ,$$

and

$$[z_1 : \ldots : z_{m+1}] \longmapsto [z_{\sigma(1)} : \ldots : z_{\sigma(m+1)}] ,$$

where $\sigma \in \mathcal{S}_{m+1}$ is any permutation of $\{1, \ldots, m+1\}$. We consider the set of fixed points of K

$$p_1 := [1 : 0 : \ldots : 0], \quad \ldots, \quad p_{m+1} := [0 : \ldots : 0 : 1] .$$

In this case, the space $\mathfrak{h} = \{0\}$ and, as a consequence of the result of Theorem 1, we obtain constant scalar curvature Kähler metrics on the blow up of $\mathbb{P}^m$ at the points $p_1, \ldots, p_n$, for any $n = 1, \ldots, m+1$ whose associated Kähler form ω_ε lies in the class

$$\pi^*[\omega_{FS}] - \varepsilon^2 \left(PD[E_1] + \ldots + PD[E_n]\right) .$$

2.2. The obstructed case. Another important application of Theorem 1 is when g is a constant scalar curvature Kähler metric and when $K = \{Id\}$. In this case $\mathfrak{h}' = \{0\}$ and $\mathfrak{h}'' = \mathfrak{h}$. The following results show that there is a large set of possible applications of Theorem 1.

Lemma 2. *[2] Assume that $n \geq dim\,\mathfrak{h}$ then, the set of points $p_1, \ldots, p_n \in M$ (all distinct) such that condition (i) is fulfilled is an open and dense subset of M^n.*

When $n \geq \dim \mathfrak{h}$, it is well known that, for a choice of blow up points $p_1, \ldots, p_n$ in some open and dense subset of M^n, the group of automorphisms of M blown up at $p_1, \ldots, p_n$ is trivial (observe that $\dim \mathfrak{h}$ is also equal to the dimension of $\mathrm{Aut}_0(M)$). In view of all these results, one is tempted to conjecture that condition (i) is equivalent to the fact that the group of automorphisms of M blown up at $p_1, \ldots, p_n$ is trivial. However, this is not the case since these two conditions turn out to be of a different nature. For example, let us assume that $\mathfrak{h} = \mathrm{Span}\,\{X\}$ for $X \neq 0$. If we denote by $f := \langle \xi_\omega, X \rangle$, it is enough to choose $p_1, \ldots, p_n$ not all in the zero set of f for condition (i) to hold, while the group of automorphisms of M blown up at $p_1, \ldots, p_n$ is trivial if and only if one of the p_j is chosen away from the zero set of X, which corresponds to the set of critical points of function f.

Condition (ii) is more subtle and more of a nonlinear nature. It can be proven that this condition is always fulfilled for some careful choice of the points, provided their number is chosen to be larger than some value $n_g \geq \dim \mathfrak{h} + 1$.

Lemma 3. *[2] Assume that $n \geq dim\,\mathfrak{h}+1$, then the set of points $p_1, \ldots, p_n \in M$, all distinct, for which (i) and (ii) hold is an open (possibly empty) subset of M^n. Moreover, there exists $n_g \geq dim\,\mathfrak{h} + 1$ such that, for all $n \geq n_g$ the set of points $p_1, \ldots, p_n \in M$ (all distinct) for which (i) and (ii) hold is a nonempty open subset of M^n.*

In contrast with condition (ii), it is easy to convince oneself that condition (i) does not hold for generic choice of the points. For example, assume that $\mathfrak{h} = \mathrm{Span}\,\{X\}$ for $X \neq 0$, we denote by $f := \langle \xi_\omega, X \rangle$ and we choose $n \geq 2$. Then (ii) holds provided $f(p_1), \ldots, f(p_n)$ are not all equal to 0 and (i) holds provided $f(p_1), \ldots, f(p_n)$ do not all have the same sign. Clearly, the set of such points is a nonempty open subset of M^n.

As an application, let us again consider the projective space $\mathbb{P}^m$ endowed with the Kähler form ω_{FS} associated to a Fubini-Study metric. In this case $\dim \mathfrak{h} = m^2 + 2\,m$ and it is proven in [2] that the set of points $p_1, \ldots, p_n \in \mathbb{P}^m$ (all distinct) for which (i) and (ii) hold is a nonempty open subset of $(\mathbb{P}^m)^n$ provided $n \geq 2\,m\,(m+1)$ (Let us point out that this last result is certainly not sharp).

2.3. Extremal versus constant scalar curvature metrics. The proof of Theorem 1 is based on a perturbation argument and, if the initial

manifold has an extremal metric with non-constant scalar curvature, the extremal metrics we construct on the blown up manifold still have non-constant scalar curvature. In the case where the manifolds we start with have constant scalar curvature metrics, it might well be that the extremal metrics we obtain have in fact constant scalar curvature. There is a simple criterion involving the points $p_1, \ldots, p_n$ and the parameters $a_1, \ldots, a_n$, which ensures that this is not the case.

Proposition 4. *Under the assumptions of Theorem 1, further assume that the metric g has constant scalar curvature. If the points and weights are chosen so that*

$$\sum_{j=1}^{n} a_j^{m-1}\, \xi_\omega(p_j) \neq 0\,,$$

then the metrics g_ϵ we obtain on $\tilde{M}$ are extremal with non-constant scalar curvature.

2.4. The case of projective spaces. We now emphasize the consequences of the above results for projective spaces.

As above, we consider the projective space $\mathbb{P}^m$ endowed with the Kähler form ω_{FS} associated to a Fubini-Study metric and we let let $(z_1, \ldots, z_{m+1})$ be complex coordinates in $\mathbb{C}^{m+1}$. We now consider the group $K = S^1 \times \ldots \times S^1$, to be the maximal compact subgroup of $PGL(m+1)$, whose action is given by

$$\begin{array}{ccc} K \times \mathbb{P}^m & \longrightarrow & \mathbb{P}^m \\ ((\alpha_1, \ldots, \alpha_{m+1}), [z_1 : \ldots : z_{m+1}]) & \longmapsto & [\alpha_1 z_1 : \ldots : \alpha_{m+1} z_{m+1}]\,, \end{array}$$

and we consider the set of fixed points of K

$$p_1 := [1 : 0 : \ldots : 0], \quad \ldots, \quad p_{m+1} := [0 : \ldots : 0 : 1]\,.$$

In this case, the space $\mathfrak{h}$ is spanned by vector fields of the form

$$\Re\left(z_a\, \partial_{z_a} - z_b\, \partial_{z_b}\right)\,,$$

for $a, b \in \{1, \ldots, m+1\}$ and we have $\mathfrak{k} = \mathfrak{h} = \mathfrak{h}'$ and $\mathfrak{h}'' = \{0\}$. As a consequence of the result of Theorem 1, we obtain extremal Kähler metrics on the blow up of $\mathbb{P}^m$ at the points $p_1, \ldots, p_n$, for any $n = 1, \ldots, m+1$.

It is worth emphasizing that the special structure of the points which can be blown up on $\mathbb{P}^m$ has its origin in the fact that we are starting from a specific choice of a Fubini-Study metric and hence, away from the blow up points the extremal Kähler metric ω_ε is close to ω_{FS}. This example shows the *Riemannian* nature of our results. Now, if $q_1, \ldots, q_n \in \mathbb{P}^m$ are linearly independent one can find extremal metrics on the blow up of $\mathbb{P}^m$ at $q_1, \ldots, q_n$ but this time the

metric will be close to $\psi^*\omega_{FS}$ away from the blow up points, where ψ is an automorphism of the projective space such that

$$\psi(p_j) = q_j \, .$$

Since $[\psi^*\omega_{FS}]$ is independent of ψ and of the choice of the Fubini-Study metric, we have obtained the following:

Corollary 1. *[2], [4] Fix $1 \leq n \leq m+1$. Given $q_1, \ldots, q_n \in \mathbb{P}^m$ linearly independent points and $a_1, \ldots, a_n > 0$, there exists $\varepsilon_0 > 0$ and for all $\varepsilon \in (0, \varepsilon_0)$ there exists an extremal Kähler metric g_ε on the blow up of $\mathbb{P}^m$ at $q_1, \ldots, q_n$ whose associated Kähler form ω_ε lies in the class*

$$\pi^*[\omega_{FS}] - \varepsilon^2 \, (a_1 \, PD[E_1] + \ldots + a_n \, PD[E_n]) \, .$$

In addition, the Kähler metrics g_ε do not have constant scalar curvature unless $n = m+1$ and $a_1 = \ldots = a_{m+1}$.

The case corresponding to $n = 1$ in Corollary 1 was already obtained by E. Calabi in more generality (i.e. for all Kähler classes) [6] and the case where $\mathbb{P}^m$ is blown up at $m+1$ linearly independent points $q_1, \ldots, q_{m+1}$ and $a_1 = \ldots = a_{m+1}$ was already mentioned in Section 2.1 where constant scalar curvature metrics were obtained [2].

In the case where $\mathbb{P}^m$ is blown up at more than $m+1$ points in general position the resulting manifolds do not have nonzero holomorphic vector fields, hence extremal metrics are forced to have constant scalar curvature and the existence of some constant scalar curvature Kähler metrics follows from [2] and [28].

The conditions $n = m+1$ and $a_1 = \ldots = a_{m+1}$ being necessary and sufficient to get constant scalar curvature metrics among our family of extremal ones fits exactly into the more familiar picture of the Futaki invariants. E. Calabi has in fact proved that an extremal metric has constant scalar curvature if and only if its Futaki invariant vanishes [7], and, using T. Mabuchi's result [24] relating the Futaki invariant to the coordinates of the barycenter of the convex polytope of a toric variety, one can show that the above conditions are indeed equivalent to the vanishing of the Futaki invariants for blow ups of $\mathbb{P}^m$.

2.5. The blow up of $\mathbb{P}^2$ at two points. Applying Theorem 1 to $\mathbb{P}^2$, $\mathbb{P}^1 \times \mathbb{P}^1$ and $Bl_p\mathbb{P}^2$ as base manifolds leads interesting examples of extremal metrics. To begin with, we can apply Theorem 1 to the blow up of $\mathbb{P}^2$, endowed with a Fubini-Study metric, at two points. This shows that:

Corollary 2. *On $Bl_{p_1,p_2}\mathbb{P}^2$, the Kähler classes*

$$\pi^*[\omega_{FS}] - \varepsilon \, (a_1 \, PD[E_1] + a_2 \, PD[E_2]) \, ,$$

have extremal representatives provided $a_1, a_2 > 0$ are fixed and $\varepsilon \in (0, \varepsilon_0)$, where $\varepsilon_0 > 0$ is small enough.

Next, Theorem 1 can be applied to the blow up of $Bl_p\mathbb{P}^2$, endowed with E. Calabi's extremal metric, at one point. This shows that:

Corollary 3. *On $Bl_{p_1,p_2}\mathbb{P}^2$, the Kähler classes*

$$\pi^*[\omega_{FS}] - \left(a_1\, PD[E_1] + \varepsilon\, PD[E_2]\right) ,$$

have extremal representatives provided $a_1 \in (0,1)$ is fixed and $\varepsilon \in (0, \varepsilon_0)$, where $\varepsilon_0 > 0$ is small enough.

Finally, recall that, for $p_1 \neq p_2 \in \mathbb{P}^2$, $Bl_{p_1,p_2}\mathbb{P}^2$ contains three (-1)-curves, the two exceptional divisors E_1, E_2 and the proper transform L of the line in $\mathbb{P}^2$ passing though p_1 and p_2. Contracting (*blowing down*) L we get a manifold biholomorphic to $\mathbb{P}^1 \times \mathbb{P}^1$ where the rulings correspond to the pencils of lines through p_1 and p_2. So $Bl_{p_1,p_2}\mathbb{P}^2$ is biholomorphic to $Bl_q(\mathbb{P}^1 \times \mathbb{P}^1)$ for some choice of $q \in \mathbb{P}^1 \times \mathbb{P}^1$. We set

$$A_1 = [\mathbb{P}^1 \times \{pt\}], \qquad A_2 = [\{pt\} \times \mathbb{P}^1] ,$$

and we denote by E the exceptional divisor in $Bl_q(\mathbb{P}^1 \times \mathbb{P}^1)$. It is easy to check the correspondence between the class $a_1\, PD[A_1] + a_2\, PD[A_2] - \varepsilon\, PD[E]$ and the class $(a_1 + a_2 - \varepsilon)\, \pi^*[\omega_{FS}] - (a_1 - \varepsilon)\, PD[E_1] - (a_2 - \varepsilon)\, PD[E_2]$. Applying Theorem 1 to the blow up of $\mathbb{P}^1 \times \mathbb{P}^1$, endowed with a product of Fubini-Study metric, at one point, we show that:

Corollary 4. *On $Bl_{p_1,p_2}\mathbb{P}^2$, the Kähler classes*

$$\pi^*[\omega_{FS}] - \left(\tfrac{a_1-\varepsilon}{a_1+a_2-\varepsilon}\, PD[E_1] + \tfrac{a_2-\varepsilon}{a_1+a_2-\varepsilon}\, PD[E_2]\right) ,$$

have extremal representatives provided $a_1, a_2 > 0$ are fixed and $\varepsilon \in (0, \varepsilon_0)$, where $\varepsilon_0 > 0$ is small enough.

Corollary 1 has been used by X.X. Chen, C. LeBrun and M. Weber [8] and W. He [16] to prove that *all* Kähler classes on $M := \mathrm{Bl}_{p_1,p_2}(\mathbb{P}^2)$ of the form $\pi^*[\omega_{FS}] - a(PD[E_1] + PD[E_2])$ have an extremal representative. This last result implies the existence of Einstein (non-Kählerian) metrics of positive scalar curvature on M [8].

2.6. The case of toric varieties. The previous Corollary can be understood as a special case of the existence of extremal metrics on the blow up of toric varieties. If (M, J, g, ω) is a m-dimensional toric variety whose associated metric is extremal, one can take K to be the maximal torus T giving the torus action. In this case, $\mathfrak{h} = \mathfrak{k}$ and hence $\mathfrak{h}'' = \{0\}$. One can then apply Theorem 1 to get:

Corollary 5. *Assume that (M, J, g, ω) is a toric variety whose associated metric is extremal, and let K be the maximal torus T giving the torus action.*

Given $p_1, \ldots, p_n \in Fix(K)$ and $a_1, \ldots, a_n > 0$, there exists $\varepsilon_0 > 0$ and for all $\varepsilon \in (0, \varepsilon_0)$ there exists an extremal Kähler metric g_ε on the blow up of M at $p_1, \ldots, p_n$ whose associated Kähler form ω_ε lies in the class

$$\pi^*[\omega] - \varepsilon^2 \, (a_1 \, PD[E_1] + \ldots + a_n \, PD[E_n]) \; .$$

In other words, one can blow up *any set of points contained in the fixed-point set of the torus-action* and the weights $a_j > 0$ can be chosen arbitrarily.

Since blowing up a toric variety at such points preserves the toric structure, one can apply inductively the last Corollary. Therefore, we obtain extremal metrics on any such iterated blow up. This last Corollary can be applied to the one parameter family of extremal metrics found by E. Calabi on the blow up of $\mathbb{P}^m$ at one point, producing then a wealth of open subsets of classes in the Kähler cone which have extremal representatives.

Remark 2. A general existence result for constant scalar curvature metrics on toric surfaces follows from S.K. Donaldson's work on Abreu's equation [12], [14] and [13].

3. Overview of the construction

Recall that the Riemannian metric g, Kähler form ω and complex structure J are related by the relation $\omega(X, Y) = g(J\,X, Y)$. A vector field X is said to be a *hamiltonian vector field* if there exists a smooth *real valued* function f satisfying

$$X = J \, \nabla f \, .$$

In this case we will write $X = X_f$. Using the above relation between ω and g, we see that this equation is always equivalent to

$$-\mathrm{d}f = \omega \, (X_f, \cdot) \, ,$$

or also

$$-\bar{\partial} f = \tfrac{1}{2} \, \omega(\Xi_f, \cdot) \, .$$

Let us now define the second order operator

$$\begin{array}{rccc} P_\omega : & \mathcal{C}^\infty(M) & \longrightarrow & \Lambda^{0,1}(M, T^{1,0}) \\ & f & \longmapsto & \frac{1}{2} \, \bar{\partial} \, \Xi_f \, , \end{array}$$

where

$$\Xi_f := X_f - iJX_f \in T^{1,0} \, . \tag{1}$$

Observe that the operator P_ω depends on the Kähler metric g. Also, with this definition, a metric ω is extremal if and only if $P_\omega(\mathbf{s}(\omega)) = 0$.

Clearly, any smooth, complex valued function f, solution of

$$P_\omega^* \, P_\omega \, f = 0 \, ,$$

on M, gives rise to a holomorphic vector field Ξ_f defined by (1) since integration over M of this equation multiplid by $\bar f$ implies that $\|\bar\partial \, \Xi_f\|_{L^2(M)} = 0$. We recall the following important result which shows that the converse is also true:

Proposition 5. *[7], [19] A vector field $\Xi \in T^{1,0}$ is a holomorphic vector field with zeros if and only if there exists a* complex valued *function f solution of $P_\omega^* \, P_\omega \, f = 0$ such that $-\bar\partial f = \frac{1}{2} \, \omega(\Xi, \cdot)$.*

In addition, we have the following result which follows from a theorem of A. Lichnerowicz (see A. Besse [5], Corollary 2.125 and [19]):

Proposition 6. *[5], [19] A vector field X is a Killing vector field with zeros if and only if there exists a* real valued *function f solution of $P_\omega^* \, P_\omega \, f = 0$ such that $\omega(X, \cdot) = -\mathrm{d}f$.*

In other words, if Ξ is a holomorphic vector field, the function f given in Proposition 5 can be chosen to be real valued when $X = \Re \Xi$ is a Killing vector field and if X is a Killing vector field with zeros, then $\Xi = X - iJX$ is a holomorphic vector field. Finally, recall that a vector field X is real-holomorphic if and only if $X - iJX$ is a *holomorphic* section of $T^{1,0}M$. In particular, any Killing vector field is automatically real-holomorphic.

3.1. Perturbation of extremal metrics. It is proven in [19] and [5] that the linearization of the mapping

$$f \longmapsto \mathbf{s}(\omega + i\partial\bar\partial f) \, ,$$

is given by the formula

$$\mathbb{L}_\omega := -\tfrac{1}{2} \, (\Delta_g^2 + 2 \, \mathrm{Ric}_g \cdot \nabla_g^2) \, ,$$

where Ric_g stands for the Ricci tensor of the metric g associated to ω. On the other hand,

$$P_\omega^* \, P_\omega = \Delta_g^2 + 2 \, \mathrm{Ric}_g \cdot \nabla_g^2 - J \, X_{\mathbf{s}} + i X_{\mathbf{s}} \, ,$$

where $X_{\mathbf{s}}$ is the hamiltonian vector field associated to the scalar curvature $\mathbf{s}(\omega)$. Observe that, in general, this is a complex valued operator.

With these formulas, we can write:

$$\mathbb{L}_\omega = -\tfrac{1}{2} P_\omega^* \, P_\omega - \tfrac{1}{2} \, J \, X_{\mathbf{s}} + \tfrac{i}{2} \, X_{\mathbf{s}} \, .$$

We can see from this equality that, working equivariantly with respect to a compact group K whose Lie algebra contains $X_{\mathbf{s}}$ has an important consequence. Indeed, under such an assumption $X_{\mathbf{s}} \, f = 0$ for all f which are K-invariant

and the last term in (3.1) disappears. Therefore, when acting on K-invariant functions $\mathbb{L}_\omega$ is a real-valued operator. From the analytical point of view, this is the reason why we have asked that $X_{\mathbf{s}}$ belongs to the Lie algebra of K.

We consider the nonlinear map

$$\begin{array}{rccc} F: & \mathfrak{h}\times\mathcal{C}^\infty(M)^K & \longrightarrow & \mathcal{C}^\infty(M)^K, \\ & (X,f) & \longmapsto & \mathbf{s}(\omega+i\partial\bar\partial f)-\langle\xi_{\omega+i\,\partial\bar\partial f},X\rangle\,. \end{array}$$

Here the superscripts K denote the K-invariant part of the function spaces considered.

The following is due to E. Calabi [6] and C. LeBrun and S. Simanca [19].

Proposition 7. *[6], [19] Assume that ω is extremal and $X_{\mathbf{s}}\in\mathfrak{h}'$, then $D_fF|_{(X_{\mathbf{s}},0)}$, the linearization of F with respect to f at $(X_{\mathbf{s}},0)$ is equal to $-\frac{1}{2}\,P_\omega^*\,P_\omega$.*

Since this is one of the key points of our construction, let us briefly explain the proof of this result. We already know the linearization of the scalar curvature map, so we only need to know the linearization of

$$f\longmapsto\xi_{\omega+i\partial\bar\partial f}\,,$$

with respect to f. Take any $X\in\mathfrak{h}'$. Since f is K-invariant, X is also a Killing vector field (with zeros) for the Kähler form $\omega+i\partial\bar\partial f$. Hence, we can write

$$\tfrac{1}{2}\,(\omega+i\partial\bar\partial f)(\Xi,\cdot)=-\bar\partial\langle\xi_{\omega+i\partial\bar\partial f},X\rangle\,,$$

where $\Xi:=X-i\,J\,X$, and we see immediately that $\dot\xi$, the first variation of $f\longmapsto\xi_{\omega+i\partial\bar\partial f}$ with respect to f computed at $f=0$, satisfies

$$\tfrac{i}{2}\,\partial\bar\partial\,f(\Xi,\cdot)=-\bar\partial\,\langle\dot\xi,X\rangle\,.$$

Working in local coordinates and using the fact that the vector field Ξ is holomorphic we find

$$\bar\partial\,(\tfrac{i}{2}\,(\Xi\,f)+\langle\dot\xi,X\rangle)=0\,.$$

Since, by definition, the function $\langle\dot\xi,X\rangle+\frac{i}{2}\,(\Xi\,f)$ is real valued and has mean 0, we conclude that

$$\langle\dot\xi,X\rangle=-\tfrac{i}{2}\,\Xi\,f\,.$$

Now, we apply this analysis when ω is extremal, with extremal vector field $X_{\mathbf{s}}\in\mathfrak{h}'$. We obtain for any smooth function f

$$D_fF|_{(X_{\mathbf{s}},0)}(f)=\mathbb{L}_\omega f+\tfrac{i}{2}\,\Xi_{\mathbf{s}}\,f\qquad\text{with}\qquad\Xi_{\mathbf{s}}:=X_{\mathbf{s}}-i\,J\,X_{\mathbf{s}}\,.$$

Hence

$$D_fF|_{(X_{\mathbf{s}},0)}(f)=-\tfrac{1}{2}\,P_\omega^*\,P_\omega f-\tfrac{1}{2}\,J\,X_{\mathbf{s}}\,f+\tfrac{i}{2}\,X_{\mathbf{s}}\,f+\tfrac{i}{2}\,\Xi_{\mathbf{s}}\,f=-\tfrac{1}{2}\,P_\omega^*\,P_\omega\,f+i\,X_{\mathbf{s}}\,f\,.$$

Remembering that when f is K-invariant and $X_{\mathbf{s}} \in \mathfrak{h}'$, we have

$$X_{\mathbf{s}} f = 0\,,$$

and we conclude that $D_f F|_{(X_{\mathbf{s}},0)}(f) = -\frac{1}{2}\, P_\omega^*\, P_\omega\, f$. This completes the proof of the Proposition.

3.2. The origin of the constraints. We are now in a position to explain where the constraints in the statement of Theorem 1 come from.

We denote by $\Xi_{\mathbf{s}}$ the holomorphic vector field associated to $X_{\mathbf{s}}$. We also assume that we have chosen some compact subgroup of isometries $K \subset \mathrm{Isom}(M, g)$ and some finite set of points $p_1, \ldots, p_n \in \mathrm{Fix}(K_0)$. Since we want to work equivariantly with respect to K, we assume that $\{p_1, \ldots, p_n\}$ is invariant under the action of K. We note that a holomorphic vector field Ξ lifts to $\tilde{M}$, the blow up of M at $p_1, \ldots, p_n$, if and only if Ξ vanishes at each of the points p_j. As already mentioned, our construction of the extremal Kähler metric on the blow up being based on a perturbation argument, it is natural to require that $\Xi_{\mathbf{s}}$ vanishes at all points $p_1, \ldots, p_n$ and hence will lift to $\tilde{M}$.

Now, if $\tilde{\omega}$ is a putative extremal Kähler metric on $\tilde{M}$ its scalar curvature must be a sum of K-invariant potentials corresponding to vector fields which vanish at the p_j, are K-invariant and are associated to isometries of the new metric, hence they have to correspond to the lift of vector fields which are both in $\mathfrak{h}$ and $\mathfrak{k}$. Thus we introduce the lie algebra $\mathfrak{h}' \subset \mathfrak{h}$ defined by

$$\mathfrak{h}' := \mathfrak{h} \cap \mathfrak{k}\,.$$

In particular, elements of $\mathfrak{h}'$ vanish at all points $p_1, \ldots, p_n$.

We denote by $\mathfrak{h}''$ the orthogonal complement of $\mathfrak{h}'$ in $\mathfrak{h}$ with respect to the scalar product

$$(X, \tilde{X})_{\mathfrak{h}} := \int_M \langle \xi_\omega, X\rangle\, \langle \xi_\omega, \tilde{X}\rangle\, \mathrm{dvol}_\omega\,.$$

Informally, the potentials of the form $\langle \xi_\omega, X\rangle$, for $X \in \mathfrak{h}'$, will correspond to the *good potentials* which are associated to vector fields lifting to $\tilde{M}$ (since they vanish at all points $p_1, \ldots, p_n$). In particular, these can be used to deform the scalar curvature of the Kähler form. In contrast, the potentials of the form $\langle \xi_\omega, X\rangle$, when $X \in \mathfrak{h}''$, will correspond to the *bad potentials* corresponding to vector fields which do not lift to $\tilde{M}$. Hence, these are the potentials which cannot be used in the deformation of the scalar curvature of the Kähler form.

To apply a perturbation argument, we need to solve two linear problems. First, we will need to find a function Γ, a constant λ and a vector field $Y \in \mathfrak{h}'$ solutions of

$$\tfrac{1}{2}\, P_\omega^*\, P_\omega \Gamma + \langle \xi_\omega, Y\rangle + \lambda = -c_m \sum_{j=1}^{n} a_j^{m-1}\, \delta_{p_j}\,, \tag{2}$$

where the masses a_j are positive and $c_m > 0$ is a positive constant only depending on the dimension m. The solvability of this problem is equivalent to

the *relative moment condition*:

$$\sum_{j=1}^{n} a_j^{m-1}\, \xi_\omega(p_j) \in \mathfrak{h}'^{\,*}\,. \tag{3}$$

This is precisely (ii) in the statement of Theorem 1. Observe that the parameters a_j and $a_{j'}$ corresponding to points p_j and $p_{j'}$ in the same orbit with respect to the action of K should be equal to preserve the K-invariance of the metric we will construct.

Using this, we consider a first perturbation of ω, away from the points to be blown up. This perturbed Kähler form we consider is given explicitly by

$$\hat{\omega}_\varepsilon := \omega + i\,\partial\,\bar{\partial}(\varepsilon^{2m-2}\,\Gamma)\,,$$

where $\varepsilon > 0$ is a small parameter. This Kähler form is well defined away from balls of radius $c\,\varepsilon$ centered at the points p_j (provided c is fixed large enough and ε is chosen small enough) and one can check that the associated Kähler metric has scalar curvature given by

$$\mathbf{s}(\hat{\omega}_\varepsilon) = \mathbf{s}(\omega) + \varepsilon^{2m-2}\,\left(\langle \xi_\omega, Y\rangle + \lambda\right) + \mathcal{O}(\varepsilon^{4m-2})\,.$$

The final task will be to perturb this Kähler metric into an extremal metric. To this aim, given any (smooth) function f, we need to be able to find a function ϕ, a constant ν, a vector field $Z \in \mathfrak{h}'$ and parameters $b_j \in \mathbb{R}$ solutions of

$$\tfrac{1}{2}\,P_\omega^*\,P_\omega\,\phi + \nu + \langle \xi_\omega, Z\rangle + c_m \sum_{j=1}^{n} b_j\,\delta_{p_j} = f\,. \tag{4}$$

The solvability of this problem is precisely equivalent to the *genericity condition*:

$$\text{The projections of } \xi_\omega(p_1), \ldots, \xi_\omega(p_n) \text{ on } \mathfrak{h}''^{\,*} \text{ spans } \mathfrak{h}''^{\,*}\,. \tag{5}$$

This is precisely (i) in the statement of Theorem 1.

The idea is now to proceed to connected sums of M with n copies of $\tilde{\mathbb{C}}^m$, the blow up at the origin of $\mathbb{C}^m$, endowed with a rescaled copy of a scalar flat Kähler metric g_0 which we describe now. The metric g_0 is $U(m)$ invariant and was found by D. Burns [18], when $m = 2$, and S. Simanca [29], when $m \geq 3$, following a method introduced in [6]. Away from the exceptional divisor, the Kähler form η associated to this metric is given by

$$\eta = i\,\partial\,\bar{\partial}\Phi_m(v)\,,$$

where $v = (v^1, \ldots, v^m)$ are complex coordinates in $\mathbb{C}^m \setminus \{0\}$ and where the function Φ_m is explicitly given, in dimension $m = 2$, by

$$\Phi_2(v) := \tfrac{1}{2}\,|v|^2 + \log|v|^2\,,$$

while in dimension $m \geq 3$, even though there is no explicit formula for E_m we have the following expansion

$$\Phi_m(v) = \tfrac{1}{2}\,|v|^2 - |v|^{4-2m} + \mathcal{O}(|v|^{2-2m})\,,$$

as $|v|$ tends to ∞. Details can be obtained either in [6], [29] or in [1] for a proof of this expansion. The scalar flat metric g_0 which is used for the connected sum at the point p_j has to be multiplied $a_j\,\epsilon^2$ so that the asymptotics of the metrics $a_j\,\epsilon^2\,\eta$ and ω_ϵ do match. Indeed, in dimension $m \geq 3$, the Kähler form $\hat{\omega}_\varepsilon$ can be expanded near p_j as

$$\hat{\omega}_\varepsilon = i\,\partial\,\bar{\partial}\left(\tfrac{1}{2}|z|^2 - \varepsilon^{2m-2}\,a_j\,|z|^{4-2m} + \ldots\right)\,,$$

while the Kähler form $a\,\varepsilon^2\,\eta$ can be expanded as

$$\begin{aligned} a\,\varepsilon^2\,\tilde{\eta} &= i\,\varepsilon^2\,a\,\partial\,\bar{\partial}\,\left(\tfrac{1}{2}\,|v|^2 - |v|^{4-2m} + \ldots\right) \\ &= i\,\partial\,\bar{\partial}\left(\tfrac{1}{2}\,|z|^2 - \varepsilon^{2m-2}\,a\,|z|^{4-2m} + \ldots\right)\,, \end{aligned}$$

after the change of variables $z = \varepsilon\,a\,v$. And hence, the asymptotics do match provided we choose $a = a_j$.

The rest of the proof of Theorem 1 is to show that these conditions and the construction outlined above are indeed *sufficient* to guarantee that a perturbation argument implies the existence of extremal metrics in the appropriate classes. This part of the proof is rather technical and uses in a crucial way the analysis in weighted function spaces of elliptic operators on some class of complete non-compact manifolds introduced by R.B. Lockhart and R.C McOwen [23], R. Melrose [27] and R. Mazzeo [26].

References

[1] C. Arezzo and F. Pacard, *Blowing up and desingularizing Kähler orbifolds with constant scalar curvature*, Acta Mathematica 196, no 2, (2006), 179–228.

[2] C. Arezzo and F. Pacard, *Blowing up Kähler manifolds with constant scalar curvature II*, Annals of Math. (2), 170, no 2, (2009), 685–738.

[3] C. Arezzo and F. Pacard, *On the Kähler classes of constant scalar curvature metrics on blow ups*. Aspects analytiques de la géométrie riemannienne. Série Séminaires et Congrès (SMF).

[4] C. Arezzo, F. Pacard and M. Singer *On the Kähler classes of extremal metric on blow ups*, arXiv:0706.1838

[5] A. Besse, *Einstein manifolds*. Springer-Verlag, Berlin, (1987).

[6] E. Calabi, *Extremal Kähler metrics*, Annals of Math. Studies **102** , Princeton Univ. Press, (1982), 269–290.

[7] E. Calabi, *Extremal Kähler metrics II*, in Differ. Geometry and its Complex Analysis, edited by I. Chavel and H.M. Farkas, Springer, (1985).

[8] X.X. Chen, C. LeBrun and B. Weber, *On Conformally Kähler, Einstein Manifolds*, J. Amer. Math. Soc. 21, no. 4, (2008), 1137–1168.

[9] X.X. Chen and G. Tian, *Geometry of Kähler metrics and holomorphic foliation by discs*, Publ. Math. Inst. Hautes Études Sci. No. 107 (2008), 1–107.

[10] A. Della Vedova, *CMstability of blowups and canonical metrics*, arXiv:0810.5584.

[11] S.K. Donaldson, *Scalar curvature and stability of toric varieties*, J. Differential Geom. **62**, (2002), 289–349.

[12] S.K. Donaldson, *Interior estimates for Abreu's equation*, Collect. Math. **56**, (2005), 103–142.

[13] S.K. Donaldson *Extremal metrics on toric surfaces: a continuity method*, J. Differential Geom. **79** (2008), 389–432.

[14] S.K. Donaldson *Constant scalar curvature metrics on toric surfaces*, Geom. Funct. Analysis **19**, (2009), 83–136.

[15] V. Guillemin, *Moment maps and combinatorial invariants of Hamiltonian T^n-spaces*, Progress in Mathematics **122**, Birkhäuser, (1994).

[16] W. He, *Remarks on the existence of bilaterally symmetric extremal Kähler metrics on* $\mathbb{CP}^2\sharp 2\bar{\mathbb{CP}}^2$, Int. Math. Res. Not. IMRN (2007), no. 24.

[17] J. Kim, C. LeBrun and M. Pontecorvo, *Scalar-flat Kähler surfaces of all genera*, J. Reine Angew. Math. **486**, (1997), 69–95.

[18] C. LeBrun, *Counter-examples to the generalized positive action conjecture* Comm. Math. Phys. **118**, (1988), 591–596.

[19] C. LeBrun and S. Simanca, *Extremal Kähler metrics and complex deformation theory*, Geom. Funct. Anal. 4, (1994), 298–336.

[20] C. LeBrun and S. Simanca, *On the Kähler classes of extremal metrics*, in Geometry and Global Analysis (Sendai, 1993), Tohoku Univ., 255–271.

[21] C. LeBrun and M. Singer, *Existence and deformation theory for scalar flat Kähler metrics on compact complex surfaces* Invent. Math. **112**, (1993), 273–313.

[22] M. Levine, *A remark on extremal Kähler metrics*, J. Differential Geom. **21**, (1985), 73–77.

[23] R.B. Lockhart and R.C McOwen, *Elliptic differential operators on noncompact manifolds.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) 1, no. 3, (1985), 409–447 .

[24] T. Mabuchi, *Einstein-Kähler forms, Futaki invariants and convex geometry on toric Fano varieties.* Osaka J. Math. **24**, no. 4, (1987), 705–737.

[25] T. Mabuchi, *Stability of extremal Kähler manifolds.* Osaka J. Math. **41**, no. 3, (2004), 563–582.

[26] R. Mazzeo, *Elliptic theory of edge operators I.* Comm. in PDE, **10**, (1991), 1616–1664.

[27] R. Melrose, *The Atiyah-Patodi-Singer index theorem*, Research notes in Math, **4**, (1993).

[28] Y. Rollin and M. Singer, *Non-minimal scalar-flat Kaehler surfaces and parabolic stability*, Invent. Math. **162**, (2005), 235–270.

[29] S. Simanca, *Kähler metrics of constant scalar curvature on bundles over* $C\mathrm{P}^{n-1}$, Math. Ann. **291**, (1991), 239–246.

[30] J. Stoppa, *Unstable blowups.* J. Algebraic Geom. 19, no. 1, (2010), 1–17.

[31] G. Székelyhidi, *Extremal metrics and K-stability*, Bull. Lond. Math. Soc. 39, (2007), 76–84.

[32] R. Thomas, *Notes on GIT and symplectic reduction for bundles and varieties*, Surveys in differential geometry. Vol. X, 221–273, Surv. Differ. Geom., 10, Int. Press, Somerville, MA, (2006).

[33] G. Tian, *Extremal metrics and geometric stability*, Houston J. Math. **28**, (2002), 411–432.

[34] G. Tian, *Recent progress on Khler-Einstein metrics. Geometry and physics*, (Aarhus, 1995), 149–155, Lecture Notes in Pure and Appl. Math., 184, Dekker, New York, (1997).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Reconstruction of Collapsed Manifolds

Takao Yamaguchi*

Abstract

In this article, we consider the problem of reconstructing collapsed manifolds in a moduli space by means of geometric or analytic data of the limit spaces. The moduli space of our main interest is that consisting of closed Riemannian manifolds of fixed dimension with a lower sectional curvature and an upper diameter bound. In this moduli space, we can reconstruct the topology of three-dimensional or four-dimensional collapsed manifolds in terms of the singularities of the limit Alexandrov spaces. In the general dimension, we define a new covering invariant and prove the uniform boundedness of it with an application to Gromov's Betti number theorem. Finally we discuss the reconstruction and stability problems of collapsed manifolds by using analytic spectral data, where we assume an additional upper sectional curvature bound.

Mathematics Subject Classification (2010). Primary 53C20; Secondary 58J50.

Keywords. Gromov-Hausdorff convergence, collapsing, three-manifolds, four-manifolds, essential coverings, Betti numbers, inverse spectral problem

1. Introduction

The study of Gromov-Hausdorff convergence of Riemannian manifolds has been a significant subject in differential geometry. In this theory one usually considers a moduli space of closed Riemannian manifolds satisfying certain curvature bounds, and try to reconstruct geometry and topology of Riemannian manifolds in the moduli space from the information on the limit spaces.

When the absolute value of sectional curvature sec is uniformly bounded, say $|\sec| \leq 1$, Cheeger, Fukaya and Gromov [CFG] developed a general theory of collapsing, where the collapsing phenomena were described in terms of the generalized group actions by nilpotent groups, called N-structures. It should be noted that these actions are not permitted to have fixed points. Now if we turn

*Institute of Mathematics, University of Tsukuba, Tsukuba, 305-8571, Japan.
E-mail: takao@math.tsukuba.ac.jp.

attention to the case when the sectional curvature has only a uniform lower bound, $\sec \geq -1$, we recognize several kinds of essentially different collapsing phenomena in this situation. For example, any effective action on a compact manifold by a compact connected Lie group of positive dimension causes a collapsing of the manifold under $\sec \geq -1$ ([Y1]) . Thus the study of collapsing of Riemannian manifolds with $\sec \geq -1$ will enable us to understand a wider class of collapsing phenomena than the case of $|\sec| \leq 1$.

A main concern of this article is the study of collapsed Riemannian manifolds with $\sec \geq -1$, and we will survey some aspects of the development in this direction.

In lower dimensions, dimension three or four, the structure of collapsed manifolds with a lower curvature bound has become clear by Shioya-Yamaguchi [SY1], [SY2] and Yamaguchi [Y3]. In the solution of the geometrization conjecture of three-manifolds due to Perelman [P4], [P5], a related result on collapsing three-manifolds with local lower sectional curvature bounds was essentially used. Namely Perelman proved that under the Ricci flow with (well controlled) surgery on every closed three-manifold, after the passage of a long time, the three-manifold is decomposed into the two parts along incompressible tori: the non-collapsing hyperbolic parts and the collapsing parts under local lower curvature bounds. Then one can determine the topology of the latter part applying the structure result for collapsed three-manifolds.

Unfortunately, in the general dimension, it is still open to get general picture of collapsing. In stead, we will define a new geometric covering invariant of Riemannian manifolds, and show the uniform boundedness of this invariant from the view point of the Gromov-Hausdorff convergence ([Y5]). This provides a clearer view for the proof of Gromov's Betti number theorem.

Finally, we will discuss the inverse spectral problem for collapsed manifolds. This is the problem to reconstruct collapsed manifolds from certain spectral data concerning the Laplace-Beltrami operator. At this stage, we need both lower and upper sectional curvature bounds for this problem to ensure a certain regularity of the limit spaces in order to carry out some analysis.

2. Gromov-Hausdorff Convergence

For compact subsets A and B of a metric space Z, the classical Hausdorff distance $d_H^Z(A,B)$ between A and B is defined by the infimum of those $\varepsilon > 0$ such that the ε-neighborhood of A contains B and the ε-neighborhood of B contains A. Let $\mathcal{C}$ denote the set of all isometry classes of compact metric spaces. For $X, Y \in \mathcal{C}$, the Gromov-Hausdorff distance $d_{GH}(X,Y)$ between X and Y is defined as

$$d_{GH}(X,Y) = \inf_{Z,f,g} d_H^Z(f(X), g(Y)),$$

where the infimum is taken over all possible isometric embeddings $f : X \to Z$, $g : Y \to Z$ together with all possible metric spaces Z.

Given a positive integer n and $D > 0$, let us consider the set $\mathcal{M}(n, D)$ of isometry classes of n-dimensional closed Riemannian manifolds M whose sectional curvature and diameter satisfy

$$\sec(M) \geq -1, \qquad \operatorname{diam}(M) \leq D.$$

The lower curvature bound $\sec(M) \geq -1$ implies that the geometry of M, or the rate of expanding of M, is bounded by that of hyperbolic space $H^n(-1)$ of constant curvature -1 in the sense of geodesic deviation. This yields the following precompactness theorem due to Gromov [GLP]:

Theorem 2.1 ([GLP]). *$\mathcal{M}(n, D)$ is relatively compact with respect to the Gromov-Hausdorff distance.*

Thus it is quite natural to consider a sequence M_i, $i = 1, 2, \ldots$, in $\mathcal{M}(n, D)$ which converges to a compact metric space X in the closure $\overline{\mathcal{M}}(n, D)$ with respect to the Gromov-Hausdorff distance. The boundedness of the geometry of the manifolds in $\mathcal{M}(n, D)$ yields that X is an *Alexandrov space with curvature* ≥ -1 having dimension $\leq n$. In such an Alexandrov space, every geodesic triangle is thicker than a comparison triangle in the hyperbolic plane $H^2(-1)$ having the same side-lengths.

Problem 2.2. Let a sequence M_i in $\mathcal{M}(n, D)$ converge to an Alexandrov space X. Then find topological, geometrical or analytical relations between M_i and X for sufficiently large i.

In other words, this is a problem to reconstruct manifolds M_i using geometric data containing the singularities of X or analytic data of X.

3. Basic Results

Some answers to Problem 2.2 are known in several cases as stated in the following. Fibration Theorem 3.1 was established by Yamaguchi [Y1], Fukaya-Yamaguchi [FY], and Stability Theorem 3.2 was established by Perelman [P1] (see also Kapovitch [Kp]). Both play fundamental roles in the study of the Gromov-Hausdorff convergence in $\mathcal{M}(n, D)$.

Theorem 3.1 ([Y1], [FY]). *If X has no 'singular points', then there exists a fibration $F_i \hookrightarrow M_i \to X$, where the fiber F_i satisfies*

(1) *the first Betti number $b_1(F_i)$ is not greater than the dimension $\dim F_i$;*

(2) *the fundamental group $\pi_1(F_i)$ contains a nilpotent subgroup of finite index.*

Theorem 3.2 ([P1], cf. [Kp]). *If $\dim X = n$, then M_i is homeomorphic to X.*

Now we shortly recall the geometry of Alexandrov spaces with curvature bounded below that was mainly established by Burago, Gromov and Perelman ([BGP]).

Let X be a finite-dimensional Alexandrov space with curvature bounded below. For every point $p \in X$, the notion of the *space of directions at* p, denoted by Σ_p, is defined. The Euclidean cone K_p over Σ_p is called the *tangent cone at* p and coincides with the blow-up limit:

$$K_p = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(X, p).$$

Let $k := \dim X$. Since Σ_p becomes a $(k-1)$-dimensional compact Alexandrov space with curvature ≥ 1, it is a smaller than or equal to the unit sphere $S^{k-1}(1)$ as metric spaces. If $\Sigma_p = S^{k-1}(1)$, p is called a *regular* point. Otherwise it is called a *singular* point. Let $S(X)$ denote the set of all singular points of X. The Hausdorff dimension of $S(X) \cap \partial X$ is at most $k-2$ (see also Otsu and Shioya [OS]), where the boundary ∂X of X is defined inductively in terms of the spaces of directions.

One of the difficulties of the study of Alexandrov spaces is that $S(X)$ could be dense in X. If one considers the set of 'almost regular' points, denoted by $R_0(X)$, then it is an open subset of full measure, and each point of $\mathcal{R}_0$ has a neighborhood almost isometric to an open subset of $\mathbb{R}^k$.

The following result is related with Stability Theorem 3.2.

Theorem 3.3 ([P1],[P2], cf. [Kp]). *The local structure of X is described as follows:*

(1) *A small metric ball around every point $p \in X$ is homeomorphic to K_p;*

(2) *X has a topological stratification, namely a sequence of closed subsets,*

$$X = S_0(X) \supset S_1(X) \supset \cdots \supset S_\ell(X) = \phi,$$

such that each $S_j(X) \setminus S_{j+1}(X)$ is a topological manifolds.

Finally we define a few notions concerning singular points of X. A point $p \in X$ is called an *extremal point* if the diameter $\mathrm{diam}(\Sigma_p)$ is at most $\pi/2$, and p is called *an essential singular point* if the radius $rad(\Sigma_p)$ defined by

$$\mathrm{rad}(\Sigma_p) := \min_{\xi \in \Sigma_p} \left(\max_{\eta \in \Sigma_p} d(\xi, \eta) \right),$$

is at most $\pi/2$. Note that extremal points are essential singular points and isolated.

Theorem 3.4 ([P3]). *If X has no 'bad singularities' called extremal subsets, then there is an isomorphism $\pi_k(M_i, F_i) \simeq \pi_k(X)$ for homotopy groups, where F_i is a general fiber and i is large enough compared with k.*

4. Reconstruction of Low-dimensional Collapsed Manifolds

In this paragraph, we discuss topological reconstruction of three-dimensional or four-dimensional collapsed manifolds. In view of Fibration Theorem 3.1, we only have to focus the case when the limit space X has singular points. If X is one-dimensional, this is the case when X is an arc.

We denote by D^k, S^k, T^2, P^2, K^2 and I, a disk and a sphere of dimension k, a torus, a projective plane, a Klein bottle and a closed interval respectively.

The topology of collapsed three-manifolds in $\mathcal{M}(3, D)$ can be reconstructed by the following result due to Shioya-Yamaguchi [SY1]:

Theorem 4.1 ([SY1]). *Suppose that a sequence of closed orientable three-manifolds M_i in $\mathcal{M}(3, D)$ collapses to a space X with* $\dim X \in \{1, 2\}$.

(1) *Assume* $\dim X = 2$. *If X has no boundary, then M_i is homeomorphic to a Seifert fibered space over X. If X has nonempty boundary, then M_i is a gluing of a Seifert fibered space over X and $\partial X \times D^2$ glued along their boundaries.*

(2) *Assume that X is one-dimensional and an arc. Then M_i is homeomorphic to a gluing of D^3, $P^2 \widetilde{\times} I$, or a gluing of $S^1 \times D^2$, $K^2 \widetilde{\times} I$, along their boundaries (spheres or tori), where $\widetilde{\times}$ denotes the twisted product.*

A closed three-manifold is called a *graph manifold* if it is a finite gluing of Seifert fibered spaces along their boundary tori. If one drops the diameter bound, the topology of collapsed three-manifold under a lower curvature bound can be reconstructed as follows:

Theorem 4.2 ([SY2]). *There exist small positive numbers ϵ_0 and δ_0 such that if an orientable three-manifold M has a complete Riemannian metric whose sectional curvature and volume satisfy* $\sec(M) \geq -1$ *and* $\mathrm{vol}(M) < \epsilon_0$, *then one of the following holds*:

(1) *M is homeomorphic to a graph manifold*;

(2) $\mathrm{diam}(M) < \delta_0$ *and M has finite fundamental group.*

In [P4],[P5], Perelman states that the conclusion of Theorem 4.2 holds under a weaker assumption of collapse with local lower curvature bounds (see [SY2], and also recent works [MT], [CaGe]).

Next we turn to the reconstruction problem of collapsed four-manifolds in $\mathcal{M}(4, D)$:

Theorem 4.3 ([Y3]). *Suppose that a sequence of closed orientable four-manifolds M_i in $\mathcal{M}(4, D)$ collapses to a space X with* $1 \leq \dim X \leq 3$.

Then M_i has a singular fiber structure over X in a generalized sense. More precisely,

(1) *If* $\dim X = 3$*, then there exists a locally smooth, local S^1-action on M_i such that the orbit space M_i/S^1 is homeomorphic to X;*

(2) *Suppose* $\dim X = 2$*. If X has no boundary, then M_i is homeomorphic to either an S^2-bundle, or a Seifert T^2-bundle over X. If X has non-empty boundary, then we have a singular fibration $f_i : M_i \to X$ such that f_i has the same fiber structure over* $\operatorname{int} X = X \setminus \partial X$ *as above and the fibers over ∂X are ones of $\{point, S^1, S^2, S^3, P^2, K^2\}$;.*

(3) *Suppose X is one-dimensional and an arc. Then M_i is the result of a gluing of at least two and at most four disk-bundles.*

Remark 4.4. In Theorem 4.3, the fiber type does not change along each stratum of a stratification of X, but may change when the stratum changes. The singularity of a singular fiber over a point $p \in X$ can be sharply estimated by the singularity at p. In particular, in the case of $\dim X = 3$, the singular locus in X of the local S^1-action consists of ∂X extremal points and quasigeodesics consisting of essential singular points in $\operatorname{int} X$. Quasigeodesics are generalization of geodesics. See Perelman and Petrunin [PP], Petrunin [P] for the construction and basic properties of quasigeodesics. In any case, M_i is a fiber bundle over X if X has no essential singular points.

As a very special example, let us suppose the case when X is homeomorphic to D^3 and there exists a unique essential singular point p in the interior of X which is an extremal point. Then M_i is one of the following:

$$S^4 = D^3 \times S^1 \cup S^2 \times D^2, \qquad \mathbb{C}P^2 = D^4 \cup_{S^3} S^2 \widetilde{\times} D^2.$$

As a conclusion of Theorem 4.3 together with Stability Theorem 3.2, we have a description of the homeomorphism classes in $\mathcal{M}(4, D)$ as follows:

Corollary 4.5 ([Y3]). *For a given $D > 0$, there exist finitely many elements $N_1, \dots, N_k$ of $\mathcal{M}(4, D)$, where $k = k(D)$, such that for any element M of $\mathcal{M}(4, D)$ one of the following holds:*

(1) *M is homeomorphic to one of $\{N_1, \dots, N_k\}$;*

(2) *M is homeomorphic to a closed 4-manifold as described in Theorem 4.3.*

Remark 4.6. Theorem 4.3 is stated in terms of homeomorphism classes since we essentially use Stability Theorem 3.2 in the proof. In dimension four, there are big differences between homeomorphism classes and diffeomorphism classes. If one could have the Lipschitz version of Stability Theorem 3.2, then Theorem 4.3 would be stated in terms of bi-Lipschitz homeomorphism classes, at least.

The above results support the following conjecture in the general dimension.

Conjecture 4.7. Suppose a sequence M_i in $\mathcal{M}(n, D)$ collapses to X. Then there is a singular fibration $f_i : M_i \to X$ in some generalized sense.

5. Essential Coverings

There are some relations between a covering and the topology of a manifold. In particular, if one considers coverings by contractible metric balls of a closed Riemannian manifold, the minimal number of such balls seems to represent a topological complexity of the manifold. Actually it was an underlying essential idea in the proofs of finiteness theorems of Riemannian manifolds.

For given n, $D, v > 0$, let $\mathcal{M}(n, D, v)$ denote the family of n-dimensional closed Riemannian manifolds M satisfying

$$\sec(M) \geq -1, \ \operatorname{diam}(M) \leq D, \ \operatorname{vol}(M) \geq v.$$

After the pioneering works due to Cheeger [C] and Weinstein [W], Grove, Peterson and Wu [GPW] and finally Perelman [P1] proved the following finiteness theorem, which is a direct consequence of Precompactness Theorem 2.1 and Stability Theorem 3.2

Theorem 5.1. *The set of homeomorphism classes of the manifolds in $\mathcal{M}(n, D, v)$ is finite.*

The basic idea behind the proof of Theorem 5.1 is to find a uniform bound for the minimal number of contractible metric balls needed to cover the manifolds in the family. This becomes possible because we work with the non-collapsing family $\mathcal{M}(n, D, v)$.

If one considers the collapsing family $\mathcal{M}(n, D)$, where we have no lower volume bound, obviously it is impossible to find such a uniform bound. In stead, we define a *system* of metric balls to cover a collapsed manifold in an efficient way in place of just one covering by contractible balls. This will lead us to the notion of a contractible essential covering.

To illustrate the notion of contractible essential covering, let us take the flat torus $T^2_\epsilon = S^1(1) \times S^1(\epsilon)$ for a small $\epsilon > 0$. The torus T^2_ϵ can be covered by two thin metric balls B_α, $\alpha \in \{1, 2\}$. Each ball B_α is isotopic to a much smaller concentric metric ball $\widehat{B}_\alpha$ of radius, say 2ϵ. If one tries to cover B_α by contractible metric balls, we need too many, about $[1/\epsilon]$-pieces of such balls. In stead, we take a covering of $\widehat{B}_\alpha$. It is possible to cover $\widehat{B}_\alpha$ by two contractible metric balls $\{B_{\alpha\beta}\}_{\beta\in\{1,2\}}$. Thus we have a collection of four contractible metric balls $\{B_{\alpha\beta}\}_{\alpha\beta\in\{1,2\}}$, which will be called a contractible essential covering of T^2_ϵ. Although it is not a usual covering of T^2_ϵ, deforming and enlarging $B_{\alpha\beta}$ by isotopies, we obtain a covering $\{\widetilde{B}_{\alpha\beta}\}_{\alpha\beta\in\{1,2\}}$ of T^2_ϵ by contractible open

subsets. In that sense, the contractible essential covering seems to contain an essential feature of T^2_ϵ.

Now we describe the general definition of a contractible essential covering.

Definition 5.2. Let M be a closed n-dimensional Riemannian manifold. We first begin with a covering of M by metric balls $\{B_{\alpha_1}\}_{\alpha_1=1}^N$ that are not necessary contractible. For each non-contractible ball B_{α_1}, we try to find a smaller concentric ball $\widehat{B}_{\alpha_1}$ isotopic to B_{α_1}, and consider a covering of $\widehat{B}_{\alpha_1}$ in stead of B_{α_1} by much smaller metric balls $\{B_{\alpha_1\alpha_2}\}_{\alpha_2}$. We repeat this procedure for these small metric balls $B_{\alpha_1\alpha_2}$. After finitely many repeats, we will get to contractible balls. Suppose that in this way we get a finite collection of metric balls,

$$\mathcal{B} = \{B_{\alpha_1\cdots\alpha_k}\},$$

consisting of balls $B_{\alpha_1\cdots\alpha_k}$ of M, such that

(1) $B_{\alpha_1} \supset B_{\alpha_1\alpha_2} \supset \cdots \supset B_{\alpha_1\cdots\alpha_k} \supset \cdots$;

(2) $B_{\alpha_1\cdots\alpha_k}$ is isotopic to a smaller concentric ball $\widehat{B}_{\alpha_1\cdots\alpha_k}$;

(3) $\{B_{\alpha_1\cdots\alpha_k\alpha_{k+1}}\}_{\alpha_{k+1}}$ covers $\widehat{B}_{\alpha_1\cdots\alpha_k}$;

(4) At every terminal of a sequence $\alpha_1 \to \alpha_1\alpha_2 \to \cdots \to \alpha_1\alpha_2\cdots\alpha_k \to \cdots$ satisfying the above (1), (2) and (3), we have a contractible ball.

Note that the size of the balls becomes smaller and smaller. The collection $\widehat{\mathcal{B}}$ of those contractible balls appearing at the ends of the sequences in the above (4) is called a *contractible essential covering* of the manifold M.

From construction, there is a tree T associated with the system $\mathcal{B}$. The maximal number of edges in the simple paths from the top vertex of T (corresponding to the manifold M) to the bottom terminal points of T (corresponding to the contractible balls which appear in the above (4)) is called the *depth* of the contractible essential covering $\widehat{\mathcal{B}}$.

We denote by $\tau_m(M)$ the minimal number of contractible balls in all the essential coverings of M of depth at most m. This provides a new geometric invariant of M.

Theorem 5.3 ([Y5]). *For given n and D, there is a positive constant $C_n(D)$ such that $\tau_n(M) \le C_n(D)$ for all M in $\mathcal{M}(n, D)$.*

Corollary 5.4 ([Y5]). *For given n, there is a positive constant C_n such that if M has nonnegative sectional curvature, then $\tau_n(M) \le C_n$.*

For instance, let us take the n-dimensional flat torus

$$T^n(\varepsilon) = S^1(1) \times S^1(\varepsilon) \times S^1(\varepsilon^2) \times \cdots \times S^1(\varepsilon^{n-1}).$$

In a way similar to the previous example of $n = 2$, we easily have $\tau_n(T^n(\varepsilon)) \leq 2^n$ and $\tau_{n-1}(T^n(\varepsilon)) \to \infty$ as $\varepsilon \to 0$. There are examples of n-dimensional nilmanifolds (N, g_ε) with almost flat metrics ([G1]) having similar estimates: $\tau_n(N, g_\varepsilon) \leq 2^n$ and $\tau_{n-1}(N, g_\varepsilon) \to \infty$ as $\varepsilon \to 0$.

Conjecture 5.5. If M has nonnegative sectional curvature, then $\tau_n(M) \leq 2^n$.

In general, collapsed manifolds are expected to have singular fiber structure over the limit spaces in a generalized sense (Conjecture 4.7). The fiber may shrink to a point with different order in different independent directions like $T^n(\varepsilon)$ or (N, g_ε). This explains an essential reason why we need the depth n for the uniform bound.

Together with Gromov's topological lemma in [G2], Theorem 5.3 yields the following uniform bound on the total Betti numbers, which was originally proved by Gromov ([G2]).

Corollary 5.6. *For given n and D, there is a positive integer $C(n, D)$ such that if M is in $\mathcal{M}(n, D)$, then*

$$\sum_{i=0}^{n} \operatorname{rank} H_i(M; F) \leq C(n, D),$$

where F is any field.

In the original work [G2], Gromov developed the critical point theory ([GS]) for distance functions to obtain an explicit bound on the total Betti numbers. Unfortunately our bound is not explicit. However our approach provides a conceptually clearer view showing what the essence of Corollary 5.6 is like.

Concerning the total Betti numbers of nonnegatively curved manifolds, the following conjecture is known: If an n-dimensional closed Riemannian manifold M has nonnegative sectional curvature, then

$$\sum_{i=0}^{n} \operatorname{rank} H_i(M; F) \leq 2^n.$$

Probably there will be some relation between this conjecture and Conjecture 5.5.

6. Reconstruction by Spectral Data

In this paragraph, we discuss the reconstruction problem of collapsed manifolds by spectral data. This is recent joint works [KLY1], [KLY2] with Y. Kurylev and M. Lassas.

Let Δ be the Laplace-Beltrami operator on a compact Riemannian manifold M, and

$$0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots,$$

be the set of eigenvalues of Δ counted multiplicities with for instance the Neumann problem if M has nonempty boundary. Let $\{\phi_k\}$ be corresponding eigenfunctions forming a complete orthonormal system of $L^2(M, \mu_M)$, where $\mu_M = dV/\operatorname{vol}(M)$ is the normalized Riemannian measure.

We are concerned with the inverse spectral problem which asks the influence of spectral data of Δ on the the geometry of M. Our interest is to reconstruct the manifold from certain spectral data. It is well known that the spectrum $\{\lambda_k\}$ is not sufficient for this purpose. There are several formulations for the setting of spectral data to settle the problem. A typical one is the boundary spectral data consisting of the spectrum $\{\lambda_k\}$ and the eigenfunctions $\{\phi_k|_{\partial M}\}$ restricted to the boundary ∂M if it is not empty (see the monograph [KKL] for details). In this direction, Anderson, Katsuda, Kurylev, Lassas and Taylor [AKKLT] discussed the stability of the inverse boundary spectral problem in a certain moduli space of Riemannian manifolds with boundary such that no collapse occurs there.

We consider the following moduli space of n-dimensional closed Riemannian manifolds with bounded sectional curvature and diameter. Let $\mathcal{N}_m(n, D)$ be the set of (M, μ_M), where M is an n-dimensional closed Riemannian manifold with the normalized Riemannian measure μ_M which satisfies

$$|\sec(M)| \leq 1, \ \ \operatorname{diam}(M) \leq D.$$

We equip $\mathcal{N}_m(n, D)$ with the measured Gromov-Hausdorff topology introduced by Fukaya [F2]. Let a sequence (M_i, μ_{M_i}) in $\mathcal{N}_m(n, D)$ converge to a metric measure space (X, μ) with respect to the measured Gromov-Hausdorff topology. This means that there exists a measurable map $\varphi_i : M_i \to X$ such that for some $\varepsilon_i \to 0$

(1) $|d(\varphi_i(x), \varphi_i(y)) - d(x, y)| < \varepsilon_i$ for all $x, y \in M_i$;

(2) the ε_i-neighborhood of $\varphi_i(M_i)$ coincides with X;

(3) the puchforward measure $(\varphi_i)_*(\mu_{M_i})$ converges to μ for the weak*-topology.

Theorem 6.1 ([F2]). *Under the above situation, there exists a self-adjoint operator $\Delta_{(X,\mu)}$ on $L^2(X, \mu)$ with discrete spectrum $\lambda_1(\Delta_{X,\mu}) \leq \lambda_2(\Delta_{X,\mu}) \leq \cdots$ such that*

(1) *$\lambda_k(\Delta_{M_i})$ converges to $\lambda_k(\Delta_{X,\mu})$ as $i \to \infty$ for each $k = 0, 1, 2, \ldots$;*

(2) *a normalized eigenfunction of Δ_{M_i} is close to an eigenfunction of $\Delta_{(X,\mu)}$ in some L^2-sense.*

The limit measure μ on X has the expression $\mu = \rho dX$, where dX denotes the Hausdorff measure and ρ is the density function on X. It was proved in [F1]

that the regular part X^{reg} of X is a smooth manifold with $C^{1,\alpha}$-metric tensor. Kasue ([Ks]) proved that ρ is of class $C^{1,\alpha}$ on X^{reg}. One can express the limit operator as

$$\Delta_{(X,\mu)}u = \frac{1}{\rho(x)\sqrt{G(x)}}\frac{\partial}{\partial x_j}\left(\rho(x)\sqrt{G(x)}g^{jk}(x)\frac{\partial}{\partial x_k}u\right),$$

on X^{reg}, where $G(x) = \det(g_{ij}(x))$.

Now we discuss the inverse problem in the closure $\overline{\mathcal{N}}_m(n,D)$ of $\mathcal{N}_m(n,D)$ with respect to the measured Gromov-Hausdorff topology. We consider the heat data on a region of the space as spectral data. First to treat the uniqueness, we concentrate on a space (X,μ) in the boundary $\partial\mathcal{N}_m(n,D) = \overline{\mathcal{N}}_m(n,D) \setminus \mathcal{N}_m(n,D)$.

Let $\lambda_1 \le \lambda_2 \le \cdots$ and $\{\phi_k\}_{k=1}^{\infty}$ be the eigenvalues of $\Delta_{(X,\mu)}$ and a complete orthonormal system of $L^2(X,\mu)$ consisting of corresponding eigenfunctions. Consider the heat kernel of (X,μ):

$$h_{(X,\mu)}(x,y,t) = \sum_{i=1}^{\infty} e^{-\lambda_i t}\phi_i(x)\phi_i(y).$$

Let Ω be an open domain of X, where we are going to measure point heat data. In view of the actual application to stability discussed later, we take countable dense subsets

$$\Omega_0 = \{z_j\}_{j=1}^{\infty}, \qquad I_0 = \{t_\ell\}_{\ell=1}^{\infty}$$

of Ω and the interval $I = [1,2]$ respectively. On these countable sets, we measure the point heat data defined as :

$$(PHD)(\Omega_0) := h_{(X,\mu)}|_{\Omega_0\times\Omega_0\times I_0} = \{h_{(X,\mu)}(z_j,z_k,t_\ell)\}_{j,k,\ell\in\mathbb{Z}_+}.$$

We ask if one can determine the metric measure space (X,μ) from the point heat data $(PHD)(\Omega_0)$.

For $(X,\mu),(Y,\mu') \in \partial\mathcal{N}_m(n,D)$ and for open domains $\Omega \subset X$, $\Omega' \subset Y$, let Ω_0 and Ω_0' be countable dense subsets of Ω and Ω' respectively.

Definition 6.2. We say that (X,μ) and (Y,μ') have the *same PHD* on Ω and Ω' respectively, and write

$$PHD(\Omega_0) = PHD(\Omega_0')$$

if there is a bijection $\Omega = \{z_j\}_{j=1}^{\infty} \to \Omega' = \{z_j'\}_{j=1}^{\infty}$ sending z_j to z_j' such that

$$h_{(X,\mu)}(z_j,z_k,t_\ell) = h_{(Y,\mu')}(z_j',z_k',t_\ell)$$

for all $(j,k,\ell) \in \mathbb{Z}_+\times\mathbb{Z}_+\times\mathbb{Z}_+$.

In the above definition, we do not require the continuity of $z_j \to z_j'$ in advance.

Theorem 6.3 ([KLY1],[KLY2]). *Under the above situation if $PHD(\Omega_0) = PHD(\Omega_0')$, then there exists an isometry $\Phi : X \to Y$ satisfying $\Phi_*(\mu) = \mu'$.*

Remark 6.4. (1) In Theorem 6.3, we not only have the uniqueness but also can construct the isometry class of (X, μ) as a metric measure space from $PHD(\Omega_0)$.

(2) When X is an orbifold (this happens for instance if $\dim X = n - 1$), we can also reconstruct the algebraic isomorphism class of X as an orbifold ([KLY1]). In [KLY1], we set up a certain sub-moduli space $\mathcal{SN}_m(n, D)$ of $\mathcal{N}_m(n, D)$ consisting of M satisfying a volume growth condition, and show that any collapsing limit in $\mathcal{SN}_m(n, D)$ becomes an orbifold.

We denote by $\overline{\mathcal{N}}_{m,p}(n, D)$ the set of all pointed measure space (X, μ, x_0) with $(X, \mu) \in \overline{\mathcal{N}}_m(n, D)$ and $x_0 \in X$. To discuss the stability of the inverse problem in $\overline{\mathcal{N}}_{m,p}(n, D)$, we need to measure point heat data on some balls of a *fixed* radius r_0, the *observation radius*.

Definition 6.5. We say that (X, μ, x_0) and (Y, μ', y_0) in $\overline{\mathcal{N}}_{m,p}(n, D)$ have *δ-close PHD with scale r_0* if and only if there are δ-dense subsets $\{z_j\}_{j=1}^N \subset B(x_0, r_0)$, $\{z_j'\}_{j=1}^N \subset B(y_0, r_0)$ and $\{t_\ell\}_{\ell=1}^T \subset I$ such that

$$|h_{(X,\mu)}(z_j, z_k, t_\ell) - h_{(Y,\mu')}(z_j', z_k', t_\ell)| < \delta$$

for all $1 \le j, k \le N$ and $1 \le \ell \le T$.

Theorem 6.6 ([KLY1], [KLY2]). *For given n, D, r_0 and $\varepsilon > 0$, there exists a positive number δ such that if (X, μ, x_0) and (Y, μ', y_0) in $\overline{\mathcal{N}}_{m,p}(n, D)$ have δ-close PHD with scale r_0, then $d_{GH}(X, Y) < \varepsilon$.*

Corollary 6.7 ([KLY2]). *For given $(X, \mu, x_0) \in \partial\mathcal{N}_{m,p}(n, D)$ and r_0, there exists a positive number δ such that if (X, μ, x_0) and $(M, \mu, p) \in \mathcal{N}_{m,p}(n, D)$ have δ-close PHD with scale r_0, then there exists a singular fibration $\pi : M \to X$ in the sense of Fukaya [F1].*

For any $\delta > 0$, consider the direct map

$$PHD : \overline{\mathcal{N}}_{m,p}(n, D) \to \mathbb{R}^{N_\delta \times N_\delta \times T_\delta}$$

defined by

$$PHD(X, \mu, x_0) := \{h_{(x,\mu)}(z_j, z_k, t_\ell)\},$$

where $\{z_j\}_{j=1}^{N_\delta} \subset B(x_0, r_0)$ and $\{t_\ell\}_{\ell=1}^{T_\delta} \subset I$ are δ-dense subsets. We can prove Stability Theorem 6.6 by using Uniqueness Theorem 6.3 and the continuity of the direct map with respect to the pointed measured Gromov-Hausdorff topology.

In [KMS], Kuwae, Machigashira and Shioya discussed the Laplacian on Alexandrov spaces with curvature bounded below.

Problem 6.8. Extend the results in this paragraph to the moduli space $\mathcal{M}(n, D)$ equipped with the pointed measured Gromov-Hausdorff topology.

References

[AKKLT] M. Anderson, A. Katsuda, Y. Kurylev, M. Lassas and M. Taylor, *Boundary regularity for the Ricci equation, Geometric Convergence, and Gel'fand's Inverse Boundary Problem*, Invent. Math., **158**, (2004), 261–321.

[BGP] Yu. Burago, M. Gromov G. Perel'man, *A. D. Aleksandrov spaces with curvatures bounded below*, Uspekhi Mat. Nauk, **47**, (1992), 3–51. translation in Russian Math. Surveys **47** (1992), 1–58.

[CaGe] J. Cao and J. Ge, *A simple proof of Perelman's collapsing theorem for 3-manifolds*, arXiv:1003.2215

[C] J. Cheeger, *Finiteness theorems for Riemannian manifolds*, Amer. J. Math., **92**, (1970), 61–74.

[CFG] J. Cheeger and K. Fukaya and M. Gromov, *Nilpotent structures and invariant metrics on collapsed manifolds*, J. Amer. Math. Soc., **5**, (1992), 327–372.

[F1] K. Fukaya, *A boundary of the set of the Riemannian manifolds with bounded curvatures and diameters*, J. Differential Geom. 28 (1988), 1–21.

[F2] K. Fukaya, *Collapsing of Riemannian manifolds and eigenvalues of Laplace operator*, Invent. Math. 87 (1987), 517–547.

[FY] K. Fukaya and T. Yamaguchi, *The Fundamental Groups of Almost Nonnegatively Curved Manifolds*, Ann. of Math., **36**, (1992), 253–333.

[GLP] M. Gromov, *Structures métriques pour les variétés riemanniennes*, Edited by J. Lafontaine and P. Pansu, Textes Mathématiques [Mathematical Texts], **1**, CEDIC, Paris, 1981.

[G1] M. Gromov, *Almost flat manifolds*, J. Diff. Geometry, **13**, (1978), 231–241.

[G2] M. Gromov, *Curvature, diameter and Betti numbers*, Comment. Math. Helv., **56**, (1981), 179–195.

[GPW] K. Grove, P. Petersen and J.-Y. Wu, *Geometric finiteness theorems via controlled topology*, Invent. Math., **99**, (1990), 205–213.

[GS] K. Grove and K. Shiohama, *A generalized sphere theorem*, Ann. of Math., **106**, (1977), 201–211.

[Kp] V. Kapovitch, *Perelman's stability theorem*, Surveys in differential geometry. **11**, 103–136, Int. Press, Somerville, MA, 2007.

[Ks] A. Kasue, *Measured Hausdorff convergence of Riemannian manifolds and Laplace operators*, Osaka J. Math., **30**, (1993), 613–651.

[KKL] A. Katchalov, Y. Kurylev, M. Lassas, *Inverse Boundary Spectral Problems*, Chapman Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, **123**, 2001.

[KLY1] Y. Kurylev, M. Lassas and T. Yamaguchi, *Uniqueness and Stability in Inverse spectral problems for collapsing manifolds, I:Orbifold Case*, in preparation.

[KLY2] Y. Kurylev, M. Lassas and T. Yamaguchi, *Uniqueness and Stability in Inverse spectral problems for collapsing manifolds, II:General Case*, in preparation.

[KMS] K. Kuwae, Y. Machigashira and T. Shioya, *Sobolev spaces, Laplacian, and heat kernel on Alexandrov spaces*, Math. Z. **238**, (2001), 269–316.

[MT] J. Morgan, G. Tian, *Completion of the Proof of the Geometrization Conjecture*, arXiv:0809.4040

[OS] Y. Otsu and T. Shioya, *The Riemannian structure of Alexandrov spaces*, J. Differential Geom., **39**, (1994), 629–658.

[P1] G. Perelman, *A. D. Alexandrov's spaces with Curvatures Bounded from Below II*, preprint.

[P2] G. Perelman, *Elements of Morse Theory on Alexandrov spaces*, St. Petersburg Math. Jour., **5**, (1994), 207–214.

[P3] G. Perelman, *Collapsing with No Proper Extremal Subsets*, Comparison geometry (Berkeley, CA, 1993–94), 149–155, Math. Sci. Res. Inst. Publ., **30**, Cambridge Univ. Press, Cambridge, 1997.

[P4] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications*, arXiv:math. DG / 0211159

[P5] G. Perelman, *Ricci flow with surgery on three-manifolds*, arXiv:math. DG / 0303109.

[PP] G. Perelman and A. Petrunin, *Quasigeodesics and gradient curves in Alexandrov spaces*, preprint.

[P] A. Petrunin, *Semiconcave functions in Alexandrov's geometry*, Surveys in differential geometry. **XI**, 137–201, Int. Press, Somerville, MA, 2007.

[SY1] T. Shioya and T. Yamaguchi, *Collapsing three-manifolds under a lower curvature bound*, J. Differential Geometry, **56**, (2000), 1–66.

[SY2] T. Shioya and T. Yamaguchi, *Volume collapsed three-manifolds with a lower curvature bound*, Math. Ann., **333**, (2005), 131—155.

[Y1] T. Yamaguchi, *Collapsing and Pinching under a lower curvature bound*, Ann. of Math., **133**, (1991), 317–357.

[Y2] T. Yamaguchi, *A convergence theorem in the geometry of Alexandrov spaces*, Actes de la Table Ronde de Géométrie Différentielle (Luminy, 1992), 601–642, Sémin. Congr., **1**, Soc. Math. France, Paris, 1996.

[Y3] T. Yamaguchi, *Collapsing* 4*-manifolds under a lower curvature bound*, preprint.

[Y4] T. Yamaguchi, *Collapsing Riemannian 4-manifolds*, (Japanese) Sugaku **52** (2000), 172–186.

[Y5] T. Yamaguchi, *Collapsing and essential covering*, preprint.

[W] A. Weinstein, *On the homotopy type of positively-pinched manifolds* , Arch. Math., **18**, (1967), 523–524.

Section 6

Topology

Denis Auroux
Fukaya Categories and Bordered Heegaard-Floer Homology 917

Kevin Costello
A Geometric Construction of the Witten Genus, I 942

David Gabai
Hyperbolic 3-manifolds in the 2000's 960

Jesper Grodal
The Classification of p–compact Groups and Homotopical Group Theory 973

Ursula Hamenstädt
Actions of the Mapping Class Group 1002

Michael Hutchings
Embedded Contact Homology and Its Applications 1022

Marc Lackenby
Finite Covering Spaces of 3-manifolds 1042

Wolfgang Lück
K- and L-theory of Group Rings 1071

Jacob Lurie
Moduli Problems for Ring Spectra 1099

Maryam Mirzakhani
On Weil-Petersson Volumes and Geometry of Random Hyperbolic Surfaces 1126

Jongil Park
A New Family of Complex Surfaces of General Type with $p_g = 0$ 1146

András I. Stipsicz
Ozsváth-Szabó Invariants and 3-dimensional Contact Topology 1159

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Fukaya Categories and Bordered Heegaard-Floer Homology

Denis Auroux*

Abstract

We outline an interpretation of Heegaard-Floer homology of 3-manifolds (closed or with boundary) in terms of the symplectic topology of symmetric products of Riemann surfaces, as suggested by recent work of Tim Perutz and Yankı Lekili. In particular we discuss the connection between the Fukaya category of the symmetric product and the bordered algebra introduced by Robert Lipshitz, Peter Ozsváth and Dylan Thurston, and recast bordered Heegaard-Floer homology in this language.

Mathematics Subject Classification (2010). 53D40 (53D37, 57M27, 57R58)

Keywords. Bordered Heegaard-Floer homology, Fukaya categories

1. Introduction

In its simplest incarnation, Heegaard-Floer homology associates to a closed 3-manifold Y a graded abelian group $\widehat{HF}(Y)$. This invariant is constructed by considering a Heegaard splitting $Y = Y_1 \cup_\Sigma Y_2$ of Y into two genus g handlebodies, each of which determines a product torus in the g-fold symmetric product of the Heegaard surface $\Sigma = \partial Y_1 = -\partial Y_2$. Deleting a marked point z from Σ to obtain an open surface, $\widehat{HF}(Y)$ is then defined as the Lagrangian Floer homology of the two tori T_1, T_2 in $\mathrm{Sym}^g(\Sigma \setminus \{z\})$, see [9].

It is natural to ask how more general decompositions of 3-manifolds fit into this picture, and whether Heegaard-Floer theory can be viewed as a TQFT (at least in some partial sense). From the point of view of symplectic geometry, a natural answer is suggested by the work of Tim Perutz and Yankı Lekili. Namely, an elementary cobordism between two connected Riemann surfaces

*Department of Mathematics, UC Berkeley, Berkeley CA 94720-3840, USA.
E-mail: auroux@math.berkeley.edu.

Σ_1, Σ_2 given by attaching a single handle determines a Lagrangian correspondence between appropriate symmetric products of Σ_1 and Σ_2 [10]. By composing these correspondences, one can associate to a 3-manifold with connected boundary Σ of genus g a *generalized Lagrangian submanifold* (cf. [16]) of the symmetric product $\mathrm{Sym}^g(\Sigma \setminus \{z\})$. Recent work of Lekili and Perutz [4] shows that, given a decomposition $Y = Y_1 \cup_\Sigma Y_2$ of a closed 3-manifold, the (quilted) Floer homology of the generalized Lagrangian submanifolds of $\mathrm{Sym}^g(\Sigma \setminus \{z\})$ determined by Y_1 and Y_2 recovers $\widehat{HF}(Y)$.

From a more combinatorial perspective, the *bordered Heegaard-Floer homology* of Robert Lipshitz, Peter Ozsváth and Dylan Thurston [6] associates to a parameterized Riemann surface F with connected boundary a finite dimensional differential algebra $\mathcal{A}(F)$ over $\mathbb{Z}_2$, and to a 3-manifold Y with boundary $\partial Y = F \cup D^2$ a right A_∞-module $\widehat{CFA}(Y)$ over $\mathcal{A}(F)$, as well as a left dg-module $\widehat{CFD}(Y)$ over $\mathcal{A}(-F)$. The main result of [6] shows that, given a decomposition of a closed 3-manifold $Y = Y_1 \cup Y_2$ with $\partial Y_1 = -\partial Y_2 = F \cup D^2$, $\widehat{HF}(Y)$ can be computed in terms of the A_∞-tensor product of the modules associated to Y_1 and Y_2, namely

$$\widehat{HF}(Y) \simeq H_*(\widehat{CFA}(Y_1) \otimes_{\mathcal{A}(F)} \widehat{CFD}(Y_2)).$$

In order to connect these two approaches, we consider a *partially wrapped* version of Floer theory for product Lagrangians in symmetric products of open Riemann surfaces. Concretely, given a Riemann surface with boundary F, a finite collection Z of marked points on ∂F, and an integer $k \geq 0$, we consider a partially wrapped Fukaya category $\mathcal{F}(\mathrm{Sym}^k(F), Z)$, which differs from the usual (compactly supported) Fukaya category by the inclusion of additional objects, namely products of disjoint properly embedded arcs in F with boundary in $\partial F \setminus Z$. A nice feature of these categories is that they admit explicit sets of generating objects:

Theorem 1. *Let F be a compact Riemann surface with non-empty boundary, Z a finite subset of ∂F, and $\underline{\alpha} = \{\alpha_1, \dots, \alpha_n\}$ a collection of disjoint properly embedded arcs in F with boundary in $\partial F \setminus Z$. Assume that $F \setminus (\alpha_1 \cup \dots \cup \alpha_n)$ is a union of discs, each of which contains at most one point of Z. Then for $0 \leq k \leq n$, the partially wrapped Fukaya category $\mathcal{F}(\mathrm{Sym}^k(F), Z)$ is generated by the $\binom{n}{k}$ Lagrangian submanifolds $D_s = \prod_{i \in s} \alpha_i$, where s ranges over all k-element subsets of $\{1, \dots, n\}$.*

To a decorated surface $\mathbb{F} = (F, Z, \underline{\alpha} = \{\alpha_i\})$ we can associate an A_∞-algebra

$$\mathcal{A}(\mathbb{F}, k) = \bigoplus_{s,t} \hom_{\mathcal{F}(\mathrm{Sym}^k(F), Z)}(D_s, D_t). \tag{1}$$

The following special case is of particular interest:

Theorem 2. *Assume that F has a single boundary component, $|Z| = 1$, and the arcs $\alpha_1, \dots, \alpha_n$ ($n = 2g(F)$) decompose F into a single disc. Then $\mathcal{A}(\mathbb{F}, k)$ coincides with Lipshitz-Ozsváth-Thurston's* bordered algebra [6].

(The result remains true in greater generality, the only key requirement being that every component of $F \setminus (\alpha_1 \cup \cdots \cup \alpha_n)$ should contain at least one point of Z.)

Now, consider a *sutured* 3-manifold Y, i.e. a 3-manifold Y with non-empty boundary, equipped with a decomposition $\partial Y = (-F_-) \cup F_+$, where $F_\pm$ are oriented surfaces with boundary. Assume moreover that ∂Y and $F_\pm$ are connected, and denote by $g_\pm$ the genus of $F_\pm$. Given two integers $k_\pm$ such that $k_+ - k_- = g_+ - g_-$ and a suitable Morse function on Y, Perutz's construction associates to Y a generalized Lagrangian correspondence (i.e. a formal composition of Lagrangian correspondences) $\mathbb{T}_Y$ from $\mathrm{Sym}^{k_-}(F_-)$ to $\mathrm{Sym}^{k_+}(F_+)$. By the main result of [4] this correspondence is essentially independent of the chosen Morse function.

Picking a finite set of marked points $Z \subset \partial F_- = \partial F_+$, and two collections of disjoint arcs $\underline{\alpha}_-$ and $\underline{\alpha}_+$ on F_- and F_+, we have two decorated surfaces $\mathbb{F}_\pm = (F_\pm, Z, \underline{\alpha}_\pm)$, and collections of product Lagrangian submanifolds $D_{\pm,s}$ ($s \in \mathcal{S}_\pm$) in $\mathrm{Sym}^{k_\pm}(F_\pm)$ (namely, all products of $k_\pm$ of the arcs in $\underline{\alpha}_\pm$). By a Yoneda-style construction, the correspondence $\mathbb{T}_Y$ then determines an A_∞-bimodule

$$\mathcal{Y}(\mathbb{T}_Y) = \bigoplus_{(s,t)\in\mathcal{S}_-\times\mathcal{S}_+} \hom(D_{-,s}, \mathbb{T}_Y, D_{+,t}) \in \mathcal{A}(\mathbb{F}_-, k_-)\text{-mod-}\mathcal{A}(\mathbb{F}_+, k_+), \quad (2)$$

where $\hom(D_{-,s}, \mathbb{T}_Y, D_{+,t})$ is defined in terms of quilted Floer complexes [8, 16, 17] after suitably perturbing $D_{-,s}$ and $D_{+,t}$ by partial wrapping along the boundary. A slightly different but equivalent definition is as follows. With quite a bit of extra work, via the Ma'u-Wehrheim-Woodward machinery the correspondence $\mathbb{T}_Y$ defines an A_∞-functor Φ_Y from $\mathcal{F}(\mathrm{Sym}^{k_-}(F_-), Z)$ to a suitable enlargement of $\mathcal{F}(\mathrm{Sym}^{k_+}(F_+), Z)$; with this understood, $\mathcal{Y}(\mathbb{T}_Y) \simeq \bigoplus_{(s,t)} \hom(\Phi_Y(D_{-,s}), D_{+,t})$.

The A_∞-bimodules $\mathcal{Y}(\mathbb{T}_Y)$ are expected to obey the following gluing property:

Conjecture 3. *Let F, F', F'' be connected Riemann surfaces and Z a finite subset of $\partial F \simeq \partial F' \simeq \partial F''$. Let Y_1, Y_2 be two sutured manifolds with $\partial Y_1 = (-F) \cup F'$ and $\partial Y_2 = (-F') \cup F''$, and let $Y = Y_1 \cup_{F'} Y_2$ be the sutured manifold obtained by gluing Y_1 and Y_2 along F'. Equip F, F', F'' with collections of disjoint properly embedded arcs $\underline{\alpha}, \underline{\alpha}', \underline{\alpha}''$, and assume that $\underline{\alpha}'$ decomposes F' into a union of discs each containing at most one point of Z. Then*

$$\mathcal{Y}(\mathbb{T}_Y) \simeq \mathcal{Y}(\mathbb{T}_{Y_1}) \otimes_{\mathcal{A}(\mathbb{F}', k')} \mathcal{Y}(\mathbb{T}_{Y_2}). \quad (3)$$

In its most general form this statement relies on results in Floer theory for generalized Lagrangian correspondences which are not yet fully established, hence we state it as a conjecture; however, we believe that a proof should be within reach of standard techniques.

As a special case, let F be a genus g surface with connected boundary, decorated with a single point $z \in \partial F$ and a collection of $2g$ arcs cutting F into a disc. Then to a 3-manifold Y_1 with boundary $\partial Y_1 = F \cup D^2$ we can associate a generalized Lagrangian submanifold $\mathbb{T}_{Y_1}$ of $\mathrm{Sym}^g(F)$, and an A_∞-module $\mathcal{Y}(\mathbb{T}_{Y_1}) = \bigoplus_s \hom(\mathbb{T}_{Y_1}, D_s) \in \text{mod-}\mathcal{A}(\mathbb{F}, g)$. Viewing $\mathbb{T}_{Y_1}$ as a generalized correspondence from $\mathrm{Sym}^g(-F)$ to $\mathrm{Sym}^0(D^2) = \{\mathrm{pt}\}$ instead, we obtain a left A_∞-module over $\mathcal{A}(-\mathbb{F}, g)$. However, $\mathcal{A}(-\mathbb{F}, g) = \mathcal{A}(\mathbb{F}, g)^{op}$, and the two constructions yield the same module. If now we have another 3-manifold Y_2 with $\partial Y_2 = -F \cup D^2$, we can associate to it a generalized Lagrangian submanifold $\mathbb{T}_{Y_2}$ in $\mathrm{Sym}^g(-F)$ or, after orientation reversal, $\mathbb{T}_{-Y_2}$ in $\mathrm{Sym}^g(F)$. This yields A_∞-modules $\mathcal{Y}(\mathbb{T}_{Y_2}) \in \text{mod-}\mathcal{A}(-\mathbb{F}, g) \simeq \mathcal{A}(\mathbb{F}, g)\text{-mod}$, and $\mathcal{Y}(\mathbb{T}_{-Y_2}) \in \text{mod-}\mathcal{A}(\mathbb{F}, g)$.

Theorem 4. *With this understood, and denoting by Y the closed 3-manifold obtained by gluing Y_1 and Y_2 along their boundaries, we have quasi-isomorphisms*

$$\begin{aligned}\widehat{CF}(Y) \simeq \hom_{\mathcal{F}^\#(\mathrm{Sym}^g(F))}(\mathbb{T}_{Y_1}, \mathbb{T}_{-Y_2}) &\simeq \hom_{mod\text{-}\mathcal{A}(\mathbb{F},g)}(\mathcal{Y}(\mathbb{T}_{-Y_2}), \mathcal{Y}(\mathbb{T}_{Y_1})) \\ &\simeq \mathcal{Y}(\mathbb{T}_{Y_1}) \otimes_{\mathcal{A}(\mathbb{F},g)} \mathcal{Y}(\mathbb{T}_{Y_2}). \qquad (4)\end{aligned}$$

In fact, $\mathcal{Y}(\mathbb{T}_{Y_i})$ is quasi-isomorphic to the bordered A_∞-module $\widehat{CFA}(Y_i)$. In light of this, it is instructive to compare Theorem 4 with the pairing theorem obtained by Lipshitz, Ozsváth and Thurston [6]: even though $\widehat{CFA}(Y_i)$ and $\widehat{CFD}(Y_i)$ seem very different at first glance (and even at second glance), our result suggests that they can in fact be used interchangeably.

The rest of this paper is structured as follows: first, in section 2 we explain how Heegaard-Floer homology can be understood in terms of Lagrangian correspondences, following the work of Perutz and Lekili [10, 4]. Then in section 3 we introduce partially wrapped Fukaya categories of symmetric products, and sketch the proofs of Theorems 1 and 2. In section 4 we briefly discuss Yoneda embedding as well as Conjecture 3 and Theorem 4. Finally, in section 5 we discuss the relation with bordered Heegaard-Floer homology.

The reader will not find detailed proofs for any of the statements here, nor a general discussion of partially wrapped Fukaya categories. Some of the material is treated in greater depth in the preprint [2], the rest will appear in a future paper.

Acknowledgements

I am very grateful to Mohammed Abouzaid, Sheel Ganatra, Yankı Lekili, Robert Lipshitz, Peter Ozsváth, Tim Perutz, Paul Seidel and Dylan Thurston for many helpful discussions. I would also like to thank Ivan Smith for useful comments on the exposition. This work was partially supported by NSF grants DMS-0600148 and DMS-0652630.

2. Heegaard-Floer Homology from Lagrangian Correspondences

2.1. Lagrangian correspondences. A *Lagrangian correspondence* between two symplectic manifolds (M_1, ω_1) and (M_2, ω_2) is, by definition, a Lagrangian submanifold of the product $M_1 \times M_2$ equipped with the product symplectic form $(-\omega_1) \oplus \omega_2$. Lagrangian correspondences can be thought of as a far-reaching generalization of symplectomorphisms (whose graphs are examples of correspondences); in particular, under suitable transversality assumptions we can consider the *composition* of two correspondences $L_{01} \subset M_0 \times M_1$ and $L_{12} \subset M_1 \times M_2$,

$$L_{01} \circ L_{12} = \{(x, z) \in M_0 \times M_2 \,|\, \exists\, y \in M_1 \text{ s.t. } (x, y) \in L_{01} \text{ and } (y, z) \in L_{12}\}.$$

The image of a Lagrangian submanifold $L_1 \subset M_1$ by a Lagrangian correspondence $L_{12} \subset M_1 \times M_2$ is defined similarly, viewing L_1 as a correspondence from $\{pt\}$ to M_1. Unfortunately, in general the geometric composition is not a smooth embedded Lagrangian. Nonetheless, we can enlarge symplectic geometry by considering *generalized Lagrangian correspondences*, i.e. sequences of Lagrangian correspondences (interpreted as formal compositions), and *generalized Lagrangian submanifolds*, i.e. generalized Lagrangian correspondences from $\{pt\}$ to a given symplectic manifold.

The work of Ma'u, Wehrheim and Woodward (see e.g. [16, 17, 8]) shows that Lagrangian Floer theory behaves well with respect to (generalized) correspondences. Given a sequence of Lagrangian correspondences $L_{i-1,i} \subset M_{i-1} \times M_i$ $(i = 1, \ldots, n)$, with $M_0 = M_n = \{pt\}$, the *quilted Floer complex* $CF(L_{0,1}, \ldots, L_{n-1,n})$ is generated by *generalized intersections*, i.e. tuples $(x_1, \ldots, x_{n-1}) \in M_1 \times \cdots \times M_{n-1}$ such that $(x_{i-1}, x_i) \in L_{i-1,i}$ for all i, and carries a differential which counts "quilted pseudoholomorphic strips" in $M_1 \times \cdots \times M_{n-1}$. Under suitable technical assumptions (e.g., monotonicity), Lagrangian Floer theory carries over to this setting.

Thus, Ma'u, Wehrheim and Woodward associate to a monotone symplectic manifold (M, ω) its *extended Fukaya category* $\mathcal{F}^{\#}(M)$, whose objects are monotone generalized Lagrangian submanifolds and in which morphisms are given by quilted Floer complexes. Composition of morphisms is defined by counting quilted pseudoholomorphic discs, and as in usual Floer theory, it is only associative up to homotopy, so $\mathcal{F}^{\#}(M)$ is an A_∞-category. The key property of these extended Fukaya categories is that a monotone (generalized) Lagrangian correspondence L_{12} from M_1 to M_2 induces an A_∞-functor from $\mathcal{F}^{\#}(M_1)$ to $\mathcal{F}^{\#}(M_2)$, which on the level of objects is simply concatenation with L_{12}. Moreover, composition of Lagrangian correspondences matches with composition of A_∞-functors [8].

Remark. By construction, the usual Fukaya category $\mathcal{F}(M)$ admits a fully faithful embedding as a subcategory of $\mathcal{F}^{\#}(M)$. In fact, $\mathcal{F}^{\#}(M)$ embeds into the

category of A_∞-modules over the usual Fukaya category, so although generalized Lagrangian correspondences play an important conceptual role in our discussion, they only enlarge the Fukaya category in a fairly mild manner.

2.2. Symmetric products. As mentioned in the introduction, work in progress of Lekili and Perutz [4] shows that Heegaard-Floer homology can be understood in terms of quilted Floer homology for Lagrangian correspondences between symmetric products. The relevant correspondences were introduced by Perutz in his thesis [10].

Let Σ be an open Riemann surface (with infinite cylindrical ends, i.e., the complement of a finite set in a compact Riemann surface), equipped with an area form σ. We consider the symmetric product $\mathrm{Sym}^k(\Sigma)$, equipped with the product complex structure J, and a Kähler form ω which coincides with the product Kähler form on Σ^k away from the diagonal strata. Following Perutz we choose ω so that its cohomology class is negatively proportional to $c_1(T\mathrm{Sym}^k(\Sigma))$.

Let γ be a non-separating simple closed curve on Σ, and Σ_γ the surface obtained from Σ by deleting a tubular neighborhood of γ and gluing in two discs. Equip Σ_γ with a complex structure which agrees with that of Σ away from γ, and equip $\mathrm{Sym}^k(\Sigma)$ and $\mathrm{Sym}^{k-1}(\Sigma_\gamma)$ with Kähler forms ω and ω_γ chosen as above.

Theorem 5 (Perutz [10]). *The simple closed curve γ determines a Lagrangian correspondence T_γ in the product* $(\mathrm{Sym}^{k-1}(\Sigma_\gamma) \times \mathrm{Sym}^k(\Sigma), -\omega_\gamma \oplus \omega)$*, canonically up to Hamiltonian isotopy.*

Given r disjoint simple closed curves $\gamma_1, \dots, \gamma_r$, linearly independent in $H_1(\Sigma)$, we can consider the sequence of correspondences that arise from successive surgeries along $\gamma_1, \dots, \gamma_r$. The main properties of these correspondences (see Theorem A in [10]) imply immediately that their composition defines an embedded Lagrangian correspondence $T_{\gamma_1,\dots,\gamma_r}$ in $\mathrm{Sym}^{k-r}(\Sigma_{\gamma_1,\dots,\gamma_r}) \times \mathrm{Sym}^k(\Sigma)$.

When $r = k = g(\Sigma)$, this construction yields a Lagrangian torus in $\mathrm{Sym}^k(\Sigma)$, which by [10, Lemma 3.20] is smoothly isotopic to the product torus $\gamma_1 \times \cdots \times \gamma_k$; Lekili and Perutz show that these two tori are in fact Hamiltonian isotopic [4].

Remark. We are not quite in the setting considered by Ma'u, Wehrheim and Woodward, but Floer theory remains well behaved thanks to two key properties of the Lagrangian submanifolds under consideration: their relative π_2 is trivial (which prevents bubbling), and they are *balanced*. (A Lagrangian submanifold in a monotone symplectic manifold is said to be balanced if the holonomy of a fixed connection 1-form with curvature equal to the symplectic form vanishes on it; this is a natural analogue of the notion of exact Lagrangian submanifold in an exact symplectic manifold). The balancing condition is closely related to admissibility of Heegaard diagrams, and ensures that the symplectic area of a pseudo-holomorphic strip connecting two given intersection points is

determined *a priori* by its Maslov index (cf. [17, Lemma 4.1.5]). This property is what allows us to work over $\mathbb{Z}_2$ rather than over a Novikov field.

2.3. Heegaard-Floer homology. Consider a closed 3-manifold Y, and a Morse function $f : Y \to \mathbb{R}$ (with only one minimum and one maximum, and with distinct critical values). Then the complement Y' of a ball in Y (obtained by deleting a neighborhood of a Morse trajectory from the maximum to the minimum) can be decomposed into a succession of elementary cobordisms Y_i' $(i = 1, \dots, r)$ between connected Riemann surfaces with boundary $\Sigma_0, \Sigma_1, \dots, \Sigma_r$ (where $\Sigma_0 = \Sigma_r = D^2$, and the genus increases or decreases by 1 at each step). By Theorem 5, each Y_i' determines a Lagrangian correspondence $L_i \subset \mathrm{Sym}^{g_{i-1}}(\Sigma_{i-1}) \times \mathrm{Sym}^{g_i}(\Sigma_i)$ between the relevant symmetric products (here g_i is the genus of Σ_i, and we implicitly complete Σ_i by attaching to it an infinite cylindrical end). By the work of Lekili and Perutz [4], the quilted Floer homology of the sequence $(L_1, \dots, L_r)$ is independent of the choice of the Morse function f and isomorphic to $\widehat{HF}(Y)$.

More generally, consider a sutured 3-manifold Y, i.e. a 3-manifold whose boundary is decomposed into a union $(-F_-) \cup_\Gamma F_+$, where $F_\pm$ are connected oriented surfaces of genus $g_\pm$ with boundary $\partial F_- \simeq \partial F_+ \simeq \Gamma$. Shrinking $F_\pm$ slightly within ∂Y, it is advantageous to think of the boundary of Y as consisting actually of *three* pieces, $\partial Y = (-F_-) \cup (\Gamma \times [0,1]) \cup F_+$. By considering a Morse function $f : Y \to [0,1]$ with index 1 and 2 critical points only, with $f^{-1}(1) = F_-$ and $f^{-1}(0) = F_+$, we can view Y as a succession of elementary cobordisms between connected Riemann surfaces with boundary, starting with F_- and ending with F_+. As above, Perutz's construction associates a Lagrangian correspondence to each of these elementary cobordisms. Thus we can associate to Y a generalized Lagrangian correspondence $\mathbb{T}_Y = \mathbb{T}_{Y,k_\pm}$ from $\mathrm{Sym}^{k_-}(F_-)$ to $\mathrm{Sym}^{k_+}(F_+)$ whenever $k_+ - k_- = g_+ - g_-$. The generalized correspondence $\mathbb{T}_Y$ can be viewed either as an object of the extended Fukaya category $\mathcal{F}^\#(\mathrm{Sym}^{k_-}(-F_-) \times \mathrm{Sym}^{k_+}(F_+))$, or as an A_∞-functor from $\mathcal{F}^\#(\mathrm{Sym}^{k_-}(F_-))$ to $\mathcal{F}^\#(\mathrm{Sym}^{k_+}(F_+))$.

Theorem 6 (Lekili-Perutz [4]). *Up to quasi-isomorphism the object $\mathbb{T}_Y$ is independent of the choice of Morse function on Y.*

Given two sutured manifolds Y_1 and Y_2 ($\partial Y_i = (-F_{i,-}) \cup F_{i,+}$) and a diffeomorphism $\phi : F_{1,+} \to F_{2,-}$, gluing Y_1 and Y_2 by identifying the positive boundary of Y_1 with the negative boundary of Y_2 via ϕ yields a new sutured manifold Y'. As a cobordism from $F_{1,-}$ to $F_{2,+}$, Y' is simply the concatenation of the cobordisms Y_1 and Y_2. Hence, the generalized Lagrangian correspondence $\mathbb{T}_{Y'}$ associated to Y' is just the (formal) composition of $\mathbb{T}_{Y_1}$ and $\mathbb{T}_{Y_2}$.

The case where Y_1 is a cobordism from the disc D^2 to a genus g surface F (with a single boundary component) and Y_2 is a cobordism from F to D^2 (so $\partial Y_1 \simeq -\partial Y_2 \simeq F \cup_{S^1} D^2$) is of particular interest. In that case, we associate to Y_1 a generalized correspondence from $\mathrm{Sym}^0(D^2) = \{pt\}$ to $\mathrm{Sym}^g(F)$, i.e. an

object $\mathbb{T}_{Y_1}$ of $\mathcal{F}^{\#}(\mathrm{Sym}^g(F))$, and to Y_2 a generalized correspondence $\mathbb{T}_{Y_2}$ from $\mathrm{Sym}^g(F)$ to $\mathrm{Sym}^0(D^2) = \{pt\}$, i.e. a generalized Lagrangian submanifold of $\mathrm{Sym}^g(-F)$. Reversing the orientation of Y_2, i.e. viewing $-Y_2$ as the opposite cobordism from D^2 to F, we get a generalized Lagrangian submanifold $\mathbb{T}_{-Y_2}$ in $\mathrm{Sym}^g(F)$, which differs from $\mathbb{T}_{Y_2}$ simply by orientation reversal. Denoting by Y $(= Y' \cup B^3)$ the closed 3-manifold obtained by gluing Y_1 and Y_2 along their entire boundary, the result of [4] now says that

$$\widehat{CF}(Y) \simeq CF(\mathbb{T}_{Y_1}, \mathbb{T}_{Y_2}) \simeq \hom_{\mathcal{F}^{\#}(\mathrm{Sym}^g(F))}(\mathbb{T}_{Y_1}, \mathbb{T}_{-Y_2}). \tag{5}$$

3. Partially Wrapped Fukaya Categories of Symmetric Products

3.1. Positive perturbations and partial wrapping. Let F be a connected Riemann surface with non-empty boundary, and Z a finite subset of ∂F. Assume for now that every connected component of ∂F contains at least one point of Z. Then the components of $\partial F \setminus Z$ are open intervals, and carry a natural orientation induced by that of F.

Definition 7. *Let $\underline{\lambda} = (\lambda_1, \dots, \lambda_k)$, $\underline{\lambda}' = (\lambda'_1, \dots, \lambda'_k)$ be two k-tuples of disjoint properly embedded arcs in F, with boundary in $\partial F \setminus Z$. We say that the pair $(\underline{\lambda}, \underline{\lambda}')$ is* positive*, and write $\underline{\lambda} > \underline{\lambda}'$, if along each component of $\partial F \setminus Z$ the points of $\partial(\bigcup_i \lambda'_i)$ all lie* before *those of $\partial(\bigcup_i \lambda_i)$.*

Similarly, given tuples $\underline{\lambda}^j = (\lambda^j_1, \dots, \lambda^j_k)$ $(j = 0, \dots, \ell)$, we say that the sequence $(\underline{\lambda}^0, \dots, \underline{\lambda}^\ell)$ is positive if each pair $(\underline{\lambda}^j, \underline{\lambda}^{j+1})$ is positive, i.e. $\underline{\lambda}^0 > \dots > \underline{\lambda}^\ell$.

Given two tuples $\underline{\lambda} = (\lambda_1, \dots, \lambda_k)$ and $\underline{\lambda}' = (\lambda'_1, \dots, \lambda'_k)$, we can perturb each arc λ_i (resp. λ'_i) by an isotopy that pushes it in the positive (resp. negative) direction along ∂F, without crossing Z, to obtain new tuples $\tilde{\underline{\lambda}} = (\tilde{\lambda}_1, \dots, \tilde{\lambda}_k)$ and $\tilde{\underline{\lambda}}' = (\tilde{\lambda}'_1, \dots, \tilde{\lambda}'_k)$ with the property that $\tilde{\underline{\lambda}} > \tilde{\underline{\lambda}}'$. Similarly, any sequence $(\underline{\lambda}^0, \dots, \underline{\lambda}^\ell)$ can be made into a positive sequence by means of suitable isotopies supported near ∂F (again, the isotopies are not allowed to cross Z).

Example. Let $\underline{\alpha} = (\alpha_1, \dots, \alpha_{2g})$ be the tuple of arcs represented on Figure 1 left: then the perturbed tuples $\tilde{\underline{\alpha}}^j = (\tilde{\alpha}^j_1, \dots, \tilde{\alpha}^j_{2g})$ (Figure 1 right) satisfy $\tilde{\underline{\alpha}}^0 > \tilde{\underline{\alpha}}^1$, i.e. the pair $(\tilde{\underline{\alpha}}^0, \tilde{\underline{\alpha}}^1)$ is a positive perturbation of $(\underline{\alpha}, \underline{\alpha})$.

Next, consider a sequence $(L_0, \dots, L_\ell)$ of Lagrangian submanifolds in the symmetric product $\mathrm{Sym}^k(F)$, each of which is either a closed submanifold contained in the interior of $\mathrm{Sym}^k(F)$ or a product of disjoint properly embedded arcs $L_j = \lambda^j_1 \times \dots \times \lambda^j_k$. Then we say that the sequence $(L_0, \dots, L_\ell)$ is positive if, whenever L_i and L_j are products of disjointly embedded arcs for $i < j$, the corresponding k-tuples of arcs satisfy $\underline{\lambda}^i > \underline{\lambda}^j$. (There is no condition on the closed Lagrangians). Modifying the arcs $\lambda^j_1, \dots, \lambda^j_k$ by suitable isotopies supported near ∂F (without crossing Z) as above, given any sequence $(L_0, \dots, L_\ell)$ we can

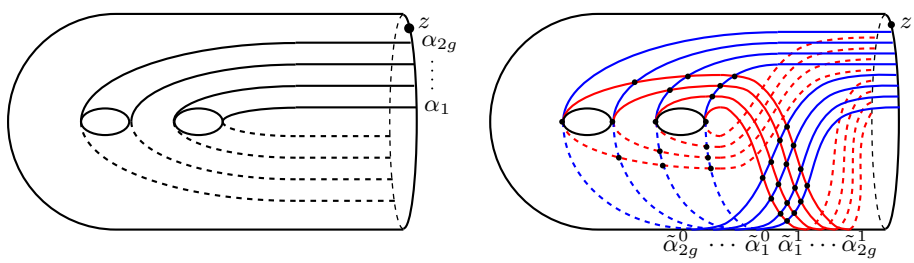


Figure 1. The arcs α_i and $\tilde{\alpha}_i^j$ on $(F, \{z\})$

construct Lagrangian submanifolds $\tilde{L}_0, \ldots, \tilde{L}_\ell$ such that: (1) $\tilde{L}_i$ is Hamiltonian isotopic to L_i, and either contained in the interior of $\mathrm{Sym}^k(F)$ or a product of disjoint properly embedded arcs; and (2) the sequence $(\tilde{L}_0, \ldots, \tilde{L}_\ell)$ is positive. We call $(\tilde{L}_0, \ldots, \tilde{L}_\ell)$ a *positive perturbation* of the sequence $(L_0, \ldots, L_\ell)$.

With this understood, we can now give an informal (and imprecise) definition of the partially wrapped Fukaya category of the symmetric product $\mathrm{Sym}^k(F)$ relative to the set Z; we are still assuming that every component of ∂F contains at least one point of Z. The reader is referred to [2] for a more precise construction.

Definition 8. *The partially wrapped Fukaya category $\mathcal{F} = \mathcal{F}(\mathrm{Sym}^k(F), Z)$ is an A_∞-category with objects of two types:*

1. *closed balanced Lagrangian submanifolds lying in the interior of $\mathrm{Sym}^k(F)$;*
2. *properly embedded Lagrangian submanifolds of the form $\lambda_1 \times \cdots \times \lambda_k$, where λ_i are disjoint properly embedded arcs with boundary contained in $\partial F \setminus Z$.*

Morphism spaces and compositions are defined by perturbing objects of the second type in a suitable manner near the boundary so that they form positive sequences. Namely, we set $\hom_{\mathcal{F}}(L_0, L_1) = CF(\tilde{L}_0, \tilde{L}_1)$ (i.e., the $\mathbb{Z}_2$-vector space generated by points of $\tilde{L}_0 \cap \tilde{L}_1$, with a differential counting rigid holomorphic discs), where $(\tilde{L}_0, \tilde{L}_1)$ is a suitably chosen positive perturbation of the pair (L_0, L_1). The composition $m_2 : \hom_{\mathcal{F}}(L_0, L_1) \otimes \hom_{\mathcal{F}}(L_1, L_2) \to \hom_{\mathcal{F}}(L_0, L_2)$ and higher products $m_\ell : \hom_{\mathcal{F}}(L_0, L_1) \otimes \cdots \otimes \hom_{\mathcal{F}}(L_{\ell-1}, L_\ell) \to \hom_{\mathcal{F}}(L_0, L_\ell)$ are similarly defined by perturbing $(L_0, \ldots, L_\ell)$ to a positive sequence $(\tilde{L}_0, \ldots, \tilde{L}_\ell)$ and counting rigid holomorphic discs with boundary on the perturbed Lagrangians.

The extended category $\mathcal{F}^\# = \mathcal{F}^\#(\mathrm{Sym}^k(F), Z)$ is defined similarly, but also includes closed balanced generalized Lagrangian submanifolds of $\mathrm{Sym}^k(F)$ (i.e., formal images of Lagrangians under sequences of balanced Lagrangian correspondences) of the sort introduced in §2.

To be more precise, the construction of the partially wrapped Fukaya category involves the completion $\hat{F} = F \cup (\partial F \times [1, \infty))$, and its symmetric

product $\mathrm{Sym}^k(\hat{F})$. Arcs in F can be completed to properly embedded arcs in $\hat{F}$, translation-invariant in the cylindrical ends, and hence the objects of $\mathcal{F}(\mathrm{Sym}^k(F), Z)$ can be completed to properly embedded Lagrangian submanifolds of $\mathrm{Sym}^k(\hat{F})$ which are cylindrical at infinity. The Riemann surface $\hat{F}$ carries a Hamiltonian vector field supported away from the interior of F and whose positive (resp. negative) time flow rotates the cylindrical ends of $\hat{F}$ in the positive (resp. negative) direction and accumulates towards the rays $Z \times [1, \infty)$. (In the cylindrical ends $\partial F \times [1, \infty)$, the generating Hamiltonian function h is of the form $h(x, r) = \rho(x)r$ where $\rho : \partial F \to [0, 1]$ satisfies $\rho^{-1}(0) = Z$). This flow on $\hat{F}$ can be used to construct a Hamiltonian flow on $\mathrm{Sym}^k(\hat{F})$ which preserves the product structure away from the diagonal (namely, the generating Hamiltonian is given by $H(\{z_1, \ldots, z_k\}) = \sum_i h(z_i)$ away from the diagonal). The A_∞-category $\mathcal{F}(\mathrm{Sym}^k(F), Z)$ is then constructed in essentially the same manner as the wrapped Fukaya category defined by Abouzaid and Seidel [1]: namely, morphism spaces are limits of the Floer complexes upon long-time perturbation by the Hamiltonian flow. (In general various technical issues could arise with this construction, but product Lagrangians in $\mathrm{Sym}^k(\hat{F})$ are fairly well-behaved, see [2]).

When a component of ∂F does not contain any point of Z, the Hamiltonian flow that we consider rotates the corresponding cylindrical end of $\hat{F}$ by arbitrarily large amounts. Hence the perturbation causes properly embedded arcs in $\hat{F}$ to wrap around the cylindrical end infinitely many times, which typically makes the complex $\hom_{\mathcal{F}}(L_0, L_1)$ infinitely generated when L_0 and L_1 are non-compact objects of $\mathcal{F}(\mathrm{Sym}^k(F), Z)$. For instance, when $Z = \emptyset$ the category we consider is simply the wrapped Fukaya category of $\mathrm{Sym}^k(\hat{F})$ as defined in [1].

3.2. The algebra of a decorated surface

Definition 9. *A* decorated surface *is a triple $\mathbb{F} = (F, Z, \underline{\alpha})$ where F is a connected compact Riemann surface with non-empty boundary, Z is a finite subset of ∂F, and $\underline{\alpha} = \{\alpha_1, \ldots, \alpha_n\}$ is a collection of disjoint properly embedded arcs in F with boundary in $\partial F \setminus Z$.*

Given a decorated surface $\mathbb{F} = (F, Z, \underline{\alpha})$, an integer $k \leq n$, and a k-element subset $s \subseteq \{1, \ldots, n\}$, the product $D_s = \prod_{i \in s} \alpha_i$ is an object of $\mathcal{F} = \mathcal{F}(\mathrm{Sym}^k(F), Z)$. The endomorphism algebra of the direct sum of these objects is an A_∞-algebra naturally associated to $\mathbb{F}$.

Definition 10. *For $k \leq n$, denote by $\mathcal{S}_k^n$ the set of all k-element subsets of $\{1, \ldots, n\}$. Then to a decorated surface $\mathbb{F} = (F, Z, \underline{\alpha} = \{\alpha_1, \ldots, \alpha_n\})$ and an integer $k \leq n$ we associate the A_∞-algebra*

$$\mathcal{A}(\mathbb{F}, k) = \bigoplus_{s,t \in \mathcal{S}_k^n} \hom_{\mathcal{F}}(D_s, D_t), \quad \textit{where } D_s = \prod_{i \in s} \alpha_i,$$

with differential and products defined by those of $\mathcal{F} = \mathcal{F}(\mathrm{Sym}^k(F), Z)$.

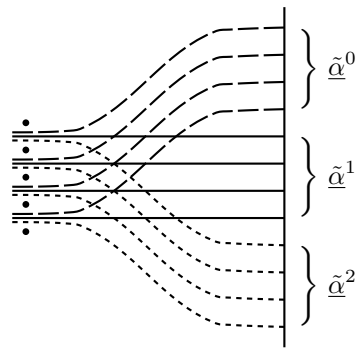


Figure 2. Positive perturbations of $\underline{\alpha}$ near ∂F

In the rest of this section, we focus on a special case where $\mathcal{A}(\mathbb{F}, k)$ can be expressed in purely combinatorial terms, and is in fact isomorphic to (the obvious generalization of) the *bordered algebra* introduced by Lipshitz, Ozsváth and Thurston [6]. The following proposition implies Theorem 2 as a special case:

Proposition 11. *Let $\mathbb{F} = (F, Z, \underline{\alpha})$ be a decorated surface, and assume that every connected component of $F \setminus (\alpha_1 \cup \cdots \cup \alpha_n)$ contains at least one point of Z. For $i, j \in \{1, \ldots, n\}$, denote by χ_i^j the set of* chords *from $\partial\alpha_i$ to $\partial\alpha_j$ in $\partial F \setminus Z$, i.e. homotopy classes of immersed arcs $\gamma : [0,1] \to \partial F \setminus Z$ such that $\gamma(0) \in \partial\alpha_i$, $\gamma(1) \in \partial\alpha_j$, and the tangent vector $\gamma'(t)$ is always oriented in the positive direction along ∂F. Moreover, denote $\bar{\chi}_i^i$ the set obtained by adjoining to χ_i^i an extra element $\mathbf{1}_i$, and let $\bar{\chi}_i^j = \chi_i^j$ for $i \neq j$. Then the following properties hold:*

- *Given $s, t \in \mathcal{S}_k^n$, let $s = \{i_1, \ldots, i_k\}$, and denote by $\Phi(s,t)$ the set of bijective maps from s to t. Then the $\mathbb{Z}_2$-vector space* $\hom_{\mathcal{F}(\mathrm{Sym}^k(F),Z)}(D_s, D_t)$ *admits a basis indexed by the elements of*

$$\bar{\chi}_s^t := \bigsqcup_{f \in \Phi(s,t)} \left(\bar{\chi}_{i_1}^{f(i_1)} \times \cdots \times \bar{\chi}_{i_k}^{f(i_k)} \right).$$

- *The differential and product in $\mathcal{A}(\mathbb{F}, k)$ are determined by explicit combinatorial formulas as in [6].*

- *The higher products $\{m_\ell\}_{\ell \geq 3}$ vanish identically, i.e. the A_∞-algebra $\mathcal{A}(\mathbb{F}, k)$ is in fact a differential algebra.*

Sketch of proof (see also [2]). For $\ell \geq 1$, we construct perturbations $\underline{\tilde{\alpha}}^0, \ldots, \underline{\tilde{\alpha}}^\ell$ of $\underline{\alpha}$, with $\underline{\tilde{\alpha}}^0 > \cdots > \underline{\tilde{\alpha}}^\ell$, in such a way that the diagram formed by the $\ell + 1$ collections of n arcs $\tilde{\alpha}_i^j$ on F enjoys properties similar to those of "nice" diagrams in Heegaard-Floer theory (cf. [13]). Namely, we ask that for each i

the arcs $\tilde{\alpha}_i^0, \dots, \tilde{\alpha}_i^\ell$ remain close to α_i in the interior of F, where any two of them intersect transversely exactly once; the total number of intersections in the diagram is minimal; and all intersections between the arcs of $\underline{\tilde{\alpha}}^j$ and those of $\underline{\tilde{\alpha}}^{j'}$ are transverse and occur with the same oriented angle $(j-j')\theta$ (for a fixed small $\theta > 0$) between the two arcs at the intersection point. Hence the local picture near any interval component of $\partial F \setminus Z$ is as shown in Figure 2. (At a component of ∂F which does not carry a point of Z, we need to consider arcs which wrap infinitely many times around the cylindrical end of the completed surface $\hat{F}$, but the situation is otherwise unchanged).

For $j < j'$ and $i, i' \in \{1, \dots, n\}$ we have a natural bijection between the points of $\tilde{\alpha}_i^j \cap \tilde{\alpha}_{i'}^{j'}$ and the elements of $\bar{\chi}_i^{i'}$. Hence, passing to the symmetric product, the intersections of $\tilde{D}_s^j = \prod_{i \in s} \tilde{\alpha}_i^j$ and $\tilde{D}_t^{j'} = \prod_{i' \in t} \tilde{\alpha}_{i'}^{j'}$ are transverse and in bijection with the elements of $\bar{\chi}_s^t$. The first claim follows.

The rest of the proposition follows from a calculation of the Maslov index of a holomorphic disc in $\mathrm{Sym}^k(F)$ with boundary on $\ell+1$ product Lagrangians $\tilde{D}_{s_0}^0, \dots, \tilde{D}_{s_\ell}^\ell$. Namely, let ϕ be the homotopy class of such a holomorphic disc contributing to the order ℓ product in $\mathcal{A}(\mathbb{F}, k)$. Projecting from the symmetric product to F, we can think of ϕ as a 2-chain in F with boundary on the arcs of the diagram, staying within the bounded regions of the diagram (i.e., those which do not intersect ∂F). Then the Maslov index $\mu(\phi)$ and the intersection number $i(\phi)$ of ϕ with the diagonal divisor in $\mathrm{Sym}^k(F)$ are related to each other by the following formula due to Sarkar [12]:

$$\mu(\phi) = i(\phi) + 2e(\phi) - (\ell - 1)k/2, \tag{6}$$

where $e(\phi)$ is the *Euler measure* of the 2-chain ϕ, characterized by additivity and by the property that the Euler measure of an embedded m-gon with convex corners is $1 - \frac{m}{4}$. On the other hand, since every component of $F \setminus (\alpha_1 \cup \dots \cup \alpha_n)$ contains a point of Z, the regions of the diagram corresponding to those components remain unbounded after perturbation. In particular, the regions marked by dots in Figure 2 are all unbounded, and hence not part of the support of ϕ.

This implies that the support of ϕ is contained in a union of regions which are either planar (as in Figure 2) or cylindrical (in the case of a component of ∂F which does not carry any point of Z), and within which the Euler measure of a convex polygonal region can be computed by summing contributions from its vertices, namely $\frac{1}{4} - \frac{\vartheta}{2\pi}$ for a vertex with angle ϑ. Considering the respective contributions of the $(\ell+1)k$ corners of the chain ϕ (and observing that the contributions from any other vertices traversed by the boundary of ϕ cancel out), we conclude that $e(\phi) = (\ell-1)k/4$, and $\mu(\phi) = i(\phi) \geq 0$.

On the other hand, m_ℓ counts rigid holomorphic discs, for which $\mu(\phi) = 2-\ell$. This immediately implies that $m_\ell = 0$ for $\ell \geq 3$. For $\ell = 1$, the diagram we consider is "nice", i.e. all the bounded regions are quadrilaterals; as observed by Sarkar and Wang, this implies that the Floer differential on $CF(\tilde{D}_s^0, \tilde{D}_t^1)$ counts empty embedded rectangles [13, Theorems 3.3 and 3.4]. Finally, for

$\ell = 2$, the Maslov index formula shows that the product counts discs which are disjoint from the diagonal strata in $\mathrm{Sym}^k(F)$. By an argument similar to that in [5] (see also [2, Proposition 3.5]), this implies that m_2 counts k-tuples of immersed holomorphic triangles in F which either are disjoint or overlap head-to-tail (cf. [5, Lemma 2.6]).

Finally, these combinatorial descriptions of m_1 and m_2 in terms of diagrams on F can be recast in terms of Lipshitz, Ozsváth and Thurston's definition of differentials and products in the bordered algebra [6]. Namely, the dictionary between points of $\tilde{D}^0_s \cap \tilde{D}^1_t$ proceeds by matching intersections of $\tilde{\alpha}^0_i$ with $\tilde{\alpha}^1_j$ near ∂F with chords from α_i to α_j (pictured as upwards strands in the notation of [6]), and the intersection of $\tilde{\alpha}^0_i$ with $\tilde{\alpha}^1_i$ in the interior of F with a pair of horizontal dotted lines in the graphical notation of [6]. See [2, section 3] for details. □

3.3. Generating the partially wrapped Fukaya category.

In this section, we outline the proof of Theorem 1. The main ingredients are Lefschetz fibrations on the symmetric product, their Fukaya categories as defined and studied by Seidel [14, 15], and acceleration A_∞-functors between partially wrapped Fukaya categories.

3.3.1. Lefschetz fibrations on the symmetric product. Let $\hat{F}$ be an open Riemann surface (with infinite cylindrical ends), and let $\pi : \hat{F} \to \mathbb{C}$ be a branched covering map. Assume that the critical points $q_1, \dots, q_n$ of π are non-degenerate (i.e., the covering π is *simple*), and that the critical values $p_1, \dots, p_n \in \mathbb{C}$ are distinct, lie in the unit disc, and satisfy $\mathrm{Im}\,(p_1) < \cdots < \mathrm{Im}\,(p_n)$.

Each critical point q_j of π determines a properly embedded arc $\hat{\alpha}_j \subset \hat{F}$, namely the union of the two lifts of the half-line $\mathbb{R}_{\geq 0} + p_j$ which pass through q_j.

We consider the k-fold symmetric product $\mathrm{Sym}^k(\hat{F})$ $(1 \leq k \leq n)$, equipped with the product complex structure J, and the holomorphic map $f_{n,k} : \mathrm{Sym}^k(\hat{F}) \to \mathbb{C}$ defined by $f_{n,k}(\{z_1, \dots, z_k\}) = \pi(z_1) + \cdots + \pi(z_k)$.

Proposition 12. *$f_{n,k} : \mathrm{Sym}^k(\hat{F}) \to \mathbb{C}$ is a holomorphic map with isolated non-degenerate critical points (i.e., a Lefschetz fibration); its $\binom{n}{k}$ critical points are the tuples consisting of k distinct points in $\{q_1, \dots, q_n\}$.*

Proof. Given $\underline{z} \in \mathrm{Sym}^k(\hat{F})$, denote by $z_1, \dots, z_r$ the distinct elements in the k-tuple $\underline{z}$, and by $k_1, \dots, k_r$ the multiplicities with which they appear. The tangent space $T_{\underline{z}}\mathrm{Sym}^k(\hat{F})$ decomposes into the direct sum of the $T_{\{z_i,\dots,z_i\}}\mathrm{Sym}^{k_i}(\hat{F})$, and $df_{n,k}(\underline{z})$ splits into the direct sum of the differentials $df_{n,k_i}(\{z_i, \dots, z_i\})$. Thus $\underline{z}$ is a critical point of $f_{n,k}$ if and only if $\{z_i, \dots, z_i\}$ is a critical point of f_{n,k_i} for each $i \in \{1, \dots, r\}$.

By considering the restriction of f_{n,k_i} to the diagonal stratum, we see that $\{z_i, \dots, z_i\}$ cannot be a critical point of f_{n,k_i} unless z_i is a critical point of

π. Assume now that z_i is a critical point of π, and pick a local complex coordinate w on $\hat{F}$ near z_i, in which $\pi(w) = w^2 + \text{constant}$. Then a neighborhood of $\{z_i, \dots, z_i\}$ in $\mathrm{Sym}^{k_i}(\hat{F})$ identifies with a neighborhood of the origin in $\mathrm{Sym}^{k_i}(\mathbb{C})$, with coordinates given by the elementary symmetric functions $\sigma_1, \dots, \sigma_{k_i}$. The local model for f_{n,k_i} is then

$$f_{n,k_i}(\{w_1, \dots, w_{k_i}\}) = w_1^2 + \cdots + w_{k_i}^2 + \text{constant} = \sigma_1^2 - 2\sigma_2 + \text{constant}.$$

Thus, for $k_i \geq 2$ the point $\{z_i, \dots, z_i\}$ is never a critical point of f_{n,k_i}. We conclude that the only critical points of $f_{n,k}$ are tuples of distinct critical points of π; moreover these critical points are clearly non-degenerate. $\square$

For $s \in \mathcal{S}_k^n$, we denote by $Q_s = \{q_i,\ i \in s\}$ the corresponding critical point of $f_{n,k}$, and by $P_s = \sum_{i \in s} p_i$ the associated critical value.

As in §2, equip $\mathrm{Sym}^k(\hat{F})$ with a Kähler form ω which is of product type away from the diagonal strata, and the associated Kähler metric. This allows us to associate to each critical point Q_s a properly embedded Lagrangian disc $\hat{D}_s$ in $\mathrm{Sym}^k(\hat{F})$ (called *Lefschetz thimble*), namely the set of those points in $f_{n,k}^{-1}(\mathbb{R}_{\geq 0} + P_s)$ for which the gradient flow of $\mathrm{Re}\, f_{n,k}$ converges to the critical point Q_s. A straightforward calculation shows that $\hat{D}_s = \prod_{i \in s} \hat{\alpha}_i$.

More generally, one can associate a Lefschetz thimble to any properly embedded arc γ connecting P_s to infinity: namely, the set of points in $f_{n,k}^{-1}(\gamma)$ for which symplectic parallel transport converges to the critical point Q_s. We will only consider the case where γ is a straight half-line. Given $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$, the thimble associated to the half-line $e^{i\theta}\mathbb{R}_{\geq 0} + P_s$ is again a product $\hat{D}_s(\theta) = \prod_{i \in s} \hat{\alpha}_i(\theta)$, where $\hat{\alpha}_i(\theta)$ is the union of the two lifts of the half-line $e^{i\theta}\mathbb{R}_{\geq 0} + p_j$ through q_j.

3.3.2. A special case of Theorem 1. In the same setting as above, consider the Riemann surface with boundary $F = \pi^{-1}(D^2)$, i.e. the preimage of the unit disc, and let $Z = \pi^{-1}(-1) \subset \partial F$. Let $\alpha_i = \hat{\alpha}_i \cap F$, and $D_s = \hat{D}_s \cap \mathrm{Sym}^k(F) = \prod_{i \in s} \alpha_i$. Then we can reinterpret the partially wrapped Fukaya category $\mathcal{F}(\mathrm{Sym}^k(F), Z)$ and the algebra $\mathcal{A}(\mathbb{F}, k)$ associated to the arcs $\alpha_1, \dots, \alpha_n$ in different terms.

Seidel associates to the Lefschetz fibration $f_{n,k}$ a Fukaya category $\mathcal{F}(f_{n,k})$, whose objects are compact Lagrangian submanifolds of $\mathrm{Sym}^k(\hat{F})$ on one hand, and Lefschetz thimbles associated to admissible arcs connecting a critical value of $f_{n,k}$ to infinity on the other hand [15]. Here we say that an arc is admissible with slope $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ if outside of a compact set it is a half-line of slope θ. (Seidel considers the case of an exact symplectic form, and defines things somewhat differently; however our setting does not pose any significant additional difficulties).

Morphisms between thimbles in $\mathcal{F}(f_{n,k})$ (and compositions thereof) are defined by means of suitable perturbations. Namely, given two admissible

arcs γ_0, γ_1 and the corresponding thimbles $D_0, D_1 \subset \mathrm{Sym}^k(\hat{F})$, one sets $\hom_{\mathcal{F}(f_{n,k})}(D_0, D_1) = CF(\tilde{D}_0, \tilde{D}_1)$, where $\tilde{D}_0, \tilde{D}_1$ are thimbles obtained by suitably perturbing (γ_0, γ_1) to a positive pair $(\tilde{\gamma}_0, \tilde{\gamma}_1)$, i.e. one for which the slopes satisfy $\theta_0' > \theta_1'$.

Restricting ourselves to the special case of straight half-lines, and observing that for sufficiently small $\theta_0 > \cdots > \theta_\ell$ the collections of arcs $\alpha_i(\theta_j) = \hat{\alpha}_i(\theta_j) \cap F$ form a positive sequence in the sense of §3.1, it is not hard to see that we have an isomorphism of A_∞-algebras

$$\bigoplus_{s,t \in \mathcal{S}_k^n} \hom_{\mathcal{F}(\mathrm{Sym}^k(F),Z)}(D_s, D_t) \simeq \bigoplus_{s,t \in \mathcal{S}_k^n} \hom_{\mathcal{F}(f_{n,k})}(\hat{D}_s, \hat{D}_t).$$

A key result due to Seidel is the following:

Theorem 13 (Seidel [15], Theorem 18.24). *The A_∞-category $\mathcal{F}(f_{n,k})$ is generated by the exceptional collection of thimbles $\hat{D}_s$, $s \in \mathcal{S}_k^n$.*

In other terms, every object of $\mathcal{F}(f_{n,k})$ is quasi-isomorphic to a twisted complex built out of the objects $\hat{D}_s$, $s \in \mathcal{S}_k^n$.

This implies Theorem 1 in the special case where F is a simple branched cover of the disc with n critical points, Z is the preimage of -1, and the arcs $\alpha_1, \ldots, \alpha_n$ are lifts of half-lines connecting connecting the critical values to the boundary of the disc along the real positive direction. (More precisely, in view of the relation between $\mathcal{F}(f_{n,k})$ and $\mathcal{F}(\mathrm{Sym}^k(F), Z)$, Seidel's result directly implies that the compact objects of $\mathcal{F}(\mathrm{Sym}^k(F), Z)$ are generated by the D_s. On the other hand, arbitrary products of properly embedded arcs cannot be viewed as objects of $\mathcal{F}(f_{n,k})$, but by performing sequences of arc slides we can express them explicitly as iterated mapping cones involving the generators D_s, see below.)

3.3.3. Acceleration functors. Consider a fixed surface F, and two subsets $Z \subseteq Z' \subset \partial F$. Then there exists a natural A_∞-functor from $\mathcal{F}(\mathrm{Sym}^k(F), Z')$ to $\mathcal{F}(\mathrm{Sym}^k(F), Z)$, called "acceleration functor". This functor is identity on objects, and in the present case it is simply given by an inclusion of morphism spaces. In general, it is given by the Floer-theoretic continuation maps that arise when comparing the Hamiltonian perturbations used to define morphisms and compositions in $\mathcal{F}(\mathrm{Sym}^k(F), Z')$ and $\mathcal{F}(\mathrm{Sym}^k(F), Z)$.

Consider two products $\Delta = \delta_1 \times \cdots \times \delta_k$ and $L = \lambda_1 \times \cdots \times \lambda_k$ of disjoint properly embedded arcs in F with boundary in $\partial F \setminus Z'$. Perturbing the arcs $\delta_1, \ldots, \delta_k$ and $\lambda_1, \ldots, \lambda_k$ near ∂F if needed (without crossing Z'), we can assume that the pair (Δ, L) is positive with respect to Z'. On the other hand, achieving positivity with respect to the smaller subset Z may require a further perturbation of the arcs δ_i (resp. λ_i) in the positive (resp. negative) direction along ∂F, to obtain product Lagrangians $\tilde{\Delta} = \tilde{\delta}_1 \times \cdots \times \tilde{\delta}_k$ and $\tilde{L} = \tilde{\lambda}_1 \times \cdots \times \tilde{\lambda}_k$. This perturbation can be performed in such a way as to only *create* new intersection points. The local picture is as shown on Figure 3. The key observation is that

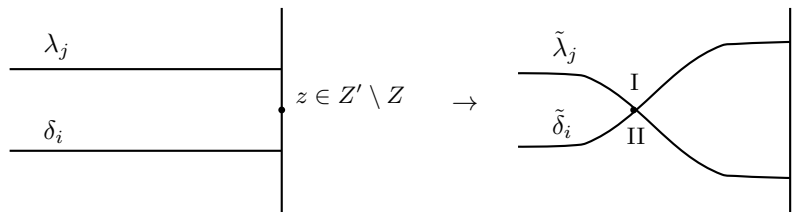


Figure 3. Perturbation and the acceleration functor

none of the intersection points created in the isotopy can be the outgoing end of a holomorphic strip in $\mathrm{Sym}^k(F)$ with boundary on $\tilde{\Delta} \cup \tilde{L}$ and whose incoming end is a previously existing intersection point (i.e., one that arises by deforming a point of $\Delta \cap L$). Indeed, considering Figure 3 right, locally the projection of this holomorphic strip to F would cover one of the two regions labelled I and II; but then by the maximum principle it would need to hit ∂F, which is not allowed. This implies that $CF(\Delta, L)$ is naturally a subcomplex of $CF(\tilde{\Delta}, \tilde{L})$. The same argument also holds for products and higher compositions, ensuring that the acceleration functor is well-defined.

In particular, given a collection $\underline{\alpha} = (\alpha_1, \ldots, \alpha_n)$ of disjoint properly embedded arcs in F, and setting $\mathbb{F} = (F, Z, \underline{\alpha})$ and $\mathbb{F}' = (F, Z', \underline{\alpha}')$, we obtain that $\mathcal{A}(\mathbb{F}', k)$ is naturally an A_∞-subalgebra of $\mathcal{A}(\mathbb{F}, k)$ for all k.

Finally, one easily checks that the acceleration functor is unital (at least on cohomology), and surjective on (isomorphism classes of) objects. Hence, if the $\binom{n}{k}$ objects $D_s = \prod_{i \in s} \alpha_i$ ($s \in \mathcal{S}_k^n$) generate $\mathcal{F}(\mathrm{Sym}^k(F), Z')$, then they also generate $\mathcal{F}(\mathrm{Sym}^k(F), Z)$. (Indeed, the assumption means that any object L of $\mathcal{F}(\mathrm{Sym}^k(F), Z')$ is quasi-isomorphic to a twisted complex built out of the D_s; since A_∞-functors are exact, this implies that L is also quasi-isomorphic to the corresponding twisted complex in $\mathcal{F}(\mathrm{Sym}^k(F), Z)$).

3.3.4. Eliminating generators by arc slides. We now consider a general decorated surface $\mathbb{F} = (F, Z, \underline{\alpha})$. The arcs $\alpha_1, \ldots, \alpha_n$ on F might not be a full set of Lefschetz thimbles for any simple branched covering map, but they are always a subset of the thimbles of a more complicated covering (with m critical points, $m \geq n$). Namely, after a suitable deformation (which does not affect the symplectic topology of the completed symmetric product $\mathrm{Sym}^k(\hat{F})$), we can always assume that F projects to the disc by a simple branched covering map π with critical values $p_1, \ldots, p_m$, in such a way that the arcs $\alpha_1, \ldots, \alpha_n$ are lifts of n of the half-lines $\mathbb{R}_{\geq 0} + p_j$, while each point of Z projects to -1. Hence, taking the remaining critical values of π and elements of $\pi^{-1}(-1)$ into account, there exists a subset $Z' \supseteq Z$ of ∂F, and a collection $\underline{\alpha}'$ of $m \geq n$ disjoint properly embedded arcs (including the α_i), such that $\mathbb{F}' = (F, Z', \underline{\alpha}')$ is as in §3.3.2. Then, as seen above, the partially wrapped Fukaya category $\mathcal{F}(\mathrm{Sym}^k(F), Z')$ is generated by the $\binom{m}{k}$ product objects $D'_s = \prod_{i \in s} \alpha'_i$ ($s \in \mathcal{S}_k^m$).

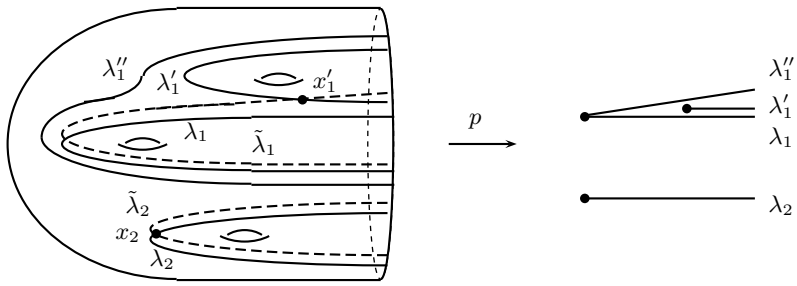


Figure 4. Sliding λ_1 along λ_1', and the auxiliary covering p

Moreover, by considering the acceleration functor as in §3.3.3, we conclude that $\mathcal{F}(\mathrm{Sym}^k(F), Z)$ is also generated by the objects D_s', $s \in \mathcal{S}_k^m$. Thus, Theorem 1 follows if, assuming that each component of $F \setminus (\alpha_1 \cup \cdots \cup \alpha_n)$ is a disc containing at most one point of Z, we can show that the $\binom{m}{k} - \binom{n}{k}$ additional objects we have introduced can be expressed in terms of the others. This is done by eliminating the additional arcs α_i' one at a time.

Consider $k+1$ disjoint properly embedded arcs $\lambda_1, \ldots, \lambda_k, \lambda_1'$ in F, with boundary in $\partial F \setminus Z$, and such that an end point of λ_1' lies immediately after an end point of λ_1 along a component of $\partial F \setminus Z$. Let λ_1'' be the arc obtained by sliding λ_1 along λ_1'. Finally, denote by $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k$ a collection of arcs obtained by slightly perturbing $\lambda_1, \ldots, \lambda_k$ in the positive direction, with each $\tilde{\lambda}_i$ intersecting λ_i in a single point $x_i \in U$, and $\tilde{\lambda}_1$ intersecting λ_1' in a single point x_1' which lies near the boundary; see Figure 4. Let $L = \lambda_1 \times \cdots \times \lambda_k$, $L' = \lambda_1' \times \lambda_2 \times \cdots \times \lambda_k$, and $L'' = \lambda_1'' \times \lambda_2 \times \cdots \times \lambda_k$. Then the point $(x_1', x_2, \ldots, x_k) \in (\tilde{\lambda}_1 \times \cdots \times \tilde{\lambda}_k) \cap (\lambda_1' \times \lambda_2 \times \cdots \times \lambda_k)$ determines (via the appropriate continuation map between Floer complexes, to account for the need to further perturb L) an element of $\hom(L, L')$, which we call u. The following result is essentially Lemma 5.2 of [2].

Lemma 14 ([2]). *In the A_∞-category of twisted complexes $Tw\,\mathcal{F}(\mathrm{Sym}^k(F), Z)$, L'' is quasi-isomorphic to the mapping cone of u.*

The main idea is to consider an auxiliary simple branched covering $p : \hat{F} \to \mathbb{C}$ for which the arcs $\lambda_1, \lambda_1', \ldots, \lambda_k$ are Lefschetz thimbles (i.e., lifts of half-lines), with the critical value for λ_1' lying immediately next to that for λ_1 and so that the monodromies at the corresponding critical values are transpositions with one common index (see Figure 4 right). The objects L, L', L'' can be viewed as Lefschetz thimbles for the Lefschetz fibration induced by p on the symmetric product; in the corresponding Fukaya category, the statement that $L'' \simeq \mathrm{Cone}(u)$ follows from a general result of Seidel [15, Proposition 18.23]. The lemma then follows from exactness of the relevant acceleration functor. See §5 of [2] for details.

The other useful fact is that sliding one factor of L over another factor of L only affects L by a Hamiltonian isotopy. For instance, in the above situation,

$\lambda_1 \times \lambda_1' \times \lambda_3 \times \cdots \times \lambda_k$ and $\lambda_1'' \times \lambda_1' \times \lambda_3 \times \cdots \times \lambda_k$ are Hamiltonian isotopic. (This is an easy consequence of the main result in [11]).

Returning to the collection of arcs $\underline{\alpha}'$ on the surface F, assume that α_m' can be erased without losing the property that every component of the complement is a disc carrying at most one point of Z. Then one of the connected components of $F \setminus (\alpha_1' \cup \cdots \cup \alpha_m')$ is a disc Δ which contains no point of Z, and whose boundary consists of portions of ∂F and the arcs $\alpha_m', \alpha_{i_1}', \ldots, \alpha_{i_r}'$ (with $i_1, \ldots, i_r$ distinct from m, but not necessarily pairwise distinct) in that order. Then the arc obtained by sliding α_{i_1}' successively over $\alpha_{i_2}', \ldots, \alpha_{i_r}'$ is isotopic to α_m'. Hence, by Lemma 14, for $m \in s$ the object D_s' can be expressed as a twisted complex built from the objects D_{s_j}', where $s_j = (s \cup \{i_j\}) \setminus \{m\}$, for $j \in \{1, \ldots, r\}$ such that $i_j \notin s$.

4. Yoneda Embedding and Invariants of Bordered 3-manifolds

Let $\mathbb{F} = (F, Z, \underline{\alpha})$ be a decorated surface, and assume that every component of $F \setminus (\bigcup \alpha_i)$ is a disc carrying at most one point of Z. By Theorem 1 the partially wrapped Fukaya category $\mathcal{F}(\mathrm{Sym}^k(F), Z)$ is generated by the product objects D_s, $s \in \mathcal{S}_k^n$. In fact Theorem 1 continues to hold if we consider the extended category $\mathcal{F}^{\#}(\mathrm{Sym}^k(F), Z)$ instead of $\mathcal{F}(\mathrm{Sym}^k(F), Z)$; see Proposition 6.3 of [2]. (The key point is that the only generalized Lagrangians we consider are compactly supported in the interior of $\mathrm{Sym}^k(F)$, and Seidel's argument for generation of compact objects by Lefschetz thimbles still applies to them.)

To each object $\mathbb{T}$ of $\mathcal{F}^{\#}(\mathrm{Sym}^k(F), Z)$ we can associate a right A_∞-module over the algebra $\mathcal{A} = \mathcal{A}(\mathbb{F}, k)$,

$$\mathcal{Y}(\mathbb{T}) = \mathcal{Y}^r(\mathbb{T}) = \bigoplus_{s \in \mathcal{S}_k^n} \hom(\mathbb{T}, D_s) \in \text{mod-}\mathcal{A},$$

where the module maps $m_\ell : \mathcal{Y}(\mathbb{T}) \otimes \mathcal{A}^{\otimes(\ell-1)} \to \mathcal{Y}(\mathbb{T})$ are defined by products and higher compositions in $\mathcal{F}^{\#}(\mathrm{Sym}^k(F), Z)$. Moreover, given two objects $\mathbb{T}_0, \mathbb{T}_1$, compositions in the partially wrapped Fukaya category yield a natural map from $\hom(\mathbb{T}_0, \mathbb{T}_1)$ to $\hom_{\text{mod-}\mathcal{A}}(\mathcal{Y}(\mathbb{T}_1), \mathcal{Y}(\mathbb{T}_0))$, as well as higher order maps. Thus, we obtain a contravariant A_∞-functor $\mathcal{Y} : \mathcal{F}^{\#}(\mathrm{Sym}^k(F), Z) \to \text{mod-}\mathcal{A}(\mathbb{F}, k)$: the *right Yoneda embedding.*

Proposition 15. *Under the assumptions of Theorem 1, $\mathcal{Y}$ is a cohomologically full and faithful (contravariant) embedding.*

Indeed, the general Yoneda embedding into $\text{mod-}\mathcal{F}^{\#}(\mathrm{Sym}^k(F), Z)$ is cohomologically full and faithful (see e.g. [15, Corollary 2.13]), while Theorem 1 (or rather its analogue for the extended Fukaya category) implies that the natural functor from $\text{mod-}\mathcal{F}^{\#}(\mathrm{Sym}^k(F), Z))$ to $\text{mod-}\mathcal{A}(\mathbb{F}, k)$ given by restricting an arbitrary A_∞-module to the subset of objects $\{D_s,\ s \in \mathcal{S}_k^n\}$ is an equivalence.

We can similarly consider the *left Yoneda embedding* to left A_∞-modules over $\mathcal{A}(\mathbb{F},k)$, namely the (covariant) A_∞-functor $\mathcal{Y}^\ell : \mathcal{F}^\#(\mathrm{Sym}^k(F),Z) \to \mathcal{A}(\mathbb{F},k)$-mod which sends the object $\mathbb{T}$ to $\mathcal{Y}^\ell(\mathbb{T}) = \bigoplus_{s\in\mathcal{S}_k^n} \hom(D_s,\mathbb{T})$.

Lemma 16. *Denote by $-\mathbb{F} = (-F,Z,\underline{\alpha})$ the decorated surface obtained by orientation reversal. Then $\mathcal{A}(-\mathbb{F},k)$ is isomorphic to the opposite A_∞-algebra $\mathcal{A}(\mathbb{F},k)^{op}$.*

Proof. Given $s_0,\dots,s_\ell \in \mathcal{S}_k^n$, and any positive perturbation $(\tilde{D}_{s_0},\dots,\tilde{D}_{s_\ell})$ of the sequence $(D_{s_0},\dots,D_{s_\ell})$ in $\mathrm{Sym}^k(F)$ relatively to Z, the reversed sequence $(\tilde{D}_{s_\ell},\dots,\tilde{D}_{s_0})$ is a positive perturbation of $(D_{s_\ell},\dots,D_{s_0})$ in $\mathrm{Sym}^k(-F)$. Thus, the holomorphic discs in $\mathrm{Sym}^k(-F)$ which contribute to the product operation $m_\ell : \hom(D_{s_\ell},D_{s_{\ell-1}})\otimes\cdots\otimes\hom(D_{s_1},D_{s_0}) \to \hom(D_{s_\ell},D_{s_0})$ in $\mathcal{A}(-\mathbb{F},k)$ are exactly the complex conjugates of the holomorphic discs in $\mathrm{Sym}^k(F)$ which contribute to $m_\ell : \hom(D_{s_0},D_{s_1})\otimes\cdots\otimes\hom(D_{s_{\ell-1}},D_{s_\ell}) \to \hom(D_{s_0},D_{s_\ell})$ in $\mathcal{A}(\mathbb{F},k)$. $\square$

Hence, left A_∞-modules over $\mathcal{A} = \mathcal{A}(\mathbb{F},k)$ can be interchangeably viewed as right A_∞-modules over $\mathcal{A}^{op} = \mathcal{A}(-\mathbb{F},k)$; more specifically, given a generalized Lagrangian $\mathbb{T}$ in $\mathrm{Sym}^k(F)$ and its conjugate $-\mathbb{T}$ in $\mathrm{Sym}^k(-F)$, the left Yoneda module $\mathcal{Y}^\ell(\mathbb{T}) \in \mathcal{A}$-mod is the same as the right Yoneda module $\mathcal{Y}^r(-\mathbb{T}) \in$ mod-$\mathcal{A}^{op}$.

Moreover, the left and right Yoneda embeddings are dual to each other:

Lemma 17. *For any object $\mathbb{T}$, the modules $\mathcal{Y}^\ell(\mathbb{T}) \in \mathcal{A}$-mod and $\mathcal{Y}^r(\mathbb{T}) \in$ mod-$\mathcal{A}$ satisfy $\mathcal{Y}^r(\mathbb{T}) \simeq \hom_{\mathcal{A}\text{-}mod}(\mathcal{Y}^\ell(\mathbb{T}),\mathcal{A})$ and $\mathcal{Y}^\ell(\mathbb{T}) \simeq \hom_{mod\text{-}\mathcal{A}}(\mathcal{Y}^r(\mathbb{T}),\mathcal{A})$ (where $\mathcal{A}$ is viewed as an A_∞-bimodule over itself).*

Proof. By definition, $\mathcal{Y}^r(\mathbb{T}) = \hom(\mathbb{T},\bigoplus_s D_s)$ (working in an additive enlargement of $\mathcal{F}^\#(\mathrm{Sym}^k(F),Z)$), with the right A_∞-module structure coming from right composition (and higher products) with endomorphisms of $\bigoplus_s D_s$. However, the left Yoneda embedding functor is full and faithful, and maps $\mathbb{T}$ to $\mathcal{Y}^\ell(\mathbb{T})$ and $\bigoplus_s D_s$ to $\mathcal{A}$. Hence, as chain complexes $\hom(\mathbb{T},\bigoplus_s D_s) \simeq \hom_{\mathcal{A}\text{-mod}}(\mathcal{Y}^\ell(\mathbb{T}),\mathcal{A})$. Moreover this quasi-isomorphism is compatible with the right module structures (by functoriality of the left Yoneda embedding). The other statement is proved similarly, by applying the right Yoneda functor (contravariant, full and faithful) to prove that $\mathcal{Y}^\ell(\mathbb{T}) = \hom(\bigoplus_s D_s,\mathbb{T}) \simeq \hom_{\text{mod-}\mathcal{A}}(\mathcal{Y}^r(\mathbb{T}),\mathcal{A})$. $\square$

All the ingredients are now in place for the proof of Theorem 4 (and other similar pairing results). Consider as in the introduction a closed 3-manifold Y obtained by gluing two 3-manifolds Y_1, Y_2 with $\partial Y_1 = -\partial Y_2 = F \cup D^2$ along their common boundary, and equip the surface F with boundary marked points Z and a collection $\underline{\alpha}$ of disjoint properly embedded arcs such that the decorated surface $\mathbb{F} = (F,Z,\underline{\alpha})$ satisfies the assumption of Theorem 1.

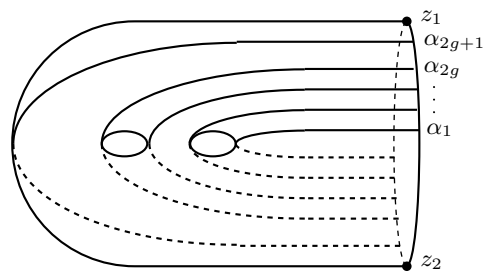


Figure 5. Decorating F with two marked points and $2g+1$ arcs

Remark. The natural choice in view of Lipshitz-Ozsváth-Thurston's work on bordered Heegaard-Floer homology [6] is to equip F with a single marked point and a collection of $2g$ arcs that decompose it into a single disc, e.g. as in Figure 1. However, one could also equip F with *two* boundary marked points and $2g+1$ arcs, by viewing F as a double cover of the unit disc with $2g+1$ branch points and proceeding as in §3.3.2; see Figure 5. While this yields a larger generating set, with $\binom{2g+1}{k}$ objects instead of $\binom{2g}{k}$, the resulting algebra remains combinatorial in nature (by Proposition 11) and it is more familiar from the perspective of symplectic geometry, since we are now dealing with the Fukaya category of a Lefschetz fibration on the symmetric product. Among other nice features, the generators are exceptional objects, and the algebra is directed.

Let us now return to our main argument. As explained in §2.3, the work of Lekili and Perutz [4] associates to the 3-manifolds Y_1 and $-Y_2$ (viewed as sutured cobordisms from D^2 to F) two generalized Lagrangian submanifolds $\mathbb{T}_{Y_1}$ and $\mathbb{T}_{-Y_2}$ of $\mathrm{Sym}^g(F)$, with the property that $\widehat{CF}(Y)$ is quasi-isomorphic to $\hom_{\mathcal{F}^\#(\mathrm{Sym}^g(F))}(\mathbb{T}_{Y_1}, \mathbb{T}_{-Y_2})$. However, by Proposition 15 we have

$$\hom_{\mathcal{F}^\#(\mathrm{Sym}^g(F))}(\mathbb{T}_{Y_1}, \mathbb{T}_{-Y_2}) \simeq \hom_{\text{mod-}\mathcal{A}(\mathbb{F},g)}(\mathcal{Y}(\mathbb{T}_{-Y_2}), \mathcal{Y}(\mathbb{T}_{Y_1}))$$

where $\mathcal{Y} = \mathcal{Y}^r$ denotes the right Yoneda embedding functor. Moreover, using Lemma 17, and setting $\mathcal{A} = \mathcal{A}(\mathbb{F}, g)$, we have:

$$\begin{aligned}
\mathcal{Y}^r(\mathbb{T}_{Y_1}) \otimes_{\mathcal{A}} \mathcal{Y}^\ell(\mathbb{T}_{-Y_2}) &\simeq \mathcal{Y}^r(\mathbb{T}_{Y_1}) \otimes_{\mathcal{A}} \hom_{\text{mod-}\mathcal{A}}(\mathcal{Y}^r(\mathbb{T}_{-Y_2}), \mathcal{A}) \\
&\simeq \hom_{\text{mod-}\mathcal{A}}(\mathcal{Y}^r(\mathbb{T}_{-Y_2}), \mathcal{Y}^r(\mathbb{T}_{Y_1}) \otimes_{\mathcal{A}} \mathcal{A}) \\
&\simeq \hom_{\text{mod-}\mathcal{A}}(\mathcal{Y}^r(\mathbb{T}_{-Y_2}), \mathcal{Y}^r(\mathbb{T}_{Y_1})).
\end{aligned}$$

Finally, by the discussion after Lemma 16, we can identify the left $\mathcal{A}(\mathbb{F}, g)$-module $\mathcal{Y}^\ell(\mathbb{T}_{-Y_2})$ with the right module $\mathcal{Y}^r(\mathbb{T}_{Y_2}) \in \text{mod-}\mathcal{A}(-\mathbb{F}, g)$. This completes the proof of Theorem 4.

Turning to the case of more general cobordisms, recall that the construction of Lekili and Perutz associates to a sutured manifold Y with $\partial Y = (-F_-) \cup F_+$ a generalized Lagrangian correspondence $\mathbb{T}_Y$ from $\mathrm{Sym}^{k_-}(F_-)$ to $\mathrm{Sym}^{k_+}(F_+)$ (where $k_+ - k_- = g(F_+) - g(F_-)$), i.e. an object of $\mathcal{F}^\#(\mathrm{Sym}^{k_-}(-F_-) \times \mathrm{Sym}^{k_+}(F_+))$.

Equip the surfaces F_+ and F_- with sets of boundary marked points $Z_\pm$ and two collections $\underline{\alpha}_\pm$ of properly embedded arcs such that the decorated surfaces $\mathbb{F}_\pm = (F_\pm, Z_\pm, \underline{\alpha}_\pm)$ satisfy the assumption of Theorem 1. Considering products of $k_\pm$ of the arcs in $\underline{\alpha}_\pm$, we have two collections of product Lagrangian submanifolds $D_{\pm,s}$ ($s \in \mathcal{S}_\pm$) in $\mathrm{Sym}^{k_\pm}(F_\pm)$. By a straightforward generalization of Theorem 1, the partially wrapped Fukaya category $\mathcal{F}^\#(\mathrm{Sym}^{k_-}(-F_-) \times \mathrm{Sym}^{k_+}(F_+), Z_- \sqcup Z_+)$ is generated by the product objects $(-D_{-,s}) \times D_{+,t}$ for $(s,t) \in \mathcal{S}_- \times \mathcal{S}_+$. Indeed, $\mathrm{Sym}^{k_-}(-F_-) \times \mathrm{Sym}^{k_+}(F_+)$ is a connected component of $\mathrm{Sym}^{k_-+k_+}((-F_-) \sqcup F_+)$, and the proof of Theorem 1 applies without modification to the disconnected decorated surface $(-\mathbb{F}_-) \sqcup \mathbb{F}_+ = ((-F_-) \sqcup F_+, Z_- \sqcup Z_+, \underline{\alpha}_- \sqcup \underline{\alpha}_+)$. Hence, as before, the Yoneda construction

$$\mathcal{Y}(\mathbb{T}_Y) = \bigoplus_{(s,t)\in\mathcal{S}_-\times\mathcal{S}_+} \hom(\mathbb{T}_Y, (-D_{-,s}) \times D_{+,t})$$

defines a cohomologically full and faithful embedding into the category of right A_∞-bimodules over $\mathcal{A}(-\mathbb{F}_-, k_-)$ and $\mathcal{A}(\mathbb{F}_+, k_+)$, or equivalently, the category of A_∞-bimodules $\mathcal{A}(\mathbb{F}_-, k_-)$-mod-$\mathcal{A}(\mathbb{F}_+, k_+)$. This property is the key ingredient that makes it possible to relate compositions of generalized Lagrangian correspondences (i.e., gluing of sutured cobordisms) to algebraic operations on A_∞-bimodules, as in Conjecture 3 for instance.

5. Relation to Bordered Heegaard-Floer Homology

Consider a sutured 3-manifold Y, with $\partial Y = (-F_-) \cup (\Gamma \times [0,1]) \cup F_+$, and pick decorations $\mathbb{F}_\pm = (F_\pm, Z_\pm, \underline{\alpha}_\pm)$ of $F_\pm$. Assume for simplicity that $Z_+ = Z_-$. Denote by $g_\pm$ the genus of $F_\pm$, and by $n_\pm$ the number of arcs in $\underline{\alpha}_\pm$. Choose a Morse function $f : Y \to [0,1]$ with index 1 and 2 critical points only, such that $f^{-1}(1) = F_-$ and $f^{-1}(0) = F_+$. Assume that all the index 1 critical points lie in $f^{-1}((0, \frac{1}{2}))$ and all the index 2 critical points lie in $f^{-1}((\frac{1}{2}, 1))$. Also pick a gradient-like vector field for f, tangent to the boundary along $\Gamma \times [0,1]$, and equip the level sets of f with complex structures such that the gradient flow induces biholomorphisms away from the critical locus. The above data determines a *bordered Heegaard diagram* on the surface $\Sigma = f^{-1}(\frac{1}{2})$ of genus $\bar{g} = g(\Sigma)$, consisting of:

- $\bar{g} - g_+$ simple closed curves $\alpha_1^c, \ldots, \alpha_{\bar{g}-g_+}^c$, where α_i^c is the set of points of Σ from which the downwards gradient flow converges to the i-th index 1 critical point;

- n_+ properly embedded arcs $\alpha_1^a, \ldots, \alpha_{n_+}^a$, where α_i^a is the set of points of Σ from which the downwards gradient flow ends at a point of $\alpha_{+,i} \subset F_+$;

- $\bar{g} - g_-$ simple closed curves $\beta_1^c, \dots, \beta_{\bar{g}-g_-}^c$, where β_i^c is the set of points of Σ from which the upwards gradient flow converges to the i-th index 2 critical point;
- n_- properly embedded arcs $\beta_1^a, \dots, \beta_{n_-}^a$, where β_i^a is the set of points of Σ from which the upwards gradient flow ends at a point of $\alpha_{-,i} \subset F_-$;
- a finite set Z of boundary marked points (which match with $Z_\pm$ under the gradient flow).

Given integers $\bar{k}, k_+, k_-$ satisfying $\bar{k} - \bar{g} = k_+ - g_+ = k_- - g_-$, we can view the generalized Lagrangian correspondence $\mathbb{T}_Y$ associated to Y as the composition of the correspondence $T_\beta \subset \mathrm{Sym}^{k_-}(-F_-) \times \mathrm{Sym}^{\bar{k}}(\Sigma)$ determined by $f^{-1}([\frac{1}{2}, 1])$ and the correspondence $T_\alpha \subset \mathrm{Sym}^{\bar{k}}(-\Sigma) \times \mathrm{Sym}^{k_+}(F_+)$ determined by $f^{-1}([0, \frac{1}{2}])$.

The A_∞-bimodule $\mathcal{Y}(\mathbb{T}_Y) \in \mathcal{A}(\mathbb{F}_-, k_-)$-mod-$\mathcal{A}(\mathbb{F}_+, k_+)$ associated to Y can then be understood entirely in terms of the symmetric product $\mathrm{Sym}^{\bar{k}}(\Sigma)$. Namely, denote by $\bar{\mathcal{F}}^\# = \bar{\mathcal{F}}^\#(\mathrm{Sym}^{\bar{k}}(\Sigma), Z)$ a partially wrapped Fukaya category defined similarly to the construction in §3, except we also allow objects which are products of mutually disjoint simple closed curves and properly embedded arcs in Σ.

The Lagrangian correspondences $-T_\alpha$ and T_β induce A_∞-functors Φ_α and Φ_β from $\mathcal{F}^\#(\mathrm{Sym}^{k_\pm}(F_\pm), Z_\pm)$ to $\bar{\mathcal{F}}^\#$. Considering the product Lagrangians $D_{-,s}$ for $s \in \mathcal{S}_- = \mathcal{S}_{k_-}^{n_-}$, the description of the geometry of the correspondence T_β away from the diagonal [10] (or the result of [4]) implies that $\Phi_\beta(D_{-,s})$, i.e., the composition of $D_{-,s}$ with the correspondence T_β, is Hamiltonian isotopic to

$$\Delta_{\beta,s} = \prod_{i \in s} \beta_i^a \times \prod_{j=1}^{\bar{g}-g_-} \beta_j^c \subset \mathrm{Sym}^{\bar{k}}(\Sigma).$$

Similarly, for $t \in \mathcal{S}_+ = \mathcal{S}_{k_+}^{n_+}$ the image of $D_{+,t}$ under the correspondence $(-T_\alpha)$ is Hamiltonian isotopic to the product

$$\Delta_{\alpha,t} = \prod_{i \in t} \alpha_i^a \times \prod_{j=1}^{\bar{g}-g_+} \alpha_j^c \subset \mathrm{Sym}^{\bar{k}}(\Sigma).$$

This implies the following result:

Proposition 18. *The A_∞-bimodule $\mathcal{Y}(\mathbb{T}_Y) \in \mathcal{A}(\mathbb{F}_-, k_-)$-mod-$\mathcal{A}(\mathbb{F}_+, k_+)$ is quasi-isomorphic to $\bigoplus_{s,t} \hom_{\bar{\mathcal{F}}^\#}(\Delta_{\beta,s}, \Delta_{\alpha,t})$.*

To clarify this statement, observe that Φ_α induces an A_∞-homomorphism from $\mathcal{A}(\mathbb{F}_+, k_+) = \bigoplus_{s,t} \hom(D_{+,s}, D_{+,t})$ to $\mathcal{A}_\alpha = \bigoplus_{s,t} \hom_{\bar{\mathcal{F}}^\#}(\Delta_{\alpha,s}, \Delta_{\alpha,t})$. In fact, suitable choices in the construction ensure that $\mathcal{A}_\alpha \simeq \mathcal{A}(\mathbb{F}_+, k_+) \otimes H^*(T^{\bar{g}-g_+}, \mathbb{Z}_2)$ and the map from $\mathcal{A}(\mathbb{F}_+, k_+)$ to $\mathcal{A}_\alpha$ is simply given by $x \mapsto x \otimes 1$. In any case, via Φ_α we can view any right A_∞-module over $\mathcal{A}_\alpha$ as a

right A_∞-module over $\mathcal{A}(\mathbb{F}_+, k_+)$. Similarly, Φ_β induces an A_∞-homomorphism from $\mathcal{A}(\mathbb{F}_-, k_-)$ to $\mathcal{A}_\beta = \bigoplus_{s,t} \hom_{\bar{\mathcal{F}}^\#}(\Delta_{\beta,s}, \Delta_{\beta,t})$, through which any left A_∞-module over $\mathcal{A}_\beta$ can be viewed as a left A_∞-module over $\mathcal{A}(\mathbb{F}_-, k_-)$.

With this understood, Proposition 18 essentially follows from the fact that the A_∞-functors induced by the correspondences T_α and $(-T_\alpha)$ on one hand, and T_β and $(-T_\beta)$ on the other hand, are adjoint to each other; see Proposition 6.6 in [2] for the case of A_∞-modules.

The case where one of $k_\pm$ vanishes, say $k_- = 0$, is of particular interest; then the β-arcs play no role whatsoever, and we only need to consider the product torus $T_\beta = \beta_1^c \times \cdots \times \beta_{\bar{g}-g_-}^c \subset \mathrm{Sym}^{\bar{k}}(\Sigma)$. This happens for instance when F_- is a disc, i.e. when Y is a 3-manifold with boundary $\partial Y = F_+ \cup D^2$ viewed as a sutured cobordism from D^2 to F_+. (This corresponds to the situation considered in [6]; in this case we have $\bar{k} = \bar{g}$ and $k_+ = g_+$).

In this situation, the statement of Proposition 18 becomes that the right A_∞-module $\mathcal{Y}(\mathbb{T}_Y) \in \text{mod-}\mathcal{A}(\mathbb{F}_+, k_+)$ is quasi-isomorphic to $\bigoplus_{t \in \mathcal{S}_+} \hom_{\bar{\mathcal{F}}^\#}(T_\beta, \Delta_{\alpha,t})$. Then we have the following result (Proposition 6.5 of [2]):

Proposition 19. *The right A_∞-modules over $\mathcal{A}(\mathbb{F}_+, k_+)$ constructed by Yoneda embedding, $\mathcal{Y}(\mathbb{T}_Y) \simeq \bigoplus_{t \in \mathcal{S}_+} \hom_{\bar{\mathcal{F}}^\#}(T_\beta, \Delta_{\alpha,t})$, and by bordered Heegaard-Floer homology, $\widehat{CFA}(Y)$, are quasi-isomorphic.*

The fact that $\bigoplus_t \hom_{\bar{\mathcal{F}}^\#}(T_\beta, \Delta_{\alpha,t})$ and $\widehat{CFA}(Y)$ are quasi-isomorphic (in fact isomorphic) as chain complexes is a straightforward consequence of the definitions. Comparing the module structures requires a comparison of the moduli spaces of holomorphic curves which determine the module maps; this can be done via a neck-stretching argument, see [2, Proposition 6.5].

Remark. Another special case worth mentioning is when $k_+ = k_- = 0$, which requires the sutured manifold Y to be balanced in the sense of [3]. Then we can discard all the arcs from the Heegaard diagram, and $\mathcal{Y}(\mathbb{T}_Y) \simeq \hom_{\bar{\mathcal{F}}^\#}(T_\beta, T_\alpha)$ is simply the chain complex which defines the sutured Floer homology of [3]. In this sense, bordered Heegaard-Floer homology and our constructions can be viewed as natural generalizations of Juhász's sutured Floer homology. (An even greater level of generality is considered in [18].)

In light of the relation between $\mathcal{Y}(\mathbb{T}_Y)$ and $\widehat{CFA}(Y)$, it is interesting to compare Theorem 4 with the pairing theorem obtained by Lipshitz, Ozsváth and Thurston for bordered Heegaard-Floer homology [6]. In particular, a side-by-side comparison suggests that the modules $\widehat{CFA}(Y)$ and $\widehat{CFD}(Y)$ might be quasi-isomorphic.

Another surprising aspect, about which we can only offer speculation, is the seemingly different manners in which bimodules arise in the two stories. In our case, bimodules arise from sutured 3-manifolds viewed as cobordisms between decorated surfaces, i.e. from bordered Heegaard diagrams where both

α- and β-arcs are simultaneously present; and pairing results arise from "top-to-bottom" stacking of cobordisms. On the other hand, the work of Lipshitz, Ozsváth and Thurston [6, 7] provides a different construction of bimodules associated to cobordisms between decorated surfaces, involving diagrams in which there are no β-arcs; and pairing results arise from "side-by-side" gluing of bordered Heegaard diagrams.

As a possible way to understand "side-by-side" gluing in our framework, observe that given two decorated surfaces $\mathbb{F}_i = (F_i, Z_i, \underline{\alpha}_i)$ for $i = 1, 2$, and given two points $z_1 \in Z_1$ and $z_2 \in Z_2$, we can form the *boundary connected sum* $F = F_1 \cup_\partial F_2$ of F_1 and F_2 by attaching a 1-handle (i.e., a band) to small intervals of ∂F_1 and ∂F_2 containing z_1 and z_2 respectively. The surface F can be equipped with the set of marked points $Z = (Z_1 \setminus \{z_1\}) \cup (Z_2 \setminus \{z_2\}) \cup \{z_-, z_+\}$, where z_- and z_+ lie on either side of the connecting handle, and the collection of properly embedded arcs $\underline{\alpha} = \underline{\alpha}_1 \cup \underline{\alpha}_2$. Assume moreover that $\mathbb{F}_1$ and $\mathbb{F}_2$ satisfy the conditions of Proposition 11, so that the associated algebras are honest differential algebras. Denoting by $\mathbb{F}$ the decorated surface $(F, Z, \underline{\alpha})$, it is then easy to check that $\mathcal{A}(\mathbb{F}, k) \simeq \bigoplus_{k_1+k_2=k} \mathcal{A}(\mathbb{F}_1, k_1) \otimes \mathcal{A}(\mathbb{F}_2, k_2)$.

Now, given two 3-manifolds Y_1, Y_2 with boundary $\partial Y_i \simeq F_i \cup D^2$, we can form their boundary connected sum $Y = Y_1 \cup_\partial Y_2$ by attaching a 1-handle at the points z_1, z_2; then $\partial Y = F \cup D^2$, and the bordered Heegaard diagram representing Y is simply the boundary connected sum of the bordered Heegaard diagrams representing Y_1 and Y_2. Accordingly, the right A_∞-module associated to Y is the tensor product (over the ground field $\mathbb{Z}_2$!) of the right A_∞-modules associated to Y_1 and Y_2. In the case where $\mathbb{F}_1 \simeq -\mathbb{F}_2$, we can glue a standard handlebody to Y in order to obtain a closed 3-manifold $\bar{Y}$, namely the result of gluing Y_1 and Y_2 along their entire boundaries rather than just at small discs near the points z_1, z_2. However, because the decorated surface $\mathbb{F}$ never satisfies the assumption of Theorem 1 (the two new marked points $z_\pm$ lie in the same component), the Yoneda functor to A_∞-modules over $\mathcal{A}(\mathbb{F}, g)$ is not guaranteed to be full and faithful, so our gluing result does not apply.

References

[1] M. Abouzaid, P. Seidel, *An open string analogue of Viterbo functoriality*, Geom. Topol. **14** (2010), 627–718.

[2] D. Auroux, *Fukaya categories of symmetric products and bordered Heegaard-Floer homology*, arXiv:1001.4323v1.

[3] A. Juhász, *Holomorphic discs and sutured manifolds*, Alg. Geom. Topol. **6** (2006), 1429–1457.

[4] Y. Lekili, T. Perutz, in preparation.

[5] R. Lipshitz, C. Manolescu, J. Wang, *Combinatorial cobordism maps in hat Heegaard Floer theory*, Duke Math. J. **145** (2008), 207–247.

[6] R. Lipshitz, P. Ozsváth, D. Thurston, *Bordered Heegaard Floer homology: invariance and pairing*, arXiv:0810.0687.

[7] R. Lipshitz, P. Ozsváth, D. Thurston, *Bimodules in bordered Heegaard Floer homology*, arXiv:1003.0598.

[8] S. Ma'u, K. Wehrheim, C. Woodward, *A_∞-functors for Lagrangian correspondences*, preprint.

[9] P. Ozsváth, Z. Szabó, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. (2) **159** (2004), 1027–1158.

[10] T. Perutz, *Lagrangian matching invariants for fibred four-manifolds: I*, Geom. Topol. **11** (2007), 759–828.

[11] T. Perutz, *Hamiltonian handleslides for Heegaard Floer homology*, Proc. 14th Gökova Geometry-Topology Conference, 2008, pp. 15–35, arXiv:0801.0564.

[12] S. Sarkar, *Maslov index formulas for Whitney n-gons*, arXiv:math.GT/0609673.

[13] S. Sarkar, J. Wang, *An algorithm for computing some Heegaard Floer homologies*, to appear in Ann. Math., arXiv:math.GT/0607777.

[14] P. Seidel, *Vanishing cycles and mutation*, Proc. 3rd European Congress of Mathematics (Barcelona, 2000), Vol. II, Progr. Math. **202**, Birkhäuser, Basel, 2001, pp. 65–85, arXiv:math.SG/0007115.

[15] P. Seidel, *Fukaya categories and Picard-Lefschetz theory*, Zurich Lect. in Adv. Math., European Math. Soc., Zürich, 2008.

[16] K. Wehrheim, C. Woodward, *Functoriality for Lagrangian correspondences in Floer theory*, arXiv:0708.2851.

[17] K. Wehrheim, C. Woodward, *Quilted Floer cohomology*, arXiv:0905.1370.

[18] R. Zarev, *Bordered Floer homology for sutured manifolds*, arXiv:0908.1106.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

A Geometric Construction of the Witten Genus, I

Kevin Costello*

Abstract

I describe how the Witten genus of a complex manifold X can be seen from a rigorous analysis of a certain two-dimensional quantum field theory of maps from a surface to X.

Mathematics Subject Classification (2010). 58J26, 81T40

Keywords. Elliptic genera, quantum field theory

1. Introduction

This paper will describe an application of my work on the foundations of quantum field theory (much of it joint with Owen Gwilliam) to topology. I will show how consideration of certain two-dimensional quantum field theories – called holomorphic Chern-Simons theories – leads to a geometric construction of the Witten genus.

Usually the Witten genus is defined by its q-expansion. In the construction presented here, however, we find *directly* a function on the moduli space of (suitable decorated) elliptic curves. It is only after careful calculation that we can compute the q-expansion of this function and identify it with the Witten class.

Hopefully, this construction will give some hints about the mysterious geometric origins of elliptic cohomology.

I am very grateful to Owen Gwilliam, Mike Hopkins, Josh Shadlen, Stefan Stolz and Peter Teichner for many helpful conversations about the material in this paper.

*Department of Mathematics, Northwestern University, Evanston, Illinois, United States of America. E-mail: costello@math.northwestern.edu.

2. Hochschild Homology and the Todd Class

Before turning to elliptic cohomology and the Witten class, I will describe the analog of my construction for the Todd class.

The most familiar way in which the Todd class occurs is, of course, in the Grothendieck-Riemann-Roch theorem. Let me recall the statement. Let X be a smooth projective variety, and let E be an algebraic vector bundle on X. Then, the Grothendieck-Riemann-Roch theorem states that

$$\sum (-1)^i \dim H^i(X, E) = \int_X \mathrm{Td}(TX)\,\mathrm{ch}(E).$$

Another (and closely related) way in which the Todd class appears is in the study of deformation quantization. There is a rich literature on algebraic and non-commutative analogs of the index theorem: see [Fed96, BNT02]. Much of this literature concerns index-type statements on quantizations of general Poisson manifolds. For the purposes of this paper, we are only interested in the relatively simple case when we are quantizing the cotangent bundle of a complex manifold X.

Let Diff_X denote the algebra of differential operators on X. Let $\mathrm{Diff}_X^\hbar$ denote the sheaf of algebras on X over the ring $\mathbb{C}[\hbar]$ obtained by forming the Rees algebra of the filtered algebra Diff_X. Explicitly,

$$\mathrm{Diff}_X^\hbar \subset \mathrm{Diff}_X \otimes \mathbb{C}[\hbar]$$

is the subalgebra consisting of those finite sums

$$\sum \hbar^i D_i$$

where D_i is a differential operator of order at most i. Thus, $\mathrm{Diff}_X^\hbar$ is a $\mathbb{C}[\hbar]$ algebra whose specialization to $\hbar = 0$ is the commutative algebra $\mathscr{O}_{T^*X}$ of functions on the cotangent bundle of X. When specialized to a non-zero value of $\hbar$, $\mathrm{Diff}_X^\hbar$ is just Diff_X.

2.1. The theorem we are interested in states that the Todd class of X appears when one computes the Hochschild homology of the algebra $\mathrm{Diff}_X^\hbar$. The index theorem concerns, ultimately, traces of differential operators. Since $HH(\mathrm{Diff}_X^\hbar)$ is the universal recipient of a trace on the algebra $\mathrm{Diff}_X^\hbar$, it is perhaps not so surprising that the Todd class should appear in this context.

2.2. Recall that the Hochschild-Kostant-Rosenberg theorem gives a quasi-isomorphism

$$\mathfrak{I}_{HKR} : HH(\mathscr{O}_X) \cong \Omega^{-*}(X).$$

Here $HH(\mathscr{O}_X)$ refers to the sheaf of Hochschild chains of $\mathscr{O}_X$, and $\Omega^{-*}(X)$ refers to the algebra of forms of X, with reversed grading. Applied to the cotangent

bundle of X, the Hochschild-Kostant-Rosenberg theorem gives an isomorphism

$$\mathcal{I}_{HKR} : HH(\mathscr{O}_{T^*X}) \cong \Omega^{-*}(T^*X).$$

The algebra $\mathrm{Diff}_X^\hbar$ is a deformation quantization of $\mathscr{O}_{T^*X}$. We will see that the Todd genus appears when we study how $HH(\mathscr{O}_{T^*X})$ changes when we replace $\mathscr{O}_{T^*X}$ by $\mathrm{Diff}_X^\hbar$.

2.3. Before we state the theorem, we need some notation. Let $\pi \in \Gamma(T^*X, \wedge^2 T(T^*X))$ denote the canonical Poisson tensor on T^*X. Let

$$L_\pi : \Omega^i(T^*X) \to \Omega^{i-1}(T^*X)$$

denote the operator of Lie derivative with respect to π. Thus, if i_π is contraction by π,

$$L_\pi = [i_\pi, \mathrm{d}_{dR}].$$

Note that L_π makes $\Omega^{-*}(T^*X)$ into a cochain complex; the cohomology of this complex is called Poisson homology.

Let

$$\mathrm{Td}(X) \in H^0(X, \Omega^{-*}(X)) = \oplus H^i(X, \Omega^i)$$

be the Todd class of X. Note that the reversal of grading in the de Rham complex means that $\mathrm{Td}(X)$ is an element of cohomological degree 0.

The first statement of the theorem is as follows.

Theorem 2.3.1 (Fedosov [Fed96], Bressler-Nest-Tsyan [BNT02]). *There is a natural quasi-isomorphism of cochain complexes*

$$HH(\mathrm{Diff}_X^\hbar) \simeq \left(\Omega^{-*}(T^*X)[\hbar], \hbar L_\pi\right)$$

sending $1 \in HH(\mathrm{Diff}_X^\hbar)$ *to*

$$\mathrm{Td}(X) \in \mathbb{R}\Gamma(X, \Omega^{-*}(X)).$$

2.4. This is a rather weak formulation of the theorem, because both sides in the quasi-isomorphism are simply cochain complexes. There is a refined version which identifies a certain algebraic structure present on both sides. It will take a certain amount of preparation to state this refined version.

The operator L_π is an order two differential operator with respect to the natural product on $\Omega^{-*}(T^*X)$. We will let $\{-,-\}_\pi$ denote the Poisson bracket on $\Omega^{-*}(T^*X)$ of cohomological degree 1 defined by the standard formula

$$\{a, b\}_\pi = L_\pi(ab) - (L_\pi a)b - (-1)^{|a|} a L_\pi b.$$

The bracket $\{-,-\}_\pi$ is of cohomological degree 1, and satisfies the standard Leibniz rule. Further, L_π is a derivation for the bracket $\{-,-\}_\pi$.

Theorem 2.4.2. *There is a quasi-isomorphism of cochain complexes*

$$HH(\mathrm{Diff}_X^{\hbar}) \simeq \left(\Omega^{-*}(T^*X)[\hbar], \hbar L_\pi + \hbar\{\log \mathrm{Td}(X), -\}\right).$$

The isomorphism in this theorem is related to that of the previous formulation by conjugating by $\mathrm{Td}(X)$.

2.5. The isomorphism appearing in this second formulation is the one that is compatible with an additional algebraic structure. The structure is that of an algebra over a certain operad, introduced by Beilinson and Drinfeld [BD04].

Definition 2.5.3. *A* Beilinson-Drinfeld algebra *A is a flat graded $\mathbb{C}[\hbar]$ module endowed with the following structures.*

1. *A commutative unital product.*
2. *A Poisson bracket $\{-,-\}$ of cohomological degree 1.*
3. *A differential $D : A \to A$ of cohomological degree 1, satisfying $D^2 = 0$ and $D1 = 0$, such that*

$$D(ab) = (Da)b + (-1)^{|a|}a(Db) + \hbar\{a, b\}.$$

The complex $HH(\mathrm{Diff}_X^{\hbar})$ is endowed with the structure of Beilinson-Drinfeld (or BD) algebra in a natural way.

The complex $\Omega^{-*}(T^*X)[\hbar])$ also has the structure of BD algebra, with product the ordinary wedge product of forms. The bracket is $\{-,-\}_\pi$, and the differential $\hbar L_\pi + \hbar\{\log \mathrm{Td}(X), -\}$.

Proposition 2.5.4. *The quasi-isomorphisms*

$$HH(\mathrm{Diff}_X^{\hbar}) \simeq \left(\Omega^{-*}(T^*X)[\hbar], \hbar L_\pi + \hbar\{\log \mathrm{Td}(X), -\}\right).$$

is a quasi-isomorphism of BD algebras.

In this lecture I will state a generalization of this result, in which the Witten class appears in place of the Todd class.

3. Factorization Algebras

Hochschild homology, K-theory and the Todd genus are all intimately concerned with the concept of associative algebra. In order to understand the Witten genus, one needs to consider a richer algebraic structure called a factorization algebra (or more precisely, a translation-invariant factorization algebra on the complex plane $\mathbb{C}$).

Factorization algebras can be defined on any smooth manifold: they can be viewed as a "multiplicative" analog of a cosheaf. In the algebro-geometric

context, factorization algebras were first considered by Beilinson and Drinfeld [BD04].

In this section, I will give the formal definition of a factorization algebra, and state a theorem (from [CG10]) which allows one to construct factorization algebras using the machinery of perturbative renormalization developed in [Cos10b].

The approach to constructing factorization algebras developed in [CG10] is a quantum field theoretic analog of the deformation quantization approach to quantum mechanics. Thus, a classical field theory yields a *commutative* factorization algebra (I will define what this means shortly). Quantizing a classical field theory amounts to replacing this commutative factorization algebra by a plain factorization algebra. Just like the Todd genus of a complex manifold X appears when one considers the deformation quantization of the cotangent bundle T^*X, we will see that the Witten genus arises when we consider the quantization of a commutative factorization algebra associated to a classical field theory whose fields are maps from a Riemann surface to T^*X.

3.1. The definition of a factorization algebra is rather straightforward to give.

Definition 3.1.5. *Let M be a manifold. A factorization algebra $\mathcal{F}$ on M consists of the following data.*

1. *For every open set $U \subset M$, a cochain complex of topological vector spaces, $\mathcal{F}(U)$.*

2. *If $U_1, \dots, U_k$ are disjoint open sets in M, all contained in a larger open set V, a continuous linear map*

$$\mathcal{F}(U_1) \otimes \cdots \otimes \mathcal{F}(U_k) \to \mathcal{F}(V)$$

   (where we use the completed projective tensor product).

3. *These maps must satisfy an evident compability condition, which says that different ways of composing these maps yield the same answer.*

4. *Finally, we need a locality axiom, saying that every element of $\mathcal{F}(V)$ can be built from elements of $\mathcal{F}(U)$ for arbitrarily small open subsets U of V.*

   Let $V \subset M$, and let $\mathcal{V} = \{V_i \mid i \in I\}$ be an open cover of V. Then, we require that

$$\mathcal{F}(V) = \operatorname{hocolim}_{U_1,\dots,U_n} F(U_1) \otimes \cdots \otimes F(U_n)$$

   where $U_1, \dots, U_n$ are disjoint subsets of V, each of which is contained in some V_j.

   If U_1, U_2 are disjoint subsets, then a particular case of this axiom says that

$$\mathcal{F}(U_1 \amalg U_2) = \mathcal{F}(U_1) \otimes \mathcal{F}(U_2).$$

This definition is reminiscent of that of an E_n algebra. In fact, Jacob Lurie has shown the following [Lur09].

Proposition 3.1.6. *There is an equivalence of $(\infty, 1)$-categories between the category of E_n algebras, and the category of factorization algebras $\mathcal{F}$ on $\mathbb{R}^n$ with the additional property that if $B \subset B'$ are balls, the map*

$$\mathcal{F}(B) \to \mathcal{F}(B')$$

is a quasi-isomorphism.

In another direction, what we call a factorization algebra is the C^∞ analog of a definition introduced by Beilinson and Drinfeld [BD04]. Beilinson and Drinfeld introduced the notion of chiral algebra in order to give a geometric formulation of the axioms of a vertex algebra. In particular, every vertex algebra yields a chiral algebra on the complex line $\mathbb{C}$, and one can turn this into a factorization algebra on $\mathbb{C}$ (considered as a Riemann surface).

3.2. As our first example of a factorization algebra, let us see how a differential graded associative algebra A gives rise to a translation-invariant factorization algebra $\mathcal{F}_A$ on $\mathbb{R}$.

We will define the value of $\mathcal{F}_A$ on the open intervals of $\mathbb{R}$; the value of $\mathcal{F}_A$ on more complicated open subsets is formally determined by this data.

Let $-\infty \leq a < b \leq \infty$, and let (a, b) be the corresponding (possibly infinite) open interval in $\mathbb{R}$. We set

$$\mathcal{F}_A((a, b)) = A.$$

If $(a, b) \subset (c, d)$, then the map

$$\mathcal{F}_A((a, b)) \to \mathcal{F}_A((c, d))$$

is the identity map on A.

If $-\infty \leq a_1 < b_1 < a_2 < b_2 < \cdots < a_n < b_n \leq \infty$, then the intervals (a_i, b_i) are disjoint. Part of the data of a factorization algebra is thus a map

$$\mathcal{F}_A((a_1, b_1)) \otimes \cdots \otimes \mathcal{F}_A((a_n, b_n)) \to \mathcal{F}_A((a_1, b_n)).$$

Once we identify each $\mathcal{F}_A((a_i, b_i))$ with A, this map is the n-fold product map

$$\begin{aligned} A^{\otimes n} &\to A \\ \alpha_1 \otimes \cdots \otimes \alpha_n &\mapsto \alpha_1 \cdot \alpha_2 \cdot \cdots \cdot \alpha_n. \end{aligned}$$

The value of $\mathcal{F}_A$ on any other open subset of $\mathbb{R}$ is determined from this data by the axioms of a factorization algebra.

4. Descent and Factorization Homology

In this paper, we are only interested in translation-invariant factorization algebras on $\mathbb{C}$. In this section, we will see that associated to such a factorization algebra $\mathcal{F}$, and to an elliptic curve $\mathcal{E}$, equipped with a never-vanishing volume element ω, one can define the *factorization homology*

$$FH(E, \mathcal{F}).$$

Factorization homology is the analog, in the world of factorization algebras, of Hochschild homology.

As motiviation, I will first explain how the Hochschild homology groups of an associative algebra A can be viewed as the factorization homology of the translation-invariant factorization algebra $\mathcal{F}_A$ on $\mathbb{R}$ associated to A.

4.1. Factorization algebras satisfy a gluing axiom. Suppose that our manifold M is written as a union $M = U \cup V$ of two open subsets. If $\mathcal{F}$ is a factorization algebra on an open subset $U \subset M$, and if $\mathcal{G}$ is a factorization algebra on V, and if

$$\phi : \mathcal{F}\mid_{U\cap V} \to \mathcal{G}\mid_{U\cap V}$$

is an isomorphism of factorization algebras on $U \cap V$, then we can construct a factorization algebra $\mathcal{H}$ on M, whose restriction to U is $\mathcal{F}$ and whose restriction to V is $\mathcal{G}$.

Similarly, factorization algebras satisfy descent. Suppose that a discrete group G acts properly discontinuously on a manifold M, and suppose that $\widetilde{\mathcal{F}}$ is a G-equivariant factorization algebra on M. Then, $\widetilde{\mathcal{F}}$ descends to a factorization algebra $\mathcal{F}$ on the quotient M/G.

4.2. Since we will be using the descent property extensively, it is worth explaining how one constructs the descended factorization algebra $\mathcal{F}$. Let us choose an open cover of M/G by connected and simply connected open subsets $\{U_i\}$. Let us choose open subsets $\widetilde{U}_i \subset M$ which map homeomorphically onto U_i.

Then, if $V \subset M/G$ is an open subset which lies in some U_i, we set

$$\mathcal{F}(V) = \widetilde{\mathcal{F}}(\widetilde{V})$$

where $\widetilde{V} \subset \widetilde{U_i}$ is the lift of V.

If $V \subset U_i \cap U_j$, then the fact that the factorization algebra $\widetilde{\mathcal{F}}$ on M is G-equivariant implies that $\mathcal{F}(V)$ is independent of the lift we choose.

If $V \subset M/G$ is an arbitrary open subset of M/G, then we set

$$\mathcal{F}(V) = \mathrm{hocolim}_{V_1,\dots,V_n} \widetilde{\mathcal{F}}(\widetilde{V}_1) \otimes \cdots \widetilde{\mathcal{F}}(\widetilde{V}_n)$$

where the homotopy colimit is over open subsets $V_1, \dots, V_n \subset M/G$ each of which lies inside one of the subsets U_j.

4.3. This descent property implies that any translation-invariant factorization algebra $\mathcal{F}$ on $\mathbb{R}$ descends to a factorization algebra $\mathcal{F}^{S^1}$ on $S^1 = \mathbb{R}/\mathbb{Z}$. We will let

$$FH(S^1, \mathcal{F}) = \mathcal{F}^{S^1}(S^1)$$

denote the complex of global sections of the factorization algebra $\mathcal{F}^{S^1}$ on S^1. We will refer to the complex $FH(S^1, \mathcal{F})$ as the factorization homology complex of S^1 with coefficients in $\mathcal{F}$.

Lemma 4.3.7. *Let $\mathcal{F}_A$ denote the factorization algebra on $\mathbb{R}$ associated to a differential graded associated algebra A. Then, there is a natural quasi-isomorphism*

$$FH(S^1, \mathcal{F}_A) \simeq HH(A)$$

between the factorization homology complex of S^1 with coefficients in $\mathcal{F}_A$, and the Hochschild complex of A.

Proof. If one analyzes the descent prescription described above, one sees that

$$FH(S^1, \mathcal{F}_A) = \mathrm{hocolim}_{I_1,\ldots,I_n} A^{\otimes n}$$

where the homotopy colimit is over disjoint unordered intervals in S^1. The maps in this homotopy colimit just arise form multiplication in A. One sees that a complex which looks like the ordinary cyclic bar complex emerges from this procedure. In [Lur09] it is proven that the result of this homotopy colimit is indeed homotopy equivalent to the cyclic bar complex. □

4.4. If $\lambda \in \mathbb{R}_{>0}$, let S^1_λ be the quotient of $\mathbb{R}$ by the lattice $\lambda\mathbb{Z}$. If $\mathcal{F}$ is a translation-invariant factorization algebra on $\mathbb{R}$, then we can descend $\mathcal{F}$ to a factorization algebra on S^1_λ, and thus define factorization homology $FH(S^1_\lambda, \mathcal{F})$. When $\lambda = 1$, this coincides with the definition given above. In principle, there is no reason that $FH(S^1_\lambda, \mathcal{F})$ should be independent of λ.

If we use the factorization algebra $\mathcal{F}_A$ arising from an associative algebra A, then all the factorization homology complexes $FH(S^1_\lambda, \mathcal{F}_A)$ are canonically isomorphic. This is because the factorization algebra $\mathcal{F}_A$ on $\mathbb{R}$ is not only translation invariant but also dilation invariant.

4.5. As I mentioned earlier, the factorization algebras relevant to the Witten genus are translation-invariant factorization algebras on $\mathbb{C}$. Let $\mathcal{F}$ be such a factorization algebra. Let E be an elliptic curve equipped with a volume element ω. We will write E as a quotient $\mathbb{C}/\Lambda$ of $\mathbb{C}$ by a lattice Λ, in such a way that form ω on E pulls back to the volume form $\mathrm{d}z$ on $\mathbb{C}$.

Since $\mathcal{F}$ is translation-invariant, it is in particular invariant under Λ. Thus, $\mathcal{F}$ descends to a factorization algebra $\mathcal{F}^E$ on E. We define the factorization homology complex of E with coefficients in $\mathcal{F}$ by

$$FH(E, \mathcal{F}) = \mathcal{F}^E(E).$$

Thus, $FH(E, \mathcal{F})$ is the global sections of $\mathcal{F}^E$ on E.

Thus, there is an analog of the Hochschild homology groups for every elliptic curve E with volume element ω.

5. Main Theorem

The main theorem states that the Witten class of a complex manifold X arises when one considers the factorization homology of a certain sheaf (on X) of translation-invariant factorization algebras on $\mathbb{C}$. Before I state this theorem, I need to recall the definition of the Witten class.

5.1. Let E be an elliptic curve, and let ω be a translation-invariant volume element on E. The Witten class

$$\mathrm{Wit}(X, E, \omega) \in \mathbb{R}\Gamma(X, \Omega^{-*}(X)) = \oplus H^i(X, \Omega^i(X))$$

is a cohomology class, defined as follows.

Let

$$E_{2k}(E, \omega) = \sum_{\lambda \in \Lambda} \lambda^{-2k}$$

be the Eisenstein series of the marked elliptic curve (E, ω). Here, as before, we are writing E as the quotient of $\mathbb{C}$ by a lattice Λ, in such a way that ω corresponds to $\mathrm{d}z$.

The Witten class of X is defined by

$$\mathrm{Wit}(X, E, \omega) = \exp\left\{ \sum_{k \geq 2} \frac{(2k-1)!}{(2\pi i)^{2k}} E_{2k}(E, \omega) ch_{2k}(TX) \right\}.$$

If τ is in the upper half-plane, let (E_τ, ω_τ) denote the elliptic curve associated to the lattice generated by $(1, \tau)$, with volume form ω_τ corresponding to $\mathrm{d}z$. Then, the Witten class has the property that

$$\lim_{\tau \to i\infty} \mathrm{Wit}(X, E_\tau, \omega_\tau) = e^{-c_1(T_X)/2} \,\mathrm{Td}(TX).$$

This follows from the identities

$$\lim_{\tau \to i\infty} E_{2k}(E_\tau, \omega_\tau) = 2\zeta(2k)$$
$$\sum_{k \geq 1} 2\zeta(2k) \frac{x^{2k}}{2k(2\pi i)^{2k}} = \log\left(\frac{x}{1 - e^{-x}}\right) - \frac{x}{2}$$

where ζ is the Riemann zeta function.

5.2. Now we can state the theorem.

Theorem 5.2.8. *Let X be a complex manifold, equipped with a trivialization of the second Chern character $ch_2(TX)$. Then, there is a sheaf $D^{\hbar}_{X,ch}$ of translation-invariant factorization algebras on $\mathbb{C}$, over the algebra $\mathbb{C}[\hbar]$, such that, for every elliptic curve E with volume element ω, there is a natural isomorphism of BD algebras*

$$FH(E, D^{\hbar}_{X,ch}) \simeq \left(\Omega^{-*}(T^*X)[\hbar], \hbar L_\pi + \hbar\{\log \mathrm{Wit}(X, E, \omega), -\}\right).$$

Alternatively, there is an isomorphism of cochain complexes

$$FH(E, D^{\hbar}_{X,ch}) \simeq \left(\Omega^{-*}(T^*X)[\hbar], \hbar L_\pi\right).$$

sending

$$1 \to \mathrm{Wit}(X, E, \omega).$$

The factorization algebra appearing in this theorem is an analytic avatar of the chiral differential operators constructed by Gorbounov, Mailkov and Schechtman [GMS00]. Note that in their work, the q-expansion of the Witten genus appears as the character of the algebra of chiral differential operators. The way the Witten genus appears in this paper is somewhat different, and has the advantage that we see the Witten genus directly as a function on the moduli space of elliptic curves, and not just as a q-expansion.

6. Factorization Algebras from Quantum Field Theory

A factorization algebra is the algebraic structure satisfied by the observables of a quantum field theory. In [CG10] we prove a theorem allowing one to construct factorization algebras using the techniques of perturbative renormalization. The factorization algebra $D^{\hbar}_{X,ch}$ encoding the Witten genus will be constructed by quantizing a certain two-dimensional quantum field theory, called holomorphic Chern-Simons theory.

Before I discuss this particular quantum field theory, let me explain, heuristically, why one would expect the observables of a quantum field theory to form a factorization algebra. Suppose we have a quantum field theory (whatever that is) on a manifold M. Then, for every open subset $U \subset M$, we would expect the set of observables on U – that is, the set of measurements that can be made by an observer in the open subset U – to form a vector space, which we call $\mathcal{F}(U)$.

If $U \subset V$, then an observable on U will, in particular, be an observable on V, so that we get a map $\mathcal{F}(U) \to \mathcal{F}(V)$.

If U_1 and U_2 are disjoint, we would expect that all obervables on $U_1 \amalg U_2$ are obtained by taking the product of an observable on U_1 with one on U_2. Thus,

we would expect that

$$\mathcal{F}(U_1 \amalg U_2) = \mathcal{F}(U_1) \otimes \mathcal{F}(U_2).$$

Together, these maps give $\mathcal{F}$ the structure of a factorization algebra.

6.1. The idea that the observables of a quantum field theory form a factorization algebra is compatible with two familiar examples.

Quantum mechanics is a quantum field theory on the real line $\mathbb{R}$. The obserables for quantum mechanics form an associative algebra. Associative algebras are a particular class of factorization algebras on $\mathbb{R}$. In [CG10], we show that the factorization algebra associated to the free field theory on $\mathbb{R}$ is, in fact, an E_1 algebra; specifically, it is the familiar Weyl algebra of observables of quantum mechanics.

A second well-understood example is conformal field theory. The observables of conformal field theory on $\mathbb{C}$ form a vertex algebra; and, as we have seen, vertex algebras are a special class of factorization algebra $\mathbb{C}$.

6.2. Let me now briefly state the results of [CG10] and [Cos10b], allowing one to construct factorization algebras.

In [Cos10b], I gave a definition of a quantum field theory on a manifold M, using a synthesis between Wilson's concept of a low-energy effective field theory and the Batalin-Vilkovisky formalism for quantizing gauge theories. Further, I developed techniques (based on the machinery of perturbative renormalization) allowing one to construct such quantum field theories from Lagrangians.

Many quantum field theories of physical and mathematical interest, such as Chern-Simons theory and Yang-Mills theory, can be put in this framework.

The most succinct way to state the main construction of [CG10] is as follows.

Theorem 6.2.9. *Any quantum field theory in the sense of [Cos10b], on a manifold M, yields a factorization algebra on M.*

We have seen that factorization algebras satisfy a descent property: if a discrete group G acts properly discontinuously on a manifold M, then a G-equivariant factorization algebra $\widetilde{\mathcal{F}}$ on M descends to the quotient M/G. Quantum field theories in the sense of [Cos10b] satisfy a similar descent property, and the construction of a factorization algebra from a quantum field theory is compatible with descent.

7. Deformation Quantization in Quantum Field Theory

In this section, I will explain a little about how one associates a factorization algebra to a classical or quantum field theory. We will see that the procedure

of quantizing a classical field theory can be interpreted in algebraic terms as a kind of deformation quantization in the world of factorization algebras.

The observables of a classical mechanical system form a commutative algebra, whereas the observables of a quantum mechanical system are only an associative algebra. We should view this commutativity as being an extra structure present on the observables of a classical system.

There is a similar story in field theory: the observables of a classical field theory on a manifold M have an extra structure, that of a *commutative* factorization algebra.

Factorization algebras form a symmetric monoidal category: if $\mathcal{F}, \mathcal{G}$ are factorization algebras, then we can define a factorization algebra $\mathcal{F} \otimes \mathcal{G}$ by the formula

$$(\mathcal{F} \otimes \mathcal{G})(U) = \mathcal{F}(U) \otimes \mathcal{G}(U)$$

for an open subset $U \subset M$.

Definition 7.0.10. *A* commutative factorization algebra *is a commutative algebra in the category of factorization algebras.*

Thus, a commutative factorization algebra assigns to every $U \subset X$ a commutative algebra $\mathfrak{F}(U)$, and to inclusion maps $U \hookrightarrow V$, a map of commutative algebras.

7.1. The main object of interest in a classical field theory is the space of solutions to the Euler-Lagrange equation. If $U \subset M$ is an open set, let $\mathcal{EL}(U)$ be this space. Sending $U \mapsto \mathcal{EL}(U)$ defines a sheaf of formal spaces on M.

This sheaf of solutions to the Euler-Lagrange equations can be encoded in the structure of a commutative factorization algebra. If $U \subset M$ is an open subset, we will let $\mathscr{O}(\mathcal{EL}(U))$ denote the space of functions on $\mathcal{EL}(U)$.

Sending $U \mapsto \mathscr{O}(\mathcal{EL}(U))$ defines a commutative factorization algebra: if $U_1, \ldots, U_n$ are disjoint open subsets of U_{n+1}, there is a restriction map

$$\mathcal{EL}(U_{n+1}) \to \mathcal{EL}(U_1) \times \cdots \times \mathcal{EL}(U_n).$$

Replacing the map of spaces by the corresponding map of algebras of functions yields the desired structure of commutative factorization algebra.

7.2. In the familiar deformation quantization story, the algebra of observables of a classical mechanical system is a commutative algebra endowed with an extra structure, namely a Poisson bracket. This extra structure is what tells us that the commutative algebra "wants" to deform into an associative algebra.

There is a similar picture in the world of factorization algebras: the commutative factorization algebra associated to a classical field theory is endowed with an extra structure, which makes it "want" to deform into a plain factorization algebra.

Ordinary Poisson algebras interpolate between commutative algebras and associative (or E_1) algebras. For us, the object describing the observables of a quantum field theory is not an E_1 algebra in a symmetric monoidal category; instead, it is an E_0 algebra. An E_0 algebra in vector spaces is simply a vector space with an element. An E_0 algebra in any symmetric monoidal category is an object of this category with a map from the unit object. An E_0 algebra in the symmetric monoidal category of factorization algebras is simply a factorization algebra, as every factorization algebra is equipped with a unit.

Thus, the analog of the Poisson operad we are searching for is an operad that interpolates between the commutative operad and the E_0 operad. Such an operad was constructed by Beilinson and Drinfeld [BD04]; we will call it the BD operad [1].

Definition 7.2.11. *Let P_0 be the graded operad over $\mathbb{C}$ generated by a commutative and associatve product, $*$, and a Poisson bracket $\{-,-\}$ of cohomological degree $+1$.*

Let BD denote the differential graded operad over the ring $\mathbb{C}[\hbar]$ which, as a graded operad, is simply $P_0 \otimes \mathbb{C}[\hbar]$, but which is equipped with differential

$$\mathrm{d}* = \hbar\{-,-\}.$$

If we specialize to $\hbar = 0$, we find the BD operad becomes the operad P_0. If we specialize to $\hbar = 1$, however, the BD operad becomes the E_0 operad. Thus, we find that the operad P_0 bears the same relationship to the operad E_0 as the usual Poisson operad bears to the associative operad E_1.

Definition 7.2.12. *A* Poisson factorization algebra *on M is a P_0 algebra in the category of factorization algebras on M. A* quantization *of a Poisson factorization algebra $\mathcal{F}_{cl}$ is a BD algebra $\mathcal{F}_q$ in the category of factorization algebras on M, together with an isomorphism*

$$\mathcal{F}_q \otimes_{\mathbb{C}[\hbar]} \mathbb{C} \cong \mathcal{F}_{cl}$$

of Poisson factorization algebras.

7.3. Now we can restate the main results of [CG10].

Theorem 7.3.13. *Every classical field theory on M gives rise to a Poisson factorization algebra on M. A quantization of this classical field theory (in the sense of [Cos10b]) gives rise to a quantization of this Poisson factorization algebra.*

[1] Beilinson and Drinfeld called this operad the Batalin-Vilkovisky operad. However, in the literature, the Batalin-Vilkovisky operad has, unfortunately, come to refer to a different object.

What I mean by a classical field theory on M is detailed in [Cos10b], but it is something rather familiar. There is a space of fields, which is taken to be the space $\mathcal{E}$ of sections of some vector bundle E on M, or more generally some space of maps $M \to N$ to some other manifold N. In addition, there is an action functional $S : \mathcal{E} \to \mathbb{R}$ (or to $\mathbb{C}$), which is taken to be the integral of some Lagrangian density. When dealing with theories with gauge symmetry, this basic picture needs to be modified by the introduction of fields which possess a cohomological degree. This more sophisticated picture is known as the Batalin-Vilkovisky formalism.

In [Cos10b, CG10] we always work in the Batalin-Vilkovisky formalism. Thus, our space of classical fields is equipped with a symplectic form of cohomological degree -1. The fact that the factorization algebra of observables of a classical field theory is equipped with a Poisson bracket of degree $+1$ is simply a version of the familiar statement that the algebra of functions on a symplectic manifold has a natural Poisson bracket.

8. Holomorphic Chern-Simons Theory

As we have seen, when we work in the BV formalism, the space of classical fields is a (typically infinite dimensional) differential graded manifold equipped with a symplectic form of cohomological degree -1. The action functional is a secondary object in this approach. The differential on the space of fields preserves the symplectic form, and thus, at least locally, is given by Poisson bracket with some Hamiltonian function S, of cohomological degree zero. This function is the classical action.

In the paper [AKSZ97], Alexandrov, Kontsevich, Schwartz and Zabronovsky introduced a beautiful and general method for constructing classical field theories in the BV formalism. Many quantum field theories studied in mathematics arise from the AKSZ construction. For example, Chern-Simons theory, Rozansky-Witten theory and the Poisson σ model all fit very naturally into this framework.

For us, the relevance of the AKSZ construction is that the classical field theory related to the Witten genus arises most naturally from the AKSZ construction.

Before I introduce the AKSZ construction, we need some notation.

Definition 8.0.14. *A* differential graded manifold *is a smooth manifold X equipped with a sheaf $\mathscr{O}_X$ of differential graded commutative algebras over $\mathbb{C}$, with the property that $\mathscr{O}_X$ is locally isomorphic as a graded algebra to $C^\infty_X[[x_1, \ldots, x_n]]$, where x_i are formal variables of cohomological degree $d_i \in \mathbb{Z}$.*

In this definition, C^∞_X refers to the sheaf of complex-valued smooth functions on X. One can talk about geometric structures – such as Poisson or symplectic structures – on a differential graded manifold.

If X is a smooth manifold, we will let X_{dR} denote the dg manifold whose underlying smooth manifold is X, and whose sheaf of functions is the complexified de Rham complex Ω^*_X of X. If X is a complex manifold, we will let $X_{\overline{\partial}}$ denote the dg manifold whose underlying smooth manifold is X, and whose sheaf of functions is the Dolbeault complex $\Omega^{0,*}_X$.

8.1. Now we can explain the AKSZ construction. Suppose we have a compact differential graded manifold M, equipped with volume element of cohomological degree k. Let N be a differential graded manifold with a symplectic form of cohomological degree l. Then, the infinite-dimensional differential graded manifold $\mathrm{Maps}(M,N)$ acquires a symplectic form of cohomological degree $l-k$.

If $f : M \to N$ is a map, then the tangent space to $\mathrm{Maps}(M,N)$ at f is

$$T_f \,\mathrm{Maps}(M,N) = \Gamma(M, f^*TN).$$

We define a pairing on $T_f \,\mathrm{Maps}(M,N)$ by the formula

$$\langle \alpha, \beta \rangle = \int_M \langle \alpha, \beta \rangle_N .$$

Since the integration map $\int_M : C^\infty(M) \to \mathbb{R}$ is of cohomological degree $-k$, and the symplectic pairing on TN is of cohomological degree m, the pairing on $T_f \,\mathrm{Maps}(M,N)$ is of cohomological degree $m-k$. The case of interest in the Batalin-Vilkovisky formalism is when $m-k=-1$.

There is a variation of this construction which applies when the source manifold M is non-compact. In this situation, the space $\mathrm{Maps}(M,N)$ has a natural integrable distribution given by the subspace

$$\Gamma_c(M, f^*TN) \subset T^c_f \,\mathrm{Maps}(M,N) \subset T_f \,\mathrm{Maps}(M,N)$$

consisting of compactly supported tangent vector fields. In this situation, instead of having a symplectic pairing on $T_f \,\mathrm{Maps}(M,N)$, we only have one on the distribution $T^c_f \,\mathrm{Maps}(M,N)$. The action functional, instead of being a closed one-form on $\mathrm{Maps}(M,N)$, is a closed one-form on the leaves of the foliation.

8.2. There are two broad classes of AKSZ theories which are commonly considered. These are the theories of Chern-Simons type, and the theories of holomorphic Chern-Simons type.

The two classes of theories are distinguished by the nature of the source dg manifold M. In theories of Chern-Simons type, the source differential graded ringed space is X_{dR}, where X is an oriented manifold. The orientation on X gives rise to a volume element on X_{dR} of cohomological degree $\dim(X)$.

The target manifolds for Chern-Simons theories of dimension k are dg symplectic manifolds of dimension $k-1$. For example, perturbative Chern-Simons theory arises when we take the target to be the dg manifold whose underlying manifold is a point, and whose algebra of functions is the algebra $C^*(\mathfrak{g})$

of cochains on a semi-simple Lie algebra $\mathfrak{g}$. The Killing form endows this dg manifold with a symplectic form of cohomological degree 2. This theory is perturbative, because maps $M_{dR} \to C^*(\mathfrak{g})$ are the same as connections on the trivial principal G bundle which are infinitesimally close to the trivial connection.

Non-perturbative Chern-Simons theory arises from a genearlized form of the AKSZ construction whcih takes the stack BG as the target manifold. Vector bundles on BG are the same as G-modules; the tangent bundle of BG is the adjoint module $\mathfrak{g}[1]$. The Killing form on $\mathfrak{g}$ is G-equivariant, and so gives rise to a symplectic form on BG of cohomological degree 2.

Rozansky-Witten theory also arises from this framework. Let $X_{\overline{\partial}}$ be a holomorphic symplectic manifold. Let us work over the base ring $\mathbb{C}[q, q^{-1}]$ where q is a parameter of degree -2. Then the symplectic form $q^{-1}\omega$ on $X_{\overline{\partial}}$ is of cohomological degree 2, and so we can define a 3-dimensional Chern-Simons type theory. The fields of this theory are maps $M_{dR} \to X_{\overline{\partial}}$, where M is a 3-manifold, and everything takes place over the base ring $\mathbb{C}[q, q^{-1}]$.

Another example is the Poisson σ-model of [Kon03, CF00]. Here, the source is Σ_{dR} where Σ is a smooth surface. The target is the differential graded manifold $T^*[1]N$, whose underlying smooth manifold is N, and whose algebra of functions is $\Gamma(N, \wedge^* TN)$. The Schouten-Nijenhuis bracket $\{-,-\}$ endows this dg manifold with a symplectic form of cohomological degree 1. The differential on $\Gamma(N, \wedge^* TN)$ is given by bracketing with the Poisson tensor π.

8.3. Let us now discuss holomorphic Chern-Simons theory, which is the only quantum field theory we will be concerned with in this paper. In holomorphic Chern-Simons theory, the source dg manifold is $X_{\overline{\partial}}$, where X is a complex manifold equipped with a never-vanishing holomorphic volume element ω (thus, X is a Calabi-Yau manifold). This volume form can be thought of as a volume element on $X_{\overline{\partial}}$ of cohomological degree $\dim_{\mathbb{C}}(X)$. Integration against this volume element is simply the map

$$\begin{aligned} \Omega^{0,\dim_{\mathbb{C}}(X)}(X) &\to \mathbb{C} \\ \alpha &\mapsto \int_X \omega \wedge \alpha. \end{aligned}$$

Theories of holomorphic Chern-Simons type on Calabi-Yau manifolds of complex dimension k can thus be constructed from dg symplectic manifolds with a symplectic form of cohomological degree $k-1$.

In this paper, we are only interested in one-dimensional holomorphic Chern-Simons theories. In these theories, the source dg manifold is $\Sigma_{\overline{\partial}}$, where Σ is a Riemann surface equipped with a never-vanishing holomorphic volume form. The target is $Y_{\overline{\partial}}$, where Y is a holomorphic symplectic manifold. (The holomorphic symplectic form can be thought of as a dg symplectic form on $Y_{\overline{\partial}}$).

8.4. We can now give a more precise statement of the theorem relating elliptic cohomology and the Witten genus.

Theorem 8.4.15. *Let X be a complex manifold. Then,*

1. *The obstruction to quantizing the holomorphic Chern-Simons theory whose fields are maps $\mathbb{C}_{\overline{\partial}} \to (T^*X)_{\overline{\partial}}$ is*
$$\mathrm{ch}_2(TX) \in H^2(X, \Omega^2_{cl}(X))$$
where $\Omega^2_{cl}(X)$ is the sheaf of closed holomorphic 2-forms on X.

2. *If this obstruction vanishes (or, more precisely, is trivialized), then we can quantize holomorphic Chern-Simons theory to yield a factorization algebra on $\mathbb{C}$ with values in quasi-coherent sheaves on $X_{dR} \times \operatorname{Spec} \mathbb{C}[\hbar]$. We will call this factorization $D^{\hbar}_{X,ch}$.*

3. *If E is an elliptic curve, then there is a quasi-isomorphism of BD algebras in quasi-coherent sheaves on $X_{dR} \times \mathbb{C}[\hbar]$*
$$FH(E, D^{\hbar}_{X,ch}) \simeq \left(\Omega^{-*}(T^*X)[\hbar], \hbar L_\pi + \hbar\{\log \mathrm{Wit}(X,E) -\}\right).$$

8.5. Recall that the factorization homology complex $FH(E, D^{\hbar}_{X,ch})$ is defined by first constructing a factorization algebra $D^{E,\hbar}_{X,ch}$ on E, using the descent property of factorization algebras; and then taking global sections.

Quantum field theories in the sense of [Cos10b] have a descent property similar to that satisfied by factorization algebras, and the construction of a factorization algebra from a quantum field theory is compatible with descent. The quantum field theory on an elliptic curve E which arises by descent from holomorphic Chern-Simons theory on $\mathbb{C}$ is simply holomorphic Chern-Simons theory on E.

Thus, one can interpret the factorization homology group $FH(E, D^{\hbar}_{X,ch})$ in terms of holomorphic Chern-Simons theory on the elliptic curve E. From this point of view, $FH(E, D^{\hbar}_{X,ch})$ is the cochain complex of global observables for the holomorphic Chern-Simons theory of maps $E \to T^*X$.

This theorem is proved using the Wilsonian approach to quantum field theory developed in [Cos10b]. The result is then translated into the language of factorization algebras. The proof appears in [Cos10a].

References

[AKSZ97] M. Alexandrov, M. Kontsevich, A. Schwarz and O. Zabronovsky, *The Geometry of the master equation and topological field theory*, Internat. J. Modern Phys. **12**(7), 1405–1429 (1997), hep-th/9502010.

[BD04] A. Beilinson and V. Drinfeld, *Chiral algebras*, volume 51 of *American Mathematical Society Colloquium Publications*, American Mathematical Society, Providence, RI, 2004.

[BNT02] P. Bressler, R. Nest and B. Tsyagn, *Riemann-Roch theorems via deformation quantization, I*, Adv. Math. **167**(1), 1–25 (2002), math.AG/9904121.

[CF00] A. Cattaneo and G. Felder, *A path-integral approach to the Kontsevich quantization formula*, Comm. Math. Phys. **212**, 591–611 (2000).

[CG10] K. Costello and O. Gwilliam, *Factorization algebras in perturbative quantum field theory*, (2010).

[Cos10a] K. Costello, *A geometric construction of the Witten genus, II*, (2010).

[Cos10b] K. Costello, *Renormalization and effective field theory*, (2010), `http://www.math.northwestern.edu/~costello/`.

[Fed96] B. Fedosov, *Deformation quantization and index theory*, Akademie Verlag, 1996.

[GMS00] V. Gorbounov, F. Malikov and V. Schechtman, *Gerbes of chiral differential operators*, Math. Res. Lett. 7(1), 55–66 (2000).

[Kon03] M. Kontsevich, *Deformation quantization of Poisson manifolds*, Lett. Math. Phys. **66**(3), 157–216 (2003), q-alg/0709040.

[Lur09] J. Lurie, *Derived algebraic geometry VI: E_k algebras.*, (2009), `http://math.mit.edu/~lurie/papers/DAG-VI.pdf`.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Hyperbolic 3-manifolds in the 2000's

David Gabai*

Abstract

The first decade of the 2000's has seen remarkable progress in the theory of hyperbolic 3-manifolds. We report on some of these developments.

Mathematics Subject Classification (2010). Primary 57M50; Secondary 20F65, 30F40, 51M10, 51M25, 57N10, 57S05.

Keywords. Hyperbolic 3-manifold, generalized Smale conjecture, tube, tameness, volume, Weeks' manifold, ending lamination

1. Introduction

Hyperbolic geometry and 3-manifold topology have long and rich histories. (See [Miln] for highlights of hyperbolic geometry from the 19th century through the 1970's and and [Gor] for highlights of 3-manifold theory from the end of the 19th century to 1960.) In the 1970's both fields were revolutionized by Thurston's work on hyperbolization and more generally geometrization. These fields were also deeply affected by Marden's earlier pioneering work introducing modern 3-manifold topology into Kleinian group theory. Many of the problems and conjectures that drove contemporary research in these fields were formulated during the 1970's. For many years, despite intense effort and incremental progress these problems seemed intractable. However, this decade has seen unusual progress with the resolution of Marden's tameness conjecture, Thurston's ending lamination conjecture, the Bers - Sullivan - Thurston density conjecture, the Smale conjecture for hyperbolic 3-manifolds, the identification of the minimal volume hyperbolic 3-orbifold and 3-manifold, and Perelman's spectacular proof of Thurston's geometrization conjecture. This paper will survey these and other developments including recent work on the

*Partially supported by NSF grants DMS-0504110 and DMS-0554374.
Department of Mathematics, Princeton University, Princeton, NJ 08544 USA.
E-mail: gabai@math.princeton.edu.

topology of ending lamination space. We will also highlight some interesting open questions.

Unless otherwise stated all manifolds in this paper will be orientable.

2. Hyperbolization Theorem

Theorem 2.1. *(Perelman) Let N be a closed, connected irreducible 3-manifold such that $|\pi_1(N)| = \infty$ and $\mathbb{Z} \oplus \mathbb{Z}$ is not a subgroup of $\pi_1(N)$, then N admits a hyperbolic structure, i.e. it supports a metric of constant -1 sectional curvature.*

Remark 2.2. More generally Perelman proved Thurston's geometrization conjecture including the Poincare conjecture. See [P1], [P2], [KL], [MT1], [MT2], [BBBMP], for expositions and elaborations of Perelman's work.

Theorem 2.1 was proven for closed Haken manifolds by Thurston in 1978. A *Haken* 3-manifold is a compact irreducible 3-manifold containing an embedded π_1-injective embedded surface which is not the 2-sphere. In the 1970's Thurston also showed that the interior of a compact irreducible 3-manifold N with boundary a non empty union of tori supports a complete finite volume hyperbolic structure if and only if every $\mathbb{Z} \oplus \mathbb{Z}$ subgroup of $\pi_1(N)$ is peripheral (i.e. is conjugate to a subgroup of π_1 of a boundary component) and $\pi_1(N)$ is not virtually abelian.

3. Generalized Smale Conjecture for Hyperbolic 3-manifolds and the Log(3)/2 Theorem

The Mostow Rigidity theorem [Most], [Ma], [Pra] asserts that a complete finite volume hyperbolic 3-manifold N has a unique hyperbolic structure, up to isometry *homotopic* to the identity. Being homotopic, rather than isotopic to the identity leaves open the possibility that the space of hyperbolic metrics is not path connected, hence the possibility that there exists a loop $\gamma \subset N$ and hyperbolic metrics ρ_0, ρ_1 such that for $i = 0, 1$ with respect to metric ρ_i, γ is homotopic to the geodesic γ_i but γ_0 is not isotopic to γ_1. In [GMT] completing a program begun in [G1] and [G2], Robert Meyerhoff, Nathaniel Thurston and the author showed that the space of hyperbolic metrics is indeed path connected. Subsequently, using those papers I proved the generalized Smale conjecture for hyperbolic 3-manifolds.

Theorem 3.1. *([G3]) If N is a closed hyperbolic 3-manifold, then the inclusion of the isometry group* $\mathrm{Isom}(N)$ *into the diffeomorphism group* $\mathrm{Diff}(N)$ *is a homotopy equivalence. Consequently, the space of hyperbolic metrics on N is contractible and the space of diffeomorphisms of N homotopic to the identity is contractible.*

Remark 3.2. The corresponding result was proven for Haken manifolds and hence non compact complete finite volume manifolds in 1976 modulo the Smale conjecture, by Hatcher [Ha1] and Ivanov [Iv1], [Iv2]. Hatcher proved the Smale conjecture in 1983 [Ha2]. See Problem 3.47 [Ki] for a precise statement of the generalized Smale conjecture, including the following

Problem 3.3. *Prove the generalized Smale conjecture for orbifolds and spherical 3-manifolds.*

The proof of Theorem 3.1 relied on the following Log(3)/2 theorem of [GMT].

Theorem 3.4. *([GMT]) Let δ be a shortest geodesic in the closed hyperbolic 3-manifold N. Then either δ is the core of an embedded tube of radius log(3)/2 = .549306... or N is finitely covered by a manifold lying in one of seven exceptional families of manifolds. Furthermore, if* $\text{tuberadius}(\delta) < \log(3)/2$*, then either $N =$ Vol3 or* $0.5295 < \text{tuberadius}(\delta) < .5476$ *and* $1.059 < \text{length}(\delta) < 1.213$.

Remark 3.5. Here Vol3 is the third smallest closed 3-manifold in the Snappea census and was shown [GMT] to be the unique manifold in one of those families. Subsequently Champanerkar, Lewis, Lipyanskiy and Meltzer [CLLM], showed that each exceptional family contains a unique manifold. They further showed that the fundamental group of each these manifolds has two generators and two relators, where the relators were the *quasi-relators* given in [GMT]. Denoting these manifolds by $N_0, \cdots, N_6$, then $N_0 =$Vol3 and they explicitly identified five of the other six manifolds. The last one was described in [Ly]. In particular they identified N_2 as s778(-3,1) in the Snappea census [We]. In the appendix of [CLLM] Reid showed that N_5 and N_6 are isometric and that N_1 and N_5 nontrivally cover no manifold. Earlier Jones and Reid showed that $N_0 =$Vol3 covers no manifold. It was also shown in [CLLM] that no exceptional manifold covers a non orientable manifold since for all i, $|H_1(N_i)| < \infty$.

Problem 3.6. *Find all the manifolds finitely covered by N_2, N_3 and N_4. Give a complete list of all closed orientable hyperbolic 3-manifolds with a shortest geodesic that does not have a log(3)/2 tube.*

Remark 3.7. In [GMT] we conjectured that no manifold is nontrivally covered by an exceptional manifold. However, very recently, while working with the Snappea computer program [We], Maria Trnkova and the author, observed that N_2 two fold covers m010(-2,3).

Theorem 3.4 plays a crucial role in obtaining lower bounds on volumes of hyperbolic 3-manifolds. It would be extremely useful to have such a theorem for non orientable 3-manifolds and for cusped manifolds. The following is a variant of Problem 10.7 from [GMM3] for non orientable 3-manifolds.

Question 3.8. *If δ is a shortest geodesic in a closed non orientable 3-manifold N, then is tuberadius$(\delta) > log(3)/2$ or is N one of a reasonably small number of explicitly described exceptional manifolds?*

4. Marden's Tameness Conjecture, Bers - Sullivan - Thurston Density Conjecture and Thurston's Ending Lamination Conjecture

Ian Agol [Ag1] and Danny Calegari and the author [CG] independently proved the Marden Tameness Conjecture. In fact we both proved the following.

Theorem 4.1. *If N is a complete hyperbolic 3-manifold with finitely generated fundamental group, then N is geometrically and topologically tame.*

Remark 4.2. *Topologically tame* means that N is homeomorphic to the interior of a compact manifold. Marden's tameness conjecture [Ma] asserts the topological tameness conclusion of Theorem 4.1.

Remark 4.3. We [CG] proved Theorem 4.1 by first showing that N is geometrically tame. In 1990 Canary [Ca1] proved that topological tameness implies geometric tameness.

Theorem 4.1 has many applications to the theory of hyperbolic 3-manifolds.

Theorem 4.4. *(Ahlfors' measure conjecture [A]) If Γ is a finitely generated Kleinian group, then the limit set L_Γ is either S^2_∞ or has Lebesgue measure zero. If $L_\Gamma = S^2_\infty$, then Γ acts ergodically on S^2_∞.*

Remark 4.5. Thurston reduced the Ahlfors' Conjecture to showing that $N = \mathbb{H}^3/\Gamma$ is geometrically tame. His proof was clarified in the work of Bonahon and Canary. See Corollary 8.12.4 [T1], [Bo] and [Ca1].

Theorem 4.1 completed the proof of the following monumental Theorem 4.6. The last part of Theorem 4.6 requires the ending lamination theorem of Minsky [Mi] and Brock - Canary - Minsky [BCM] which was proven modulo Theorem 4.1. Various other chunks (some enormous) are due to Ahlfors, Bers, Kra, Marden, Maskit, Mostow, Prasad, Sullivan, Thurston, Masur - Minsky, Ohshika, Kleineidam - Souto, Lecuire, Kim - Lecuire - Ohshika, Hossein - Souto and Rees.

Theorem 4.6. *(Classification theorem) If N is a complete hyperbolic 3-manifold with finitely generated fundamental group, then N is determined up to isometry, by its topological type, the conformal boundary of its geometrically finite ends and the ending laminations of its geometrically infinite ends.*

The following is the main special case of the celebrated ending lamination theorem. The first part of the conclusion was conjectured by Thurston and is what is needed for Theorem 4.6.

Theorem 4.7. *[Mi], [BCM] Let $N = \mathbb{H}^3/\Gamma$ be a complete hyperbolic 3-manifold homeomorphic to $S \times \mathbb{R}$, where S is a closed surface of genus ≥ 2. Suppose that the limit set of Γ is the whole 2-sphere. Then N is determined up to isometry by the ending laminations of its geometrically infinite ends. Furthermore, N is bi-Lipshitz homeomorphic to a* model *3-manifold that is determined from the ending laminations.*

Remark 4.8. Theorem 4.1 is one of many important ingredients to the positive resolution of the following Bers - Sullivan - Thurston density conjecture, due to Ahlfors, Bers, Kra, Marden, Maskit, Mostow, Prasad, Sullivan, Thurston, Masur - Minsky, Ohshika, Kleineidam - Souto, Lecuire, Kim - Lecuire - Ohshika, Hossein - Souto, Rees, Bromberg and Brock - Bromberg.

Theorem 4.9. *(Density Theorem) If $N = \mathbb{H}^3/\Gamma$ is a complete hyperbolic 3-manifold with finitely generated fundamental group, then Γ is the algebraic limit of geometrically finite Kleinian groups.*

Remark 4.10. Theorem 4.1 is also one of many results needed to prove the following interesting and recent theorems.

Theorem 4.11. *(Culler - Shalen [CS]) If M is a closed hyperbolic 3-manifold with $Vol(M) \leq 3.44$, then $rank_{\mathbb{Z}_2} H_1(M, \mathbb{Z}_2) \leq 7$.*

Theorem 4.12. *(Long - Reid [LR]) If M is a closed hyperbolic 3-manifold, then $\pi_1(M)$ is Grothendieck rigid. I.e. there does not exist a proper finitely generated subgroup $H \subset \pi_1(M)$ such that $\hat{u} : \hat{H} \to \hat{\pi}_1(M)$ is an isomorphism, where $u : H \to \pi_1(M)$ is inclusion and "hat" denotes the profinite topology.*

In [Ca2] Canary establishes the following result by showing that it is a consequence of Theorem 4.9, [ACM] and [CaMc].

Theorem 4.13. *(Canary [Ca2]) If M is a compact 3-manifold whose interior supports a hyperbolic structure such that $\pi_1(M)$ is freely indecomposible, then $AH(M)$ has infinitely many components if and only if M has double trouble.*

Remark 4.14. Here $\mathrm{AH}(M)$ is the space of marked hyperbolic structures homotopy equivalent to M and *double trouble* means that there exists a simple closed curve γ lying in a torus component T of ∂M which is homotopic to two non isotopic (in ∂M) simple closed curves lying in $\partial M \setminus T$.

5. Volumes of Hyperbolic 3-manifolds

Background By Mostow rigidity [Most], [Ma], [Pra], the volume of a complete hyperbolic 3-manifold is a topological invariant. Thurston, building on work of

Gromov and Jorgenson, showed [T1], [T2] that the set of volumes of a complete finite volume hyperbolic 3-manifold is a well ordered closed subset of $\mathbb{R}$ of order type ω^ω and that there are only finitely many manifolds with a given volume.

This decade has seen much progress on understanding volumes of hyperbolic 3-manifolds. Among cusped manifolds, Cao and Meyerhoff [CM] showed that the figure-8 knot complement and its sister are exactly the 1-cusp manifolds of least volume. (Almost 15 years earlier Adams [Ad] showed that the non orientable Gieseking manifold is the non compact manifold, non orientable or not, of least volume. It is double covered by the figure-8 knot complement.) Subsequently, Robert Meyerhoff, Peter Milley and the author [GMM2], [Mill2] found the set of 1-cusped manifolds with volume below volume 2.848. These are the first 10 1-cusped manifolds in the Snappea census [We]. Agol [Ag2] showed that the Whitehead link and its sister are exactly the 2-cusp manifolds of least volume.

Here are three of the many interesting open problems in this area.

Problem 5.1. *Determine the set of lowest volume n-cupsed hyperbolic 3-manifolds.*

Problem 5.2. *(First open stem problem) Find all the 1-cusped hyperbolic 3-manifolds with volume at most that of the Whitehead link.*

Remark 5.3. By Thurston's Dehn surgery theorem [T1],[T2], the set of such manifolds is infinite. However, by Thurston's proof of the ω^ω theorem and Agol's smallest 2-cusped manifold theorem there are only finitely many 1-cusped manifolds with volume at most that of the Whitehead link that are not obtained by filling either the Whitehead link or its sister.

In analogy to Theorem 3.4 we have the following problem whose solution would be useful for understanding low volume manifolds as well as other problems such as ones involving Dehn surgery. E.g. see [LM].

Problem 5.4. *(Small cusp problem) [GMM3] Find all the cusped hyperbolic 3-manifolds having a maximal cusp of volume at most 2.5. In particular find all 1-cusped manifolds with cusp area at most 5.0*

Remark 5.5. Given a cusp C of a complete finite volume hyperbolic 3-manifold, there is a maximal region $M(C)$, called the *maximal cusp* of C, whose interior is foliated by horotori that cut off that cusp. The second part of Problem 5.4 is the restriction of the first to 1-cusped manifolds, since $\text{area}(\partial M(C)) = 2\,\text{volume}(M(C))$ and by *cusp area* we mean the area of the boundary of the maximal cusp.

It is known that there are infinitely many solutions to Problem 5.4. A good solution to the 1-cusped problem would be to exhibit a finite list of 1, 2, and 3-cusped manifolds with a maximal cusp of volume at most 2.5 such that any 1-cusped manifold with cusp area at most 5.0 is either one of the listed 1-cusped manifolds or is obtained by filling one of the listed 2 or 3-cusped manifolds.

In 1943 C. L. Siegel posed the problem (in modern language) of finding the smallest volume hyperbolic n-dimensional orbifold. The three dimensional version was solved a few years ago in a tour de force, by Gehring, Marshall and Martin.

Theorem 5.6. *[GM],[MM] If* Γ *is a discrete Kleinian group, then* volume$(\mathbb{H}/\Gamma) \geq 0.03905$ *and the minimum is uniquely achieved by the orientation-preserving subgroup of the* $\mathbb{Z}_2$ *extension of the tetrahedral reflection group with Coxeter diagram 3-5-3. The latter group gives the minimal volume nonorientable orbifold.*

Recently, Robert Meyerhoff, Peter Milley and the author solved the long standing problem of finding the smallest volume closed manifold. This result was the culmination of many contributions made over the previous 30 years. See [GMM3] for a detailed history of the problem.

Theorem 5.7. *[GMM1], [GMM2], [Mill2] The Matveev - Fomenko - Weeks manifold is the unique smallest volume closed hyperbolic 3-manifold. It has volume* $0.9427\cdots$.

Problem 5.8. *Find the smallest closed non orientable hyperbolic 3-manifold(s).*

Remark 5.9. In Snappea notation [We], the manifold m131(3,1) is the smallest known closed non orientable hyperbolic 3-manifold. According to Snappea, it has volume $2.02988\cdots$. Nathan Dunfield [D2] observed that this manifold fibers over S^1 with fiber the surface of genus-2 and with orientation reversing monodromy.

See [Mill1] for partial results towards this problem.

Our proof of Theorem 5.7 introduced the idea of *Mom technology*. The *Mom number* of a 3-manifold is a measure of the minimal complexity of certain types of handle structures that can describe M. More precisely, if M can be built by starting with a torus$\times I$, then attaching n 1-handles, then n valence-3 2-handles (i.e. the 2-handles run exactly three times over 1-handles) all to the torus$\times 1$ side, then attaching solid tori and cusps, and n is the minimal such number, then we say M has *Mom number* n. Meyerhoff, Milley and the author conjecture that all low volume hyperbolic 3-manifolds can be obtained by filling Mom-n manifolds with small n. See [GMM3]. Part of the challenge is to quantify low and small. We believe that this gives a viable approach to the following open ended conjecture from the early 1980's, where topological complexity is measured by the Mom number.

Hyperbolic Complexity Conjecture (Thurston, Hodgson - Weeks, Matveev - Fomenko) *The complete low-volume hyperbolic 3-manifolds can be obtained by filling cusped hyperbolic 3-manifolds of small topological complexity.*

See the expository paper [GMM3] for an introduction to Mom technology and an outline of the proof of Theorem 5.7. This paper also contains many open problems (e.g. Problem 5.2, 5.3, 5.8) on volumes of hyperbolic 3-manifolds, including several involving number theoretic issues contributed by Walter Neumann and Alan Reid. See also Problem 3.60 [Ki] for other volumes problems including the well known Problems 5.1 and 5.8.

6. Ending Lamination Space

Let S be a finite type hyperbolic surface. The *ending lamination space* $\mathcal{EL}(S)$ is the space of minimal fillling geodesic laminations on S. Here *minimal* means every leaf is dense and *filling* means that complementary regions are either open discs or once punctured open discs. Any ending lamination supports a measure of full support, thus $\mathcal{EL}(S)$ can be topologized as the quotient space of the subspace $\mathcal{FPML}(S) \subset \mathcal{PML}(S)$ of filling measured laminations, where forgetting the measure induces the map $\phi : \mathcal{FPML}(S) \to \mathcal{EL}(S)$. An equivalent topology on $\mathcal{EL}(S)$ is the *coarse Hausdorff topology* [Ha]. A sequence of ending laminations $\mathcal{L}_1, \mathcal{L}_2, \cdots$ converges to $\mathcal{L} \in \mathcal{EL}(S)$ with respect to the coarse Hausdorff topology if with respect to the Hausdorff topology on closed subsets of S, it converges to a diagonal extension of $\mathcal{L}$.

When endowed with the coarse Hausdorff topology, $\mathcal{EL}(S)$ is a fascinating and important space. To start with, Masur and Minsky [MaMi] showed that the curve complex $C(S)$ is δ-hyperbolic and Klarreich [Kl] showed that the Gromov boundary of $C(S)$ is homeomorphic to $\mathcal{EL}(S)$. (See Hammenstadt [Ha] for a more direct proof.) Ending lamination space is also central to hyperbolic 3-manifold theory because of the following result (and similar type results) which depends on the ending lamination Theorem 4.7.

Theorem 6.1. *(Leininger - Schleimer [LS]) The map* $\mathcal{E} : DD(S, \partial S) \to (\mathcal{EL}(S) \times \mathcal{EL}(S)) \setminus \Delta$ *is a homeomorphism.*

Remark 6.2. Here $DD(S, \partial S)$ is the space of doubly degenerate Kleinian surface groups and Δ denotes the diagonal. These are the Kleinian groups Γ whose associated manifolds $\mathbb{H}^3/\Gamma$ satisfy the hypothesis of Theorem 4.7.

Theorem 6.3. *[G4] If S is a finite type orientable surface of negative Euler characteristic which is not the 3-holed sphere, 4-holed sphere or 1-holed torus, then $\mathcal{EL}(S)$ is path connected, locally path connected, cyclic and has no cut points.*

Remark 6.4. The ending lamination spaces of the 3-holed sphere, 4-holed sphere and 1-holed torus are respectively, $\emptyset, \mathbb{R} \setminus \mathbb{Q}$ and $\mathbb{R} \setminus \mathbb{Q}$. Theorem 6.3 gave a positive solution to a conjecture of Peter Storm (circa. 2000) that if S is not one of the three exceptional manifolds, then $\mathcal{EL}(S)$ is connected.

Interestingly, with respect to the Hausdorff topology on compact subsets of S, Thurston (Theorem 10.2 [T3]) proved that $\mathcal{EL}(S)$ has Hausdorff dimension

zero. In fact, Zhu - Bonahon [ZB] show that with respect to the Hausdorff topology, the larger space of geodesic laminations $\mathcal{L}(S)$ has Hausdorff dimension zero.

Theorem 6.3 and the recent characterization of Nobeling manifolds, independently by Levin and Nagorko [Le], [Na], play central roles in the very recent proof of the remarkable

Theorem 6.5. *(Hensel - Przytycki [HP]) The ending lamination space of the five-punctured sphere and the twice-punctured torus is the Nobeling curve.*

Remark 6.6. The n-dimensional Nobeling manifold is the subspace of $\mathbb{R}^{2n+1}$ consisting of all points with at most n rational coordinates. The Nobeling curve is the 1-dimensional Nobeling manifold.

Hensel and Przytycki made the following bold conjecture.

Conjecture 6.7. *[HP] $\mathcal{EL}(S)$ is homeomorphic to the n-dimensional Nobeling manifold, where $n = 3g + p - 4$, g=genus(S) and p is the number of punctures.*

If true, this would give a positive answer to an earlier question of ours. We asked [G4] whether $\mathcal{EL}(S)$ was n-connected for sufficiently complicated surfaces. We suspect that the Hensel - Przytycki conjecture needs to be altered slightly.

Conjecture 6.8. *Let S be a hyperbolic surface of genus g with $p \geq 0$ punctures. Then $\mathcal{EL}(S)$ is homeomorphic to $\mathbb{R}^{2n+1}_{n+k}$ where $n = 3g + p - 4$, where*

i) $k = 0$, *if* $g = 0$

ii) $k = \text{genus}(S) - 1$, *if* $p > 0$ *and* $g > 0$ *and*

iii) $k = \text{genus}(S) - 2$, *if* $p = 0$.

Here $\mathbb{R}^p_q$ denotes the subspace of $\mathbb{R}^p$ consisting of all points with at most q rational coordinates.

Remark 6.9. Note that $2n + 1 = \dim(\mathcal{PML}(S))$. Unlike the case of $\mathbb{R}^{2n+1}_n$, the author is not aware of any characterization of the spaces $\mathbb{R}^p_q$, hence this conjecture should be viewed as very speculative. The value of k arises from Conjecture 6.12 stated below. Here are three concrete conjectures that are implied by Conjecture 6.8.

Conjecture 6.10. *Let S be a hyperbolic surface of genus g with $p \geq 0$ punctures. Then dim$(\mathcal{EL}(S)) = 3g + p - 4 + k$ where k is as in Conjecture 6.8.*

Conjecture 6.11. *Let S be a hyperbolic surface of genus g with $p \geq 0$ punctures. Then $\pi_m(\mathcal{EL}(S))$ is infinitely generated and $\mathcal{EL}(S)$ is $(m-1)$-connected and locally $(m-1)$-connected where $m = 3g + p - 4 + k$ and k is as in Conjecture 6.8.*

Conjecture 6.12. *The curve complex $\mathcal{C}(S)$ and $\mathcal{EL}(S)$ are dual in the following sense. If p (resp. q) is the minimal value such that $\pi_p(\mathcal{EL}(S)) \neq 1$ (resp. $\pi_q(\mathcal{C}(S)) \neq 1$) then dim$(\mathcal{PML}(S)) = p + q + 1$.*

Remark 6.13. A stronger form of this conjecture asserts that in an appropriate sense, elements of $\pi_q(\mathcal{C}(S))$ non trivially link elements of "$\phi^{-1}(\pi_p(\mathcal{EL}(S))$" inside of $\mathcal{PML}(S)$. Here we are identifying $\mathcal{C}(S)$ with its image under the natural injective immersion of $\mathcal{C}(S)$ into $\mathcal{PML}(S)$.

References

[Ad] C. Adams, The noncompact hyperbolic 3-manifold of minimum volume, *Proc. Amer. Math. Soc.* **100** (1987), no. 4, 601–606.

[Ag1] I. Agol, Tameness of hyperbolic 3-manifold, arXiv:math.GT/0405568.

[Ag2] I. Agol, The minimal volume orientable hyperbolic 2-cusped 3-manifolds, arXiv:0804.0043.

[AST] I. Agol, P. Storm, and W. Thurston, Lower bounds on volumes of hyperbolic Haken 3-manifolds, *J. Amer. Math. Soc.* **20** (2007), no. 4, 1053–1077.

[A] L. Ahlfors, Fundamental polyhedrons and limit point sets of Kleinian groups, *Proc. Nat. Acad. Sci. USA* **55** (1966), 251–254.

[ACM] J. Anderson, R. Canary and D. McCullough, On the topology of deformation spaces of Kleinian groups, *Annals of Math.* **152** (2000), 693–741.

[BBBMP] L. Bessieres, G. Besson, M. Boileau, S. Maillot and J. Porti, Weak collapsing and geometrization of aspherical 3-manifolds, arXiv:math.GT/0706.2065.

[Bo] F. Bonahon, *Bouts des varietes de dimension 3*, Annals of Math. **124** (1986), 71–158.

[BCM] J. Brock, R. Canary and Y. Minsky, The classification of Kleinian surface groups, II: The Ending Lamination Conjecture, arXiv:math.GT/0412006.

[CG] D. Calegari and D. Gabai, Shrinkwrapping and the taming of hyperbolic 3-manifolds, *J. AMS*, **19** (2006), 385–446.

[Ca1] R. Canary, *Ends of hyperbolic 3-manifolds*, J. Amer. Math. Soc. **6** (1993), 1–35.

[Ca2] R. Canary, Marden's tameness conjecture: history and applications, *Adv. Lect. Math.*, **6** (2008), 137–162, Int. Press.

[CaMc] R. Canary and D. McCullough, Homotopy equivalences of 3-manifolds and deformation theory of Kleinian groups, *Mem. AMS* **172**(2004), no. 812.

[CM] C. Cao and R. Meyerhoff, The orientable cusped hyperbolic 3-manifolds of minimum volume, *Invent. math.* **146** (2001) 451–478.

[CLLM] A. Champanerkar, J. Lewis, M. Lipyanskiy, S. Meltzer, Exceptional regions and associated exceptional hyperbolic 3-manifolds, with an appendix by Alan W. Reid, *Experiment. Math.* **16** (2007), 106–118.

[CS] M. Culler and P. Shalen, Four free groups and hyperbolic geometry, arXiv:math.GT/0806.1188.

[D1] N. Dunfield, Which small volume hyperbolic 3-manifolds are Haken, Lecture notes, available at http://dunfield.info/preprints.

[D2] N. Dunfield, personal communication.

[G1] D. Gabai, Homotopy hyperbolic 3-manifolds are virtually hyperbolic, *J. AMS* **7** (1994), 193–198.

[G2] D. Gabai, On the geometric and topological rigidity of hyperbolic 3-manifolds, *J. AMS* **10** (1997), 37–74.

[G3] D. Gabai, The Smale conjecture for hyperbolic 3-manifolds: $\mathrm{Isom}(\mathrm{M}^3) \simeq \mathrm{Diff}(\mathrm{M}^3)$. *J. Differential Geom.* **58** (2001), 113–149.

[G4] D. Gabai, Almost filling laminations and the connectivity of ending lamination space, *Geometry and Topology*, **13** (2009), 1017–1041.

[GMM1] D. Gabai, R. Meyerhoff and P. Milley, Mom-technology and volumes of hyperbolic 3-manifolds, arXiv:math/0606072.

[GMM2] D. Gabai, R. Meyerhoff and P. Milley, Minimum volume cusped hyperbolic three-manifolds, *J. Amer. Math. Soc.* **22** (2009), 1157–1215.

[GMM3] D. Gabai, R. Meyerhoff and P. Milley, Mom technology and hyperbolic 3-manifolds, *Cont. Math.* **510** (2010), 81–107.

[GMT] D. Gabai, R. Meyerhoff, and N. Thurston, Homotopy hyperbolic 3-manifolds are hyperbolic, *Annals of Math.* **157** (2003), 335–431.

[GM] F. Gehring and G. Martin, Minimal co-volume hyperbolic lattices, I: the spherical points of a Kleinian group, *Annals of Math.* **170** (2009), 123–161.

[Gor] C. Gordon, 3-dimensional topology up to 1960, *History of Topology*, 449-489, North-Holland, Amsterdam, 1999.

[Ha] U. Hamenstadt, Train tracks and the Gromov boundary of the complex of curves, *London Math. Soc. Lecture Notes* **329** (2006), 187–207.

[Ha1] A. Hatcher, Homeomorphisms of sufficiently large P^2-irreducible 3-manifolds, *Topology* **15** (1976), 343–347.

[Ha2] A. Hatcher, A proof of the Smale conjecture, $\mathrm{Diff}(S^3) \simeq \mathrm{O}(4)$, *Ann. of Math.* (2) **27** (1983), 553–607.

[HP] S. Hensel and P. Przytycki, The ending lamination space of the five-punctured sphere is the Nobeling curve, 2009 preprint, arXiv:math.GT/0910.3554.

[Iv1] N. Ivanov, Research in Topology II (Russian), *Notes of LOMI scientific seminars* **66** (1976), 172–176.

[Iv2] N. Ivanov, Spaces of surfaces in Waldhausen manifolds (Russian), preprint, *LOMI* P-5-80 (1980).

[JR] K. Jones and A. Reid, Vol 3 and other exceptional hyperbolic 3-manifolds, *Proc. Amer. Math. Soc.* **129** (2001), 2175–2185.

[Ki] R. Kirby, Problems in low-dimensional topology *AMS/IP Stud. Adv. Math.*, **2.2** (1997) 35–473.

[Kl] E. Klarreich, The boundary at infinity of the curve complex and the relative mapping class group, 1999 preprint.

[KL] B. Kleiner and J. Lott, Notes on Perelman's papers, *Geometry and Topology* **12** (2008), 2587–2855.

[LM] M. Lackenby and R. Meyerhoff, The maximal number of exceptional Dehn surgeries, arXiv:math.GT.0808.1176.

[LS] C. Leininger and S. Schleimer, Connectivity of the space of ending laminations, *Duke Math. J.* **150** (2009), 533–575.

[Le] M. Levin, Characterizing Nobeling spaces, arXiv:math.GT/0510571.

[LR] D. Long and A. Reid, Grothendieck's problem for 3-manifold groups, preprint.

[Ly] M. Lipyanskiy, A computer assisted application of Poincare's fundamental polyhedron theorem, preprint.

[Ma] A. Marden, The geometry of finitely generated Kleinian groups, *Annals of Math.* **99** (1974), 383–462.

[MM] T. Marshall and G. Martin, Minimal co-volume hyperbolic lattices, II: Simple torsion in Kleinian groups, preprint, 2008.

[MaMi] H. Masur and Y. Minsky, Geometry of the Complex of Curves I: Hyperbolicity, *Invent. Math.* **138** (1999), 103–149.

[MF] S. Matveev and A. Fomenko, Isoenergetic surfaces of Hamiltonian systems, the enumeration of three-dimensional manifolds in order of growth of their complexity, and the calculation of the volumes of closed hyperbolic 3-manifolds, translated in *Russian Math. Surveys* **43** (1988), 3–24.

[Mill1] P. Milley, Line arrangements in $\mathbb{H}^3$, *Proc. Amer. Math. Soc.* **133** (2005), 3115–3120.

[Mill2] P. Milley, Minimum-volume hyperbolic 3-manifolds, *J. Topol.* **2** (2009), no. 2, 181–192.

[Miln] J. Milnor, Hyperbolic geometry: The first 150 years, *Bull. Amer. Math. Soc.* **6** (1982), no. 1, 9–24.

[Mi] Y. Minsky, The classification of Kleinian surface groups, I: Models and bounds, *Annals of Math* **171** (2010) 1–107.

[MT1] J. Morgan and G. Tian, Ricci flow and the Poincare conjecture, *Clay Mathematics Monographs*, AMS - Clay Mathematics Institute, **3** (2007).

[MT2] J. Morgan and G. Tian, Completion of the proof of the geometrization conjecture, arXiv:math.GT/0809.4040.

[Most] G. Mostow, Quasi-conformal mappings in n-space and the rigidity of hyperbolic space forms, *Inst. Hautes Etudes Sci. Publ. Math.* **34** (1968), 53–104.

[Na] A. Nagorko, Characterization and topological rigidity of Nobeling manifolds, arXiv:math.GT/0602574.

[NS] S. Novikov and M. Shubin, Morse inequalities and von Neumann II_1-factors, *Dokl. Akad. Nauk SSSR* **289** (1986), 289–292.

[P1] G. Perelman, Ricci flow with surgery on three-manifolds, arXiv:math.DG/0303109.

[P2] G. Perelman, The entropy formula for the Ricci flow and its geometric applications, arXiv:math.DG/0211159.

[Pra] G. Prasad, Strong rigidity of $\mathbb{Q}$-rank 1 lattices, *Invent. math.* **21** (1973), 255–286.

[T1] W. Thurston, The geometry and topology of 3-manifolds, Princeton University lecture notes, 1980, available at http://www.msri.org/publications/books/gt3m/.

[T2] W. Thurston, Three-dimensional manifolds, Kleinian groups and hyperbolic geometry, *Bull. Amer. Math. Soc.* **6** (1982), 357–381.

[T3] W. Thurston, Minimal stretch maps between hyperbolic surfaces, unpublished preprint (1986), arXiv:math.GT/980139.

[Wa] H. Wang, Topics in totally discontinuous groups, *Symmetric Spaces*, Boothby - Weiss, New York (1972), 460–485.

[We] J. Weeks, *SnapPea*, available from the author at www.geometrygames.org.

[ZB] X. Zhu and F. Bonahon, The metric space of geodesic laminations on a surface, I, *Geom. Top.* **8** (2004), 539–564.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

The Classification of p–compact Groups and Homotopical Group Theory

Jesper Grodal*

Abstract

We survey some recent advances in the homotopy theory of classifying spaces, and homotopical group theory. We focus on the classification of p–compact groups in terms of root data over the p–adic integers, and discuss some of its consequences e.g., for finite loop spaces and polynomial cohomology rings.

Mathematics Subject Classification (2010). Primary: 55R35; Secondary: 55R37, 55P35, 20F55.

Keywords. Homotopical group theory, classifying space, p–compact group, reflection group, finite loop space, cohomology ring.

Groups are ubiquitous in real life, as symmetries of geometric objects. For many purposes in mathematics, for instance in bundle theory, it is however not the group itself but rather its classifying space, which takes center stage. The classifying space encodes the group multiplication directly in a topological space, to be studied and manipulated using the toolbox of homotopy theory. This leads to the idea of homotopical group theory, that one should try to do group theory in terms of classifying spaces.

The idea that there should be a homotopical version of group theory is an old one. The seeds were sown already in the 40s and 50s with the work of Hopf and Serre on finite H–spaces and loop spaces, and these objects were intensely studied in the 60s using the techniques of Hopf algebras, Steenrod operations, etc., in the hands of Browder, Thomas, and others. A bibliography containing 347 items was collected by James in 1970 [59]; see also [62] for a continuation.

In the same year, Sullivan, in his widely circulated MIT notes [95, 94], laid out a theory of p–completions of topological spaces, which had a profound influence on the subject. On the one hand it provided an infusion of new exotic

*Supported by an ESF EURYI award and the Danish National Research Foundation.
Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark. E-mail: jg@math.ku.dk.

examples, laying old hopes and conjectures to rest. On the other hand his theory of p–completions seemed to indicate that the dream of doing group theory on the level of classifying spaces could still be valid, if one is willing to replace real life, at least temporarily, by a p–adic existence. However, the tools for seriously digging into the world of p–complete spaces were at the time insufficient, a stumbling block being the so-called Sullivan conjecture [95, p. 179] relating fixed-points to homotopy fixed-points, at a prime p.

The impasse ended with the solution of the Sullivan conjecture by Miller [69], and the work of Carlsson [25], reported on at this congress in 1986 [70, 26], followed by the development of "Lannes theory" [63, 64] giving effective tools for calculating homotopy fixed-points and maps between classifying spaces. This led to a spate of progress. Dwyer and Wilkerson [42] defined the notion of a p–compact group, a p–complete version of a finite loop space, and showed that these objects posses much of the structure of compact Lie groups: maximal tori, Weyl groups, etc. In parallel to this, Jackowski, McClure, and Oliver [55] combined Lannes theory with space-level decomposition techniques and sophisticated homological algebra calculations to get precise information about maps between classifying spaces of compact Lie groups, that used to be out of reach. These developments were described at this congress in 1998 [36, 82].

The aim here is to report on some recent progress, building on the above mentioned achievements. In particular, a complete classification of p–compact groups has recently been obtained in collaborations involving the author [9, 8]. It states that connected p–compact groups are classified by their root data over the p–adic integers $\mathbb{Z}_p$ (once defined!), completely analogously to the classification of compact connected Lie groups by root data over $\mathbb{Z}$. It has in turn allowed for the solution of a number of problems and conjectures dating from the 60s and 70s, such as the Steenrod problem of realizing polynomial cohomology rings and the so-called maximal torus conjecture giving a completely homotopical description of compact Lie groups. By local-to-global principles the classification of p–compact groups furthermore provides a quite complete understanding of what finite loop spaces look like, integrally as well as rationally.

Homotopical group theory has branched out considerably over the last decade. There is now an expanding theory of homotopical versions of finite groups, the so-called p–local finite groups, showing signs of strong connections to deep questions in finite group theory, such as the classification of finite simple groups. There has been progress on homotopical group actions, providing in some sense a homotopical version of the "geometric representation theory" of tom Dieck [97]. And there is even evidence that certain aspects of the theory might extend to Kac–Moody groups and other classes of groups. We shall only be able to provide very small appetizers to some of these last developments, but we hope that they collectively serve as an inspiration to the reader to try to take a more homotopical approach to his or her favorite class of groups.

This paper is structured as follows: Section 1 is an algebraic prelude, discussing the theory of $\mathbb{Z}_p$–root data—the impatient reader can skip it at first,

referring back to it as needed. Section 2 gives the definition and basic properties of p–compact groups, states the classification theorem, and outlines its proof. It also presents various structural consequences for p–compact groups. Section 3 discusses applications to finite loop spaces such as an algebraic parametrization of finite loop spaces and the solution of the maximal torus conjecture. Section 4 presents the solution of the Steenrod problem of realizing polynomial cohomology rings, and finally Section 5 provides brief samples of other topics in homotopical group theory.

Notation: Throughout this paper, the word "space" will mean "topological space of the homotopy type of a CW–complex".

Acknowledgments: I would like to thank Kasper Andersen, Bill Dwyer, Haynes Miller, and Bob Oliver for providing helpful comments on a preliminary version of this paper. I take the opportunity to thank my coauthors on the various work reported on here, and in particular express my gratitude to Kasper Andersen for our mathematical collaboration and sparring through the years.

1. Root Data over the p–adic Integers

In standard Lie theory, root data classify compact connected Lie groups as well as reductive algebraic groups over algebraically closed fields. A root datum is usually packaged as a quadruple $(M, \Phi, M^\vee, \Phi^\vee)$ of roots Φ and coroots $\Phi^\vee$ in a $\mathbb{Z}$–lattice M and its dual $M^\vee$, satisfying some conditions [33]. For p–compact groups the lattices that come up are lattices over the p–adic integers $\mathbb{Z}_p$, rather than $\mathbb{Z}$, so the concept of a root datum needs to be tweaked to make sense also in this setting, and one must carry out a corresponding classification. In this section we produce a short summary of this theory, based on [79, 45, 6, 8]. In what follows R denotes a principal ideal domain of characteristic zero.

The starting point is the theory of reflection groups, surveyed e.g., in [47]. A finite R–reflection group is a pair (W, L) such that L is a finitely generated free R–module and $W \subseteq \mathrm{Aut}_R(L)$ is a finite subgroup generated by reflections, i.e., non-trivial elements σ that fix an R–submodule of corank one.

Reflection groups have been classified for several choices of R, the most well-known cases being the classification of finite real and rational reflection groups in terms of certain Coxeter diagrams [53]. Finite complex reflection groups were classified by Shephard–Todd [89] in 1954. The main (irreducible) examples in the complex case are the groups $G(m, s, n)$ of $n \times n$ monomial matrices with non-zero entries being mth roots of unity and determinant an (m/s)th root of unity, where $s|m$; in addition to this there are 34 exceptional cases usually named G_4 to G_{37}. From the classification over $\mathbb{C}$ one can obtain a classification over $\mathbb{Q}_p$ as the sublist whose character field $\mathbb{Q}(\chi)$ is embeddable in $\mathbb{Q}_p$. This was examined by Clark–Ewing [31], and we list their result in Table 1, using the original notation.

W	Order	Degrees	$\mathbb{Q}(\chi)$	Primes
Σ_{n+1} (family 1)	$(n+1)!$	$2, 3, \ldots, n+1$	$\mathbb{Q}$	all p
$G(m,s,n)$ (family 2a) $m \geq 2, n \geq 2, m \neq s$ if $n = 2$	$n! m^{n-1} \frac{m}{s}$	$m, 2m, \ldots, (n-1)m, n\frac{m}{s}$	$\mathbb{Q}(\zeta_m)$	$p \equiv 1$ (m); all p for $m = 2$
$D_{2m} = G(m,m,2)$ (family 2b) $m \geq 3$	$2m$	$2, m$	$\mathbb{Q}(\zeta_m + \zeta_m^{-1})$	$p \equiv \pm 1$ (m); all p for $m = 3, 4, 6$
$C_m = G(m,1,1)$ (family 3) $m \geq 2$	m	m	$\mathbb{Q}(\zeta_m)$	$p \equiv 1$ (m); all p for $m = 2$
G_4	24	$4, 6$	$\mathbb{Q}(\zeta_3)$	$p \equiv 1$ (3)
G_5	72	$6, 12$	$\mathbb{Q}(\zeta_3)$	$p \equiv 1$ (3)
G_6	48	$4, 12$	$\mathbb{Q}(\zeta_{12})$	$p \equiv 1$ (12)
G_7	144	$12, 12$	$\mathbb{Q}(\zeta_{12})$	$p \equiv 1$ (12)
G_8	96	$8, 12$	$\mathbb{Q}(\zeta_4)$	$p \equiv 1$ (4)
G_9	192	$8, 24$	$\mathbb{Q}(\zeta_8)$	$p \equiv 1$ (8)
G_{10}	288	$12, 24$	$\mathbb{Q}(\zeta_{12})$	$p \equiv 1$ (12)
G_{11}	576	$24, 24$	$\mathbb{Q}(\zeta_{24})$	$p \equiv 1$ (24)
G_{12}	48	$6, 8$	$\mathbb{Q}(\sqrt{-2})$	$p \equiv 1, 3$ (8)
G_{13}	96	$8, 12$	$\mathbb{Q}(\zeta_8)$	$p \equiv 1$ (8)
G_{14}	144	$6, 24$	$\mathbb{Q}(\zeta_3, \sqrt{-2})$	$p \equiv 1, 19$ (24)
G_{15}	288	$12, 24$	$\mathbb{Q}(\zeta_{24})$	$p \equiv 1$ (24)
G_{16}	600	$20, 30$	$\mathbb{Q}(\zeta_5)$	$p \equiv 1$ (5)
G_{17}	1200	$20, 60$	$\mathbb{Q}(\zeta_{20})$	$p \equiv 1$ (20)
G_{18}	1800	$30, 60$	$\mathbb{Q}(\zeta_{15})$	$p \equiv 1$ (15)
G_{19}	3600	$60, 60$	$\mathbb{Q}(\zeta_{60})$	$p \equiv 1$ (60)
G_{20}	360	$12, 30$	$\mathbb{Q}(\zeta_3, \sqrt{5})$	$p \equiv 1, 4$ (15)
G_{21}	720	$12, 60$	$\mathbb{Q}(\zeta_{12}, \sqrt{5})$	$p \equiv 1, 49$ (60)
G_{22}	240	$12, 20$	$\mathbb{Q}(\zeta_4, \sqrt{5})$	$p \equiv 1, 9$ (20)
G_{23}	120	$2, 6, 10$	$\mathbb{Q}(\sqrt{5})$	$p \equiv 1, 4$ (5)
G_{24}	336	$4, 6, 14$	$\mathbb{Q}(\sqrt{-7})$	$p \equiv 1, 2, 4$ (7)
G_{25}	648	$6, 9, 12$	$\mathbb{Q}(\zeta_3)$	$p \equiv 1$ (3)
G_{26}	1296	$6, 12, 18$	$\mathbb{Q}(\zeta_3)$	$p \equiv 1$ (3)
G_{27}	2160	$6, 12, 30$	$\mathbb{Q}(\zeta_3, \sqrt{5})$	$p \equiv 1, 4$ (15)
G_{28}	1152	$2, 6, 8, 12$	$\mathbb{Q}$	all p
G_{29}	7680	$4, 8, 12, 20$	$\mathbb{Q}(\zeta_4)$	$p \equiv 1$ (4)
G_{30}	14400	$2, 12, 20, 30$	$\mathbb{Q}(\sqrt{5})$	$p \equiv 1, 4$ (5)
G_{31}	$64 \cdot 6!$	$8, 12, 20, 24$	$\mathbb{Q}(\zeta_4)$	$p \equiv 1$ (4)
G_{32}	$216 \cdot 6!$	$12, 18, 24, 30$	$\mathbb{Q}(\zeta_3)$	$p \equiv 1$ (3)
G_{33}	$72 \cdot 6!$	$4, 6, 10, 12, 18$	$\mathbb{Q}(\zeta_3)$	$p \equiv 1$ (3)
G_{34}	$108 \cdot 9!$	$6, 12, 18, 24, 30, 42$	$\mathbb{Q}(\zeta_3)$	$p \equiv 1$ (3)
G_{35}	$72 \cdot 6!$	$2, 5, 6, 8, 9, 12$	$\mathbb{Q}$	all p
G_{36}	$8 \cdot 9!$	$2, 6, 8, 10, 12, 14, 18$	$\mathbb{Q}$	all p
G_{37}	$192 \cdot 10!$	$2, 8, 12, 14, 18, 20, 24, 30$	$\mathbb{Q}$	all p

Table 1. The irreducible $\mathbb{Q}_p$-reflection groups

The ring $\mathbb{Q}_p[L]^W$ of W–invariant polynomial functions on L is polynomial if and only if W is a reflection group, by the Shephard–Todd–Chevalley theorem [11, Thm. 7.2.1]; the column *degrees* in Table 1 lists the degrees of the generators, and the number of degrees equals the rank of (W, L). For many W, none of the primes listed in the last column divide $|W|$; in fact this can only happen in the infinite families, and in the sporadic examples 12, 24, 28, 29, 31, and 34–37. It is a good exercise to look for the Weyl groups of the various simple compact Lie groups in the table, where they have character field $\mathbb{Q}$. One may observe that for $p = 2$ and 3 there is only one *exotic* reflection group (i.e., irreducible with $\mathbb{Q}(\chi) \neq \mathbb{Q}$), namely G_{24} and G_{12} respectively, whereas for $p \geq 5$ there are always infinitely many.

The classification over $\mathbb{Q}_p$ can be lifted to a classification over $\mathbb{Z}_p$, but instead of stating this now, we proceed directly to root data.

Definition 1.1 (R–root datum). *An R–root datum $\mathbf{D}$ is a triple $(W, L, \{Rb_\sigma\})$, where (W, L) is a finite R–reflection group, and $\{Rb_\sigma\}$ is a collection of rank one submodules of L, indexed by the set of reflections σ in W, and satisfying that* $\mathrm{im}(1-\sigma) \subseteq Rb_\sigma$ *(coroot condition) and $w(Rb_\sigma) = Rb_{w\sigma w^{-1}}$ for all $w \in W$ (conjugation invariance).*

An *isomorphism* of R–root data $\varphi\colon \mathbf{D} \to \mathbf{D}'$ is defined to be an isomorphism $\varphi\colon L \to L'$ such that $\varphi W \varphi^{-1} = W'$ as subgroups of $\mathrm{Aut}(L')$ and $\varphi(Rb_\sigma) = Rb'_{\varphi\sigma\varphi^{-1}}$ for every reflection $\sigma \in W$. The element $b_\sigma \in L$, determined up to a unit in R, is called the *coroot* corresponding to σ. The coroot condition ensures that given (σ, b_σ) we can define a *root* $\beta_\sigma\colon L \to R$ via the formula

$$\sigma(x) = x - \beta_\sigma(x) b_\sigma \tag{1.1}$$

The classification of R–root data of course depends heavily on R. For $R = \mathbb{Z}$ root data correspond bijectively to classically defined root data $(M, \Phi, M^\vee, \Phi^\vee)$ via the association $(W, L, \{\mathbb{Z}b_\sigma\}) \rightsquigarrow (L^*, \{\pm\beta_\sigma\}, L, \{\pm b_\sigma\})$. One easily checks that $Rb_\sigma \subseteq \ker(N)$, where $N = 1 + \sigma + \ldots + \sigma^{|\sigma|-1}$ is the norm element, so giving an R–root datum with underlying reflection group (W, L) corresponds to choosing a cyclic R–submodule of $H^1(\langle\sigma\rangle; L)$ for each conjugacy class of reflections σ. It is hence in practice not hard to parametrize all possible R–root data supported by a given finite R–reflection group. For $R = \mathbb{Z}_p$, p odd, reflections have order dividing $p - 1$, hence prime to p, so here $\mathbb{Z}_p$–root data coincides with finite $\mathbb{Z}_p$–reflection groups. For $R = \mathbb{Z}$ or $\mathbb{Z}_2$ the difference between the two notions only occur for the root data of $\mathrm{Sp}(n)$ and $\mathrm{SO}(2n+1)$, but due to the ubiquity of $\mathrm{SU}(2)$ and $\mathrm{SO}(3)$ this distinction turns out to be an important one. Note that since a root and a coroot (β_σ, b_σ) determine the reflection σ by (1.1), one could indeed have defined a root datum as a set of pairs (β_σ, b_σ), each determined up to a unit and subject to certain conditions; see also [76].

The relationship between $\mathbb{Z}_p$–root data and $\mathbb{Z}$–root data is given as follows.

Theorem 1.2 (The classification of $\mathbb{Z}_p$–root data, splitting version).

1. *Any $\mathbb{Z}_p$–root datum $\mathbf{D}$ can be written as a product $\mathbf{D} \cong (\mathbf{D}_1 \otimes_{\mathbb{Z}} \mathbb{Z}_p) \times \mathbf{D}_2$, where $\mathbf{D}_1$ is a $\mathbb{Z}$–root datum and $\mathbf{D}_2$ is a product of exotic $\mathbb{Z}_p$–root data.*

2. *Exotic $\mathbb{Z}_p$–root data are in 1-1 correspondence with exotic $\mathbb{Q}_p$–reflection groups via $\mathbf{D} = (W, L, \{\mathbb{Z}_p b_\sigma\}) \rightsquigarrow (W, L \otimes_{\mathbb{Z}_p} \mathbb{Q}_p)$.*

Define the *fundamental group* as $\pi_1(\mathbf{D}) = L/L_0$, where $L_0 = \sum_\sigma \mathbb{Z}_p b_\sigma$ is the coroot lattice, and, with the *p–discrete torus* $\breve{T} = L \otimes \mathbb{Z}/p^\infty$, we define the *$p$–discrete center* as $\breve{\mathcal{Z}}(\mathbf{D}) = \bigcap_\sigma \ker(\breve{\beta}_\sigma \colon \breve{T} \to \mathbb{Z}/p^\infty)$; compare e.g., [15]. It turns out that $\pi_1(\mathbf{D}) = \breve{\mathcal{Z}}(\mathbf{D}) = 0$ for all exotic root data, and this plays a role in the proof of the above statement. If A is a finite subgroup of $\breve{\mathcal{Z}}(\mathbf{D})$, we can define the quotient root datum $\mathbf{D}/A$ by taking $\breve{T}_{\mathbf{D}/A} = \breve{T}/A$, and hence $L_{\mathbf{D}/A} = \mathrm{Hom}(\mathbb{Z}/p^\infty, \breve{T}/A)$, and defining the roots and coroots of $\mathbf{D}/A$ via the induced maps.

Theorem 1.3 (The classification of $\mathbb{Z}_p$–root data, structure version).

1. *Any $\mathbb{Z}_p$–root datum $\mathbf{D} = (W, L, \{\mathbb{Z}_p b_\sigma\})$ can be written as a quotient*

$$\mathbf{D} = (\mathbf{D}_1 \times \cdots \times \mathbf{D}_n \times (1, L^W, \emptyset))/A$$

   where $\pi_1(\mathbf{D}_i) = 0$ for all i, for a finite central subgroup A.

2. *Irreducible $\mathbb{Z}_p$–root data $\mathbf{D}$ with $\pi_1(\mathbf{D}) = 0$ are in 1-1 or 2-1 correspondence with non-trivial irreducible $\mathbb{Q}_p$–reflection groups via $\mathbf{D} \rightsquigarrow (W, L \otimes_{\mathbb{Z}_p} \mathbb{Q}_p)$, the sole identification being $\mathbf{D}_{\mathrm{Sp}(n)} \otimes_{\mathbb{Z}} \mathbb{Z}_2$ with $\mathbf{D}_{\mathrm{Spin}(2n+1)} \otimes_{\mathbb{Z}} \mathbb{Z}_2$, $n \geq 3$.*

A main ingredient used to derive the classification of root data from the classification of $\mathbb{Q}_p$–reflection groups is the case-by-case observation that the mod p reduction of all the exotic reflection groups remain irreducible, which ensures that any lift to $\mathbb{Z}_p$ is uniquely determined by the $\mathbb{Q}_p$–representation.

Remark 1.4. It seems that $\mathbb{Z}_p$–root data ought to parametrize some purely algebraic objects, just as $\mathbb{Z}$–root data parametrize both compact connected Lie groups and reductive algebraic groups. Similar structures come up in Lusztig's approach to the representation theory of finite groups of Lie type, as examined by Bessis, Broué, Malle, Michel, Rouquier, and others [23], involving mythical objects from the Greek island of Spetses [68].

2. p–compact Groups and their Classification

In this section we give a brief introduction to p–compact groups, followed by the statement of the classification theorem, an outline of its proof, and some

of its consequences. Additional background information on p–compact groups can be found in the surveys [36, 65, 72, 78].

The first ingredient we need is the theory of p–completions. The p–completion construction of Sullivan [95] produces for each space X a map $X \to X^\wedge_p$, which, when X is simply connected and of finite type, has the property that $\pi_i(X^\wedge_p) \cong \pi_i(X) \otimes \mathbb{Z}_p$ for all i. A space is called p–complete if this map is a homotopy equivalence. In fact, when X is simply connected and $H_*(X;\mathbb{F}_p)$ is of finite type, then X is p–complete if and only if the homotopy groups of X are finitely generated $\mathbb{Z}_p$–modules. We remark that Bousfield–Kan [16] produced a variant on Sullivan's p–completion functor, and for the spaces that occur in this paper these two constructions agree up to homotopy, so the words p–complete and p–completion can be taken in either sense.

A finite loop space is a triple (X, BX, e), where BX is a pointed connected space, X is a finite CW–complex, and $e\colon X \to \Omega BX$ is a homotopy equivalence, where Ω denotes based loops. We will return to finite loop spaces in Section 3, but now move straight to their p–complete analogs.

Definition 2.1 (p–compact group [42]). *A p–compact group is a triple (X, BX, e), where BX is a pointed, connected, p–complete space, $H^*(X;\mathbb{F}_p)$ is finite, and $e\colon X \xrightarrow{\simeq} \Omega BX$ is a homotopy equivalence.*

The loop multiplication on ΩBX is here the homotopical analog of a group structure; while standard loop multiplication does not define a group, it is equivalent in a strong sense (as an A_∞–space) to a topological group, whose classifying space is homotopy equivalent to BX. We therefore baptise BX the *classifying space*, and note that, since all structure can be derived from BX, one could equivalently have defined a p–compact group to be a space BX, subject to the above conditions. The finiteness of $H^*(X;\mathbb{F}_p)$ is to be thought of as a homotopical version of compactness, and replaces the condition that the underlying loop space be homotopy equivalent to a finite complex. We will usually refer to a p–compact group just by X or BX when there is little possibility for confusion.

Examples of p–compact groups include of course the p–completed classifying space $BG^\wedge_p$ of a compact Lie group G with $\pi_0(G)$ a p–group. However, non-isomorphic compact Lie groups may give rise to equivalent p–compact groups if they have the same p–local structure, perhaps the most interesting example being $B\mathrm{SO}(2n+1)^\wedge_p \simeq B\mathrm{Sp}(n)^\wedge_p$ for p odd [46]. Exotic examples (i.e., examples with exotic root data) are discussed in Section 2.1.

A *morphism* between p–compact groups is a pointed map $BX \to BY$; it is called a *monomorphism* if the homotopy fiber, denoted Y/X, has finite $\mathbb{F}_p$–homology. Two morphisms are called *conjugate* if they are freely homotopic, and two p–compact groups are called *isomorphic* if their classifying spaces are homotopy equivalent. A p–compact group is called *connected* if X is connected. By a standard argument $H^*(BX;\mathbb{Z}_p) \otimes \mathbb{Q}$ is seen to be a polynomial algebra

over $\mathbb{Q}_p$, and we define the *rank* $r = \mathrm{rank}(X)$ to be number of generators. The following is the main structural result of Dwyer–Wilkerson [42].

Theorem 2.2 (Maximal tori and Weyl groups of p–compact groups [42]).

1. *Any p–compact group X has a* maximal torus*: a monomorphism $i\colon BT = (BS^1{}^\wedge_p)^r \to BX$ with r the rank of X. Any other monomorphism $i'\colon BT' = (BS^1{}^\wedge_p)^s \to BX$ factors as $i' \simeq i \circ \varphi$ for some $\varphi\colon BT' \to BT$. In particular i is unique up to conjugacy.*

2. *The* Weyl space $\mathcal{W}_X(T)$*, defined as the topological monoid of self-equivalences $BT \to BT$ over i (with i made into a fibration), has contractible components.*

3. *If X is connected, the natural action of the* Weyl group $W_X(T) = \pi_0(\mathcal{W}_X(T))$ *on $L_X = \pi_2(BT)$ gives a faithful representation of W_X as a finite $\mathbb{Z}_p$–reflection group.*

A short outline of the proof can be found in [65]. The *maximal torus normalizer* is defined as the homotopy orbit space, or Borel construction, $B\mathcal{N}_X(T) = BT_{h\mathcal{W}_X(T)}$ and hence sits in a fibration sequence

$$BT \to B\mathcal{N}_X(T) \to B\mathcal{W}_X(T).$$

The normalizer is said to be *split* if the above fibration has a section. It is worth mentioning that one sees that (W_X, L_X) is a $\mathbb{Z}_p$–reflection group indirectly, by proving that

$$H^*(BX;\mathbb{Z}_p) \otimes \mathbb{Q} \cong (H^*(BT;\mathbb{Z}_p) \otimes \mathbb{Q})^{W_X}$$

and applying the Shephard–Todd–Chevalley theorem.

To define the $\mathbb{Z}_p$–root datum, one therefore needs to proceed in a non-standard way [45, 6, 8]. For p odd, the $\mathbb{Z}_p$–root datum $\mathbf{D}_X$ can be defined from the $\mathbb{Z}_p$–reflection group (W_X, L_X), by setting $\mathbb{Z}_p b_\sigma = \mathrm{im}(L_X \xrightarrow{1-\sigma} L_X)$. The definition for $p = 2$ is more complicated, and in order to give meaning to the words we first need a few extra definitions for p–compact groups. The centralizer of a morphism $\nu : BA \to BX$ is defined as $B\mathcal{C}_X(\nu) = \mathrm{map}(BA, BX)_\nu$, where the subscript denotes the component corresponding to ν. While this may look odd at first sight, it does in fact generalize the Lie group notion [35]. For a connected p–compact group X, define the derived p–compact group $\mathcal{D}X$ to be the covering space of X corresponding to the torsion subgroup of $\pi_1(X)$. Consider the *p–discrete singular torus $\breve{T}_0^{\langle\sigma\rangle}$ for σ*, i.e., the largest divisible subgroup of the fixed-points $\breve{T}^{\langle\sigma\rangle}$, with $\breve{T} = L_X \otimes \mathbb{Z}/p^\infty$, and set $X_\sigma = \mathcal{D}(\mathcal{C}_X(\breve{T}_0^{\langle\sigma\rangle}))$. Then X_σ is a connected p–compact group of rank one with p–discrete maximal torus $(1-\sigma)\breve{T}$; denote the corresponding maximal torus normalizer by $\mathcal{N}_\sigma$, called the *root subgroup* of σ. Define the coroots in $\mathbf{D}_X$ via the formula

$$\mathbb{Z}_p b_\sigma = \begin{cases} \mathrm{im}(L_X \xrightarrow{1-\sigma} L_X) & \text{if } \mathcal{N}_\sigma \text{ is split,} \\ \ker(L_X \xrightarrow{1+\sigma} L_X) & \text{if } \mathcal{N}_\sigma \text{ is not split.} \end{cases}$$

For p odd, only the first case occurs, and for $p = 2$ the split case corresponds to $BX_\sigma \simeq B\mathrm{SO}(3)^\wedge_2$ and the non-split corresponds to $BX_\sigma \simeq B\mathrm{SU}(2)^\wedge_2$. For comparison we note that when $X = G^\wedge_p$, for a reductive complex algebraic group G, $BX_\sigma \simeq B\langle U_\alpha, U_{-\alpha}\rangle^\wedge_p$, where U_α is what is ordinarily called the root subgroup of the root $\alpha = \beta_\sigma$, and the above formula can be read off from e.g., [90, Pf. of Lem. 7.3.5]. We can now state the classification theorem.

Theorem 2.3 (Classification of p–compact groups [9, 8]). *The assignment which to a connected p–compact group X associates its $\mathbb{Z}_p$–root datum $\mathbf{D}_X$ gives a one-to-one correspondence between connected p–compact groups, up to isomorphism, and $\mathbb{Z}_p$–root data, up to isomorphism.*

Furthermore the map $\Phi\colon \mathrm{Out}(BX) \to \mathrm{Out}(\mathbf{D}_X)$, given by lifting a self-homotopy equivalence of BX to BT, is an isomorphism.

Here $\mathrm{Out}(BX)$ denotes the group of free homotopy classes of self-homotopy equivalences $BX \to BX$, and $\mathrm{Out}(\mathbf{D}_X) = \mathrm{Aut}(\mathbf{D}_X)/W_X$. A stronger space-level statement about self-maps is in fact true, namely

$$B\mathrm{Aut}(BX) \xrightarrow{\simeq} ((B^2\breve{\mathcal{Z}}(\mathbf{D}_X))^\wedge_p)_{h\,\mathrm{Out}(\mathbf{D}_X)} \tag{2.1}$$

where $\mathrm{Aut}(BX)$ is the space of self-homotopy equivalences, $\breve{\mathcal{Z}}(\mathbf{D}_X)$ the p–discrete center of $\mathbf{D}_X$ as introduced in Section 1, and the action of $\mathrm{Out}(\mathbf{D}_X)$ on $(B^2\breve{\mathcal{Z}}(\mathbf{D}_X))^\wedge_p$ is the canonical one. Having control of the whole space of self-equivalences turns out to be important in the proof.

Theorem 2.3 implies, by Theorem 1.2, that any connected p–compact group splits as a product of the p–completion of a compact connected Lie group and a product of known exotic p–compact groups. For $p = 2$ it shows that there is only one exotic 2–compact group, the one corresponding to the $\mathbb{Q}_2$–reflection group G_{24}, and this 2–compact group was constructed in [41]. We will return to the construction of the exotic p–compact groups in the next subsection.

Since we understand the whole space of self–equivalences, one can derive a classification also of non-connected p–compact groups. The set of isomorphism classes of non-connected p–compact groups with root datum of the identity component $\mathbf{D}$ and group of components π, is parametrized by the components of the moduli space

$$\mathrm{map}(B\pi, ((B^2\breve{\mathcal{Z}}(\mathbf{D}))^\wedge_p)_{h\,\mathrm{Out}(\mathbf{D})})_{h\,\mathrm{Aut}(B\pi)} \tag{2.2}$$

As with the classification of compact Lie groups, the classification statement can naturally be broken up into two parts, existence and uniqueness of p–compact groups. The uniqueness statement can be formulated as an *isomorphism theorem* saying that there is a 1-1–correspondence between conjugacy classes of isomorphisms of connected p–compact groups $BX \to BX'$ and isomorphisms of root data $\mathbf{D}_X \to \mathbf{D}_{X'}$, up to $W_{X'}$–conjugation. This last statement can in fact be strengthened to an *isogeny theorem* classifying maps that are rational isomorphisms [5].

While the existence and uniqueness are separate statements, they are currently most succinctly proved simultaneously by an induction on the size of $\mathbf{D}$, since the proof of existence requires knowledge of certain facts about self-maps, and the proof of uniqueness at the last step is aided by specific facts about concrete models. We will discuss the proof of existence in Section 2.1 and of uniqueness in Section 2.2, along with some information about the history.

2.1. Construction of p–compact groups. Compact connected Lie groups can be constructed in different ways. They can be exhibited as symmetries of geometric objects, or can be systematically constructed via generators-and-relations type constructions that involve first constructing a finite dimensional Lie algebra from the root system, and then passing to the group [61, 90].

An adaptation of the above tools to p–compact groups is still largely missing, so one currently has to proceed by more ad hoc means, with the limited aim of constructing only the exotic p–compact groups. These were in fact already constructed some years ago, but we take the opportunity here to retell the tale, and outline the closest we currently get to a streamlined construction.

The first exotic p–compact groups were constructed by Sullivan [95] as the homotopy orbit space of the action of the would-be Weyl group on the would-be torus. The most basic case he observed is the following: If C_m is a cyclic group of order m, and p an odd prime such that $m|p-1$, then $C_m \leq \mathbb{Z}_p^\times$, and hence C_m acts on the Eilenberg–MacLane space $K(\mathbb{Z}_p, 2)$. The Serre spectral sequence for the fibration

$$K(\mathbb{Z}_p, 2) \to K(\mathbb{Z}_p, 2)_{hC_m} \to BC_m$$

reveals that the $\mathbb{F}_p$–cohomology of $K(\mathbb{Z}_p, 2)_{hC_m}$ is a polynomial algebra on a class in degree $2m$, using that m is prime to p. Therefore the cohomology of its loop space is an exterior algebra in degree $2m-1$ and $BX = (K(\mathbb{Z}_p, 2)_{hC_m})^\wedge_p$ is a p–compact group, with $\Omega BX \simeq (S^{2m-1})^\wedge_p$. We have just realized all exotic groups in family 3 of Table 1!

Exactly the same argument carries over to the general case of a root datum $\mathbf{D}$ where $p \nmid |W|$, just replacing C_n by W and $\mathbb{Z}_p$ by L, since $\mathbb{F}_p[L \otimes \mathbb{F}_p]^W$ is a polynomial algebra exactly when W is a reflection group, when $p \nmid |W|$, by the Shephard–Todd–Chevalley theorem used earlier. This observation was made by Clark–Ewing [31], and realizes a large number of groups in Table 1. However, the method as it stands cannot be pushed further, since the assumption that $p \nmid |W|$ is crucial for the collapse of the Serre spectral sequence.

Additional exotic p–compact groups were constructed in the 1970s by other methods. Quillen realized $G(m, 1, n)$ at all possible primes by constructing an approximation via classifying spaces of discrete groups [84, §10], and Zabrodsky [102, 4.3] realized G_{12} and G_{31} at $p = 3$ and 5 respectively, by taking homotopy fixed-points of a p'–group acting on the classifying space of a compact Lie group.

To build the remaining exotic p–compact groups one needs a far-reaching generalization of Sullivan's technique, obtained by replacing the homotopy orbit space with a more sophisticated homotopy colimit, that ensures that we

still get a collapsing spectral sequence even when p divides the order of W. The technique was introduced by Jackowski–McClure [54], as a decomposition technique in terms of centralizers of elementary abelian subgroups, and was subsequently used by Aguadé [2] $(G_{12}, G_{29}, G_{31}, G_{34})$, Dwyer–Wilkerson [41] (G_{24}), and Notbohm–Oliver [80] $(G(m,s,n))$ to finish the construction of the exotic p–compact groups.

The following is an extension of Aguadé's argument, and can be used inductively to realize all exotic p–compact groups for p odd—that this works in all cases relies on the stroke of luck, checked case-by-case, that all exotic finite $\mathbb{Z}_p$–reflection groups for p odd have $\mathbb{Z}_p[L]^W$ a polynomial algebra; cf. also [81].

Theorem 2.4 (Inductive construction of p–compact groups, p odd [9]). *Consider a finite $\mathbb{Z}_p$–reflection group (W, L), p odd, with $\mathbb{Z}_p[L]^W$ a polynomial algebra.*

Then (W_V, L) is a again a $\mathbb{Z}_p$–reflection group and $\mathbb{Z}_p[L]^{W_V}$ a polynomial algebra, for W_V the pointwise stabilizer in W of $V \leq L \otimes \mathbb{F}_p$.

Assume that, for all non-trivial V, (W_V, L) is realized by a connected p–compact group Y_V satisfying the isomorphism part of Theorem 2.3 and $H^*(Y_V; \mathbb{Z}_p) \cong \mathbb{Z}_p[L]^{W_V}$ (with L in degree 2). Then $V \mapsto Y_V$ extends to a functor $Y\colon \mathbf{A}^{op} \to \mathbf{Spaces}$, where $\mathbf{A}$ has objects non-trivial $V \leq L \otimes \mathbb{F}_p$ and morphisms given by conjugation in W, and

$$BX = (\mathrm{hocolim}_{\mathbf{A}^{op}} Y)^\wedge_p$$

is a p–compact group with Weyl group (W, L) and $H^*(BX; \mathbb{Z}_p) \cong \mathbb{Z}_p[L]^W$.

Idea of proof. The statement that $\mathbb{Z}_p[L]^{W_V}$ is a polynomial algebra is an extension of Steinberg's fixed-point theorem in the version of Nakajima [75, Lem. 1.4]. The proof uses Lannes' T–functor, together with case-by-case considerations.

The inductive construction is straightforward, given current technology, and uses only general arguments: Since we assume we know Y_V and its automorphisms for all $V \neq 1$, one easily sets up a functor $\mathbf{A}^{op} \to \mathrm{Ho}(\mathbf{Spaces})$, the homotopy category of spaces, and the task is to rigidify this to a functor in the category of spaces. The diagram can be show to be "centric", so one can use the obstruction theory developed by Dwyer–Kan in [37]. The relevant obstruction groups identify with the higher limits of a functor obtained by taking fixed-points, and in particular this is a Mackey functor whose higher limits vanish by a theorem of Jackowski–McClure [54]. We can therefore rigidify the diagram to a diagram in spaces, and the resulting homotopy colimit is easily shown to have the desired cohomology. □

We now turn to the prime 2. Here the sole exotic $\mathbb{Z}_2$–reflection group is G_{24}, and the corresponding 2–compact group was realized by Dwyer–Wilkerson [41] and dubbed DI(4), due to the fact that, for $E = (\mathbb{Z}/2)^4$,

$$H^*(B\mathrm{DI}(4); \mathbb{F}_2) \cong \mathbb{F}_2[E]^{\mathrm{GL}(E)}$$

the rank four Dickson invariants. At first glance this might look like the setup of Theorem 2.4, but note that G_{24} is a rank three $\mathbb{Z}_2$–reflection group, not four, so E is not just the elements of order 2 in the maximal torus. However by taking $\mathbf{A}$ to be the category with objects the non-trivial subgroups of E, and morphisms induced by conjugation in $\mathrm{GL}(E)$, and correctly guessing the centralizers of elementary abelian subgroups, the argument can still be pushed through; the starting point is declaring the centralizer of any element of order two to be $\mathrm{Spin}(7)^{\hat{}}_2$.

We again stress the apparent luck in being able to guess the rather uncomplicated structure of $\mathbf{A}$ and the centralizers. If one hypothetically had to construct an exotic p–compact group with a seriously complicated cohomology ring, say one would try to construct E_8 at the prime 2 by these methods, it would not be clear how to start. As a first step one would need a way to describe the p–fusion in the group, just from the root datum $\mathbf{D}$. This relates to old questions in Lie theory, which have occupied Borel, Serre, and many others [88]...

2.2. Uniqueness of p–compact groups. In this subsection, we outline the proof of the uniqueness part of the classification theorem for p–compact groups, Theorem 2.3, following [8] by Andersen and the author; it extends [9] also with Møller and Viruel. We mention that the quest for uniqueness was initiated by Dwyer–Miller–Wilkerson [38] in the 80s and in particular Notbohm [77] obtained strong partial results; a different approach for $p = 2$ using computer algebra was independently given by Møller [73, 74]. See [9, 8] for more details on the history of the proof.

From now on we consider two connected p–compact groups X and X' with the same root datum $\mathbf{D}$, and want to build a homotopy equivalence $BX \to BX'$. The proof goes by an induction on the size of (W, L).

Step 1*: (The maximal torus normalizer and its automorphisms,* [45, 6]*).* A first step is to show that X and X' have isomorphic maximal torus normalizers. Working with the maximal torus normalizer has a number of technical advantages over the maximal torus, related to the fact that the fiber of the map $B\mathcal{N} \to BX$ has Euler characteristic prime to p (one, actually).

One shows that the maximal torus normalizers are isomorphic, by giving a *construction* from the root datum. For p odd the construction is simple, since the maximal torus normalizer turns out always to be split, and hence isomorphic to $(B^2L)_{hW}$ with the canonical action. This was established in [3], by showing that the relevant extension group is zero except in one case, which can be handled by other means; cf. also [9, Rem. 2.5]. For $p = 2$, the problem is more difficult. The corresponding problem for compact Lie groups, or reductive algebraic groups, was solved by Tits [96] many years ago. A thorough reading of Tits' paper, with a cohomological rephrasing of some of his key constructions, allows his construction to be pushed through also for p–compact groups [45]. One thus algebraically constructs a maximal torus normalizer $\mathcal{N}_{\mathbf{D}}$ and show it

to be isomorphic to the topologically defined one. A problem is however that $\mathcal{N}$ in general has too large automorphism group. To correct this, it was shown in [6] that the root subgroups $\mathcal{N}_\sigma$, introduced before Theorem 2.3, can also be built algebraically, and adding this extra data give the correct automorphism group. Concretely, one has a canonical factorization

$$\Phi\colon\ \mathrm{Out}(BX) \to \mathrm{Out}(B\mathcal{N}, \{B\mathcal{N}_\sigma\}) \xrightarrow{\cong} \mathrm{Out}(\mathbf{D}_X)$$

and one can furthermore build a candidate model for the whole space $B\mathrm{Aut}(BX)$, by a slight modification of $B\mathrm{Aut}(B\mathcal{N}, \{B\mathcal{N}_\sigma\})$, the space of self-homotopy equivalences of $B\mathcal{N}$ preserving the root subgroups.

Step 2*: (Reduction to simple, center-free groups,* [8, §2]*).* This next step involves relating the p–compact group and its factors and center-free quotient via certain fibration sequences, and studying automorphisms via these fibrations. Several of the necessary tools, such as the understanding of the center of a p–compact group [43], the product splitting theorem [44], etc., were already available in the 90s. But, in particular for $p = 2$, one needs to incorporate the machinery of root data and root subgroups; we refer the reader to [8, §2] for the details.

Step 3*: (Defining a map on centralizers of elements of order p,* [8, §4]*).* We now assume that X and X' are simple, center-free p–compact groups. The next tool needed is a homology decomposition theorem, more precisely the centralizer decomposition, of Jackowski–McClure [54] and Dwyer–Wilkerson [40], already mentioned in the previous subsection. Let $\mathbf{A}(X)$ be the Quillen category of X with objects monomorphisms $\nu\colon\ BE \to BX$, where $E = (\mathbb{Z}/p)^s$ is a non-trivial elementary abelian p–group, and morphisms $(\nu\colon\ BE \to BX) \to (\nu'\colon\ BE' \to BX)$ are group monomorphisms $\varphi\colon\ E \to E'$ such that $\nu' \circ B\varphi$ is conjugate to ν. The centralizer decomposition theorem now says that for any p–compact group X, the evaluation map

$$\mathrm{hocolim}_{\nu\in\mathbf{A}(X)^{\mathrm{op}}}\, B\mathcal{C}_X(\nu) \to BX$$

is an isomorphism on $\mathbb{F}_p$–cohomology.

This opens the possibility for a proof by induction, since the centralizers will be smaller p–compact groups if X is center-free. As explained above we can assume that X and X' have common maximal torus normalizer and root subgroups $(\mathcal{N}, \{\mathcal{N}_\sigma\})$, so that we are in the situation of following diagram

$$\begin{tikzcd}
 & (B\mathcal{N}, \{B\mathcal{N}_\sigma\}) \arrow[dl, "j"'] \arrow[dr, "j'"] & \\
BX \arrow[rr, dotted] & & BX'
\end{tikzcd}$$

where the dotted arrow is the one we want to construct.

If $\nu\colon\ B\mathbb{Z}/p \to BX$ is a monomorphism, then it can be conjugated into T, uniquely up to conjugation in $\mathcal{N}$. This gives a well defined way of viewing ν as

a map $\nu\colon B\mathbb{Z}/p \to BT \to B\mathcal{N}$. Taking centralizers of this map produces a new diagram

$$\begin{tikzcd} & (B\mathcal{C}_{\mathcal{N}}(\nu), \{B\mathcal{C}_{\mathcal{N}}(\nu)_\sigma\}) \arrow[dl] \arrow[dr] & \\ B\mathcal{C}_X(\nu) \arrow[rr, dotted] & & B\mathcal{C}_{X'}(\nu) \end{tikzcd}$$

One now argues that the induction hypothesis guarantees that we can construct the dotted arrow. There is the slight twist that the centralizer will be disconnected in general, so we have to use that we inductively know the whole space of self-equivalences of the identity component.

Step 4*: (Compatibility of maps on all centralizers,* [8, §5]*).* The next step is to define the map on centralizers of arbitrary elementary abelian p–subgroups $\nu\colon BE \to BX$. This is done by restricting to a rank one subgroup $E' \leq E$ and considering the composition

$$B\mathcal{C}_X(\nu) \to B\mathcal{C}_X(\nu|_{E'}) \to B\mathcal{C}_{X'}(\nu|_{E'}) \to BX'.$$

One now has to show that these maps do not depend on the choice of E', and that they fit together to define an element in

$$\lim_{\nu \in \mathbf{A}(X)}{}^0 [B\mathcal{C}_X(\nu), BX']$$

By the induction hypothesis it turns out that one can reduce to the case where E has rank two and $\mathcal{C}_X(\nu)$ is discrete. An inspection of the classification of $\mathbb{Z}_p$–root data shows that this case only occurs for $\mathbf{D} \cong \mathbf{D}_{\mathrm{PU}(p)^\wedge_p}$, which can then be handled by direct arguments, producing the element in $\lim^0$.

In fact one can prove something slightly stronger, which will be needed in the next step: A close inspection of the whole preceeding argument reveals that all maps can be constructed over $B^2\pi_1(\mathbf{D})$, which allows one to produce an element in

$$\lim_{\nu \in \mathbf{A}(X)}{}^0 [\widetilde{B\mathcal{C}_X(\nu)}, \widetilde{BX'}]$$

where the tilde denotes covers with respect to the kernel of the map to $\pi_1(\mathbf{D})$.

With this step complete one can see that BX and BX' have the same p–fusion, i.e., that p–subgroups are conjugate in the same way, but we are left with a rigidification issue.

Step 5*: (Rigidifying the map,* [8, §6]*).* One now wants to define a map on the whole homotopy colimit, which can then easily be checked to have the correct properties, finishing the proof of the classification. Constructing such a map directly from an element in $\lim^0$ requires knowing that the higher limits of the functors $F_i\colon \mathbf{A}(X) \to \mathbb{Z}_p\text{-mod}$ given by $E \mapsto \pi_i(\mathcal{Z}\mathcal{C}_X(E))$, vanish, where $\mathcal{Z}$ denotes the center. In turn, this calculation requires knowing the structure of

$\mathbf{A}(X)$, and for this we use that X is a known p–compact group, where we can examine the structure. For the part of the functor corresponding to elementary abelian subgroups that can be conjugated into T, the higher limits can be show to vanish via a Mackey functor argument, going back to [54] and [40]. This in fact equals the whole functor for all exotic groups for p odd, and DI(4) also works via a variant on this argument, which finish off those cases.

We can hence assume that X is the p–completion of a compact Lie group G. Here the obstruction groups were computed to identically vanish in [9], for p odd, relying on detailed information about the elementary abelian p–subgroups of G, partially tabulated by Griess [48]. This is easy when there is little torsion in the cohomology, but harder for the small torsion primes, and the exceptional groups. In [8], however, we use a different argument to cover all primes, inspired by [99]. Using the above element in $\lim^0$ it turns out that one can produce an element in

$$\lim_{\tilde{G}/\tilde{P}\in\mathbf{O}_p^r(\tilde{G})^{\mathrm{op}}}{}^{0}[B\tilde{P}, B\widetilde{X'}]$$

where $\mathbf{O}_p^r(\tilde{G})$ is the subcategory of the orbit category of $\tilde{G}$ with objects the so-called p–radical subgroups. Here one again wants to show vanishing of the higher limits, in order to get a map on the homotopy colimit. Calculating higher limits over this orbit category is in many ways similar to calculating it over the Quillen category [49]. In this case, however, the relevant higher limits were in fact shown to identically vanish in earlier work of Jackowski–McClure–Oliver [55], also building on substantial case-by-case calculations. This again produces a map $B\tilde{G} \xrightarrow{\simeq} B\widetilde{X'}$, and passing to a quotient provides the sought homotopy equivalence $BG \xrightarrow{\simeq} BX'$. The statements about self-maps also fall out of this approach. □

2.3. Lie theory for p–compact groups.

We have already seen many Lie-type results for p–compact groups. Quite a few more can be proved by observing that the classical Lie result only depends on the p–completion of the compact Lie group, and verifying case-by-case that it holds for the exotic p–compact groups. We collect some theorems of this type in this section, encouraging the reader to look for more conceptual proofs, and include also a brief discussion of homotopical representation theory. Throughout this section X is a connected p–compact group with maximal torus T.

The first theorem on the list is the analog of theorems of Bott [14] from 1954.

Theorem 2.5. *$H^*(X/T;\mathbb{Z}_p)$ and $H^*(\Omega X;\mathbb{Z}_p)$ are both torsion free and concentrated in even degrees, and $H^*(X/T;\mathbb{Z}_p)$ has rank $|W_X|$ as a $\mathbb{Z}_p$–module.*

The result about ΩX was known as the loop space conjecture, and in fact proved by Lin and Kane in a series of papers in the more general setting of finite mod p H–spaces, using complicated calculations with Steenrod operations [67].

Bott's proof used Morse theory and the result may be viewed in the context of Schubert cell decompositions [71]. Rationally $H^*(X/T;\mathbb{Z}_p)\otimes\mathbb{Q} = \mathbb{Q}_p[L]\otimes_{\mathbb{Q}_p[L]^{W_X}}\mathbb{Q}_p$, so calculating the Betti numbers, given the theorem, is reduced to a question about complex reflection groups—an interpretation of these numbers in terms of length functions on the root system has been obtained for certain classes of complex reflection groups, cf. [17, 98], but the complete picture is still not clear. In general the theory of homogeneous and symmetric spaces for p–compact groups is rather unexplored, and warrants attention.

Theorem 2.5 implies that $\pi_3(X)$ is torsion free, and proving that in a conceptual way might be a good starting point. For Lie groups, Bott in fact stated the, now classical, fact that $\pi_3(G)\cong\mathbb{Z}$ for G simple. The analogous statement is *not* true for most of the exotic p–compact groups; for instance it obviously fail for the Sullivan spheres other than S^3. However, it *is* true that π_3 is non-zero for finite loop spaces, as a consequence of a celebrated theorem of Clark [30] from 1963 giving strong restrictions on the degrees of finite loop spaces. These results helped fuel the speculation that finite loop spaces should look a lot like compact Lie groups, a point we will return to in the next section.

Most of the general results about torsion in the cohomology of BX and X due to Borel, Steinberg, and others, also carry through to p–compact groups, but here again with many results relying on the classification. This fault is partly inherited from Lie groups; see Borel [13, p. 775] for a summary of the status there. In particular we mention that X has torsion free $\mathbb{Z}_p$–cohomology if and only if BX has torsion free $\mathbb{Z}_p$–cohomology if and only if every elementary abelian p–subgroup factors through a maximal torus. Likewise $\pi_1(X)$ is torsion free if and only if every elementary abelian group of rank two factors through a maximal torus; see [9, 8].

The (complex linear) homotopy representation theory of X is encoded in the semi-ring

$$\mathrm{Rep}^{\mathbb{C}}(BX) = \left[BX, \coprod_n B\mathrm{U}(n)^\wedge_p\right]$$

It is non-trivial since for any connected p–compact group X there exists a monomorphism $BX\to B\mathrm{U}(n)^\wedge_p$, for some n; the exotic groups were checked in [27, 28, 103]—indeed, as already alluded to, several exotic p–compact groups can conveniently be *constructed* as homotopy fixed-points inside a p–completed compact Lie group. The general structure of the semi-ring is however still far from understood. The classification allows one to focus on p–completed classifying spaces of compact Lie groups, but even in this case the semi-ring appears very complicated [56]; there are higher limits obstructions, related to interesting problems in group theory [49].

Weights can be constructed as usual: By the existence of a maximal torus, we can lift a homotopy representation to a map $BT_X\to BT_{\mathrm{U}(n)^\wedge_p}$, well defined up to an action of Σ_n, and produce a collection of n weights in $L_X^* = \mathrm{Hom}_{\mathbb{Z}_p}(L_X,\mathbb{Z}_p)$, invariant under the action of the Weyl group W_X. When $p\nmid |W_X|$, homotopy representations just correspond to finite W_X–invariant subsets of L_X^*, and any

homotopy representation decomposes up to conjugation uniquely into indecomposable representations given by transitive W_X–sets. When $p \mid |W_X|$ the situation is much more complicated.

Let us describe what happens in the basic case of $X = \mathrm{SU}(2)^\wedge_2$. Denote by ρ_i the irreducible complex representation of $\mathrm{SU}(2)$ with highest weight i, and use the same letter for the induced map $B\mathrm{SU}(2)^\wedge_2 \to B\mathrm{U}(i+1)^\wedge_2$. Precomposing with the self-homotopy equivalence ψ^k of $B\mathrm{SU}(2)^\wedge_2$, $k \in \mathbb{Z}_2^\times$, corresponding to multiplication by k on the root datum, gives a new representation $k \star \rho_i$ of the same dimension, but with weights multiplied by k.

Theorem 2.6. $\mathrm{Rep}^{\mathbb{C}}(B\mathrm{SU}(2)^\wedge_2)$ *has an additive generating set given by* ρ_0, $k \star \rho_1$, $k \star \rho_2$ *and* $((k+2k') \star \rho_1) \otimes ((k-2k') \star \rho_1)$, $k \in \mathbb{Z}_2^\times$, $0 \neq k' \in \mathbb{Z}_2$. *These generators are indecomposable, and two representations agree if they have the same weights.*

The reader may verify that the decomposition into indecomposables is not unique, e.g., for ρ_6. It is at present not clear how to use $\mathrm{SU}(2)^\wedge_2$ to describe the general structure, as one could have hoped—the thing to note is that homotopy representations are governed by questions of p–fusion of elements, rather than more global structure. Already for $\mathrm{SU}(2)^\wedge_2 \times \mathrm{SU}(2)^\wedge_2$ there is no upper bound on the dimension of the indecomposables, and in particular they are not always a tensor product of indecomposable $\mathrm{SU}(2)^\wedge_2$ representations. More severely, representations need not be uniquely determined by their weights, e.g., for $\mathrm{Sp}(2)^\wedge_2 \times \mathrm{Sp}(2)^\wedge_2$.

By using case-by-case arguments, there might be hope to establish a version of Weyl's theorem $R(BX) \xrightarrow{\cong} R(BT)^{W_X}$, where $R(BX) = \mathrm{Gr}(\mathrm{Rep}^{\mathbf{C}}(BX))$ is the Grothendieck group. The result is not proved even for p–completions of compact Lie groups, but the integral version is the main result in [58]. The weaker K–theoretic result $K^*(BX;\mathbb{Z}_p) \xrightarrow{\cong} K^*(BT;\mathbb{Z}_p)^W$ was established in [60] (using that $H^*(\Omega X;\mathbb{Z}_p)$ is torsion free). The ring structure of $R(BT)^W$ is also not clear, and in particular it would be interesting to exhibit some fundamental representations.

3. Finite Loop Spaces

In the 1960s and early 1970s, finite loop spaces, not p–compact groups, were the primary objects of study, and there were many conjectures about them [91]. The theory of p–compact groups enables the resolution of most of them, either in the positive or the negative, and gives what is essentially a parametrization of all connected finite loop spaces.

We already defined finite loop spaces in Section 2; let us now briefly recall their history in broad strokes. Hopf proved in 1941 [52] that the rational cohomology of any connected, finite loop space is a graded exterior algebra $H^*(X;\mathbb{Q}) \cong \bigwedge_{\mathbb{Q}}(x_1, ..., x_r)$, where $|x_i| = 2d_i - 1$, and r is called the *rank*. Serre,

ten years later [87], showed that the list of *degrees* $d_1, \dots, d_r$ uniquely determines the rational homotopy type of (X, BX, e). In those days, there were not many examples of finite loop spaces. Indeed, in the early 1960s it was speculated that perhaps every finite loop space was homotopy equivalent to a compact Lie group, a would-be variant of Hilbert's 5th problem. This was soon shown to be wrong in several different ways: Hilton–Roitberg, in 1968, exhibited a 'criminal' [51], a finite loop space (X, BX, e), of the rational homotopy type of $\mathrm{Sp}(2)$, such that the underlying space X is not homotopy equivalent to any Lie group; and Rector [85] in 1971 observed that there exists *uncountable* many finite loop spaces (X, BX, e) such that X is homotopy equivalent to $\mathrm{SU}(2)$. The first example may superficially look more benign than the second; indeed in general there are only finitely many possibilities for the homotopy type of the underlying space X, given the rational homotopy type of BX [32]. But the exact number depends on homotopy groups of finite complexes, and does not appear closely related to Lie theory, so shifting focus from loop space structures (X, BX, e) to that of homotopy types of X, does not appear desirable.

An apparently better option is, as the reader has probably sensed, to pass to p–completions, defined in Section 2. Sullivan made precise how one can recover a (simply connected) space integrally if one knows the space "at all primes and rationally, as well as how they are glued together". Along with his p–completion, he constructed a rationalization functor $X \to X_{\mathbb{Q}}$, with analogous properties, and proved that these functors fit together in the following arithmetic square.

Proposition 3.1 (Sullivan's arithmetic square [94, 34]). *Let Y be a simply connected space of finite type. Then the following diagram, with obvious maps, is a homotopy pull-back square.*

$$\begin{tikzcd} Y \arrow[r] \arrow[d] & \prod_p Y_{\hat{p}} \arrow[d] \\ Y_{\mathbb{Q}} \arrow[r] & (\prod_p Y_{\hat{p}})_{\mathbb{Q}} \end{tikzcd}$$

This parallels the usual fact that the integers $\mathbb{Z}$ is a pullback of $\hat{\mathbb{Z}} = \prod_p \mathbb{Z}_p$ and $\mathbb{Q}$ over the finite adeles $\mathbb{A}_f = \hat{\mathbb{Z}} \otimes \mathbb{Q}$. If BX is the classifying space of a connected finite loop space then, by the classification of p–compact groups, all spaces in the diagram are now understood: Each $BX_{\hat{p}}$ is the classifying space of a p–compact group, and the spaces at the bottom of the diagram are determined by numerical data, namely the degrees: $BX_{\mathbb{Q}} \simeq K(\mathbb{Q}, 2d_1) \times \cdots \times K(\mathbb{Q}, 2d_r)$ and $(\prod_p BX_{\hat{p}})_{\mathbb{Q}} \simeq K(\mathbb{A}_f, 2d_1) \times \cdots \times K(\mathbb{A}_f, 2d_r)$, by the result of Serre quoted earlier. Hence to classify connected finite loop spaces with a given list of degrees, we first have to enumerate all collections of p–compact groups with those degrees; there are a finite number of these, and they can be enumerated given the classification [8, Prop. 8.18]. The question of how many finite loop spaces with a given set of p–completions is then a question of *genus*, determined by an explicit set of double cosets.

Theorem 3.2 (Classification of finite loop spaces). *The assignment which to a finite loop space Y associates the collection of $\mathbb{Z}_p$–root data $\{\mathbf{D}_{Y^{\hat{}}_p}\}_p$ is a surjection from connected finite loop spaces to collections of $\mathbb{Z}_p$–root data, all p, with the same degrees $d_1, \ldots, d_r$. The pre-image of $\{\mathbf{D}_p\}_p$ is parametrized by the set of double cosets*

$$\mathrm{Out}(K_{\mathbb{Q}}) \backslash \mathrm{Out}^c(K_{\mathbb{A}_f}) / \prod_p \mathrm{Out}(\mathbf{D}_p)$$

where $K_R = K(R, 2d_1) \times \cdots \times K(R, 2d_r)$, $R = \mathbb{Q}$ or $\mathbb{A}_f$.

Here $\mathrm{Out}(K_{\mathbb{Q}})$ denotes the group of free homotopy classes of self-homotopy equivalences, and $\mathrm{Out}^c(K_{\mathbb{A}_f})$ denotes those homotopy classes of homotopy equivalences that induce $\mathbb{A}_f$–linear maps on homotopy groups. Since K_R is an Eilenberg–MacLane space, the set of double cosets can be completely described algebraically; see [9, §13] for a calculation of $\mathrm{Out}(\mathbf{D}_p)$.

The set of double cosets will, except for the degenerate case of tori, be uncountable. Allowing for only a single prime p everywhere above would parametrize the number of $\mathbb{Z}_{(p)}$–local finite loop spaces corresponding to a given p–compact group Y_p, and also this set is usually uncountable, with a few more exceptions, such as groups of rank one. A similar result holds when one inverts some collection of primes $\mathcal{P}$; see [7, Rem. 3.3] for more information.

Sketch of proof of Theorem 3.2. There is a natural inclusion $K_{\mathbb{Q}} \to K_{\mathbb{A}_f}$ induced by the unit map $\mathbb{Q} \to \mathbb{A}_f$, and one easily proves that the pull-back provides a space Y such that $H^*(\Omega Y; \mathbb{Z})$ is finite over $\mathbb{Z}$. That Y is actually homotopy equivalent to a finite complex follows by the vanishing of the finiteness obstruction, as proved by Notbohm [81] (see [4, Lemma 1.2] for more details). Twisting the pullback by an element in $\mathrm{Out}^c(K_{\mathbb{A}_f})$ provides a new finite loop space, and after passing to double cosets, this assignment is easily seen to be surjective and injective on homotopy types (see [94] and [101, Thm. 3.8]). □

If one assumes that the finite loop space X has a maximal torus, as defined by Rector [86], i.e., a map $(BS^1)^r \to BX$ with homotopy fiber homotopy equivalent to a finite complex, for $r = \mathrm{rank}(X)$, the above picture changes completely. The inclusion of an 'integral' maximal torus prohibits the twisting in the earlier theorem, and one obtains a proof of the classical maximal torus conjecture stated by Wilkerson [100] in 1974, giving a homotopy theoretical description of compact Lie groups as exactly the finite loop spaces admitting a maximal torus.

Theorem 3.3 (Maximal torus conjecture [8]). *The classifying space functor, which to a compact Lie group G associates the finite loop space $(G, BG, e\colon G \xrightarrow{\simeq} \Omega BG)$ gives a one-to-one correspondence between isomorphism classes of compact Lie groups and finite loop spaces with a maximal torus. Furthermore, for G connected,* $\mathrm{Out}(BG) \cong \mathrm{Out}(G) \cong \mathrm{Out}(\mathbf{D}_G)$.

The statement about automorphisms, which was not part of the original conjecture, follows from work of Jackowski–McClure–Oliver [57, Cor. 3.7].

In light of the above structural statement it is natural to further enquire how exotic finite loop spaces can be. Whether they are all manifolds was recently settled in the affirmative by Bauer–Kitchloo–Notbohm–Pedersen, answering an old question of Browder [24].

Theorem 3.4 ([10]). *For any finite loop space* (Y, BY, e)*,* Y *is homotopy equivalent to a closed, smooth, parallelizable manifold.*

The result is proved using the theory of p–compact groups, combined with classical surgery techniques, as set up by Pedersen. It shows the subtle failure of a naïve homotopical version of Hilbert's fifth problem: Every finite loop space is, by classical results, homotopy equivalent to a topological group, and homotopy equivalent to a compact smooth manifold by the above. But one cannot always achieve both properties at once. This would otherwise imply that every finite loop space was homotopy equivalent to a compact Lie group, by the solution to Hilbert's fifth problem, contradicting that many exotic finite loop spaces exist.

One can still ask if every finite loop space is *rationally* equivalent to some compact Lie group? Indeed this was conjectured in the 70s to be the case, and was verified up to rank 5. However, the answer to this question turns out to be negative as well, although counterexamples only start appearing in high rank.

Theorem 3.5 (A 'rational criminal' [4]). *There exists a connected finite loop space* X *of rank* 66*, dimension* 1254*, and degrees*

$$\{2^8, 3^2, 4^8, 5^2, 6^7, 7, 8^7, 9, 10^5, 11, 12^5, 13, 14^5, 16^3, 18^2, 20^2, 22, 24^2, 26, 28, 30\}$$

(where n^k *means that* n *is repeated* k *times) such that* $H_*(X;\mathbb{Q})$ *does not agree with* $H_*(G;\mathbb{Q})$ *for any compact Lie group* G*, as graded vector spaces.*

This example is minimal, in the sense that any connected, finite loop space of rank less than 66 *is rationally equivalent to some compact Lie group* G*.*

In [4] there is a list of which p–compact group to choose at each prime. By the preceeding discussion, the problem of finding such a space is a combinatorial problem, and one can show that in high enough rank there will be many examples.

4. Steenrod's Problem of Realizing Polynomial Rings

The 1960 "Steenrod problem" [92, 93], asks, for a given ring R, which graded polynomial algebras are realized as $H^*(Y;R)$ of some space Y, i.e., in which degrees can the generators occur? In this section we give some background on this classical problem and describe its solution in [7, 8].

Steenrod, in his original paper [92], addressed the case of polynomial rings in a single variable: For $R = \mathbb{Z}$ the only polynomial rings that occur are $H^*(\mathbb{C}P^\infty;\mathbb{Z}) \cong \mathbb{Z}[x_2]$ and $H^*(\mathbb{H}P^\infty;\mathbb{Z}) \cong \mathbb{Z}[x_4]$, as he showed by a short argument using his cohomology operations. Similarly, for $R = \mathbb{F}_p$ he showed that the generator has to sit in degree 1,2, or 4 for $p = 2$ and in degree $2n$ with $n|p-1$ for p odd, but now as a consequence of Hopf invariant one and its odd primary version (though it was not known at the time whether the p odd cases were realized when $n \neq 1, 2$).

There were attempts to use the above techniques to handle polynomial rings in several variables, but they gave only very partial results. In the 70s, however, Sullivan's method, as generalized by Clark–Ewing, realized many polynomial rings, as explained in Section 2.1. Conversely, in the 80s Adams–Wilkerson [1] and others put restrictions on the potential degrees, using categorical properties of the category of unstable algebras over the Steenrod algebra. This eventually led to the result of Dwyer–Miller–Wilkerson [39] that for p large enough the Clark–Ewing examples are exactly the possible polynomial cohomology rings over $\mathbb{F}_p$.

In order to tackle all primes, it turns out to be useful to have a space-level theory, and that is what p–compact groups provide. Namely, if Y is a space such that $H^*(Y;\mathbb{F}_p)$ is a polynomial algebra, then the Eilenberg–Moore spectral sequence shows that $H^*(\Omega Y;\mathbb{F}_p)$ is finite, and hence $Y^\wedge_p$ is a p–compact group.

Theorem 4.1 (Steenrod's problem, $\mathrm{char}(R) \neq 2$ [7]). *Let R be a commutative Noetherian ring of finite Krull dimension and let P^* be a graded polynomial R–algebra in finitely many variables, all in positive even degrees.*

Then there exists a space Y such that $P^* \cong H^*(Y;R)$ as graded algebras if and only if for each prime p not a unit in R, the degrees of P^* halved is a multiset union of the degrees lists occurring in Table 1 at that prime p, and the degree one, with the following exclusions (due to torsion): $(G(2,2,n), p = 2; n \geq 4)$, $(G(6,6,2), p = 2)$, $(G_{24}, p = 2)$, $(G_{28}, p = 2,3)$, $(G_{35}, p = 2,3)$, $(G_{36}, p = 2,3)$, and $(G_{37}, p = 2,3,5)$.

When $\mathrm{char}(R) \neq 2$, all generators are in even degrees by anti-commutativity, so the assumptions of the theorem are satisfied. The proof in [7] only relies on the general theory of p–compact groups, not on the classification. The case $R = \mathbb{F}_p$, p odd, was solved earlier by Notbohm [81], also using p–compact group theory. Taking $R = \mathbb{Z}$ gives the old conjecture that if $H^*(Y;\mathbb{Z})$ is a polynomial ring, then it is isomorphic to a tensor product of copies of $\mathbb{Z}[x_2]$, $\mathbb{Z}[x_4, x_6, \ldots, x_{2n}]$, and $\mathbb{Z}[x_4, x_8, \ldots, x_{4n}]$, the cohomology rings of $\mathbb{C}P^\infty$, $B\mathrm{SU}(n)$ and $B\mathrm{Sp}(n)$.

Theorem 4.2 (Steenrod's problem, $\mathrm{char}(R) = 2$ [8]). *Suppose that P^* is a graded polynomial algebra in finitely many variables over a commutative ring R of characteristic 2. Then $P^* \cong H^*(Y;R)$ for a space Y if and only if*

$$P^* \cong H^*(BG;R) \otimes H^*(B\mathrm{DI}(4);R)^{\otimes r} \otimes H^*(\mathbb{R}P^\infty;R)^{\otimes s} \otimes H^*(\mathbb{C}P^\infty;R)^{\otimes t}$$

as a graded algebra, for some $r, s, t \geq 0$, where G is a compact connected Lie group with finite center. In particular, if all generators of P^* are in degree ≥ 3 then P^* is a tensor product of the cohomology rings of the classifying spaces of $\mathrm{SU}(n)$, $\mathrm{Sp}(n)$, $\mathrm{Spin}(7)$, $\mathrm{Spin}(8)$, $\mathrm{Spin}(9)$, G_2, F_4, *and* $\mathrm{DI}(4)$.

The proof reduces to $R = \mathbb{F}_2$, and then uses the classification of 2–compact groups. It would be interesting to try to list all polynomial rings which occur as $H^*(BG; \mathbb{F}_2)$ for G a compact connected Lie group with finite center.

One can also determine to which extent the space is unique. The following result was proved by Notbohm [81] for p odd and in [8] for $p = 2$, as the culmination of a long series of partial results, started by Dwyer–Miller–Wilkerson [38, 39].

Theorem 4.3 (Uniqueness of spaces with polynomial $\mathbb{F}_p$–cohomology). *Suppose A^* is a finitely generated polynomial $\mathbb{F}_p$–algebra over the Steenrod algebra $\mathcal{A}_p$, with generators in degree ≥ 3. Then there exists, up to p–completion, at most one homotopy type Y with $H^*(Y; \mathbb{F}_p) \cong A^*$, as graded algebras over the Steenrod algebra.*

If P^* is a finitely generated polynomial $\mathbb{F}_p$–algebra, then there exists at most finitely many homotopy types Y, up to p–completion, such that $H^*(Y; \mathbb{F}_p) \cong P^*$ as graded $\mathbb{F}_p$–algebras.

The assumption ≥ 3 above cannot be dropped, as easy examples show, and integrally uniqueness rarely hold, as discussed in Section 3; see also [7, 8].

5. Homotopical Finite Groups, Group Actions,..

This survey is rapidly coming to an end, but we nevertheless want to briefly mention some other recent developments in homotopical group theory.

In connection with the determination of the algebraic K-theory of finite fields, Quillen and Friedlander proved the following: If G is a reductive group scheme, and q is a prime power, $p \nmid q$, then

$$BG(\mathbb{F}_q)^\wedge_p \simeq (BG(\mathbb{C})^\wedge_p)^{h\langle\psi^q\rangle}$$

where the superscript means taking homotopy fixed-points of the self-map ψ^q corresponding to multiplication by q on the root datum—it says that, at p, fixed-points and homotopy fixed-points of the Frobenius map raising to the qth power agree.

The right-hand side of the equation makes sense with $BG(\mathbb{C})^\wedge_p$ replaced by a p–compact group. Benson speculated in the mid 90s that the resulting object should be the classifying space of a "p–local finite group", and be determined by a conjugacy or fusion pattern on a finite p–group S, as axiomatized by Puig [83] (motivated by block theory), together with a certain rigidifying 2–cocycle. He even gave a candidate fusion pattern corresponding to $\mathrm{DI}(4)$, namely a

fusion pattern constructed by Solomon years earlier in connection with the classification of finite simple groups, but shown not to exist inside any finite group [12].

All this turns out to be true and more! A theory of p–local finite groups was founded and developed by Broto–Levi–Oliver in [20], and has seen rapid development by both homotopy theorists and group theorists since then. The Solomon 2–local finite groups $\mathrm{Sol}(q)$ were shown to exist in [66], and a study of Chevalley p–local finite groups, p odd, was initiated in [22]. A number of exotic p–local finite groups have been found for p odd, but the family $\mathrm{Sol}(q)$ remains the only known examples at $p = 2$, prompting the speculation that perhaps they are the only exotic simple 2–local finite groups! Even partial results in this direction could have implications for the proof of the classification of finite simple groups. A modest starting point is the result in [18] that any so-called *constrained* fusion pattern comes from a (unique constrained) finite group—the result is purely group theoretic, and, while not terribly difficult, the only known proof uses techniques of a kind hitherto foreign to the classification of finite simple groups.

One can ask for a theory more general than p–local finite groups, broad enough to encompass both p–completions of arbitrary compact Lie groups and p–compact groups, and one such theory was indeed developed in [21], the so-called p–local compact groups. One would like to identify connected p–compact groups inside p–local compact groups in some group theoretic manner. This relates to the question of describing the relationship between the classical Lie theoretic structure and the p–fusion structure, mentioned several times before in this paper; the proof of the classification of p–compact groups may offer some hints on how to proceed.

In a related direction, one may attempt to relax the condition of compactness in p–compact groups to include more general types of groups; the paper [29] shows that replacing cohomologically finite by noetherian gives few new examples. An important class of groups to understand is Kac–Moody groups, and the paper [19] shows, amongst other things, that homomorphisms from finite p–groups to Kac–Moody groups still correspond to maps between classifying spaces. This gives hope that some of the homotopical theory of maximal tori, Weyl groups, etc., may also be brought to work in this setting, but the correct general definition of a homotopy Kac–Moody group is still elusive, the Lie theoretic definition being via generators-and-relations rather than intrinsic. A good understanding of the restricted case of affine Kac–Moody groups and loop groups would already be very interesting.

Groups were historically born to act, a group action being a homomorphism from G to the group of homeomorphisms of a space X. In homotopy theory, one is however often only given X up to an equivariant map which is a homotopy equivalence. Here the appropriate notion of an action is an element in the mapping space $\mathrm{map}(BG, B\mathrm{Aut}(X))$, where as before $\mathrm{Aut}(X)$ denotes the space of self-homotopy equivalences of X (itself an interesting group!).

Homotopical group actions can also be studied one prime at a time, and assembled to global results afterwards. Of particular interest is the case where X is a sphere. Spheres are non-equivariantly determined by their dimension, and self-maps by their degree. It turns out that something similar is true for homotopical group actions of finite groups on p–complete spheres [50]. But, one has to interpret dimension as meaning dimension function, assigning to each p–subgroup of G the homological dimension of the corresponding homotopy fixed-point set, and correspondingly the degree is a degree function, viewed as an element in a certain p–adic Burnside ring. Furthermore there is hope to determine exactly which dimension functions are realizable. Understanding groups is homotopically open-ended...

References

[1] J. F. Adams and C. W. Wilkerson. Finite H-spaces and algebras over the Steenrod algebra. *Ann. of Math. (2)*, 111(1):95–143, 1980.

[2] J. Aguadé. Constructing modular classifying spaces. *Israel J. Math.*, 66(1–3):23–40, 1989.

[3] K. K. S. Andersen. The normalizer splitting conjecture for p-compact groups. *Fund. Math.*, 161(1–2):1–16, 1999.

[4] K. K. S. Andersen, T. Bauer, J. Grodal, and E. K. Pedersen. A finite loop space not rationally equivalent to a compact Lie group. *Invent. Math.*, 157(1):1–10, 2004.

[5] K. K. S. Andersen and J. Grodal. The isogeny theorem for p-compact groups. In preparation.

[6] K. K. S. Andersen and J. Grodal. Automorphisms of p-compact groups and their root data. *Geom. Topol.*, 12:1427–1460, 2008.

[7] K. K. S. Andersen and J. Grodal. The Steenrod problem of realizing polynomial cohomology rings. *J. Topol.*, 1(4):747–760, 2008.

[8] K. K. S. Andersen and J. Grodal. The classification of 2-compact groups. *J. Amer. Math. Soc.*, 22(2):387–436, 2009.

[9] K. K. S. Andersen, J. Grodal, J. M. Møller, and A. Viruel. The classification of p-compact groups for p odd. *Ann. of Math. (2)*, 167(1):95–210, 2008.

[10] T. Bauer, N. Kitchloo, D. Notbohm, and E. K. Pedersen. Finite loop spaces are manifolds. *Acta Math.*, 192(1):5–31, 2004.

[11] D. J. Benson. *Polynomial invariants of finite groups.* Cambridge Univ. Press, 1993.

[12] D. J. Benson. Cohomology of sporadic groups, finite loop spaces, and the Dickson invariants. In *Geometry and cohomology in group theory (Durham, 1994)*, LMS LNS 252, pages 10–23. Cambridge Univ. Press, 1998.

[13] A. Borel. *Œuvres: collected papers. Vol. II.* Springer, 1983.

[14] R. Bott. An application of the Morse theory to the topology of Lie-groups. *Bull. Soc. Math. France*, 84:251–281, 1956.

[15] N. Bourbaki. *Éléments de mathématique: Groupes et algèbres de Lie.* Masson, Paris, 1982. Chapitre 9. Groupes de Lie réels compacts.

[16] A. K. Bousfield and D. M. Kan. *Homotopy limits, completions and localizations.* LNM 304. Springer, 1972.

[17] K. Bremke and G. Malle. Root systems and length functions. *Geom. Dedicata*, 72(1):83–97, 1998.

[18] C. Broto, N. Castellana, J. Grodal, R. Levi, and B. Oliver. Subgroup families controllling p-local finite groups. *Proc. London Math. Soc. (3)*, 91(2):325–354, 2005.

[19] C. Broto and N. Kitchloo. Classifying spaces of Kac-Moody groups. *Math. Z.*, 240(3):621–649, 2002.

[20] C. Broto, R. Levi, and B. Oliver. The homotopy theory of fusion systems. *J. Amer. Math. Soc.*, 16(4):779–856, 2003.

[21] C. Broto, R. Levi, and B. Oliver. Discrete models for the p-local homotopy theory of compact Lie groups and p-compact groups. *Geom. Topol.*, 11:315–427, 2007.

[22] C. Broto and J. M. Møller. Chevalley p-local finite groups. *Algebr. Geom. Topol.*, 7:1809–1919, 2007.

[23] M. Broué. Reflection groups, braid groups, Hecke algebras, finite reductive groups. In *Current Developments in Mathematics 2000*, pages 1–107. Intl. Press, 2001.

[24] W. Browder. Torsion in H-spaces. *Ann. of Math. (2)*, 74:24–51, 1961.

[25] G. Carlsson. Equivariant stable homotopy and Segal's Burnside ring conjecture. *Ann. of Math. (2)*, 120(2):189–224, 1984.

[26] G. Carlsson. Segal's Burnside ring conjecture and related problems in topology. In *Proc. Intl. Congress of Mathematicians, Vol. 1 (Berkeley, 1986)*, pages 574–579, 1987.

[27] N. Castellana. *Representacions homotopiques de grups p-compactes.* PhD thesis, Universitat Autònoma de Barcelona, 1999.

[28] N. Castellana. On the p-compact groups corresponding to the p-adic reflection groups $G(q,r,n)$. *Trans. Amer. Math. Soc.*, 358(7):2799–2819, 2006.

[29] N. Castellana, J. A. Crespo, and J. Scherer. Noetherian loop spaces. *J. Eur. Math. Soc. (JEMS)*, to appear. arXiv:0903.1701.

[30] A. Clark. On π_3 of finite dimensional H-spaces. *Ann. of Math. (2)*, 78:193–196, 1963.

[31] A. Clark and J. Ewing. The realization of polynomial algebras as cohomology rings. *Pacific J. Math.*, 50:425–434, 1974.

[32] C. R. Curjel and R. R. Douglas. On H-spaces of finite dimension. *Topology*, 10:385–389, 1971.

[33] M. Demazure. Exposé XXI. Donnees Radicielles. In *Schémas en groupes. III (SGA 3)*, LNM 153, pages 85–155. Springer, 1962/1964.

[34] E. Dror, W. G. Dwyer, and D. M. Kan. An arithmetic square for virtually nilpotent spaces. *Illinois J. Math.*, 21(2):242–254, 1977.

[35] W. Dwyer and A. Zabrodsky. Maps between classifying spaces. In *Algebraic topology, Barcelona, 1986*, pages 106–119. Springer, 1987.

[36] W. G. Dwyer. Lie groups and p-compact groups. In *Proc. Intl. Congress of Mathematicians, Extra Vol. II (Berlin, 1998)*, pages 433–442, 1998.

[37] W. G. Dwyer and D. M. Kan. Centric maps and realization of diagrams in the homotopy category. *Proc. Amer. Math. Soc.*, 114(2):575–584, 1992.

[38] W. G. Dwyer, H. R. Miller, and C. W. Wilkerson. The homotopic uniqueness of BS^3. In *Algebraic topology, Barcelona, 1986*, pages 90–105. Springer, 1987.

[39] W. G. Dwyer, H. R. Miller, and C. W. Wilkerson. Homotopical uniqueness of classifying spaces. *Topology*, 31(1):29–45, 1992.

[40] W. G. Dwyer and C. W. Wilkerson. A cohomology decomposition theorem. *Topology*, 31(2):433–443, 1992.

[41] W. G. Dwyer and C. W. Wilkerson. A new finite loop space at the prime two. *J. Amer. Math. Soc.*, 6(1):37–64, 1993.

[42] W. G. Dwyer and C. W. Wilkerson. Homotopy fixed-point methods for Lie groups and finite loop spaces. *Ann. of Math. (2)*, 139(2):395–442, 1994.

[43] W. G. Dwyer and C. W. Wilkerson. The center of a p-compact group. In *The Čech centennial (Boston, 1993)*, pages 119–157. Amer. Math. Soc., 1995.

[44] W. G. Dwyer and C. W. Wilkerson. Product splittings for p-compact groups. *Fund. Math.*, 147(3):279–300, 1995.

[45] W. G. Dwyer and C. W. Wilkerson. Normalizers of tori. *Geom. Topol.*, 9:1337–1380, 2005.

[46] E. M. Friedlander. Exceptional isogenies and the classifying spaces of simple Lie groups. *Ann. Math. (2)*, 101:510–520, 1975.

[47] M. Geck and G. Malle. Reflection groups. In *Handbook of algebra. Vol. 4*, pages 337–383. Elsevier, 2006.

[48] R. L. Griess, Jr. Elementary abelian p-subgroups of algebraic groups. *Geom. Dedicata*, 39(3):253–305, 1991.

[49] J. Grodal. Higher limits via subgroup complexes. *Ann. of Math. (2)*, 155(2):405–457, 2002.

[50] J. Grodal and J. H. Smith. Classification of homotopy G-actions on spheres. In preparation.

[51] P. Hilton and J. Roitberg. On principal S^3-bundles over spheres. *Ann. of Math. (2)*, 90:91–107, 1969.

[52] H. Hopf. Über die Topologie der Gruppen-Mannigfaltigkeiten und ihre Verallgemeinerungen. *Ann. of Math. (2)*, 42:22–52, 1941.

[53] J. E. Humphreys. *Reflection groups and Coxeter groups.* Cambridge Univ. Press, 1990.

[54] S. Jackowski and J. McClure. Homotopy decomposition of classifying spaces via elementary abelian subgroups. *Topology*, 31(1):113–132, 1992.

[55] S. Jackowski, J. McClure, and B. Oliver. Homotopy classification of self-maps of BG via G-actions. I+II. *Ann. of Math. (2)*, 135(2):183–226,227–270, 1992.

[56] S. Jackowski, J. McClure, and B. Oliver. Maps between classifying spaces revisited. In *The Čech centennial (Boston, MA, 1993)*, volume 181 of *Contemp. Math.*, pages 263–298. Amer. Math. Soc., 1995.

[57] S. Jackowski, J. McClure, and B. Oliver. Self-homotopy equivalences of classifying spaces of compact connected Lie groups. *Fund. Math.*, 147(2):99–126, 1995.

[58] S. Jackowski and B. Oliver. Vector bundles over classifying spaces of compact Lie groups. *Acta Math.*, 176(1):109–143, 1996.

[59] I. M. James. Bibliography on H-spaces. In *H-Spaces (Neuchâtel, 1970)*, LNM 196, pages 137–156. Springer, 1971.

[60] A. Jeanneret and A. Osse. The K-theory of p-compact groups. *Comment. Math. Helv.*, 72(4):556–581, 1997.

[61] V. G. Kac. *Infinite-dimensional Lie algebras.* Cambridge Univ. Press, 1990.

[62] R. M. Kane. *The homology of Hopf spaces.* North-Holland, 1988.

[63] J. Lannes. Sur les espaces fonctionnels dont la source est le classifiant d'un p-groupe abélien élémentaire. *Inst. Hautes Études Sci. Publ. Math.*, (75):135–244, 1992.

[64] J. Lannes. Applications dont la source est un classifiant. In *Proc. Intl. Congress of Mathematicians, Vol. 1 (Zürich, 1994)*, pages 566–573, 1995.

[65] J. Lannes. Théorie homotopique des groupes de Lie (d'après W. G. Dwyer et C. W. Wilkerson). *Astérisque*, (227):Exp. 776, 3, 21–45, 1995. Sém. Bourbaki, 1993/94.

[66] R. Levi and B. Oliver. Construction of 2-local finite groups of a type studied by Solomon and Benson. *Geom. Topol.*, 6:917–990, 2002.

[67] J. P. Lin. Two torsion and the loop space conjecture. *Ann. of Math. (2)*, 115(1):35–91, 1982.

[68] G. Malle. Spetses. In *Proc. Intl. Congress of Mathematicians, Vol. II (Berlin, 1998)*, pages 87–96, 1998.

[69] H. Miller. The Sullivan conjecture on maps from classifying spaces. *Ann. of Math. (2)*, 120(1):39–87, 1984.

[70] H. Miller. The Sullivan conjecture and homotopical representation theory. In *Proc. Intl. Congress of Mathematicians, Vol. 1 (Berkeley, 1986)*, pages 580–589, 1987.

[71] S. A. Mitchell. Quillen's theorem on buildings and the loops on a symmetric space. *Enseign. Math. (2)*, 34(1–2):123–166, 1988.

[72] J. M. Møller. Homotopy Lie groups. *Bull. Amer. Math. Soc. (N.S.)*, 32(4):413–428, 1995.

[73] J. M. Møller. N-determined 2-compact groups. I. *Fund. Math.*, 195(1):11–84, 2007.

[74] J. M. Møller. N-determined 2-compact groups. II. *Fund. Math.*, 196(1):1–90, 2007.

[75] H. Nakajima. Invariants of finite groups generated by pseudoreflections in positive characteristic. *Tsukuba J. Math.*, 3(1):109–122, 1979.

[76] G. Nebe. The root lattices of the complex reflection groups. *J. Group Theory*, 2(1):15–38, 1999.

[77] D. Notbohm. Homotopy uniqueness of classifying spaces of compact connected Lie groups at primes dividing the order of the Weyl group. *Topology*, 33(2):271–330, 1994.

[78] D. Notbohm. Classifying spaces of compact Lie groups and finite loop spaces. In *Handbook of algebraic topology*, pages 1049–1094. North-Holland, 1995.

[79] D. Notbohm. p-adic lattices of pseudo reflection groups. In *Algebraic topology (Sant Feliu de Guíxols, 1994)*, pages 337–352. Birkhäuser, 1996.

[80] D. Notbohm. Topological realization of a family of pseudoreflection groups. *Fund. Math.*, 155(1):1–31, 1998.

[81] D. Notbohm. Spaces with polynomial mod-p cohomology. *Math. Proc. Cambridge Philos. Soc.*, 126(2):277–292, 1999.

[82] B. Oliver. Vector bundles over classifying spaces. In *Proc. Intl. Congress of Mathematicians, Vol. II (Berlin, 1998)*, pages 483–492, 1998.

[83] Ll. Puig. Frobenius categories. *J. Algebra*, 303(1):309–357, 2006.

[84] D. Quillen. On the cohomology and K-theory of the general linear groups over a finite field. *Ann. of Math. (2)*, 96:552–586, 1972.

[85] D. L. Rector. Loop structures on the homotopy type of S^3. In *Symp. on Algebraic Topology (Seattle, 1971)*, LNM 249, pages 99–105. Springer, 1971.

[86] D. L. Rector. Subgroups of finite dimensional topological groups. *J. Pure Appl. Algebra*, 1(3):253–273, 1971.

[87] J.-P. Serre. Groupes d'homotopie et classes de groupes abéliens. *Ann. of Math. (2)*, 58:258–294, 1953.

[88] J.-P. Serre. Sous-groupes finis des groupes de Lie. *Astérisque*, (266):Exp. 864, 5, 415–430, 2000. Sém. Bourbaki, 1998/99.

[89] G. C. Shephard and J. A. Todd. Finite unitary reflection groups. *Canadian J. Math.*, 6:274–304, 1954.

[90] T. A. Springer. *Linear algebraic groups*. Birkhäuser, second edition, 1998.

[91] J. D. Stasheff. H-space problems. In *H-spaces (Neuchâtel, 1970)*, LNM 196, pages 122–136. Springer, 1971.

[92] N. E. Steenrod. The cohomology algebra of a space. *Enseignement Math. (2)*, 7:153–178 (1962), 1961.

[93] N. E. Steenrod. Polynomial algebras over the algebra of cohomology operations. In *H-spaces (Neuchâtel, 1970)*, LNM 196, pages 85–99. Springer, 1971.

[94] D. Sullivan. Genetics of homotopy theory and the Adams conjecture. *Ann. of Math. (2)*, 100:1–79, 1974.

[95] D. Sullivan. *Geometric topology: localization, periodicity and Galois symmetry*, volume 8 of *K-Monographs in Math.* Springer, 2005. The 1970 MIT notes.

[96] J. Tits. Normalisateurs de tores. I. Groupes de Coxeter étendus. *J. Algebra*, 4:96–116, 1966.

[97] T. tom Dieck. Geometric representation theory of compact Lie groups. In *Proc. Intl. Congress of Mathematicians, Vol. 1 (Berkeley, 1986)*, pages 615–622, 1987.

[98] B. Totaro. Towards a Schubert calculus for complex reflection groups. *Math. Proc. Cambridge Philos. Soc.*, 134(1):83–93, 2003.

[99] A. Vavpetič and A. Viruel. On the homotopy type of the classifying space of the exceptional Lie group F_4. *Manuscripta Math.*, 107(4):521–540, 2002.

[100] C. W. Wilkerson. Rational maximal tori. *J. Pure Appl. Algebra*, 4:261–272, 1974.

[101] C. W. Wilkerson. Applications of minimal simplicial groups. *Topology*, 15(2):111–130, 1976.

[102] A. Zabrodsky. On the realization of invariant subgroups of $\pi_*(X)$. *Trans. Amer. Math. Soc.*, 285(2):467–496, 1984.

[103] K. Ziemiański. A faithful unitary representation of the 2-compact group DI(4). *J. Pure Appl. Algebra*, 213(7):1239–1253, 2009.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Actions of the Mapping Class Group

Ursula Hamenstädt*

Abstract

Let S be a closed oriented surface S of genus $g \geq 0$ with $m \geq 0$ marked points (punctures) and $3g - 3 + m \geq 2$. This is a survey of recent results on actions of the mapping class group of S which led to a geometric understanding of this group.

Mathematics Subject Classification (2010). Primary 30F60, Secondary 20F28, 20F65, 20F69

Keywords. Mapping class group, isometric actions, geometric rigidity

1. Introduction

Let S be a closed oriented surface of genus $g \geq 0$ with $m \geq 0$ marked points (punctures) and $3g - 3 + m \geq 2$. Such a surface is called *non-exceptional*. The *mapping class group* $\mathcal{MCG}(S)$ of S is the group of isotopy classes of orientation preserving homeomorphisms of S preserving the marked points.

Mapping class groups and their subgroups naturally appear in many branches of mathematics and have been extensively studied in the past from the point of view of group theory, topology and geometry. Similarities and differences to other families of groups, like irreducible lattices in semi-simple Lie groups of non-compact type, have been detected. In recent years, investigating the mapping class group through its action on various spaces related to the objects which give rise to it, namely the surfaces themselves, turned out to be particularly fruitful. The goal of this article is to survey some of these developments.

The perhaps simplest topological object defined on the surface S is a simple closed curve. Such a curve c is called *essential* if c is not contractible and not freely homotopic into a puncture. Mapping classes which are particularly easy

*Mathematisches Institut der Universität Bonn, Endenicher Allee 60, 53115 Bonn, Germany. E-mail: ursula@math.uni-bonn.de.

to understand are *Dehn twists* about essential simple closed curves. They are defined as follows.

Let $A = S^1 \times [0,1]$ and let $T : A \to A$ be the orientation preserving homeomorphism defined by

$$T(\theta, t) = (\theta + 2\pi t, t).$$

The map T fixes ∂A pointwise. As a consequence, for every orientation preserving embedding $\phi : A \to \phi(A) \subset S$ with core curve $\phi(S^1 \times \{\frac{1}{2}\})$ freely homotopic to c, the map $\phi \circ T \circ \phi^{-1}$ can be extended by the identity on $S - \phi(A)$ to an orientation preserving homeomorphism of S. Its isotopy class only depends on the free homotopy class of c. It is called a *Dehn twist* about c.

The mapping class group $\mathcal{MCG}(S)$ is finitely generated. If S does not have punctures (i.e. if $m = 0$) then there is a generating set consisting of Dehn twists about a suitable collection of $2g + 1$ simple closed non-separating curves in S (see [13] for details and references).

There are other generating sets with fewer generators. Indeed, as was pointed out by Wajnryb [45], the mapping class group of a closed surface can be generated by two elements. Korkmaz [32] showed that two torsion elements or one Dehn twist and one torsion element suffice.

Every finitely generated group G can be equipped with a distance function d which is invariant under the left action of G. Namely, let $\mathcal{G}$ be a finite symmetric generating set. Here symmetric means that $\mathcal{G}$ contains with every element g also its inverse g^{-1}. The *word norm* $|g|$ of an element $g \in G$ is the smallest length of a word in $\mathcal{G}$ which represents g. Then $d(g,h) = |g^{-1}h|$ defines a distance function on G, a so-called *word metric*, which is invariant under the left action of G. Two such distance functions d, d' defined by different symmetric generating sets are *quasi-isometric*.

Namely, for a number $L \geq 1$, an *L-quasi-isometric embedding* of a metric space (X, d) into a metric space (Z, d') is a map $F : X \to Z$ such that

$$d(x,y)/L - L \leq d'(Fx, Fy) \leq Ld(x,y) + L$$

for all $x, y \in X$. If for every $z \in Z$ there is some $x \in X$ such that $d'(Fx, z) \leq L$ then F is called an *L-quasi-isometry*. If d, d' are two word metrics on a finitely generated group G then the identity $(G, d) \to (G, d')$ is a quasi-isometry.

As a consequence, $\mathcal{MCG}(S)$ equipped with a word metric can be studied as a geometric space, uniquely determined up to quasi-isometry. The geometry of $\mathcal{MCG}(S)$ can then be related to its topology (for example its group homology or cohomology, perhaps with coefficients) and the structure of the spaces on which $\mathcal{MCG}(S)$ acts effectively.

2. The Action of the Mapping Class Group on the Curve Graph

Finitely generated groups can be divided into classes according to geometric types of the metric spaces on which they act in an interesting way as isometries. Simplicial trees are a class of metric spaces which reveal information about groups which act on them as simplicial automorphisms (or isometries). A group Γ is said to have *property FA* if every action of Γ without inversion on a simplicial tree has a fixed point [42].

A countable group Γ has *property T* if every affine isometric action of Γ on a real Hilbert space has a fixed point. Groups with property T are known to have property FA as well. Irreducible lattices in semi-simple Lie groups of non-compact type and higher rank have property T, and the same holds true for lattices in the rank-one simple Lie groups $Sp(n,1), F_4^{-20}$. However, lattices Γ in the simple rank-one Lie groups $SO(n,1), SU(n,1)$ are known to fulfill a strong negation of property T. Namely, they admit an isometric action on a real Hilbert space H which is proper in a metric sense: For every bounded set $A \subset H$ there are only finitely many $\phi \in \Gamma$ with $\phi A \cap A \neq \emptyset$. This property is called the *Haagerup property*.

The mapping class group of the space of tori is the group $SL(2,\mathbb{Z})$ which has the Haagerup property. The mapping class group of the closed surface of genus 2 does not have property T. This can for example be deduced from a result of Korkmaz [31] who showed that for any $n \geq 1$ the mapping class group of a surface S of genus $g \leq 2$ with $m \geq 0$ punctures and $3g-3+m \geq 2$ contains a subgroup Γ_n of finite index which surjects onto the free group F_n of rank n. As a consequence, Γ_n admits an isometric action on a tree with unbounded orbits and hence Γ_n does not have property T. Then same holds true for $\mathcal{MCG}(S)$.

For surfaces of higher genus, J. Andersen announced

Theorem 2.1 (J. Andersen). *Mapping class groups do not have property T.*

This leads to the following questions.

Question 1: Do mapping class groups act on simplicial trees without fixed point?

Question 2: Are there subgroups Γ of finite index with non-vanishing first cohomology group, perhaps with coefficients in some representation with infinite image?

Question 3: Do mapping class groups have the Haagerup property?

A simplicial tree equipped with any simplicial metric is a *hyperbolic geodesic metric space in the sense of Gromov* (see [9]). The class of hyperbolic spaces is much larger than the class of trees. For example, there are many *word hyperbolic* groups, i.e. groups with hyperbolic word metric (or Cayley graphs), which have property T.

Mapping class groups admit isometric actions on Gromov hyperbolic spaces without bounded orbits. The best known example of such an action is an action on a space which is not locally compact and defined as follows.

The *curve complex* $\mathcal{CC}(S)$ of S is a finite dimensional simplicial complex whose vertices are the free homotopy classes of essential simple closed curves on S. A collection $c_1, \ldots, c_m$ of $m \geq 1$ vertices spans a simplex if and only if the curves c_i can be realized disjointly. The *curve graph* $\mathcal{C}(S)$ is the one-skeleton of the curve complex. It carries a canonical simplicial metric. Masur and Minsky showed [36]

Theorem 2.2 (Masur-Minsky 99). *The curve graph is a hyperbolic geodesic metric space.*

The curve complex is not locally finite and hence not locally compact. The mapping class group admits a natural simplicial action on the curve complex. Even more is true. The following result is due to Ivanov [23] in most cases and was completed by Korkmaz [30]. For its formulation, define the *extended mapping class group* to be the group of *all* isotopy classes of homeomorphisms of S.

Theorem 2.3 (Ivanov 97). *If S is not a closed surface of genus 2 or a twice punctured torus or a six punctured sphere then the automorphism group of the curve complex $\mathcal{CC}(S)$ coincides with the extended mapping class group.*

More recently, Bestvina and Feighn [5] constructed *proper* (i.e. locally compact complete) hyperbolic geodesic metric spaces which admit isometric actions of $\mathcal{MCG}(S)$ with unbounded orbits.

Individual mapping classes act on the curve graph in the following way. A mapping class is called *reducible* if it preserves a non-trivial *multicurve*, i.e. a collection of essential simple closed curves which span a simplex in $\mathcal{CC}(S)$. A Dehn twist is reducible. The subgroup of $\mathcal{MCG}(S)$ generated by a reducible element acts on the curve graph $\mathcal{C}(S)$ with bounded orbits. An element $\phi \in \mathcal{MCG}(S)$ which is neither reducible nor of finite order is called *pseudo-Anosov*. A pseudo-Anosov element preserves a geodesic in the curve graph $\mathcal{C}(S)$ and acts on this geodesic as a nontrivial translation [37], [6].

The action of $\mathcal{MCG}(S)$ on $\mathcal{C}(S)$ can be used to construct non-trivial *second bounded cohomology classes* for subgroups of $\mathcal{MCG}(S)$ [4], also with nontrivial coefficients [15]. Since the second bounded cohomology group of an irreducible lattice in a semi-simple Lie group of non-compact type and higher rank injects into its usual second cohomology group and hence is finite dimensional, one obtains the following result [4] which was first established with a different method by Farb and Masur [12] building in an essential way on the work of Kaimanovich and Masur [26].

Theorem 2.4 (Kaimanovich-Masur, Farb-Masur). *Let Γ be an irreducible lattice in a semi-simple Lie group of non-compact type and rank at least 2. Then the image of every homomorphism $\rho : \Gamma \to \mathcal{MCG}(S)$ is finite.*

This result can also be extended to cocycles and to irreducible lattices in arbitrary groups of higher rank [16]. Here a *cocycle* for a group Γ acting on a probability space (X, ν) by measure preserving transformations with values in a group Λ is a ν-measurable map $\alpha : \Gamma \times X \to \Lambda$ such that

$$\alpha(gh, x) = \alpha(g, hx)\alpha(h, x)$$

for all $g, h \in \Gamma$ and almost all $x \in X$. Two cocycles α, β are *cohomologous* if there is a measurable map $\chi : X \to \Lambda$ such that

$$\alpha(g, x)\chi(x) = \chi(gx)\beta(g, x)$$

for all $g \in \Gamma$ and almost all x.

Theorem 2.5. *Let $n \geq 2$, let $\Gamma < G_1 \times \cdots \times G_n = G$ be an irreducible lattice, let (X, ν) be a mildly mixing Γ-space and let $\alpha : \Gamma \times X \to \mathcal{MCG}(S)$ be any cocycle. Then α is cohomologous to a cocycle α' with values in a subgroup $H = H_0 \times H_1$ of $\mathcal{MCG}(S)$ where H_0 is virtually abelian and where H_1 contains a finite normal subgroup K such that the projection of α' into H_1/K defines a continuous homomorphism $G \to H_1/K$.*

Hyperbolic behavior of subgroups of the mapping class group is closely related to large-scale properties of their action on the curve graph. To this end, note that since $\mathcal{C}(S)$ is arc connected, there is a constant $c > 0$ such that for every vertex $\gamma \in \mathcal{C}(S)$ the orbit map $\phi \to \phi\gamma$ ($\phi \in \mathcal{MCG}(S)$) is *c-Lipschitz*. However, this orbit map largely distorts distances. In fact, the orbit of an infinite cyclic subgroup of $\mathcal{MCG}(S)$ generated by a reducible element of infinite order is bounded.

On the other hand, if $\phi \in \mathcal{MCG}(S)$ is pseudo-Anosov then for any vertex $\gamma \in \mathcal{C}(S)$ the orbit map $k \in \mathbb{Z} \to \phi^k\gamma$ is a quasi-geodesic. Even more is true.

Namely, let for the moment S be a *closed* surface. If $G < \mathcal{MCG}(S)$ is any subgroup then there is an exact sequence

$$0 \to \pi_1(S) \to H \to G \to 0.$$

The group H is an extension of the surface group $\pi_1(S)$. Vice versa, for every exact sequence of this form there is a natural homomorphism $G \to \mathcal{MCG}(S)$. If $G =< \phi >$ is generated by a pseudo-Anosov element $\phi \in \mathcal{MCG}(S)$ then the extension H is the fundamental group of a closed hyperbolic 3-manifold (this is the celebrated hyperbolization result for Haken manifolds of Thurston, see [43] for the foundational cornerstone of Thurston's work). This three-manifold is just the mapping torus of ϕ. In particular, H is a word hyperbolic group.

Farb and Mosher [14] defined a geometric generalization of such subgroups of $\mathcal{MCG}(S)$. The following definition is equivalent to the one given in [14] and was introduced in [24], [19]. It also makes sense if S has punctures.

Definition 2.6. A finitely generated subgroup G of $\mathcal{MCG}(S)$ is *convex cocompact* if and only if an orbit map $g \in G \to g\gamma \in \mathcal{C}(S)$ ($\gamma \in \mathcal{C}(S)$) is a quasi-isometric embedding.

One direction of the following equivalence is due to Farb and Mosher [14], the second direction was established in [20].

Proposition 2.7. *Let S be a closed surface and let $0 \to \pi_1(S) \to H \to G \to 0$ be an exact sequence. Then the following are equivalent.*

1. *The kernel of the homomorphism $G \to \mathcal{MCG}(S)$ is finite, and the image is convex cocompact.*
2. *H is word hyperbolic.*

As indicated above, infinite cyclic groups generated by a single pseudo-Anosov element are convex cocompact. Other examples can be obtained as follows.

Two pseudo-Anosov elements $\phi, \psi \in \mathcal{MCG}(S)$ are *independent* if they are not contained in a common virtually cyclic subgroup of $\mathcal{MCG}(S)$. For such elements, it follows from a standard ping-pong argument for the action on the space of geodesic laminations that for sufficiently large $k, \ell > 0$ the subgroup of $\mathcal{MCG}(S)$ generated by ϕ^k, ψ^ℓ is free [38]. Moreover, this group is convex cocompact provided that k, ℓ are sufficiently large (see [14] for a detailed discussion). As a consequence, free convex cocompact groups with an arbitrary number of generators exist.

However, up to now no convex cocompact surface group is known. In other words, there are no known examples of surface bundles over surfaces with word hyperbolic fundamental group.

Even more, there are no known examples of finitely generated subgroups of $\mathcal{MCG}(S)$ which only contain pseudo-Anosov elements and are *not* virtually free. This leads to the following question.

Question 5: Is a finitely generated purely pseudo-Anosov subgroup of $\mathcal{MCG}(S)$ convex cocompact and virtually free?

Recent results give some evidence in this direction. In [25], Kent, Leininger and Schleimer investigate purely pseudo-Anosov subgroups of $\mathcal{MCG}(S)$ of a special form. To formulate their result, let S be a closed surface of genus $g \geq 2$ and let S^* be the surface S with one point deleted. There is an exact sequence

$$0 \to \pi_1(S) \to \mathcal{MCG}(S^*) \to \mathcal{MCG}(S) \to 0.$$

Here the homomorphism $\mathcal{MCG}(S^*) \to \mathcal{MCG}(S)$ is defined by the point closing map which deletes the marked point (puncture), and an element $\alpha \in \pi_1(S)$ defines an element in $\mathcal{MCG}(S^*)$ by dragging the puncture along a representative of α. Kent, Leininger and Schleimer show

Proposition 2.8 (Kent-Leininger-Schleimer). *If $G < \pi_1(S)$ is finitely generated and defines a purely pseudo-Anosov subgroup of $\mathcal{MCG}(S^*)$ then G is convex cocompact.*

The following is an extension by Dahmani and Fujiwara [10] to one-ended hyperbolic groups of a result of Bowditch [7] for surface groups.

Theorem 2.9 (Bowditch, Dahmani-Fujiwara). *There are only finitely many conjugacy classes of purely pseudo-Anosov one-ended hyperbolic subgroups of* $\mathcal{MCG}(S)$.

On the other hand, surface subgroups of $\mathcal{MCG}(S)$ do exist. Indeed, Leininger and Reid [33] showed

Proposition 2.10 (Leininger-Reid 06). *There are surface subgroups of* $\mathcal{MCG}(S)$ *which contain a single conjugacy class of a maximal abelian subgroup not consisting of pseudo-Anosov elements.*

3. The Action of the Mapping Class Group on Teichmüller Space

Let $\mathcal{T}(S)$ be the *Teichmüller space* of all isometry classes of complete marked hyperbolic metrics on S of finite area. Here two hyperbolic metrics g, g' define the same point in $\mathcal{T}(S)$ if there is a diffeomorphism $\psi : S \to S$ which is isotopic to the identity and such that $\psi^* g' = g$. Equivalently, $\mathcal{T}(S)$ is the set of all marked complex or conformal structures on S. Teichmüller space has the structure of a smooth manifold diffeomorphic to $\mathbb{R}^{6g-6+2m}$. The mapping class group $\mathcal{MCG}(S)$ naturally acts on $\mathcal{T}(S)$ properly discontinuously. However, this action is neither free nor cocompact.

The *Weil-Petersson metric* on $\mathcal{T}(S)$ is a smooth $\mathcal{MCG}(S)$-invariant Riemannian metric of negative sectional curvature. The metric is incomplete. Its completion can be described as follows. A *surface with nodes* is obtained from S by pinching one or several simple closed curves on S to a point. The Teichmüller space of a surface with nodes is defined. The completion of the Weil-Petersson metric is a CAT(0)-space which is the union of the Weil-Petersson spaces for surfaces with nodes. This completion is not locally compact.

An isometry ϕ of a CAT(0)-space X is called *semi-simple* if the dilation $x \to d(x, \phi(x))$ assumes a minimum on X. A semi-simple isometry is *elliptic* if it fixes a point in X. An isometry which is not semi-simple is called *parabolic*, and it is *neutral parabolic* if the infimum of the dilation vanishes.

The following observation can be found in [18] or [8].

Proposition 3.1. *The mapping class group acts on the WP-completion of* $\mathcal{T}(S)$ *by semi-simple isometries.*

Each Dehn twist acts as an elliptic element. In particular, the action is not proper in a metric sense. In fact, Bridson showed [8]

Proposition 3.2 (Bridson 09). *If* S *is a closed surface of genus* $g \geq 3$ *and if* $\mathcal{MCG}(S)$ *acts isometrically on a* CAT(0)*-space then each Dehn twist acts as an elliptic element or a neutral parabolic.*

For a closed surface S of genus $g \geq 2$ there is a natural isometric action of $\mathcal{MCG}(S)$ on a *proper* CAT(0)-space. Namely, an orientation preserving homeomorphism of S acts as an automorphism on the first homology group $H_1(S, \mathbb{Z})$ of S preserving the *homology intersection form*. Since the intersection form is non-degenerate and since $H_1(S, \mathbb{Z}) = \mathbb{Z}^{2g}$, this action defines a representation of $\mathcal{MCG}(S)$ into the integral symplectic group $Sp(2g, \mathbb{Z})$. The representation is surjective. Its kernel is called the *Torelli group* (see [13]).

The group $Sp(2g, \mathbb{Z})$ is a lattice in the isometry group $Sp(2g, \mathbb{R})$ of a symmetric space of non-compact type. As a consequence, $\mathcal{MCG}(S)$ admits an isometric action on a CAT(0)-space which factors through the inclusion $Sp(2g, \mathbb{Z}) \to Sp(2g, \mathbb{R})$. However, this action is far from effective. Indeed, the Torelli group is infinite. More precisely, it is a consequence of a result of Mess that for $g = 2$ the Torelli group is an infinitely generated free group. For $g \geq 3$ the Torelli group is finitely generated, but it is not known whether it is finitely presented (see [13]).

Question 6: Does $\mathcal{MCG}(S)$ admit a *proper* isometric action on a CAT(0)-space?

If S is a closed surface of genus $g = 2$ then there is a proper isometric action of $\mathcal{MCG}(S)$ by semi-simple isometries on a CAT(0)-space of dimension 18 [8]. By Proposition 3.2, such an action can not exist for $g \geq 3$.

4. A Geometric Model for the Mapping Class Group

A geometric model for the mapping class group is a locally compact geodesic metric space on which $\mathcal{MCG}(S)$ acts properly and cocompactly. In this section we present such a geometric model and explain how it is used to gain information on $\mathcal{MCG}(S)$.

A *train track* on S is an embedded 1-complex $\tau \subset S$ whose edges (called *branches*) are smooth arcs with well-defined tangent vectors at the endpoints. At any vertex (called a *switch*) the incident edges are mutually tangent. Through each switch there is a path of class C^1 which is embedded in τ and contains the switch in its interior. In particular, the branches which are incident on a fixed switch are divided into "incoming" and "outgoing" branches according to their inward pointing tangent at the switch. Each closed curve component of τ has a unique bivalent switch, and all other switches are at least trivalent. The complementary regions of the train track have negative Euler characteristic, which means that they are different from discs with $0, 1$ or 2 cusps at the boundary and different from annuli and once-punctured discs with no cusps at the boundary. A train track is called *generic* if every switch is at most 3-valent. Train tracks are identified if they are isotopic.

For a given complete hyperbolic metric of finite volume on the surface S, a *geodesic lamination* is a closed subset of S which can be foliated by simple geodesics. A geodesic lamination is *minimal* if every half-leaf is dense. A geodesic lamination is *maximal* if its complementary components are all ideal triangles or once punctured monogons. A geodesic lamination is *complete* if it is maximal and can be approximated in the Hausdorff topology for compact subsets of S by simple closed geodesics.

A geodesic lamination λ is *carried* by a train track τ if there is a map $F : S \to S$ of class C^1 which is homotopic to the identity and maps λ into τ so that the restriction of the differential dF of F to λ vanishes nowhere. A train track τ is called *complete* if it is generic and *transversely recurrent* (see [40] for details on this technical concept) and if it carries a complete geodesic lamination. Such a train track is necessarily *maximal*, i.e. its complementary components are all trigons and once punctured monogons.

A half-branch b of a generic train track τ is called *large* if every locally embedded path $\alpha : (-\epsilon, \epsilon) \to \tau$ of class C^1 which passes through the switch on which b is incident intersects the interior of b. A half-branch which is not large is called *small*. A branch is *large* if each of its half-branches is large. If τ is a complete train track and if b is a large branch of τ then τ can be modified with a single right or left *split* at b to a maximal train track η as shown in the figure. If λ is a complete geodesic lamination carried by τ then there is a single

choice of a right or left split at b so that the split track η carries λ and hence it is complete.

Let $\mathcal{TT}(S)$ be the locally finite directed graph whose vertices are the isotopy classes of complete train tracks on S and where two such train tracks τ, η are connected by a directed edge of length one if η can be obtained from τ by a single split. Then [16]

Proposition 4.1. *$\mathcal{TT}(S)$ is connected and $\mathcal{MCG}(S)$ acts on $\mathcal{TT}(S)$ properly and cocompactly.*

As a consequence, $\mathcal{TT}(S)$ is a geometric model for $\mathcal{MCG}(S)$ which can be used to gain some understanding of $\mathcal{MCG}(S)$. For example, it allows to give a fairly explicit description of efficient paths connecting two points in $\mathcal{MCG}(S)$. To this end, let $\tau \in \mathcal{TT}(S)$ be any vertex and let λ be a complete geodesic lamination carried by τ. As mentioned above, for every large branch b of τ there is a unique choice of a right or left split of τ at b so that the split track carries λ. Let $E(\tau, \lambda)$ be the complete subgraph of $\mathcal{TT}(S)$ whose vertices are the train tracks which can be obtained from τ by a directed edge path (also called a splitting sequence) and which carry λ. Call $E(\tau, \lambda)$ a *cubical euclidean cone*. These cubical euclidean cones can be equipped with their intrinsic path metric d_E.

Proposition 4.2. *There is a number $L > 1$ such that for every vertex $\tau \in \mathcal{TT}(S)$ and every complete geodesic lamination λ carried by τ the inclusion $(E(\tau, \lambda), d_E) \to \mathcal{TT}(S)$ is an L-quasi-isometric embedding.*

As an example of such a cubical euclidean cone, let P be a pants decomposition of S and let τ be a complete train track containing every pants curve c of P as an embedded subgraph consisting of two branches, one small branch (with two small half-branches) and one large branch. Then there is a train track which is obtained from τ by a single split at the large branch and which coincides up to isotopy with the image of τ by a positive (or negative) Dehn twist about c. As a consequence, there is a choice of a positive or negative Dehn twist about each pants curve of P such that the sub-semigroup $A \subset \mathcal{MCG}(S)$ which is generated by these Dehn twists has the following property. There is a finite complete geodesic lamination λ carried by τ (i.e. λ contains only finitely many leaves) which contains P as the union of minimal components such that the set of vertices of $E(\tau, \lambda)$ is precisely $A(\tau)$.

Now splitting sequences in a cone $E(\tau, \lambda)$ are geodesics for the intrinsic path metric d_E and hence splitting sequences are uniform quasi-geodesics in $\mathcal{TT}(S)$. These quasi-geodesics can be used to obtain a fairly explicit understanding of the geometry of $\mathcal{MCG}(S)$. It turns out that on the large scale, $\mathcal{MCG}(S)$ resembles a group which acts properly and isometrically on a CAT(0)-space.

To this end, note that groups Γ which admit a proper cocompact isometric action on a Cat(0)-space X have particularly nice properties. Namely, since in a Cat(0)-space uniqueness of geodesics holds true, these geodesics coarsely define a *bicombing* of Γ. Such a bicombing consists of a collection of discrete paths $\rho_{a,b}$ (i.e. maps $\rho_{a,b} : [0, k] \cap \mathbb{N} \to \Gamma$) connecting any pair of points $a, b \in \Gamma$. Simply fix a basepoint $x_0 \in X$ and choose a geodesic γ connecting ax_0 to bx_0. Then there is a uniform quasi-geodesic $(a_i) \subset \Gamma$ so that $a_0 = a$, $a_m = b$ and that $a_i x_0$ is contained in a uniformly bounded neighborhood of γ. It is convenient to extend the combing paths $\rho_{a,b}$ to eventually constant paths defined on all natural numbers.

By the convexity property of CAT(0)-spaces, these paths have the following fellow traveller property.

Definition 4.3. A bicombing of a group Γ consisting of discrete paths $\rho_{a,b}$ $(a, b \in \Gamma)$ is *quasi-geodesic* if there exists a constant $L \geq 1$ so that each of the combing lines $\rho_{a,b}$ is an L-quasi-geodesic. A bicombing is *bounded* if there is a number $L \geq 1$ such that

$$d(\rho_{a,b}(t), \rho_{a',b'}(t)) \leq L(d(a, a') + d(b, b')) + L$$

for all $a, a', b, b' \in \Gamma$ (where this estimate is meant to hold for the eventually constant extensions of the combing paths).

A group Γ is called *semi-hyperbolic* if it admits a bounded quasi-geodesic bicombing which is equivariant with respect to the left action of Γ. Examples

of semi-hyperbolic groups are groups which admit proper cocompact isometric actions on a CAT(0)-space. Word hyperbolic groups are semi-hyperbolic as well. Semi-hyperbolicity passes on to subgroups of finite index and finite extensions.

The following properties can be found in [9].

Theorem 4.4. *Let Γ be a semi-hyperbolic group.*

1. *Γ is finite presented.*
2. *Γ has solvable word problem, with quadratic Dehn function.*
3. *The conjugacy problem in Γ can be solved in exponential time.*
4. *Every finitely generated abelian subgroup of Γ is quasi-isometrically embedded.*
5. *Every polycyclic subgroup of Γ is virtually abelian.*

An *automatic structure* for Γ consists of a finite *alphabet* A, a (not necessarily injective) map $\pi : A \to \Gamma$ and a *regular language* L in A with the following properties.

1. The set $\pi(A)$ generates Γ as a semi-group.
2. Via concatenation, every word w in the alphabet A is mapped to a word $\pi(w)$ in the generators $\pi(A)$ of Γ. The restriction of the map π to the set of all words from the language L maps L onto Γ.
3. There is a number $\kappa > 0$ with the following property. For all $x \in A$ and each word $w \in L$ of length $k \geq 0$, the word wx defines a path $s_{wx} : [0, k+1] \to \Gamma$ connecting the unit element to $\pi(wx)$. Since $\pi(L) = \Gamma$, there is a word $w' \in L$ of length $\ell > 0$ with $\pi(w') = \pi(wx)$. Let $s_{w'} : [0, \ell] \to \Gamma$ be the corresponding path in Γ. Then $d(s_{wx}(i), s_{w'}(i)) \leq \kappa$ for every $i \leq \min\{k+1, \ell\}$.

A *biautomatic structure* for the group Γ is an automatic structure (A, L) with the following additional property. The alphabet A admits an inversion ι with $\pi(\iota a) = \pi(a)^{-1}$ for all a, and

$$d(\pi(x)s_w(i), s_{w'}(i)) \leq \kappa$$

for all $w \in L$ all $x \in A$, for any $w' \in L$ with $\pi(w') = \pi(xw)$ and all i.

Thus a biautomatic structure of a group is a semi-hyperbolic structure which can be processed with a finite state automaton. Mosher [39] showed that $\mathcal{MCG}(S)$ admits an automatic structure. This was promoted in [17] to the following result.

Theorem 4.5. *$\mathcal{MCG}(S)$ admits a biautomatic structure.*

In particular, $\mathcal{MCG}(S)$ has the properties described in Theorem 4.4. However, each of these properties besides the time-bound for solving the conjugacy problem was known before.

The automaton realizing the biautomatic structure can be computed explicitly. In fact, the cardinality and the number of its states is uniformly exponential in the complexity of S.

5. Geometry and Rigidity of $\mathcal{MCG}(S)$

The strongest possible geometric rigidity statement for a finitely generated group can be formulated as follows.

Definition 5.1. A finitely generated group Γ is *quasi-isometrically rigid* if for any group Γ' which is quasi-isometric to Γ there is a finite index subgroup Γ'_0 of Γ' and a homomorphism $\rho : \Gamma'_0 \to \Gamma$ with finite kernel and finite index image.

For arbitrary metric spaces, there is another related notion of rigidity. Namely, a metric space X is called *quasi-isometrically rigid* if every quasi-isometry of X is at bounded distance from an isometry.

The following result is due to Eskin and Farb [11] and independently to Kleiner and Leeb [29].

Theorem 5.2 (Eskin-Farb, Kleiner-Leeb 07). *Irreducible symmetric spaces of non-compact type and of rank at least two are quasi-isometrically rigid.*

A cocompact lattice Γ in a semi-simple Lie group G is quasi-isometric to G. As a consequence, any two such lattices are quasi-isometric. In contrast, Schwartz [41] showed

Theorem 5.3 (Schwartz 95). *Let Γ be a non-uniform lattice in a rank one simple Lie group which is not locally isomorphic to $SL(2,\mathbb{R})$. If Λ is any group which is quasi-isometric to Γ then Λ is a finite extension of a non-uniform lattice Γ' of G which is commensurable to Γ.*

In other words, up to passing to a finite index subgroup, there is a homomorphism $\rho : \Lambda \to \Gamma$ with finite kernel and finite index image.

Most strategies for showing that a finitely generated group Γ (or an arbitrary metric space X) is quasi-isometrically rigid evolve about the construction of asymptotic geometric invariants for such a group. Gromov proposed to construct such invariants with a renormalization (or rescaling) process. The idea is to fix a basepoint $x \in \Gamma$ and consider the sequence of pointed metric spaces $(\Gamma, x, d/m)$ where d is any distance defined by a word metric and where $m \in \mathbb{N}$. In the case that $\Gamma = \mathbb{Z}^n$ is a free abelian group of rank n, it is easily seen that the pointed metric spaces $(\mathbb{Z}^n, x, d/n)$ converge in the *pointed Gromov Hausdorff topology* to the space $\mathbb{R}^n$ equipped with some norm. However, in general convergence can not be expected.

To force convergence, Gromov uses nonprincipal ultrafilters.

A *nonprincipal ultrafilter* is a finitely additive probability measure ω on the natural numbers $\mathbb{N}$ such that $\omega(S) = 0$ or 1 for every $S \subset \mathbb{N}$ and $\omega(S) = 0$ for every finite subset $S \subset \mathbb{N}$. Given a compact metric space X and a sequence $(a_i) \subset X$ $(i \in \mathbb{N})$, there is a unique element $\omega - \lim a_i \in X$ such that for every neighborhood U of $\omega - \lim a_i$ we have $\omega\{i \mid a_i \in U\} = 1$. In particular, given any bounded sequence $(a_i) \subset \mathbb{R}$, $\omega - \lim a_i$ is a point selected by ω.

Let (X, d) be any metric space and let $x_0 \in X$. Write $X_\infty = \{(x_i) \in \prod_{i\in\mathbb{N}} X \mid d(x_i, x_0)/i \text{ is bounded}\}$. For $x = (x_i), y = (y_i) \in X_\infty$ the sequence $d(x_i, y_i)/i$ is bounded and hence we can define $\tilde{d}_\omega(x, y) = \omega - \lim d(x_i, y_i)/i$. Then $\tilde{d}_\omega$ is a pseudodistance on X_∞, and the quotient metric space X_ω equipped with the projection d_ω of the pseudodistance $\tilde{d}_\omega$ is called the *asymptotic cone* of X with respect to the non-principal ultrafilter ω and with basepoint $*$ defined by x_0. The resulting pointed metric space $(X_\omega, *)$ does not depend on the choice of $x_0 \in X$, but it may depend on the choice of ω. The asymptotic cones of two quasi-isometric spaces are bilipschitz equivalent. If the isometry group of X acts cocompactly then an asymptotic cone of X admits a transitive group of isometries whose elements can be represented by sequences in $\mathrm{Iso}(X)$. The asymptotic cone of a CAT(0)-space is a CAT(0)-space.

A choice of a word norm for the mapping class group and of a non-principal ultrafilter on $\mathbb{N}$ determines an asymptotic cone of $\mathcal{MCG}(S)$. The homological dimension of this cone, i.e. the maximal number $n \geq 0$ such that there are two open subsets $V \subset U$ with $H_n(U, U - V) \neq 0$, is independent of the choices. The following is a version of a result of Behrstock and Minsky [1], [20].

Theorem 5.4. *The homological dimension of an asymptotic cone of $\mathcal{MCG}(S)$ equals $3g - 3 + m$.*

Each cubical euclidean cone $E(\tau, \lambda) \subset \mathcal{TT}(S)$ is the one-skeleton of a Cat(0)-cubical complex, and it is of uniform polynomial growth. It turns out that the ω-asymptotic cones of these cubical euclidean cones are homeomorphic to cones in an euclidean space. Moreover, they embed with a uniform bilipschitz embedding into the ω-asymptotic cone of $\mathcal{TT}(S)$ which is bilipschitz equivalent to the asymptotic cone of $\mathcal{MCG}(S)$. As in the case of a symmetric space of higher rank, the asymptotic cones of the cubical euclidean cones and their mutual intersections define a (locally infinite) cell complex contained in $\mathcal{TT}(S)_\omega$. Pants decompositions of S define a family of asymptotic subcones of the asymptotic cone $\mathcal{TT}(S)_\omega$, one for each tuple of choices of a positive or negative Dehn twist about the pants curves. As a consequence, this cell complex contains a topological version of the curve complex as a subcomplex.

Now the main observation is as follows. Any quasi-isometry of $\mathcal{TT}(S)$ (which is viewed as a geometric model for $\mathcal{MCG}(S)$) defines a homeomorphism of the ω-asymptotic cone, and this homeomorphism induces a homeomorphism of the above cell complex. As a consequence, this homeomorphism induces an automorphism of the curve complex and hence coincides with the action of

an element of $\mathcal{MCG}(S)$ by Ivanov's result (Theorem 2.3). This leads to the following result which can be found in [20] and [3].

Theorem 5.5. *The mapping class group is quasi-isometrically rigid.*

A rigidity result in a different direction is due to Kida. Namely, call a countable group Γ *measure equivalent* to a countable group Λ if there are commuting actions of Λ, Γ on a standard Borel space preserving a locally finite measure such that both actions have a finite measure fundamental domain. Kida showed [28]

Theorem 5.6 (Kida 06). *If Γ is any group which is measure equivalent to $\mathcal{MCG}(S)$ then there is a finite index subgroup Γ' of Γ and a homomorphism $\rho : \Gamma' \to \mathcal{MCG}(S)$ with finite kernel and finite index image.*

6. Resemblance with Lattices

In a symmetric space Z of non-compact type without compact or euclidean factors, the *Weyl chambers* are totally geodesic euclidean cones of maximal dimension. The *Furstenberg boundary* of Z is the set of equivalence classes of Weyl chambers in Z where two such Weyl chambers are equivalent if their Hausdorff distance is finite. Let G be the isometry group of Z. Since the stabilizer in G of the boundary at infinity of a Weyl chamber in Z is a minimal parabolic (in particular an amenable) subgroup P of G, the Furstenberg boundary can G-equivariantly be identified with G/P. The space G/P is compact, and G acts transitively as a continuous group of transformations on G/P.

A lattice Γ in G acts continuously on $X = G/P$ as a group of homeomorphisms. This action is *topologically amenable*, which means that the following holds true. Let $\mathcal{P}(\Gamma)$ be the convex space of all Borel probability measures on Γ; note that $\mathcal{P}(\Gamma)$ can be viewed as a subset of the unit ball in the space $\ell^1(\Gamma)$ of summable functions on Γ and therefore it admits a natural norm $\| \, \|$. The group Γ acts on $(\mathcal{P}(\Gamma), \| \, \|)$ isometrically by left translation. We require that there is a sequence of weak*-continuous maps $\xi_n : X \to \mathcal{P}(\Gamma)$ with the property that $\|g\xi_n(x) - \xi_n(gx)\| \to 0$ $(n \to \infty)$ uniformly on compact subsets of $\Gamma \times X$. A countable group Γ is *boundary amenable* if it admits a topologically amenable action on a compact space.

By the work of Higson [22], for any countable group Γ which is boundary amenable and for every separable $\Gamma - C^*$-algebra A, the Baum-Connes assembly map

$$\mu : KK_*^{\Gamma}(\mathcal{E}\Gamma, A) \to KK(\mathbb{C}, C_r^*(\Gamma, A))$$

is split injective. As a consequence, the strong Novikov conjecture holds for Γ and hence the Novikov higher signature conjecture holds as well.

In the train track complex $\mathcal{TT}(S)$, two cubical euclidean cones $E(\tau, \lambda), E(\sigma, \nu)$ have bounded Hausdorff distance if $\lambda = \nu$. Thus the space

$\mathcal{CL}(S)$ of complete geodesic laminations equipped with the Hausdorff topology can be viewed as a space of equivalence classes of cubical euclidean cones where two cones are equivalent only if their Hausdorff distance is finite. However, there are cubical euclidean cones of finite Hausdorff distance which are defined by distinct complete geodesic laminations, for example by geodesic laminations which contain a common sublamination which fills up S. Then $\mathcal{CL}(S)$ is a compact $\mathcal{MCG}(S)$-space. As for lattices in semi-simple Lie groups of non-compact type, for the mapping class group the following is satisfied [16].

Theorem 6.1. *The action of $\mathcal{MCG}(S)$ on the space $\mathcal{CL}(S)$ of complete geodesic laminations on S is topologically amenable.*

As a consequence, the mapping class group is boundary amenable (which also follows from the work of Kida [27]). Since boundary amenability is passed on to subgroups one obtains as a corollary

Corollary 6.2. *The Novikov higher order signature conjecture holds for any subgroup of the mapping class group of a non-exceptional surface of finite type.*

More recently, Behrstock and Minsky [2] gave another proof of Corollary 6.2 which does not use boundary amenability.

The analogy of $\mathcal{CL}(S)$ with the Furstenberg boundary of a symmetric space goes further. Namely, the following Tits alternative due to McCarthy [38] provides a complete understanding of amenable subgroups of mapping class groups.

Theorem 6.3 (McCarthy 85). *Let Γ be any subgroup of $\mathcal{MCG}(S)$. Then either Γ contains an abelian subgroup of finite index or Γ contains a free group of rank 2.*

As a consequence, amenable sugroups of $\mathcal{MCG}(S)$ are virtually abelian (i.e. they contain an abelian subgroup of finite index). This yields the following

Corollary 6.4. *Stabilizers of points in $\mathcal{CL}(S)$ are amenable, and every amenable subgroup of $\mathcal{MCG}(S)$ is commensurable to the stabilizer of a point in $\mathcal{CL}(S)$.*

Theorem 6.1 implies that the action of the mapping class group on $\mathcal{CL}(S)$ is *universally amenable*. This means that it is amenable for every invariant Borel measure class. Such a measure class can be defined as follows.

A *measured geodesic lamination* is a geodesic lamination equipped with a tranverse translation invariant measure. The space $\mathcal{ML}(S)$ of all measured geodesic laminations can be equipped with the weak*-topology. With respect to this topology, $\mathcal{ML}(S)$ is homeomorphic to a cone over a sphere of dimension $6g-7+2m$.

A measured geodesic lamination is *carried* by a complete train track τ if its support is carried by τ. Each measured geodesic lamination carried by τ defines a non-negative weight function on the branches of τ which satisfies a system of linear equations, the so-called *switch conditions*. Vice versa, a non-negative non

vanishing solution to the switch conditions defines a measured geodesic lamination carried by τ. The space $\mathcal{ML}(\tau)$ of nonnegative solutions of the switch conditions for τ is a convex cone in a linear space and hence it admits a natural measure in the Lebesgue measure class. If τ' is obtained from τ by a single split then $\mathcal{ML}(\tau') \subset \mathcal{ML}(\tau)$. Moreover, the *carrying map* transforming solutions of the switch conditions for τ' to solutions of the switch conditions for τ is linear. In fact, it is contained in the special linear group and hence it preserves the Lebesgue measure. As a consequence, this Lebesgue measure does not depend on the train track used to define it. In other words, this construction defines a locally finite Borel measure (i.e. a Radon measure) on $\mathcal{ML}(S)$. Since $\mathcal{MCG}(S)$ acts naturally on both train tracks and measured geodesic laminations, this measure is $\mathcal{MCG}(S)$-invariant and projects to a $\mathcal{MCG}(S)$-invariant measure class on the projectivization $\mathcal{PML}(S)$ of $\mathcal{ML}(S)$.

The following result is due to Masur [35] and Veech [44]. For its formulation, a measured geodesic lamination λ is called *uniquely ergodic* if its support admits a unique transverse measure up to scale. A $\mathcal{MCG}(S)$-invariant Radon measure μ on $\mathcal{ML}(S)$ is called *ergodic* if $\mu(A) = 0$ or $\mu(\mathcal{ML}(S) - A) = 0$ for every $\mathcal{MCG}(S)$-invariant Borel set A.

Theorem 6.5 (Masur, Veech). *The Lebesgue measure λ on $\mathcal{ML}(S)$ gives full mass to the measured geodesic laminations whose support is minimal and maximal and which are uniquely ergodic. Moreover, the action of $\mathcal{MCG}(S)$ on $\mathcal{ML}(S)$ is ergodic.*

By Theorem 6.5, there is a $\mathcal{MCG}(S)$-invariant Borel subset A of $\mathcal{PML}(S)$ of full measure (namely, the set of all uniquely ergodic projective measured geodesic laminations whose support is maximal and which are uniquely ergodic) which admits an equivariant homeomorphism (the map which associates to a measured geodesic lamination its support) onto an invariant Borel subset of $\mathcal{CL}(S)$. Via this map, the Lebesgue measure induces an invariant ergodic measure class on $\mathcal{CL}(S)$. By Theorem 6.1, the action of $\mathcal{MCG}(S)$ with respect to this measure class is amenable.

The isometry group $PSL(2,\mathbb{R})$ of the hyperbolic plane $\mathbf{H}^2$ acts simply transitively on the unit tangent bundle of $\mathbf{H}^2$. The quotient $PSL(2,\mathbb{R})/PSL(2,\mathbb{Z})$ can naturally be identified with the unit tangent bundle of the modular surface $\mathbf{H}^2/PSL(2,\mathbb{Z})$. The unipotent group of all upper triangular matrices with diagonal 1 acts from the left as the *horocycle flow*. There are two types of invariant ergodic Borel probability measures for this flow. The first kind is supported on a periodic orbit. There is also the Lebesgue measure. This is a complete classification of invariant ergodic probability measures.

The classification problem of invariant measures for the horocycle flow admits a second description. Namely, such measures correspond to invariant Radon measures (locally finite Borel measures) for the standard linear action of the group $SL(2,\mathbb{Z})$ on $\mathbb{R}^2$. There are precisely two types of invariant ergodic Radon measures. The first type is supported on the orbit of a point

whose coordinates are rationally dependent. Namely, such an orbit intersects any compact subset of $\mathbb{R}^2$ in a finite set and hence the sum of the Dirac masses on the orbit points is an invariant Radon measure. There is also the Lebesgue measure.

The classification problem of $SL(2,\mathbb{Z})$-invariant Radon measures on $\mathbb{R}^2$ has an analog for the mapping class group: the classification of $\mathcal{MCG}(S)$-invariant Radon measures on the space $\mathcal{ML}(S)$ of measured geodesic laminations. Such measures correspond to finite Borel measures on the unit cotangent bundle of moduli space $\mathcal{T}(S)/\mathcal{MCG}(S)$ which are invariant under "the stable foliation". Besides the Lebesgue measure and rational measures support on the orbit of a multi-curve, there are some additional types of invariant Radon measures. Namely, a *proper bordered subsurface* S_0 of S is a union of connected components of the space which we obtain from S by cutting S open along a collection of disjoint simple closed geodesics. Then S_0 is a surface with non-empty geodesic boundary and of negative Euler characteristic. If two boundary components of S_0 correspond to the same closed geodesic γ in S then we require that $S - S_0$ contains a connected component which is an annulus with core curve γ. Let $\mathcal{ML}(S_0) \subset \mathcal{ML}(S)$ be the space of all measured geodesic laminations on S which are contained in the interior of S_0. The space $\mathcal{ML}(S_0)$ can naturally be identified with the space of measured geodesic laminations on the surface $\hat{S}_0$ of finite type which we obtain from S_0 by collapsing each boundary circle to a puncture. The stabilizer in $\mathcal{MCG}(S)$ of the subsurface S_0 is the direct product of the group of all elements which can be represented by diffeomorphisms leaving S_0 pointwise fixed and a group which is naturally isomorphic to a subgroup G of finite index of the mapping class group $\mathcal{MCG}(\hat{S}_0)$ of $\hat{S}_0$.

Let c be a weighted geodesic multi-curve on S which is disjoint from the interior of S_0. Then for every $\zeta \in \mathcal{ML}(S_0)$ the union $c \cup \zeta$ is a measured geodesic lamination on S which we denote by $c \times \zeta$. Let $\mu(S_0)$ be an $G < \mathcal{MCG}(\hat{S}_0)$-invariant Radon measure on $\mathcal{ML}(S_0)$ which is contained in the Lebesgue measure class. The measure $\mu(S_0)$ can be viewed as a Radon measure on $\mathcal{ML}(S)$ which gives full measure to the laminations of the form $c \times \zeta$ ($\zeta \in \mathcal{ML}(S_0)$) and which is invariant and ergodic under the stabilizer of $c \cup S_0$ in $\mathcal{MCG}(S)$. The translates of this measure under the action of $\mathcal{MCG}(S)$ define an $\mathcal{MCG}(S)$-invariant ergodic wandering measure on $\mathcal{ML}(S)$ which is called a *standard subsurface measure*. If the weighted geodesic multi-curve c contains the boundary of S_0 then the standard subsurface measure defined by $\mu(S_0)$ and c is a Radon measure on $\mathcal{ML}(S)$.

The following is shown in [17] and [34].

Theorem 6.6. *1. An invariant ergodic non-wandering Radon measure for the action of $\mathcal{MCG}(S)$ on $\mathcal{ML}(S)$ coincides with the Lebesgue measure up to scale.*

2. An invariant ergodic wandering Radon measure for the action of $\mathcal{MCG}(S)$ on $\mathcal{ML}(S)$ is either rational or a standard subsurface measure.

References

[1] J. Behrstock, Y. Minsky, *Dimension and rank for mapping class groups*, Ann. of Math. (2) 167 (2008), no. 3, 1055–1077.

[2] J. Behrstock, Y. Minsky, *Centroids and the Rapid Decay property in mapping class groups*, arXiv:0810.1969.

[3] J. Behrstock, B. Kleiner, Y. Minsky, L. Mosher, *Geometry and rigidity of mapping class groups*, arXiv:0801.2006.

[4] M. Bestvina, K. Fujiwara, *Bounded cohomology of subgroups of mapping class groups*, Geom. Topol. 6 (2002), 69–89.

[5] M. Bestvina, M. Feighn, *A hyperbolic* $\mathrm{Out}(F_n)$*-complex*, Groups, Geom. Dyn. 4 (2010), 31–58.

[6] B. Bowditch, *Tight geodesics in the curve complex*, Invent. Math. 171 (2008), 281–300.

[7] B. Bowditch, *Atoroidal surface bundles over surfaces*, preprint October 2007.

[8] M. Bridson, *Semisimple actions of mapping class groups on CAT(0)-spaces*, arXiv:0908.0685, to appear in "The geometry of Riemann surfaces", London Math. Soc.Lecture Notes 368.

[9] M. Bridson, A, Haefliger, *Metric spaces of non-positive curvature*, Springer Grundlehren 319, Springer 1999.

[10] F. Dahmani, K. Fujiwara, *Copies of a one-ended group in a mapping class group*, Groups Geom. Dyn. 3 (2009), no. 3, 359–377.

[11] A. Eskin, B. Farb, *Quasi-flats and rigidity in higher rank symmetric spaces*, J. Amer. Math. Soc. 10 (1997), no. 3, 653–692.

[12] B. Farb, H. Masur, *Superrigidity and mapping class groups*, Topology 37 (1998), 1169–1176.

[13] B. Farb, D. Margalit, *A primer on mapping class groups*, working draft 2009.

[14] B. Farb, L. Mosher, *Convex cocompact subgroups of mapping class groups*, Geom. Topol. 6 (2002), 91–152.

London Math. Soc. Lec. Notes 329 (2006), 187–207.

[15] U. Hamenstädt, *Bounded cohomology and isometry groups of hyperbolic spaces*, J. Eur. Math. Soc. 10 (2008), 315–349.

[16] U. Hamenstädt, *Geometry of the mapping class groups I: Boundary amenability*, Invent. Math. 175 (2009), 545–609.

[17] U. Hamenstädt, *Invariant Radon measures on measured lamination space*, Invent. Math. 176 (2009), 223–273.

[18] U. Hamenstädt, *Dynamical properties of the Weil-Petersson metric*, Contemp. Math. 510 (2010), 109–127.

[19] U. Hamenstädt, *Word hyperbolic extensions of surface groups*, arXiv:math.GT/0505244.

[20] U. Hamenstädt, *Geometry of the mapping class groups III: Quasi-isometric rigidity*, arXiv:math.GT/0512429.

[21] U. Hamenstädt, *Geometry of the mapping class group II: A biautomatic structure*, arXiv:0912.0137.

[22] N. Higson, *Biinvariant K-theory and the Novikov conjecture*, Geom. Funct. Anal. 10 (2000), 563–581.

[23] N. Ivanov, *Automorphism of complexes of curves and of Teichmüller spaces*, Internat. Math. Res. Notices 1997, no. 14, 651–666.

[24] R. Kent, C. Leininger, *Shadows of mapping class groups: capturing convex cocompactness*, Geom. Funct. Anal. 18 (2008), no. 4, 1270–1325.

[25] R. Kent, C. Leininger, S. Schleimer, *Trees and mapping class groups*, to appear in J. reine angew. Math.

[26] V. Kaimanovich, H. Masur, *The Poisson boundary of the mapping class group*, Invent. Math. 125 (1996), 221–264.

[27] Y. Kida, *The mapping class group from the viewpoint of measure equivalence theory*, Mem. Amer. Math. Soc. 196 (2008), no. 916.

[28] Y. Kida, *Measure equivalence rigidity of the mapping class group*, arXiv:math/0607600, to appear in Ann. of Math.

[29] B. Kleiner, B. Leeb, *Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings*, Inst. Hautes Etudes Sci. Publ. Math. No. 86 (1997), 115–197.

[30] M. Korkmaz, *Automorphisms of complexes of curves on punctured spheres and on punctured tori*, Topology Appl. 95 (1999), no. 2, 85–111.

[31] M. Korkmaz, *On cofinite subgroups of mapping class groups*, Turkish J. Math. 27 (2003), no. 1, 115–123.

[32] M. Korkmaz, *Generating the surface mapping class group by two elements*, Trans. AMS 357 (2004), 3299–3310.

[33] C. Leininger, A. Reid, *A combination theorem for Veech subgroups of the mapping class group*, Geom. Funct. Anal. 16 (2006), no. 2, 403–436.

[34] E. Lindenstrauss, M. Mirzakhani, *Ergodic theory of the space of measured laminations*, Int. Math. Res. Not. IMRN 4, 49pp, (2008).

[35] H. Masur, *Interval exchange transformations and measured foliations*, Ann. Math. 115 (1982), 169–201.

[36] H. Masur, Y. Minsky, *Geometry of the complex of curves I: Hyperbolicity*, Invent. Math. 138 (1990), 103–149.

[37] H. Masur, Y. Minsky, *Geometry of the complex of curves II: Hierarchical structures*, Geom. Func. Anal. 10 (2000), 902–974.

[38] J. McCarthy, *A "Tits-alternative" for subgroups of surface mapping class groups*, Trans. Amer. Math. Soc. 291 (1985), no. 2, 583–612.

[39] L. Mosher, *Mapping class groups are automatic*, Ann. Math. 142 (1995), 303–384.

[40] R. Penner with J. Harer, *Combinatorics of train tracks*, Ann. Math. Studies 125, Princeton University Press, Princeton 1992.

[41] R.E. Schwartz, *The quasi-isometry classification of rank one lattices*, Inst. Hautes Etudes Sci. Publ. Math. No. 82 (1995), 133–168.

[42] J.P. Serre, *Trees*, Springer monographs in mathematics, Springer 1980.

[43] W. Thurston, *Three-dimensional geometry and topology*, unpublished manuscript, 1979.

[44] W. Veech, *The Teichmüller geodesic flow*, Ann. Math. 124 (1986), 441–530.

[45] B. Wajnreb, *Mapping class group of a surface is generated by two elements*, Topology 35 (1996), 377–383.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Embedded Contact Homology and Its Applications

Michael Hutchings*

Abstract

Embedded contact homology (ECH) is a kind of Floer homology for contact three-manifolds. Taubes has shown that ECH is isomorphic to a version of Seiberg-Witten Floer homology (and both are conjecturally isomorphic to a version of Heegaard Floer homology). This isomorphism allows information to be transferred between topology and contact geometry in three dimensions. In this article we first give an overview of the definition of embedded contact homology. We then outline its applications to generalizations of the Weinstein conjecture, the Arnold chord conjecture, and obstructions to symplectic embeddings in four dimensions.

Mathematics Subject Classification (2010). Primary 57R58; Secondary 57R17.

Keywords. Embedded contact homology, contact three-manifolds, Weinstein conjecture, chord conjecture

1. Embedded Contact Homology

1.1. Floer homology of 3-manifolds. There are various kinds of Floer theory that one can associate to a closed oriented 3-manifold with a spin-c structure. In this article we regard a *spin-c structure* on a closed oriented 3-manifold Y as an equivalence class of oriented 2-plane fields on Y (i.e. oriented rank 2 subbundles of the tangent bundle TY), where two oriented 2-plane fields are considered equivalent if they are homotopic on the complement of a ball in Y. The set of spin-c structures on Y is an affine space over $H^2(Y;\mathbb{Z})$. A spin-c structure $\mathfrak{s}$ has a well-defined first Chern class $c_1(\mathfrak{s}) \in 2H^2(Y;\mathbb{Z})$. A spin-c structure $\mathfrak{s}$ is called *torsion* if $c_1(\mathfrak{s})$ is torsion in $H^2(Y;\mathbb{Z})$.

*Partially supported by NSF grant DMS-0806037.
Mathematics Department, 970 Evans Hall, University of California, Berkeley CA 94720 USA. E-mail: hutching@math.berkeley.edu.

One version of Floer theory for spin-c 3-manifolds is the *Seiberg-Witten Floer cohomology*, or *monopole Floer cohomology*, as defined by Kronheimer-Mrowka [26]. There are two basic variants of this theory, which are different only for torsion spin-c structures; the variant relevant to our story is denoted by $\widehat{HM}^*(Y,\mathfrak{s})$. Very roughly, this is the homology of a chain complex which is generated by $\mathbb{R}$-invariant solutions to the Seiberg-Witten equations on $\mathbb{R}\times Y$, and whose differential counts non-$\mathbb{R}$-invariant solutions to the Seiberg-Witten equations on $\mathbb{R}\times Y$ which converge to two different $\mathbb{R}$-invariant solutions as the $\mathbb{R}$ coordinate goes to $+\infty$ or $-\infty$. This cohomology is a relatively $\mathbb{Z}/d$-graded $\mathbb{Z}$-module, where d denotes the divisibility of $c_1(\mathfrak{s})$ in $H^2(Y;\mathbb{Z})/\operatorname{Torsion}$.

The Seiberg-Witten Floer cohomology $\widehat{HM}^*(Y,\mathfrak{s})$ is conjecturally isomorphic to a second kind of Floer theory, the *Heegaard Floer homology* $HF^+_*(-Y,\mathfrak{s})$ defined by Ozsváth-Szabó [35]. The latter, roughly speaking, is defined by taking a Heegaard splitting of Y, with Heegaard surface Σ of genus g, and setting up a version of Lagrangian Floer homology in $\operatorname{Sym}^g\Sigma$ for two Lagrangians determined by the Heegaard splitting. Although the definitions of Seiberg-Witten Floer theory and Heegaard Floer theory appear very different, there is extensive evidence that they are isomorphic, and a program for proving that they are isomorphic is outlined in [29].

Seiberg-Witten and Heegaard Floer homology have had a wealth of applications to three-dimensional topology. The present article is concerned with a third kind of Floer homology, called "embedded contact homology" (ECH), which is defined for contact 3-manifolds. Because ECH is defined directly in terms of contact geometry, it is well suited to certain applications in this area.

1.2. Contact geometry preliminaries. Let Y be a closed oriented 3-manifold. A *contact form* on Y is a 1-form λ on Y such that $\lambda\wedge d\lambda>0$ everywhere. The contact form λ determines a 2-plane field $\xi=\operatorname{Ker}(\lambda)$, oriented by $d\lambda$; an oriented 2-plane field obtained in this way is called a *contact structure*. The contact form λ also determines a vector field R, called the *Reeb vector field*, characterized by $d\lambda(R,\cdot)=0$ and $\lambda(R)=1$.

Two basic questions are: First, given a closed oriented 3-manifold Y, what is the classification of contact structures on Y (say, up to homotopy through contact structures)? Second, given a contact structure ξ, what can one say about the dynamics of the Reeb vector field for a contact form λ with $\operatorname{Ker}(\lambda)=\xi$? The first question is a subject of active research which we will not say much about here, except to note that a fundamental theorem of Eliashberg [10] implies that every closed oriented 3-manifold has a contact structure, in fact a unique "overtwisted" contact structure in every homotopy class of oriented 2-plane fields. (A contact structure ξ on a 3-manifold Y is called *overtwisted* if there is an embedded disk $D\subset Y$ such that $TD|_{\partial D}=\xi|_{\partial D}$.) For more on this topic see e.g. [12].

To discuss the second question, we need to make some definitions. Given a closed oriented 3-manifold with a contact form, a *Reeb orbit* is a periodic orbit

of the Reeb vector field R, i.e. a map $\gamma : \mathbb{R}/T\mathbb{Z} \to Y$ for some $T > 0$, such that $\gamma'(t) = R(\gamma(t))$. Two Reeb orbits are considered equivalent if they differ by reparametrization. If $\gamma : \mathbb{R}/T\mathbb{Z} \to Y$ is a Reeb orbit and if k is a positive integer, the kth *iterate* of γ is defined to be the pullback of γ to $\mathbb{R}/kT\mathbb{Z}$. Every Reeb orbit is either embedded in Y, or the kth iterate of an embedded Reeb orbit for some $k > 1$. Given a contact structure ξ, one can ask: What is the minimum number of embedded Reeb orbits that a contact form λ with $\mathrm{Ker}(\lambda) = \xi$ can have? Must there exist Reeb orbits with particular properties? Some questions of this nature are discussed in §2.1 below.

Continuing with the basic definitions, if γ is a Reeb orbit as above, then the linearization of the Reeb flow near γ defines the *linearized return map* P_γ, which is an automorphism of the two-dimensional symplectic vector space $(\xi_{\gamma(0)}, d\lambda)$. The Reeb orbit γ is called *nondegenerate* if P_γ does not have 1 as an eigenvalue. If γ is nondegenerate, then either P_γ has eigenvalues on the unit circle, in which case γ is called *elliptic*; or else P_γ has real eigenvalues, in which case γ is called *hyperbolic*. These notions do not depend on the parametrization of γ. We say that the contact form λ is *nondegenerate* if all Reeb orbits are nondegenerate. For a given contact structure ξ, this property holds for "generic" contact forms λ.

To a nondegenerate Reeb orbit γ and a trivialization τ of $\gamma^*\xi$, one can associate an integer $\mathrm{CZ}_\tau(\gamma)$ called the *Conley-Zehnder index*. Roughly speaking this measures the rotation of the linearized Reeb flow around γ with respect to τ. In particular, if γ is elliptic, then the trivialization τ is homotopic to one with respect to which the linearized Reeb flow around γ rotates by angle $2\pi\theta$ for some $\theta \in \mathbb{R} \setminus \mathbb{Z}$, and

$$\mathrm{CZ}_\tau(\gamma) = 2\lfloor\theta\rfloor + 1,$$

where $\lfloor\cdot\rfloor$ denotes the greatest integer function.

1.3. The ECH chain complex. With the above preliminaries out of the way, we can now define the embedded contact homology of a closed oriented 3-manifold Y with a nondegenerate contact form λ.

To start, define an *orbit set* to be a finite set of pairs $\alpha = \{(\alpha_i, m_i)\}$, where the α_i's are distinct embedded Reeb orbits, and the m_i's are positive integers, which one can regard as "multiplicities". The orbit set is called *admissible* if $m_i = 1$ whenever the Reeb orbit α_i is hyperbolic. The homology class of the orbit set α_i is defined by $[\alpha] := \sum_i m_i\alpha_i \in H_1(Y)$. Given $\Gamma \in H_1(Y)$, we define the ECH chain complex $C_*(Y, \lambda, \Gamma)$ to be the free $\mathbb{Z}$-module generated by admissible orbit sets α with $[\alpha] = \Gamma$. As explained in §1.4.3 below, this chain complex has a relative $\mathbb{Z}/d$-grading, where d denotes the divisibility of $c_1(\xi) + 2\,\mathrm{PD}(\Gamma)$ in $H^2(Y;\mathbb{Z})/\,\mathrm{Torsion}$. We sometimes write a generator α as above using the multiplicative notation $\alpha = \prod_i \alpha_i^{m_i}$, although the chain complex grading and differential that we will define below are not well behaved with respect to this sort of "multiplication".

To define the differential on the chain complex, choose an almost complex structure J on $\mathbb{R}\times Y$ such that J sends ∂_s to the Reeb vector field R, where s denotes the $\mathbb{R}$ coordinate; J is $\mathbb{R}$-invariant; and J sends the contact structure ξ to itself, rotating positively with respect to $d\lambda$. For our purposes a *J-holomorphic curve* in $\mathbb{R}\times Y$ is a map $u:\Sigma\to\mathbb{R}\times Y$ where (Σ, j) is a punctured compact (not necessarily connected) Riemann surface, and $du\circ j = J\circ du$. If γ is a Reeb orbit, a *positive end* of u at γ is an end of Σ on which u is asymptotic to $\mathbb{R}\times\gamma$ as $s\to+\infty$; a *negative end* is defined analogously with $s\to-\infty$. If $\alpha=\{(\alpha_i,m_i)\}$ and $\beta=\{(\beta_j,n_j)\}$ are two orbit sets with $[\alpha]=[\beta]\in H_1(Y)$, let $\mathcal{M}^J(\alpha,\beta)$ denote the moduli space of J-holomorphic curves as above with positive ends at covers of α_i with total covering multiplicity m_i, negative ends at covers of β_j with total covering multiplicity n_j, and no other ends. We declare two such J-holomorphic curves to be equivalent if they represent the same *current* in $\mathbb{R}\times Y$. For this reason we can identify an element of $\mathcal{M}^J(\alpha,\beta)$ with the corresponding current in $\mathbb{R}\times Y$, which we typically denote by C. Note that since J is assumed to be $\mathbb{R}$-invariant, it follows that $\mathbb{R}$ acts on $\mathcal{M}^J(\alpha,\beta)$ by translation of the $\mathbb{R}$ coordinate on $\mathbb{R}\times Y$.

To each J-holomorphic curve $C\in\mathcal{M}^J(\alpha,\beta)$ one can associate an integer $I(C)$, called the "ECH index", which is explained in §1.4 below. The differential on the ECH chain complex counts J-holomorphic curves with ECH index 1, modulo the $\mathbb{R}$ action by translation. Curves with ECH index 1 have various special properties (assuming that J is generic). Among other things, we will see in Proposition 1.2 below that if $I(C)=1$ then C is embedded in $\mathbb{R}\times Y$ (except that C may contain multiply covered $\mathbb{R}$-invariant cylinders), hence the name "embedded" contact homology. In addition, one can use Proposition 1.2 to show that if J is generic, then the subset of $\mathcal{M}^J(\alpha,\beta)$ consisting of J-holomorphic curves C with $I(C)=1$ has finitely many components, each an orbit of the $\mathbb{R}$ action.

Now fix a generic almost complex structure J. One then defines the differential ∂ on the ECH chain complex $C_*(Y,\lambda,\Gamma)$ as follows: If α is an admissible orbit set with $[\alpha]=\Gamma$, then

$$\partial\alpha := \sum_{\beta}\ \sum_{\{C\in\mathcal{M}^J(\alpha,\beta)/\mathbb{R}\,|\,I(C)=1\}} \varepsilon(C)\cdot\beta.$$

Here the first sum is over admissible orbit sets β with $[\beta]=\Gamma$. Also $\varepsilon(C)\in\{\pm1\}$ is a sign, explained in [22, §9]; the signs depend on some orientation choices, but the chain complexes for different orientation choices are canonically isomorphic to each other. It is shown in [21, 22] that $\partial^2=0$. The homology of this chain complex is the *embedded contact homology* $ECH(Y,\lambda,\Gamma)$.

Although the differential ∂ depends on the choice of J, the homology of the chain complex does not. This is a consequence of the following much stronger theorem of Taubes [42, 43], which was conjectured in [20], and which relates ECH to Seiberg-Witten Floer cohomology. To state the theorem, observe that

the contact structure ξ, being an oriented 2-plane field, determines a spin-c structure $\mathfrak{s}_\xi$, see §1.1. We then have:

Theorem 1.1 (Taubes). *There is an isomorphism of relatively $\mathbb{Z}/d$-graded $\mathbb{Z}$-modules*

$$ECH_*(Y,\lambda,\Gamma) \simeq \widehat{HM}^{-*}(Y, \mathfrak{s}_\xi + \mathrm{PD}(\Gamma)). \tag{1.1}$$

Here d denotes the divisibility of

$$c_1(\xi) + 2\,\mathrm{PD}(\Gamma) = c_1(\mathfrak{s}_\xi + \mathrm{PD}(\Gamma))$$

in $H^2(Y;\mathbb{Z})/\operatorname{Torsion}$. Note that both sides of (1.1) are conjecturally isomorphic to the Heegaard Floer homology $HF_*^+(-Y, \mathfrak{s}_\xi + \mathrm{PD}(\Gamma))$.

Remark. Both sides of the isomorphism (1.1) in fact have absolute gradings by homotopy classes of oriented 2-plane fields [17, 26], and it is reasonable to conjecture that the isomorphism (1.1) respects these absolute gradings.

Remark. In particular, Theorem 1.1 implies that, except for possible grading shifts, ECH depends only on the 3-manifold Y and not on the contact structure. This is in sharp contrast to the symplectic field theory of Eliashberg-Givental-Hofer [11] which, while also defined in terms of Reeb orbits and holomorphic curves, is highly sensitive to the contact structure. In particular, the basic versions of symplectic field theory are trivial for overtwisted contact structures in three dimensions, see [47, 6]. On the other hand, while ECH itself does not depend on the contact structure, it contains a canonical element which does distinguish some contact structures, see §1.6.4.

1.4. The ECH index.

To complete the description of the ECH chain complex, we now outline the definition of the ECH index I; full details may be found in [16, 17]. This is the subtle part of the definition of ECH, and we will try to give some idea of its origins. Meanwhile, on a first reading one may wish to skip ahead to the examples and applications.

1.4.1. Four-dimensional motivation. To motivate the definition of the ECH index, recall that Taubes's "SW=Gr" theorem [38] relates the Seiberg-Witten invariants of a closed symplectic 4-manifold (X,ω), which count solutions to the Seiberg-Witten equations on X, to a "Gromov invariant" which counts certain J-holomorphic curves in X. Here J is an ω-compatible almost complex structure on X. The definition of ECH is an analogue of Taubes's Gromov invariant for a contact manifold (Y,λ). Thus Theorem 1.1 above is an analogue of SW=Gr for the noncompact symplectic 4-manifold $\mathbb{R}\times Y$.

For guidance on which J-holomorphic curves in $\mathbb{R}\times Y$ to count, let us recall which J-holomorphic curves are counted by Taubes's Gromov invariant of a closed symplectic 4-manifold (X,ω). Let C be a J-holomorphic curve in (X,ω), and assume that C is not multiply covered. If J is generic, then the moduli space

of J-holomorphic curves near C is a manifold, whose dimension is a topological quantity called the *Fredholm index* of C, which is given by

$$\operatorname{ind}(C) = -\chi(C) + 2c_1(C). \tag{1.2}$$

Here $c_1(C)$ denotes $\langle c_1(TX), [C]\rangle$, where TX is regarded as a rank 2 complex vector bundle via J. In addition, we have the *adjunction formula*

$$c_1(C) = \chi(C) + C \cdot C - 2\delta(C). \tag{1.3}$$

Here $C \cdot C$ denotes the self-intersection number of the homology class $[C] \in H_2(X)$. In addition $\delta(C)$ is a count of the singularities of C with positive integer weights, see [32, §7], so that $\delta(C) \geq 0$ with equality if and only if C is embedded. Now let us define an integer

$$I(C) := c_1(C) + C \cdot C. \tag{1.4}$$

Then equations (1.2), (1.3), and (1.4) above imply that

$$\operatorname{ind}(C) \leq I(C), \tag{1.5}$$

with equality if and only if C is embedded. Taubes's Gromov invariant counts holomorphic currents C with $I(C) = 0$, which are allowed to be multiply covered (but which are not allowed to contain multiple covers of spheres of negative self-intersection). Using (1.5), one can show that if J is generic, then each such C is a disjoint union of embedded curves of Fredholm index zero, except that torus components may be multiply covered. (Multiply covered tori are counted in a subtle manner explained in [39].)

1.4.2. The three-dimensional story. We now consider analogues of the above formulas (1.2), (1.3), and (1.4) in $\mathbb{R} \times Y$, where (Y, λ) is a contact 3-manifold. These necessarily include "boundary terms" arising from the ends of the J-holomorphic curves.

Let $C \in \mathcal{M}^J(\alpha, \beta)$ be a J-holomorphic curve as in §1.3, and assume that C is not multiply covered. It follows from the main theorem in [9] that if J is generic, then $\mathcal{M}^J(\alpha, \beta)$ is a manifold near C, whose dimension can be expressed, similarly to (1.2), as

$$\operatorname{ind}(C) = -\chi(C) + 2c_1(C, \tau) + \mathrm{CZ}^0_\tau(C). \tag{1.6}$$

Here $c_1(C, \tau)$ denotes the "relative first Chern class" of ξ over C with respect to a trivialization τ of ξ over the Reeb orbits α_i and β_j. This is defined by algebraically counting the zeroes of a generic section of ξ over C which on each end is nonvanishing and has winding number zero with respect to the trivialization τ. The relative first Chern class $c_1(C, \tau)$ depends only on τ and on the relative

homology class of C. Also CZ^0_τ denotes the sum, over all the positive ends of C, of the Conley-Zehnder index with respect to τ of the corresponding (possibly multiply covered) Reeb orbit, minus the analogous sum over the negative ends of C.

Second, the adjunction formula (1.3) is now replaced by the *relative adjunction formula*

$$c_1(C,\tau) = \chi(C) + Q_\tau(C) + w_\tau(C) - 2\delta(C). \tag{1.7}$$

Here $Q_\tau(C)$ is a "relative intersection pairing" defined in [16, 17], which is an analogue of the integer $C \cdot C$ in the closed case, and which depends only on τ and the relative homology class of C. Roughly speaking, it is defined by algebraically counting interior intersections of two generic surfaces in $[-1,1]\times Y$ with boundary $\{1\}\times\alpha - \{-1\}\times\beta$ which both represent the relative homology class of C and which near the boundary have a special form with respect to the trivialization τ. As before, $\delta(C)$ is a count of the singularities of C with positive integer weights (which is shown in [37] to be finite in this setting). Finally, $w_\tau(C)$ denotes the *asymptotic writhe* of C; to calculate it, take the intersection of C with $\{s\}\times Y$ where $s >> 0$ to obtain a disjoint union of closed braids around the Reeb orbits α_i, use the trivializations τ to draw these braids in $\mathbb{R}^3$, and count the crossings with appropriate signs; then subtract the corresponding count for $s << 0$.

Next we need a new ingredient, which is the following bound on the asymptotic writhe:

$$w_\tau(C) \le \mathrm{CZ}_\tau(\alpha) - \mathrm{CZ}_\tau(\beta) - \mathrm{CZ}^0_\tau(C). \tag{1.8}$$

Here

$$\mathrm{CZ}_\tau(\alpha) := \sum_i \sum_{k=1}^{m_i} \mathrm{CZ}_\tau(\alpha_i^k),$$

where $\mathrm{CZ}_\tau(\gamma^k)$ denotes the Conley-Zehnder index with respect to τ of the kth iterate of γ. To prove the writhe bound (1.8), one first needs to understand the structure of the braids that can arise from the ends of a holomorphic curve; roughly speaking these are iterated nested cablings of torus braids, with certain bounds on the winding numbers. One then needs some combinatorics to bound the writhes of these braids in terms of the Conley-Zehnder indices. The writhe bound was proved in an analytically simpler situation in [16]; the asymptotic analysis needed to carry over the proof to the present setting was carried out by Siefring [37], and an updated proof is given in [17].

Finally, by analogy with (1.4), define the *ECH index*

$$I(C) := c_1(C,\tau) + Q_\tau(C) + \mathrm{CZ}_\tau(\alpha) - \mathrm{CZ}_\tau(\beta). \tag{1.9}$$

One can check that this formula, like the formulas above, does not depend on

the choice of trivialization τ. It now follows from (1.6), (1.7), (1.8), and (1.9) that the *index inequality*

$$\operatorname{ind}(C) \leq I(C) \tag{1.10}$$

holds, with equality only if C is embedded.

Recall that we have been assuming in the preceding discussion that C is not multiply covered. Without this assumption, one still has the following proposition, which describes the $I = 1$ curves which the ECH differential counts.

Proposition 1.2. *[20, Cor. 11.5] Suppose J is generic, and let C be any J-holomorphic curve in $\mathcal{M}^J(\alpha, \beta)$, possibly multiply covered. Then:*

(a) $I(C) \geq 0$, with equality if and only if C is $\mathbb{R}$-invariant (as a current).

(b) If $I(C) = 1$, then $C = C_0 \sqcup C_1$, where $I(C_0) = 0$, and C_1 is embedded and has $\operatorname{ind}(C_1) = I(C_1) = 1$.

It may be illuminating to recall the proof here. As a current, C consists of distinct, irreducible, non-multiply-covered holomorphic curves $C_1, \ldots, C_k$, covered with positive integer multiplicities $d_1, \ldots, d_k$. For simplicity let us restrict attention to the case when none of the curves C_i is an $\mathbb{R}$-invariant cylinder. Let C' be the holomorphic curve consisting of the union, over $i = 1, \ldots, k$, of d_i different $\mathbb{R}$-translates of C_i. We then have

$$\sum_{i=1}^{k} d_i \operatorname{ind}(C_i) = \operatorname{ind}(C') \leq I(C') = I(C), \tag{1.11}$$

with equality only if the holomorphic curves C_i are embedded and disjoint. Here the equality on the left holds because the Fredholom index is additive under unions, the inequality in the middle is the index inequality (1.10) applied to the non-multiply-covered curve C', and the equality on the right holds because the ECH index of a holomorphic curve depends only on its relative homology class. Now since J is generic, and since we made the simplifying assumption that C_i is not $\mathbb{R}$-invariant, we have $\operatorname{ind}(C_i) > 0$ for each i. We can then read off the conclusions of the proposition in this case from the inequality (1.11).

1.4.3. Grading. The ECH index is also used to define the relative grading on the ECH chain complex, as follows. As noted above, the ECH index $I(C)$ depends only on the relative homology class of C, and indeed it makes perfect sense to define $I(Z)$ as in (1.9) where Z is any relative homology class of 2-chain in Y (not necessarily arising from a J-holomorphic curve) with $\partial Z = \sum_i m_i \alpha_i - \sum_j n_j \beta_j$. If Z' is another such relative homology class, then $Z - Z' \in H_2(Y)$, and one has the *index ambiguity formula* [16, Prop. 1.6(d)]

$$I(Z) - I(Z') = \langle c_1(\xi) + 2\,\mathrm{PD}(\Gamma), Z - Z' \rangle.$$

We now define the grading difference between two generators α and β to be the class of $I(Z)$ in $\mathbb{Z}/d$, where Z is any relative homology class as above. The

index ambiguity formula shows that this is well defined, and by definition the differential decreases the relative grading by 1.

1.4.4. Incoming and outgoing partitions and admissibility. We now make some technical remarks which will not be needed in the rest of this article, but which address some frequently asked questions regarding the definition of ECH.

The first remark is that embeddedness of C is not sufficient for equality to hold in (1.10), unless all of the multiplicities m_i and n_j equal 1. A curve C in $\mathcal{M}^J(\alpha,\beta)$ has positive ends at covers of α_i with some multiplicities $q_{i,k}$ whose sum is $\sum_k q_{i,k} = m_i$. If equality holds in (1.10), then the unordered list of multiplicities $(q_{i,1}, q_{i,2}, \ldots)$ is uniquely determined by α_i and m_i, and is called the "outgoing partition" $P^{\mathrm{out}}_{\alpha_i}(m_i)$. Likewise the covering multiplicities associated to the ends of C at covers of β_j must comprise a partition called the "incoming partition" $P^{\mathrm{in}}_{\beta_j}(n_j)$. See e.g. [17, §4] for details. To give the simplest example, if γ is an embedded elliptic Reeb orbit such that the linearized Reeb flow around γ with respect to some trivialization rotates by an angle in the interval $(0,\pi)$, then $P^{\mathrm{out}}_\gamma(2) = (1,1)$, while $P^{\mathrm{in}}_\gamma(2) = (2)$.

In general, if γ is an embedded elliptic Reeb orbit and if $m > 1$, then the incoming and outgoing partitions $P^{\mathrm{in}}_\gamma(m)$ and $P^{\mathrm{out}}_\gamma(m)$ are always different. This fact makes the proof that $\partial^2 = 0$ quite nontrivial.

On the other hand, suppose γ is a hyperbolic embedded Reeb orbit. If the linearized return map has positive eigenvalues then

$$P^{\mathrm{in}}_\gamma(m) = P^{\mathrm{out}}_\gamma(m) = (1,\ldots,1). \tag{1.12}$$

If the linearized return map has negative eigenvalues then

$$P^{\mathrm{in}}_\gamma(m) = P^{\mathrm{out}}_\gamma(m) = \begin{cases} (2,\ldots,2), & m \text{ even}, \\ (2,\ldots,2,1), & m \text{ odd}. \end{cases} \tag{1.13}$$

This is one reason why the generators of the ECH chain complex in §1.3 are required to be *admissible* orbit sets: one can show using (1.12) and (1.13) that if one tries to glue two $I = 1$ holomorphic curves along an inadmissible orbit set, then there are an even number of ways to glue, which by [5] count with cancelling signs. Thus one must disallow inadmissible orbit sets in order to obtain $\partial^2 = 0$. A similar issue arises in the definition of symplectic field theory [11], where "bad" Reeb orbits must be discarded.

1.5. Example: the ECH of an ellipsoid.

We now illustrate the above definitions with what is probably the simplest example of ECH. Consider $\mathbb{C}^2 = \mathbb{R}^4$ with coordinates $z_j = x_j + iy_j$ for $j = 1,2$. Let a, b be positive real numbers with a/b irrational, and consider the ellipsoid

$$E(a,b) := \left\{ (z_1,z_2) \in \mathbb{C}^2 \;\middle|\; \frac{\pi|z_1|^2}{a} + \frac{\pi|z_2|^2}{b} \le 1 \right\}. \tag{1.14}$$

We now compute the embedded contact homology of $Y = \partial E(a,b)$, with the contact form

$$\lambda := \frac{1}{2}\sum_{j=1}^{2}(x_j dy_j - y_j dx_j) \tag{1.15}$$

(and of course with $\Gamma = 0$).

The Reeb vector field on Y is given by

$$R = \frac{2\pi}{a}\frac{\partial}{\partial\theta_1} + \frac{2\pi}{b}\frac{\partial}{\partial\theta_2}$$

where $\partial/\partial\theta_j := x_j\partial_{y_j} - y_j\partial_{x_j}$. Since a/b is irrational, it follows that there are just two embedded Reeb orbits γ_1 and γ_2, given by the circles where $z_2 = 0$ and $z_1 = 0$ respectively. These Reeb orbits, as well as their iterates, are nondegenerate and elliptic. Indeed there is a natural trivialization τ of ξ over each γ_i induced by an embedded disk bounded by γ_i. With respect to this trivialization, the linearized Reeb flow around γ_1 is rotation by angle $2\pi a/b$, while the linearized Reeb flow around γ_2 is rotation by angle $2\pi b/a$.

The generators of the ECH chain complex have the form $\alpha = \gamma_1^{m_1}\gamma_2^{m_2}$ where m_1, m_2 are nonnegative integers. We now compute the grading. The relative $\mathbb{Z}$-grading has a distinguished refinement to an absolute grading in which the empty set of Reeb orbits (given by $m_1 = m_2 = 0$ above) has grading 0. An arbitrary generator α as above then has grading

$$I(\alpha) = c_1(\alpha,\tau) + Q_\tau(\alpha) + \mathrm{CZ}_\tau(\alpha),$$

where $c_1(\alpha,\tau)$ denotes the relative first Chern class of ξ over a surface bounded by α, and $Q_\tau(\alpha)$ denotes the relative intersection pairing of such a surface. Computing using the above trivialization τ, one finds, see [23, §4.2], that

$$\begin{aligned} c_1(\alpha,\tau) &= m_1 + m_2, \\ Q_\tau(\alpha) &= 2m_1m_2, \\ \mathrm{CZ}_\tau(\alpha) &= \sum_{k=1}^{m_1}(2\lfloor ka/b\rfloor + 1) + \sum_{k=1}^{m_2}(2\lfloor kb/a\rfloor + 1). \end{aligned}$$

Therefore

$$I(\alpha) = 2\left(m_1 + m_2 + m_1m_2 + \sum_{k=1}^{m_1}\lfloor ka/b\rfloor + \sum_{k=1}^{m_2}\lfloor kb/a\rfloor\right). \tag{1.16}$$

In particular, all generators have even grading, so the differential vanishes, and to determine the homology we just have to count the number of generators with each grading.

Now if the ECH of $\partial E(a,b)$ is to agree with $\widehat{HM}^{-*}$ and HF_*^+ of S^3, then we should get

$$ECH_*(\partial E(a,b),\lambda,0) \simeq \begin{cases} \mathbb{Z}, & * = 0,2,4,\ldots, \\ 0, & \text{otherwise.} \end{cases} \tag{1.17}$$

It is perhaps not immediately obvious how to deduce this from (1.16). The trick is to interpret the right hand side of (1.16) as a count of lattice points as follows. Let T denote the triangle in $\mathbb{R}^2$ bounded by the coordinate axes, together with the line L through the point (m_1, m_2) of slope $-a/b$. Then we observe that

$$I(\alpha) = 2\left(\left|T \cap \mathbb{Z}^2\right| - 1\right).$$

Now if one moves the line L up and to the right, keeping its slope fixed, then one hits all of the lattice points in the nonnegative quadrant in succession, each time increasing the number of lattice points in the triangle T by 1. It follows that the ECH chain complex has one generator in each nonnegative even grading, so (1.17) holds.

Usually direct calculations of ECH are not so easy because there are more Reeb orbits, and one has to understand the holomorphic curves. But for certain simple contact manifolds this is possible; for example the ECH of standard contact forms on T^3 is computed in [20], and these calculations are generalized to T^2-bundles over S^1 in [28].

Remark. For some mysterious reason, lattice point counts such as the one in equation (1.16) arise repeatedly in ECH in different contexts. For example one lattice point count comes up in the combinatorial part of the proof of the writhe bound (1.8), and in determining the "partition conditions" in §1.4.4, see [17, §4.6]. Another lattice point count appears in the combinatorial description of the ECH chain complex for T^3 in [20, §1.3].

1.6. Some additional structures on ECH.

ECH has various additional structures on it. We now describe those structures that are relevant elsewhere in this article.

1.6.1. The U map. On the ECH chain complex there is a degree -2 chain map

$$U : C_*(Y, \lambda, \Gamma) \longrightarrow C_{*-2}(Y, \lambda, \Gamma),$$

see e.g. [23, §2.5]. This is defined similarly to the differential ∂, but instead of counting $I = 1$ curves modulo translation, one counts $I = 2$ curves that are required to pass through a fixed, generic point $z \in \mathbb{R} \times Y$. This induces a well-defined map on homology

$$U : ECH_*(Y, \lambda, \Gamma) \longrightarrow ECH_{*-2}(Y, \lambda, \Gamma).$$

Taubes [44] has shown that this map agrees with an analogous map on $\widehat{HM}^{-*}$, and it conjecturally agrees with the U map on HF^+_*. The U map plays a crucial role in the applications to generalizations of the Weinstein conjecture discussed in §2.1 below.

1.6.2. Filtered ECH. If $\alpha = \{(\alpha_i, m_i)\}$ is a generator of the ECH chain complex, define its *symplectic action*

$$\mathcal{A}(\alpha) := \sum_i m_i \int_{\alpha_i} \lambda.$$

It follows from Stokes's theorem and the conditions on the almost complex structure J that the differential ∂ decreases the symplectic action, i.e. if $\langle \partial\alpha, \beta\rangle \neq 0$ then $\mathcal{A}(\alpha) > \mathcal{A}(\beta)$. Given $L \in \mathbb{R}$, we then define $ECH^L(Y, \lambda, \Gamma)$ to be the homology of the subcomplex of $C_*(Y, \lambda, \Gamma)$ spanned by generators with symplectic action less than L. We call this *filtered ECH*; it is shown in [24] that this does not depend on the choice of almost complex structure J. However, unlike the usual ECH, filtered ECH is not invariant under deformation of the contact form; see §2.3 for some examples. Filtered ECH has no obvious direct counterpart in Seiberg-Witten or Heegaard Floer homology, but it plays an important role in the applications in §2.2 and §2.3 below.

1.6.3. Cobordism maps. Let (Y_+, λ_+) and (Y_-, λ_-) be closed oriented 3-manifolds with nondegenerate contact forms. An *exact symplectic cobordism* from (Y_+, λ_+) to (Y_-, λ_-) is a compact symplectic 4-manifold (X, ω) with boundary $\partial X = Y_+ - Y_-$, such that there exists a 1-form λ on X with $d\lambda = \omega$ on X and $\lambda|_{Y_\pm} = \lambda_\pm$. It is shown in [24] that an exact symplectic cobordism as above induces maps on filtered ECH,

$$\Phi^L(X, \omega) : ECH^L(Y_+, \lambda_+; \mathbb{Z}/2) \longrightarrow ECH^L(Y_-, \lambda_-; \mathbb{Z}/2),$$

satisfying various axioms. Here $ECH(Y_\pm, \lambda_\pm; \mathbb{Z}/2)$ denotes ECH with $\mathbb{Z}/2$ coefficients, summed over all $\Gamma \in H_1(Y)$, and regarded as an ungraded $\mathbb{Z}/2$-module. One axiom is that if $L < L'$ then the diagram

$$\begin{array}{ccc} ECH^L(Y_+, \lambda_+; \mathbb{Z}/2) & \xrightarrow{\Phi^L(X,\omega)} & ECH^L(Y_-, \lambda_-; \mathbb{Z}/2) \\ \downarrow & & \downarrow \\ ECH^{L'}(Y_+, \lambda_+; \mathbb{Z}/2) & \xrightarrow{\Phi^{L'}(X,\omega)} & ECH^{L'}(Y_-, \lambda_-; \mathbb{Z}/2) \end{array}$$

commutes, where the vertical arrows are induced by inclusion of chain complexes. Thus the direct limit

$$\Phi(X, \omega) := \lim_{L\to\infty} \Phi^L(X, \omega) : ECH(Y_+, \lambda_+; \mathbb{Z}/2) \longrightarrow ECH(Y_-, \lambda_-; \mathbb{Z}/2) \tag{1.18}$$

is well-defined. Another axiom is that this direct limit agrees with the map $\widehat{HM}^*(Y_+; \mathbb{Z}/2) \to \widehat{HM}^*(Y_-; \mathbb{Z}/2)$ on Seiberg-Witten Floer cohomology induced by X, under the isomorphism (1.1). Here we are considering Seiberg-

Witten Floer cohomology with $\mathbb{Z}/2$ coefficients, summed over all spin-c structures.

Remark. The cobordism maps $\Phi^L(X,\omega)$ are defined in [24] using Seiberg-Witten theory and parts of the isomorphism (1.1). It would be natural to try to give an alternate, more direct definition of the cobordism maps $\Phi^L(X,\omega)$ by counting $I=0$ holomorphic curves in the "completion" of X obtained by attaching symplectization ends. Note that by Stokes's theorem and the exactness of the cobordism, such a map would automatically respect the symplectic action filtrations. Moreover, one of the axioms we did not state is that $\Phi^L(X,\omega)$ is induced by a (noncanonical) chain map whose components are nonzero only in the presence of appropriate (possibly broken) holomorphic curves. However there are technical difficulties with defining cobordism maps by counting holomorphic curves, because the compactified $I=0$ moduli spaces can include broken holomorphic curves which contain multiply covered components with negative ECH index. (There is no analogue in this setting of Proposition 1.2, whose proof made essential use of the $\mathbb{R}$-invariance of J.) Examples show that such broken curves must sometimes make contributions to the cobordism map, but it is not known how to define the contribution in general. Fortunately, the Seiberg-Witten definition of $\Phi^L(X,\omega)$ is sufficient for the applications considered here.

1.6.4. The contact invariant. The *empty set* is a legitimate generator of the ECH chain complex. By the discussion in §1.6.2 it is a cycle, and we denote its homology class by

$$c(\xi) \in ECH_0(Y,\lambda,0).$$

This depends only on the contact structure, although not just on the 3-manifold Y. Indeed the cobordism maps in §1.6.3 can be used to show that $c(\xi)$ is nonzero if there is an exact symplectic cobordism from (Y,ξ) to the empty set, e.g. if (Y,ξ) is Stein fillable. On the other hand the argument in the appendix to [47] implies that $c(\xi)=0$ if ξ is overtwisted. Some new families of contact 3-manifolds with vanishing ECH contact invariant are introduced by Wendl [46]. It is shown by Taubes [44] that $c(\xi)$ agrees with an analogous contact invariant in Seiberg-Witten Floer cohomology, and both conjecturally agree with the contact invariant in Heegaard Floer homology [36].

1.7. Analogues of ECH in other contexts.

One can also define a version of ECH for sutured 3-manifolds with contact structures adapted to the sutures, see [7]. This conjecturally agrees with the sutured Floer homology of Juhász [25] and with the sutured version of Seiberg-Witten Floer homology defined by Kronheimer-Mrowka [26].

There is also an analogue of ECH, called "periodic Floer homology", for mapping tori of area-preserving surface diffeomorphisms, see e.g. [19, 30].

We remark that unlike SFT, which is defined for contact manifolds of any odd dimension, no analogue of ECH is currently known for contact manifolds

of dimension greater than three. In higher dimensions one expects that if J is generic then all non-multiply-covered J-holomorphic curves are embedded, see [34]. In addition no good analogue of Seiberg-Witten theory is known in higher dimensions.

2. Applications

Currently all applications of ECH make use of Taubes's isomorphism (1.1), together with known properties of Seiberg-Witten Floer homology, to deduce certain properties of ECH which then have implications for contact geometry. It is an interesting open problem to establish the relevant properties of ECH without using Seiberg-Witten theory.

2.1. Generalizations of the Weinstein conjecture. The *Weinstein conjecture* in three dimensions asserts that for any contact form λ on a closed oriented 3-manifold Y, there exists a Reeb orbit. Many cases of this were proved by various authors, see e.g. [13, 2, 8], and the general case was proved by Taubes [40]. Indeed the three-dimensional Weinstein conjecture follows immediately from the isomorphism (1.1), together with a theorem of Kronheimer-Mrowka [26, Cor. 35.1.4] asserting that $\widehat{HM}^*$ is always infinitely generated for torsion spin-c structures. The reason is that if there were no Reeb orbit, then the ECH would have just one generator: the empty set of Reeb orbits. However to prove the Weinstein conjecture one does not need to use the full force of the isomorphism (1.1); one just needs a way of passing from Seiberg-Witten Floer generators to ECH generators, which is what [40] establishes.

In [23] we make heavier use of the isomorphism (1.1) to prove some stronger results. For example:

Theorem 2.1. *Let λ be a nondegenerate contact form on a closed oriented connected 3-manifold Y such that all Reeb orbits are elliptic. Then there are exactly two embedded Reeb orbits, Y is a sphere or a lens space, and the two embedded Reeb orbits are the core circles of a genus* 1 *Heegaard splitting of Y.*

The idea of the proof is as follows. Since all Reeb orbits are elliptic, a general property of the ECH index [16, Prop. 1.6(c)] implies that all ECH generators have even grading, so the ECH differential vanishes. Since $\widehat{HM}^*$ is nonvanishing for only finitely many spin-c structures, it follows that all Reeb orbits represent torsion homology classes. Estimating the number of ECH generators in a given index range then shows that there are exactly two embedded Reeb orbits; otherwise there would be either too few or too many generators to be consistent with the linear growth rate of $\widehat{HM}^*$. Indeed, known properties of $\widehat{HM}^*$ imply that the U map is an isomorphism when the grading is sufficiently large. This provides a large supply of $I = 2$ holomorphic curves in $\mathbb{R} \times Y$. By careful use of

the adjunction formula (1.7) one can show that at least one of these holomorphic curves includes a non-$\mathbb{R}$-invariant holomorphic cylinder. By adapting ideas from [14], one can show that this holomorphic cylinder projects to an embedded surface in Y which generates a foliation by cylinders of the complement in Y of the Reeb orbits. This foliation then gives rise to the desired Heegaard splitting.

Theorem 2.1 is used in [23] to extend the Weinstein conjecture to "stable Hamiltonian structures" (a certain generalization of contact forms) on 3-manifolds that are not torus bundles over S^1.

In addition, a slight refinement of the proof of Theorem 2.1 in [23] establishes:

Theorem 2.2. *Let Y be a closed oriented 3-manifold with a nondegenerate contact form λ. If Y is not a sphere or a lens space, then there are at least* 3 *embedded Reeb orbits.*

In fact, examples of contact forms with only finitely many embedded Reeb orbits are hard to come by, and to our knowledge the following question is open:

Question 2.3. Is there any example of a contact form on a closed connected oriented 3-manifold with only finitely many embedded Reeb orbits, other than contact forms on S^3 and lens spaces with exactly two embedded Reeb orbits?

It is shown in [15] that for a large class of contact forms on S^3 there are either two or infinitely many embedded Reeb orbits. It is shown in [8], using linearized contact homology, that many contact structures on 3-manifolds (namely those supported by an open book decomposition with pseudo-Anosov monodromy satisfying a certain inequality) have the property that for any contact form, there are infinitely many free homotopy classes of loops that must contain an embedded Reeb orbit.

2.2. The Arnold chord conjecture.

A *Legendrian knot* in a contact 3-manifold (Y, λ) is a knot $K \subset Y$ such that $TK \subset \xi|_K$. A *Reeb chord* of K is a Reeb trajectory starting and ending on K, i.e. a path $\gamma : [0, T] \to Y$ for some $T > 0$ such that $\gamma'(t) = R(\gamma(t))$ and $\gamma(0), \gamma(T) \in K$. The following theorem, proved in [24], is a version of the Arnold chord conjecture [3]. (This was previously known in some cases from [1, 33].)

Theorem 2.4. *Let Y_0 be a closed oriented 3-manifold with a contact form λ_0, and let K be a Legendrian knot in (Y_0, λ_0). Then K has a Reeb chord.*

The idea of the proof is as follows. Following Weinstein [45], one can perform a "Legendrian surgery" along K to obtain a new contact manifold (Y_1, λ_1), together with an exact symplectic cobordism (X, ω) from (Y_1, λ_1) to (Y_0, λ_0). If K has no Reeb chord, and if λ_0 is nondegenerate, then one can carry out the Legendrian surgery construction so that λ_1 is nondegenerate and, up to a given action, the Reeb orbits of λ_1 are the same as those of λ_0. Using this observation,

one can show that if K has no Reeb chord and if λ_0 is nondegenerate, then the cobordism map

$$\Phi(X,\omega) : ECH(Y_1,\lambda_1;\mathbb{Z}/2) \longrightarrow ECH(Y_0,\lambda_0;\mathbb{Z}/2) \tag{2.1}$$

from §1.6.3 is an isomorphism. Note that this is what one would expect by analogy with a very special case of the work of Bourgeois-Ekholm-Eliashberg [4], which studies the behavior of linearized contact homology under Legendrian surgery, possibly in the presence of Reeb chords.

But the map (2.1) cannot be an isomorphism. The reason is that as shown in [26, Thm. 42.2.1], the corresponding map on Seiberg-Witten Floer cohomology fits into an exact triangle

$$\cdots \to \widehat{HM}^*(Y_0;\mathbb{Z}/2) \to \widehat{HM}^*(Y_1;\mathbb{Z}/2) \to \widehat{HM}^*(Y_2;\mathbb{Z}/2) \to \widehat{HM}^*(Y_0;\mathbb{Z}/2) \to \cdots$$

where Y_2 is obtained from Y_0 by a different surgery along K. However, as noted before, Kronheimer-Mrowka showed that $\widehat{HM}^*(Y_2;\mathbb{Z}/2)$ is infinitely generated. This contradiction proves the chord conjecture when λ_0 is nondegenerate.

To deal with the case where λ_0 is degenerate, one can use filtered ECH to show that in the nondegenerate case, there exists a Reeb chord with an upper bound on the length, in terms of a quantitative measure of the failure of the map (2.1) to be an isomorphism. For example, if λ_0 is nondegenerate and if (2.1) is not surjective, then there exists a Reeb chord of action at most A, where A is the infimum over $L \in \mathbb{R}$ such that the image of $ECH^L(Y_0,\lambda_0;\mathbb{Z}/2)$ in $ECH(Y_0,\lambda_0;\mathbb{Z}/2)$ is not contained in the image of the map (2.1). One can show that this upper bound on the length of a Reeb chord is suitably "continuous" as one changes the contact form. A compactness argument then finds a Reeb chord in the degenerate case.

2.3. Obstructions to symplectic embeddings. ECH also gives obstructions to symplectically embedding one compact symplectic 4-manifold with boundary into another. We now explain how this works in the case of ellipsoids as in (1.14), with the standard symplectic form $\omega = \sum_{j=1}^2 dx_j dy_j$ on $\mathbb{R}^4$.

Given positive real numbers a, b, and given a positive integer k, define $(a,b)_k$ to be the kth smallest entry in the array $(ma+nb)_{m,n\in\mathbb{N}}$. Here in the definition of "k^{th} smallest" we count with repetitions. For example if $a = b$ then

$$((a,a)_1,(a,a)_2,\ldots) = (0,a,a,2a,2a,2a,3a,3a,3a,3a,\ldots).$$

We then have:

Theorem 2.5. *If there is a symplectic embedding of $E(a,b)$ into $E(c,d)$, then*

$$(a,b)_k \le (c,d)_k \tag{2.2}$$

for all positive integers k.

To prove this, one can assume without loss of generality that a/b and c/d are irrational and that there is a symplectic embedding $\varphi : E(a,b) \to \mathrm{int}(E(c,d))$. Now consider the 4-manifold $X = E(c,d) \setminus \mathrm{int}(\varphi(E(a,b)))$. One can show that X defines an exact symplectic cobordism from $\partial E(c,d)$ to $\partial E(a,b)$, where the latter two 3-manifolds are endowed with the contact form (1.15). Since X is diffeomorphic to the product $[0,1] \times S^3$, the induced map from the Seiberg-Witten Floer cohomology of $\partial E(c,d)$ to that of $\partial E(a,b)$ must be an isomorphism. Recall from (1.18) that this map is the direct limit of maps on filtered ECH. Since the ECH differentials vanish, it follows that for each $L \in \mathbb{R}$, the number of ECH generators of $\partial E(c,d)$ with action less than L does not exceed the number of ECH generators of $\partial E(a,b)$ with action less than L. Since the embedded Reeb orbits in $\partial E(a,b)$ have action a and b, and the embedded Reeb orbits in $\partial E(c,d)$ have action c and d, it follows that

$$\left|\left\{(m,n) \in \mathbb{N}^2 \mid cm + dn < L\right\}\right| \leq \left|\left\{(m,n) \in \mathbb{N}^2 \mid am + bn < L\right\}\right|. \tag{2.3}$$

The statement that the above inequality holds for all $L \in \mathbb{R}$ is equivalent to (2.2).

For example, if L is large with respect to a, b, c, d, then the inequality (2.3) implies that

$$\frac{L^2}{2cd} \leq \frac{L^2}{2ab} + O(L).$$

We conclude that $ab \leq cd$, which is simply the condition that the volume of $E(a,b)$ is less than or equal to the volume of $E(c,d)$, which of course is necessary for the existence of a symplectic embedding. But taking suitable small L often gives stronger conditions.

The amazing fact is that, at least for the problem of embedding ellipsoids into balls, the obstruction in Theorem 2.5 is sharp. Namely, for each positive real number a, define $f(a)$ to be the infimum over all $c \in \mathbb{R}$ such that $E(a,1)$ symplectically embeds into the ball $E(c,c)$. It follows from Theorem 2.5 that

$$f(a) \geq \sup_{k=2,3,\ldots} \frac{(a,1)_k}{(1,1)_k}. \tag{2.4}$$

On the other hand, McDuff-Schlenk [31] computed the function f explicitly, obtaining a complicated answer involving Fibonacci numbers. Using the result of this calculation, they checked that the opposite inequality in (2.4) holds.

Question 2.6. Is there a direct explanation for this? Does this generalize? For example, does $E(a,b)$ symplectically embed into $E(c+\varepsilon, d+\varepsilon)$ for all $\varepsilon > 0$ if $(a,b)_k \leq (c,d)_k$ for all positive integers k?

By more involved calculations, one can use ECH to find explicit (but subtle, number-theoretic) obstructions to symplectic embeddings involving other simple shapes such as four-dimensional polydisks. A systematic treatment of the symplectic embedding obstructions arising from ECH is given in [18].

References

[1] C. Abbas, *The chord problem and a new method of filling by pseudoholomorphic curves*, Int. Math. Res. Not. **2004**, 913–927.

[2] C. Abbas, K. Cieliebak, and H. Hofer, *The Weinstein conjecture for planar contact structures in dimension three*, Comment. Math. Helv. **80** (2005), 771–793.

[3] V. I. Arnold, *The first steps of symplectic topology*, Russian Math. Surveys **41** (1986), no. 6, 1–21.

[4] F. Bourgeois, T. Ekholm, and Y. Eliashberg, *Effect of Legendrian surgery*, arXiv:0911.0026.

[5] F. Bourgeois and K. Mohnke, *Coherent orientations in symplectic field theory*, Math. Z. **248** (2004), 123–146.

[6] F. Bourgeois and K. Niederkrüger, *Towards a good definition of algebraically overtwisted*, Expositiones Mathematicae **28** (2010), 85-100.

[7] V. Colin, P. Ghiggini, K. Honda, and M. Hutchings, *Sutures and contact homology I*, in preparation.

[8] V. Colin and K. Honda, *Reeb vector fields and open book decompositions*, arXiv:0809.5088.

[9] D. Dragnev, *Fredholm theory and transversality for noncompact pseudoholomorphic maps in symplectizations*, Comm. Pure Appl. Math. **57** (2004), 726–763.

[10] Y. Eliashberg, *Classification of overtwisted contact structures on 3-manifolds*, Invent. Math. **98** (1989), 623–637,

[11] Y. Eliashberg, A. Givental, and H. Hofer, *Introduction to symplectic field theory*, GAFA 2000, Special Volume, Part II, 560–673.

[12] H. Geiges, *An introduction to contact topology*, Cambridge University Press, 2008.

[13] H. Hofer, *Pseudoholomorphic curves in symplectizations with applications to the Weinstein conjecture in dimension three*, Invent. Math. **114** (1993), 515–563.

[14] H. Hofer, K. Wysocki, and E. Zehnder, *Properties of pseudoholomorphic curves in symplectizations. III. Fredholm theory*, Topics in nonlinear analysis, 381–475, Progr. Nonlinear Differential Equations Appl., **35**, Basel: Birkhäuser 1999.

[15] H. Hofer, K. Wysocki, and E. Zehnder, *Finite energy foliations of tight three-spheres and Hamiltonian dynamics*, Ann. of Math. **157** (2003), 125–255.

[16] M. Hutchings, *An index inequality for embedded pseudoholomorphic curves in symplectizations*, J. Eur. Math. Soc. **4** (2002), 313–361.

[17] M. Hutchings, *The embedded contact homology index revisited*, New perspectives and challenges in symplectic field theory, 263–297, CRM Proc. Lecture Notes, 49, AMS, 2009.

[18] M. Hutchings, *Quantitative embedded contact homology*, in preparation.

[19] M. Hutchings and M. Sullivan, *The periodic Floer homology of a Dehn twist*, Alg. and Geom. Topol. **5** (2005), 301–354.

[20] M. Hutchings and M. Sullivan, *Rounding corners of polygons and the embedded contact homology of T^3*, Geometry and Topology **10** (2006), 169–266.

[21] M. Hutchings and C. H. Taubes, *Gluing pseudoholomorphic curves along branched covered cylinders I*, J. Symplectic Geom. **5** (2007), 43–137.

[22] M. Hutchings and C. H. Taubes, *Gluing pseudoholomorphic curves along branched covered cylinders II*, J. Symplectic Geom. **7** (2009), 29–133.

[23] M. Hutchings and C. H. Taubes, *The Weinstein conjecture for stable Hamiltonian structures*, Geometry and Topology **13** (2009), 901–941.

[24] M. Hutchings and C. H. Taubes, *Proof of the Arnold chord conjecture in three dimensions*, in preparation.

[25] A. Juhász, *Holomorphic disks and sutured manifolds*, Alg. Geom. Topol. **6** (2006), 1429–1457.

[26] P.B. Kronheimer and T.S. Mrowka, *Monopoles and three-manifolds*, Cambridge University Press, 2008.

[27] P.B. Kronheimer and T.S. Mrowka, *Knots, sutures and excision*, arXiv:0807.4891.

[28] E. Lebow, *Embedded contact homology of 2-torus bundles over the circle*, UC Berkeley thesis, 2007.

[29] Y-J. Lee, *Heegaard splittings and Seiberg-Witten monopoles*, Geometry and topology of manifolds, 173–202, Fields Inst. Commun. **47**, AMS, 2005.

[30] Y-J. Lee and C.H. Taubes, *Periodic Floer homology and Seiberg-Witten Floer cohomology*, arXiv:0906.0383.

[31] D. McDuff and F. Schlenk, *The embedding capacity of 4-dimensional symplectic ellipsoids*, arXiv:0912.0532, v2.

[32] M. Micallef and B. White, *The structure of branch points in minimal surfaces and in pseudoholomorphic curves*, Ann. Math. **141** (1995), 35–85.

[33] K. Mohnke, *Holomorphic disks and the chord conjecture*, Ann. of Math. **154** (2001), 219–222.

[34] Y-G. Oh and K. Zhu, *Embedding property of J-holomorphic curves in Calabi-Yau manifolds for generic J*, arXiv:0805.3581.

[35] P. Ozsváth and Z. Szabó, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. **159** (2004), 1027–1158.

[36] P. Ozsváth and Z. Szabó, *Heegaard Floer homology and contact structures*. Duke Math. J. **129** (2005), 39–61.

[37] R. Siefring, *The relative asymptotic behavior of pseudoholomorphic half-cylinders*, Comm. Pure Appl. Math. **61** (2008), 1631–1684.

[38] C.H. Taubes, *Seiberg-Witten and Gromov invariants for symplectic 4-manifolds*, First International Press Lecture Series 2, International Press, 2000.

[39] C.H. Taubes, *Counting pseudo-holomorphic submanifolds in dimension* 4, in loc. cit.

[40] C.H. Taubes, *The Seiberg-Witten equations and the Weinstein conjecture*, Geom. Topol. **11** (2007), 2117–2202.

[41] C. H. Taubes, *The Seiberg-Witten equations and the Weinstein conjecture. II. More closed integral curves of the Reeb vector field*, Geom. Topol. **13** (2009), 1337–1417.

[42] C. H. Taubes, *Embedded contact homology and Seiberg-Witten Floer homology I*, arXiv:0811.3985.

[43] C. H. Taubes, *Embedded contact homology and Seiberg-Witten Floer homology II–IV*, preprints, 2008.

[44] C. H. Taubes, *Embedded contact homology and Seiberg-Witten Floer homology V*, preprint, 2008.

[45] A. Weinstein, *Contact surgery and symplectic handlebodies*, Hokkaido Math. J. **20** (1991), 241–251.

[46] C. Wendl, *Holomorphic curves in blown up open books*, arXiv:1001.4109.

[47] M-L. Yau, *Vanishing of the contact homology of overtwisted contact 3-manifolds*, Bull. Inst. Math. Acad. Sin. **1** (2006), 211–229.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Finite Covering Spaces of 3-manifolds

Marc Lackenby*

Abstract

Following Perelman's solution to the Geometrisation Conjecture, a 'generic' closed 3-manifold is known to admit a hyperbolic structure. However, our understanding of closed hyperbolic 3-manifolds is far from complete. In particular, the notorious Virtually Haken Conjecture remains unresolved. This proposes that every closed hyperbolic 3-manifold has a finite cover that contains a closed embedded orientable π_1-injective surface with positive genus.

I will give a survey on the progress towards this conjecture and its variants. Along the way, I will address other interesting questions, including: What are the main types of finite covering space of a hyperbolic 3-manifold? How many are there, as a function of the covering degree? What geometric, topological and algebraic properties do they have? I will show how an understanding of various geometric and topological invariants (such as the first eigenvalue of the Laplacian, the rank of mod p homology and the Heegaard genus) can be used to deduce the existence of π_1-injective surfaces, and more.

Mathematics Subject Classification (2010). Primary 57N10, 57M10; Secondary 57M07.

Keywords. Covering space; hyperbolic 3-manifold; incompressible surface; subgroup growth; Cheeger constant; Heegaard splitting; Property (τ)

Acknowledgement: Partially supported by the Leverhulme trust.

1. Introduction

In recent years, there have been several huge leaps forward in 3-manifold theory. Most notably, Perelman [50, 51, 52] has proved Thurston's Geometrisation Conjecture [62], and, as a consequence, a 'generic' closed orientable 3-manifold is known to admit a hyperbolic structure. However, our understanding of closed hyperbolic 3-manifolds is far from complete. In particular, finite covers of hyperbolic 3-manifolds remain rather mysterious. Here, the primary goal is the

*Mathematical Institute, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, United Kingdom. E-mail: lackenby@maths.ox.ac.uk.

search for closed embedded orientable π_1-injective surfaces (which are known as *incompressible*). The following conjectures remain notoriously unresolved. Does every closed hyperbolic 3-manifold have a finite cover that

1. is *Haken*, in other words, contains a closed embedded orientable incompressible surface (other than a 2-sphere)?
2. has positive first Betti number?
3. fibres over the circle?
4. has arbitrarily large first Betti number?
5. has fundamental group with a non-abelian free quotient? (When a group has a finite-index subgroup with a non-abelian free quotient, it is known as *large*.)

The obvious relationships between these problems are shown in Figure 1. This figure also includes the Surface Subgroup Conjecture, which proposes that a closed hyperbolic 3-manifold contains a closed orientable π_1-injective surface (other than a 2-sphere), which need not be embedded. While this is not strictly a question about finite covers, one might hope to lift this surface to an embedded one in some finite cover of the 3-manifold.

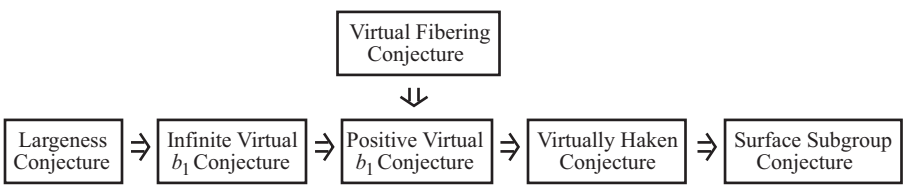


Figure 1.

There are many reasons why these questions are interesting. One source of motivation is that Haken manifolds are very well-understood, and so one might hope to use their highly-developed theory to probe general 3-manifolds. But the main reason for studying these problems is an aesthetic one. Embedded surfaces, particularly those that are π_1-injective, play a central role in low-dimensional topology and these conjectures assert that they are ubiquitous. In addition, these problems relate to many other interesting areas of mathematics, as we will see.

In order to tackle these conjectures, one is immediately led to the following questions, which are also interesting in their own right. How many finite covers does a hyperbolic 3-manifold have, as a function of the covering degree? How do natural geometric, topological and algebraic invariants behave in finite-sheeted covers, for example:

1. the spectrum of the Laplacian;
2. their Heegaard genus;

3. the rank of their fundamental group;
4. the order of their first homology, possibly with coefficients modulo a prime?

In this survey, we will outline some progress on these questions, and will particularly emphasise how an understanding of the geometric, topological and algebraic invariants of finite covers can be used to deduce the existence of incompressible surfaces, and more.

Many of the methods that we will discuss work only in dimension 3. However, many apply more generally to arbitrary finitely presented groups. We will also explain some of these group-theoretic applications.

An outline of this paper is as follows. In Section 2, we will give a summary of the progress to date (March 2010) towards the above conjectures. In Sections 3 and 4, we will explain the two main classes of covering space of a hyperbolic 3-manifold: congruence covers and abelian covers. In Section 5, we will give the best known lower bounds on the number of covers of a hyperbolic 3-manifold. In Sections 6 and 7, we will analyse the behaviour of various invariants in finite covers. In Section 8, we will explain how an understanding of this behaviour may lead to some approaches to the Virtually Haken Conjecture. In Section 9, we will examine arithmetic 3-manifolds and hyperbolic 3-orbifolds with non-empty singular locus, as these appear to be particularly tractable. In Section 10, we will consider some group-theoretic generalisations. And finally in Section 11, we will briefly give some other directions in theory of finite covers of 3-manifolds that have emerged recently.

2. The State of Play

While the techniques developed to study finite covers of 3-manifolds are interesting and important, they have not yet solved the Virtually Haken Conjecture or its variants. In fact, our understanding of these conjectures is still quite limited. We focus, in this section, on the known unconditional results, and the known interconnections between the various conjectures.

The manifolds that are most well understood, but for which our knowledge is still far from complete, are the *arithmetic* hyperbolic 3-manifolds. We will not give their definition here, but instead refer the reader to Maclachlan and Reid's excellent book on the subject [48].

We start with the Surface Subgroup Conjecture. Here, we have the following result, due to the author [33].

Theorem 2.1. *Any arithmetic hyperbolic 3-manifold contains a closed orientable immersed π_1-injective surface with positive genus.*

In fact, a proof of the Surface Subgroup Conjecture for all closed hyperbolic 3-manifolds has recently been announced by Kahn and Markovic [25]. The

details of this are still being checked. But the proof of Theorem 2.1 is still relevant because it falls into a general programme of the author for proving the Virtually Haken Conjecture.

The Virtually Haken Conjecture remains open at present, and is only known to hold in certain situations. Several authors have examined the case where the manifold is obtained by Dehn filling a one-cusped finite-volume hyperbolic 3-manifold. The expectation is that all but finitely many Dehn fillings of this manifold should be virtually Haken. This is not known at present. However, the following theorem, which is an amalgamation of results due to Cooper and Long [16] and Cooper and Walsh [19, 20], goes some way to establishing this.

Theorem 2.2. *Let X be a compact orientable 3-manifold with boundary a single torus, and with interior admitting a finite-volume hyperbolic structure. Then, infinitely many Dehn fillings of X are virtually Haken.*

One might hope to use the existence of a closed orientable immersed π_1-injective surface in a hyperbolic 3-manifold to find a finite cover of the 3-manifold which is Haken. However, the jump from the Surface Subgroup Conjecture to the Virtually Haken Conjecture is a big one. Currently, the only known method is the use of the group-theoretic condition called subgroup separability (see Section 11 for a definition). This is a powerful property, but not many 3-manifold groups are known to have it (see [4] for some notable examples). Indeed, the condition is so strong that when the fundamental group of a closed orientable 3-manifold contains the fundamental group of a closed orientable surface with positive genus that is separable, then this manifold virtually fibres over the circle or has large fundamental group. Nevertheless, subgroup separability is a useful and interesting property. For example, an early application of the condition is the following, due to Long [38].

Theorem 2.3. *Any finite-volume hyperbolic 3-manifold that contains a closed immersed totally geodesic surface has large fundamental group.*

The progress towards the Positive Virtual b_1 Conjecture is also quite limited. An experimental analysis [21] by Dunfield and Thurston of the 10986 manifolds in the Hodgson-Weeks census has found, for each manifold, a finite cover with positive b_1. This is encouraging, but the only known general results apply to certain classes of arithmetic 3-manifolds. The following is due to Clozel [15]. (We refer the reader to [48] for the definitions of the various terms in this theorem.)

Theorem 2.4. *Let M be an arithmetic hyperbolic 3-manifold, with invariant trace field k and quaternion algebra B. Assume that for every finite place ν where B ramifies, the completion k_ν contains no quadratic extension of $\mathbb{Q}_p$, where p is a rational prime and ν divides p. Then M has a finite cover with positive b_1.*

Again, the jump from positive virtual b_1 to infinite virtual b_1 is not known in general. However, it is known for arithmetic 3-manifolds, via the following

result, which was first proved by Cooper, Long and Reid [18], but shortly afterwards, alternative proofs were given by Venkataramana [63] and Agol [2].

Theorem 2.5. *Suppose that an arithmetic hyperbolic 3-manifold M has $b_1 > 0$. Then M has finite covers with arbitrarily large b_1.*

The step from infinite virtual b_1 to largeness is known to hold in some circumstances. The following result is due to the author, Long and Reid [35].

Theorem 2.6. *Let M be an arithmetic hyperbolic 3-manifold. Suppose that M has a finite cover with $b_1 \geq 4$. Then $\pi_1(M)$ is large.*

Thus, combining the above three theorems, many arithmetic hyperbolic 3-manifolds are known to have large fundamental groups. And by Theorem 2.1, they all contain closed orientable immersed π_1-injective surfaces with positive genus.

The remaining problem is the Virtual Fibering Conjecture. For a long time, this seemed to be rather less likely than the others, simply because there were very few manifolds that were known to be virtually fibred that were not already fibred. (Examples were discovered by Reid [54], Leininger [37] and Agol-Boyer-Zhang [3].) However, this situation changed recently, with work of Agol [1], which gives a useful sufficient condition for a 3-manifold to be virtually fibred. It has the following striking consequence.

Theorem 2.7. *Let M be an arithmetic hyperbolic 3-manifold that contains a closed immersed totally geodesic surface. Then M is virtually fibred.*

These manifolds were already known to have large fundamental group, by Theorem 2.3. However, virtual fibration was somewhat unexpected here.

Finally, we should mention that the Virtually Haken Conjecture and its variants are mostly resolved in the case when M is a compact orientable irreducible 3-manifold with non-empty boundary. Indeed it is a fundamental fact that $b_1(M) \geq b_1(\partial M)/2$. Hence, M trivially satisfies the Positive Virtual b_1 Conjecture. In fact, much more is true, by the following results of Cooper, Long and Reid [17].

Theorem 2.8. *Let M be a compact orientable irreducible 3-manifold with non-empty boundary, that is not an I-bundle over a disc, annulus, torus or Klein bottle. Then $\pi_1(M)$ is large.*

Theorem 2.9. *Let M be a compact orientable irreducible 3-manifold with non-empty boundary. Then $\pi_1(M)$ is trivial or free or contains the fundamental group of a closed orientable surface with positive genus.*

The main unsolved problem for 3-manifolds with non-empty boundary is therefore the Virtual Fibering Conjecture for finite-volume hyperbolic 3-manifolds.

3. Congruence Covers

We start with some obvious questions. How can one construct finite covers of a hyperbolic 3-manifold? Do they come in different 'flavours'? Of course, any finite regular cover of a 3-manifold M is associated with a surjective homomorphism from $\pi_1(M)$ onto a finite group. But there is no systematic theory for such homomorphisms in general. There are currently just two classes of finite covering spaces of general hyperbolic 3-manifolds which are at all well-understood: congruence covers and abelian covers. We will examine these in more detail in this section and the one that follows it.

If Γ is the fundamental group of an orientable hyperbolic 3-manifold M, then the hyperbolic structure determines a faithful homomorphism $\Gamma \to \mathrm{Isom}^+(\mathbb{H}^3) \cong \mathrm{PSL}(2,\mathbb{C})$. When M has finite volume, one may in fact arrange that the image lies in $\mathrm{PSL}(2,R)$, where R is obtained from the ring of integers of a number field by inverting finitely many prime ideals. This permits the use of number theory. Specifically, one can take any proper non-zero ideal I in R, and consider the composite homomorphsim

$$\Gamma \to \mathrm{PSL}(2,R) \to \mathrm{PSL}(2,R/I)$$

which is termed the *level I congruence homomorphism*. We denote it by ϕ_I. The kernel of such a homomorphism is called a *principal congruence subgroup*, and any subgroup that contains a principal congruence subgroup is *congruence*. We term the corresponding cover of M a *congruence cover*. Now, R/I is a finite ring, and in fact if I is prime, then R/I is a finite field. Hence, congruence covers always have finite degree.

There is an alternative approach to this theory, involving quaternion algebras, which is in many ways superior. It leads to the same definition of a congruence subgroup, but the congruence homomorphisms are a little different. However, we do not follow this approach here because it requires too much extra terminology.

Congruence subgroups are extremely important. They are used to prove the following foundational result.

Theorem 3.1. *The fundamental group Γ of a finite-volume orientable hyperbolic 3-manifold is residually finite. In fact, for all primes p with at most finitely many exceptions, Γ is virtually residually p-finite.*

The residual finiteness of Γ is established as follows. Let γ be any non-trivial element of Γ, and let $\hat{\gamma}$ be an inverse image of γ in $\mathrm{SL}(2,R)$. Then neither $\hat{\gamma}-1$ nor $\hat{\gamma}+1$ is the zero matrix, and so each has a non-zero matrix entry. Let x be the product of these entries. Since x is a non-zero element of R, it lies in only finitely many ideals I. Therefore $\hat{\gamma}-1$ and $\hat{\gamma}+1$ both have non-zero image in $\mathrm{SL}(2,R/I)$ for almost all ideals I. For each such I, the images of γ and the identity in $\mathrm{PSL}(2,R/I)$ are distinct, which proves residual finiteness.

To establish virtual residual p-finiteness, for some integral prime p, one works with the principal ideals (p^n) in R, where $n \in \mathbb{N}$. Provided p does not

lie in any of the prime ideals that were inverted in the definition of R, (p^n) is a proper ideal of R. Let $\Gamma(p^n)$ denote the kernel of the level (p^n) congruence homomorphism. Then, by the above argument, for any non-trivial element γ of Γ, γ does not lie in $\Gamma(p^n)$ for all sufficiently large n. In particular, this is true of all non-trivial γ in $\Gamma(p)$. Now, the image of $\Gamma(p)$ under the level (p^n) congruence homomorphism lies in the subgroup of $\mathrm{PSL}(2, R/(p^n))$ consisting of elements that are congruent to the identity mod (p). This is a finite p-group. Hence, we have found, for each non-trivial element γ of $\Gamma(p)$, a homomorphism onto a finite p-group for which the image of γ is non-trivial, thereby proving Theorem 3.1.

The conclusions of Theorem 3.1 in fact hold more generally for any finitely generated group that is linear over a field of characteristic zero, with essentially the same proof. In fact, when studying congruence homomorphisms, one is led naturally to the extensive theory of linear groups. Here, the Strong Approximation Theorem of Nori and Weisfeiler [64] is particularly important. This deals with the images of the congruence homomorphisms $\phi_I : \Gamma \to \mathrm{PSL}(2, R/I)$, as I ranges over all the proper non-zero ideals of R, simultaneously. We will not give the precise statement here, because it also requires too much extra terminology. However, we note the following consequence, which has, in fact, a completely elementary proof.

Theorem 3.2. *There is a finite set S of prime ideals I in R with following properties.*

1. *For each prime non-zero ideal I of R that is not in S, $\mathrm{Im}(\phi_I)$ is isomorphic to $\mathrm{PSL}(2, q^n)$ or $\mathrm{PGL}(2, q^n)$, where q is the characteristic of R/I and $n > 0$.*

2. *For any finite set of prime non-zero ideals $I_1, \dots, I_m$ in R, none of which lies in S, and for which the characteristics of the fields R/I_i are all distinct, the product homomorphism*

$$\prod_{i=1}^{m} \phi_{I_i} : \Gamma \to \prod_{i=1}^{m} \mathrm{PSL}(2, R/I_i)$$

has image equal to

$$\prod_{i=1}^{m} \mathrm{Im}(\phi_{I_i}).$$

This has the following important consequence for homology modulo a prime p, due to Lubotzky [41]. Given any prime p and group or space X, let $d_p(X)$ denote the dimension of $H_1(X; \mathbb{F}_p)$, as a vector space over the field $\mathbb{F}_p$.

Theorem 3.3. *Let Γ be the fundamental group of a finite-volume orientable hyperbolic 3-manifold. Let p be any prime integer, and let m be any natural number. Then Γ has a congruence subgroup $\tilde{\Gamma}$ such that $d_p(\tilde{\Gamma}) \geq m$.*

The proof runs as follows. For almost all integral primes q, there is a prime ideal I in R such that R/I is a field of characteristic q. Moreover, by Theorem 3.2, we may assume that $\mathrm{Im}(\phi_I)$ is isomorphic to $\mathrm{PSL}(2, q^n)$ or $\mathrm{PGL}(2, q^n)$, where $n \geq 1$. Inside $\mathrm{PSL}(2, q^n)$ or $\mathrm{PGL}(2, q^n)$, there is the subgroup consisting of diagonal matrices, which is abelian with order $(q^n - 1)/2$ or $(q^n - 1)$. We now want to restrict to certain primes q, and to do this, we use Dirichlet's theorem, which asserts that there are infinitely many primes q such that $q \equiv 1 \pmod{p}$. When $p = 2$, we also require that $q \equiv 1 \pmod{4}$. For these q, p divides the order of the subgroup of diagonal matrices, and so there is a subgroup of order p. We may find a set of m such primes $q_1, \ldots, q_m$ so that each q_i is the characteristic of R/I_i, where I_i is a prime ideal avoiding the finite set S described above. Then we have an inclusion of groups

$$(\mathbb{Z}/p)^m \leq \mathrm{Im}(\phi_{I_1}) \times \cdots \times \mathrm{Im}(\phi_{I_m}).$$

Now, Γ surjects onto the right-hand group. Let $\tilde{\Gamma}$ be the inverse image of the left-hand group. This is a congruence subgroup, and by construction, it surjects onto $(\mathbb{Z}/p)^m$. Hence, $d_p(\tilde{\Gamma}) \geq m$, thereby proving Theorem 3.3.

Since $d_p(\tilde{\Gamma})$ is positive, the covering space corresponding to $\tilde{\Gamma}$ has a non-trivial regular cover with covering group that is an elementary abelian p-group (in other words is isomorphic to $(\mathbb{Z}/p)^m$ for some m). Thus, we are led to the following type of covering space.

4. Abelian Covers

A covering map is *abelian* (respectively, *cyclic*) provided it is regular and the group of covering transformations is abelian (respectively, cyclic). A large amount of attention has been focused on the homology of abelian covers. Indeed, one of the earliest topological invariants, the Alexander polynomial, can be interpreted this way. However, the Alexander polynomial is only defined when the covering group is free abelian, and so we leave the realm of finite covering spaces. We therefore will not dwell too long on the Alexander polynomial, but to omit mention of it entirely would be remiss, especially as it has consequences also for certain finite cyclic covers, via the following result, due to Silver and Williams [60] (see also [55, 22]).

Theorem 4.1. *Let M be a compact orientable 3-manifold, let $\tilde{M}$ be an infinite cyclic cover and let $\Delta(t) \in \mathbb{Z}[t, t^{-1}]$ be the resulting Alexander polynomial. Its Mahler measure is defined by*

$$M(\Delta) = |c| \prod_i \max\{1, |\alpha_i|\},$$

as α_i ranges over all roots of $\Delta(t)$, and c is the coefficient of the highest order

term. Let M_n be the degree n cyclic cover of M that is covered by $\tilde{M}$. Then

$$\frac{\log |H_1(M_n)_{\mathrm{tor}}|}{n} \to \log M(\Delta),$$

as $n \to \infty$, where $H_1(M_n)_{\mathrm{tor}}$ denotes the torsion part of $H_1(M_n)$.

Thus, provided $\Delta(t)$ has at least one root off the unit circle, $|H_1(M_n)_{\mathrm{tor}}|$ has exponential growth as a function of n.

There are more sophisticated versions of this result, dealing for example with the case when $\tilde{M}$ is a regular cover with a free abelian group of covering transformations [60]. However, the theory only applies when $b_1(M)$ is positive, and so, in the absence of the solution to the Positive Virtual b_1 Conjecture, methods using the Alexander polynomial are not yet universally applicable in 3-manifold theory.

There is another important direction in the theory of abelian covers, which deals with homology modulo a prime p. Let Γ be the fundamental group of a compact orientable 3-manifold, and suppose that $\tilde{\Gamma}$ is a normal subgroup such that $\Gamma/\tilde{\Gamma}$ is an elementary abelian p-group. Then an important result of Shalen and Wagreich [59] gives a lower bound for the mod p homology of $\tilde{\Gamma}$. For simplicity, we will deal only with the case where $\Gamma/\tilde{\Gamma}$ is as big as possible, which is when $\tilde{\Gamma} = [\Gamma, \Gamma]\Gamma^p$.

Theorem 4.2. *Let Γ be the fundamental group of a compact orientable 3-manifold, and let $\tilde{\Gamma} = [\Gamma, \Gamma]\Gamma^p$. Then,*

$$d_p(\tilde{\Gamma}) \geq \binom{d_p(\Gamma)}{2}.$$

This has the consequence that when $d_p(\Gamma) \geq 3$, then also $d_p(\tilde{\Gamma}) \geq 3$. Hence, we may repeat the argument with $\tilde{\Gamma}$ in place of Γ. It is therefore natural to consider the *derived p-series* of Γ, which is defined by setting $\Gamma_0 = \Gamma$, and $\Gamma_{i+1} = [\Gamma_i, \Gamma_i](\Gamma_i)^p$ for $i \geq 0$. We deduce that when $d_p(\Gamma) \geq 3$, then the derived p-series is always strictly descending. Moreover, when $d_p(\Gamma) > 3$, then $d_p(\Gamma_i)$ tends to infinity. Note that $d_p(\Gamma_i)$ need not tend to infinity when $d_p(\Gamma) = 3$, as the example of the 3-torus demonstrates.

The original proof of Theorem 4.2 by Shalen and Wagreich used the following exact sequence of Stallings:

$$H_2(\Gamma; \mathbb{F}_p) \to H_2(\Gamma/\tilde{\Gamma}; \mathbb{F}_p) \to \frac{\tilde{\Gamma}}{[\tilde{\Gamma}, \Gamma](\tilde{\Gamma})^p} \to H_1(\Gamma; \mathbb{F}_p) \to H_1(\Gamma/\tilde{\Gamma}; \mathbb{F}_p) \to 0.$$

Let $d = d_p(\Gamma/\tilde{\Gamma}) = d_p(\Gamma)$. Then, $H_2(\Gamma/\tilde{\Gamma}; \mathbb{F}_p)$ is an elementary abelian p-group of rank $d(d+1)/2$ by the Künneth formula. However, by Poincaré duality, $H_2(\Gamma; \mathbb{F}_p)$ is an elementary abelian p-group of rank at most d. Thus, by exactness of the above sequence, $\tilde{\Gamma}/([\tilde{\Gamma}, \Gamma](\tilde{\Gamma})^p)$ has rank at least $d(d-1)/2$. But this

is a quotient of $\tilde{\Gamma}/([\tilde{\Gamma},\tilde{\Gamma}](\tilde{\Gamma})^p)$, which equals $H_1(\tilde{\Gamma};\mathbb{F}_p)$. Hence, one obtains the required lower bound on $d_p(\tilde{\Gamma})$.

Although this argument is short, it is not an easy one for a geometric topologist to digest. In an attempt to try to understand it, the author found an alternative topological proof, which then led to a considerable strengthening of the theorem. The proof runs roughly as follows, focusing on the case $p = 2$ for simplicity.

Pick a generating set $\{x_1, \ldots, x_n\}$ for Γ such that the first d elements $x_1, \ldots, x_d$ form a basis for $H_1(\Gamma;\mathbb{F}_2)$, and so that each x_i is trivial in $H_1(\Gamma;\mathbb{F}_2)$ for $i > d$. Let K be a 2-complex with fundamental group Γ, with a single 0-cell and with 1-cells corresponding to the above generating set. Let $\tilde{K}$ be the covering space corresponding to $\tilde{\Gamma}$. We are trying to find a lower bound on $d_2(\tilde{\Gamma})$, which is the rank of the cohomology group $H^1(\tilde{K};\mathbb{F}_2)$. Now, $H^1(\tilde{K};\mathbb{F}_2)$ is equal to $\{$1-cocycles on $\tilde{K}\}/\{$1-coboundaries on $\tilde{K}\}$. Each 1-cocycle is, by definition, a 1-cochain that evaluates to zero on the boundary of each 2-cell of $\tilde{K}$. However, instead of examining all cochains, we only consider special ones, which are defined as follows. For each integer $1 \leq j \leq d$, each vertex of $\tilde{K}$ has a well-defined x_j-value, which is an integer mod 2. For $1 \leq i \leq j \leq d$, define the cochain $x_i \wedge x_j$ to have support equal to the x_i-labelled edges which start (and end) at vertices with x_j-value 1. The space spanned by these cochains clearly has dimension $d(d+1)/2$.

These cochains have the key property that if two closed loops in $\tilde{K}$ differ by a covering transformation, then their evaluations under one of these cochains are equal. Hence, when we consider the space spanned by these cochains, and determine whether any element of this space is a cocycle, we only need to consider one copy of each defining relation of Γ. It turns out that none of the cocycles in this space is a coboundary, except the zero cocycle, and so $d_2(\tilde{\Gamma}) \geq d(d+1)/2 - r$, where r is the number of 2-cells of K. In fact, by modifying these cocycles a little, the number of conditions that we must check can be reduced from r to $b_2(\Gamma;\mathbb{F}_2)$. So, we deduce that, if Γ is any finitely presented group and $\tilde{\Gamma} = [\Gamma,\Gamma]\Gamma^2$ and $d = d_2(\Gamma)$, then

$$d_2(\tilde{\Gamma}) \geq d(d+1)/2 - b_2(\Gamma;\mathbb{F}_2).$$

And when Γ is the fundamental group of a compact orientable 3-manifold, Poincaré duality again gives that $b_2(\Gamma;\mathbb{F}_2) \leq d$, proving Theorem 4.2.

This is not the end of the story, because it turns out that these cochains are just the first in a whole series of cochains, each with an associated integer, which is its 'level' ℓ. The ones above are those with level $\ell = 1$. An example of a level 2 cochain has support equal to the x_1-labelled edges which start at the vertices for which the x_2-value and x_3-value are both 1. By considering cochains at different levels, we can considerably strengthen Theorem 4.2, as follows. Again, there are more general versions which deal with the general case where $\Gamma/\tilde{\Gamma}$ is an elementary abelian p-group, but we focus on the case where $\tilde{\Gamma} = [\Gamma,\Gamma]\Gamma^2$.

Theorem 4.3. *Let Γ be the fundamental group of a compact orientable 3-manifold, and let $\tilde{\Gamma} = [\Gamma, \Gamma]\Gamma^2$. Then, for each integer ℓ between 1 and $d_2(\Gamma)$,*

$$d_2(\tilde{\Gamma}) \geq d_2(\Gamma) \binom{d_2(\Gamma)}{\ell} - \sum_{j=1}^{\ell+1} \binom{d_2(\Gamma)}{j}.$$

Setting $\ell = \lfloor d_2(\Gamma)/2 \rfloor$ and using Stirling's formula to estimate factorials, we deduce the following [31].

Theorem 4.4. *Let Γ be the fundamental group of a compact orientable 3-manifold such that $d_2(\Gamma) > 3$. Let $\{\Gamma_i\}$ be the derived 2-series of Γ. Then, for each $\lambda < \sqrt{2/\pi}$,*

$$d_2(\Gamma_{i+1}) \geq \lambda 2^{d_2(\Gamma_i)} \sqrt{d_2(\Gamma_i)},$$

for all sufficiently large i.

This is not far off the fastest possible growth of homology of a finitely generated group. By comparison, when $\{\Gamma_i\}$ is the derived 2-series of a non-abelian free group, then

$$d_2(\Gamma_{i+1}) = 2^{d_2(\Gamma_i)}(d_2(\Gamma_i) - 1) + 1.$$

Theorem 4.4 can be used to produce strong lower bounds on the number of covering spaces of a hyperbolic 3-manifold, as we will see in the following section.

5. Counting Finite Covers

How many finite covers does a 3-manifold have? This question lies in the field of subgroup growth [45], which deals with the behaviour of the following function. For a finitely generated group Γ and positive integer n, let $s_n(\Gamma)$ be the number of subgroups of Γ with index at most n.

The fastest possible growth rate of $s_n(\Gamma)$, as a function of n, is clearly achieved when Γ is a non-abelian free group. In this case, $s_n(\Gamma)$ grows slightly faster than exponentially: it grows like $2^{n \log n}$. More generally, any large finitely generated group has this rate of subgroup growth.

By comparison, the subgroup growth of the fundamental group of a hyperbolic 3-manifold group has a lower bound that grows slightly slower than exponentially, as the following result [31] of the author demonstrates.

Theorem 5.1. *Let Γ be the fundamental group of a finite-volume hyperbolic 3-manifold. Then,*

$$s_n(\Gamma) > 2^{n/(\sqrt{\log(n)} \log\log n)}$$

for infinitely many n.

The proof is a rapid consequence of Theorems 4.4 and 3.3. Theorem 3.3 gives a finite index subgroup $\tilde{\Gamma}$ of Γ with $d_2(\tilde{\Gamma}) > 3$. Then Theorem 4.4 implies that the mod 2 homology of the derived 2-series of $\tilde{\Gamma}$ grows rapidly. And if Γ_i is a subgroup of Γ with index n, then clearly

$$s_{2n}(\Gamma) \geq 2^{d_2(\Gamma_i)}.$$

Thus, in the landscape of finite covers of a hyperbolic 3-manifold, abelian covers appear to play a major role. Certainly, there are far more of them than there are congruence covers, by the following result of Lubotzky [43], which estimates $c_n(\Gamma)$, which is the number of congruence subgroups of Γ with index at most n.

Theorem 5.2. *Let Γ be the fundamental group of an orientable finite-volume hyperbolic 3-manifold. Then, there are positive constants a and b such that*

$$n^{a \log n / \log\log n} \leq c_n(\Gamma) \leq n^{b \log n / \log\log n},$$

for all n.

The lower bound provided by Theorem 5.1 is not sharp in general, because there are many examples where Γ is large. Indeed, the Largeness Conjecture asserts that this should always be the case.

There is another important situation when we know that the lower bound of Theorem 5.1 can be improved upon, due to the following result of the author [30].

Theorem 5.3. *Let Γ be the fundamental group of either an arithmetic hyperbolic 3-manifold or a finite-volume hyperbolic 3-orbifold with non-empty singular locus. Then, there is a real number $c > 1$ such that $s_n(\Gamma) \geq c^n$ for infinitely many n.*

Like Theorem 5.1, this is proved by finding lower bounds on the rank of the mod p homology of certain finite covers. We will give more details in Section 9, where the covering spaces of 3-orbifolds and arithmetic 3-manifolds will be examined more systematically.

6. The Behaviour of Algebraic Invariants in Finite Covers

As we have seen, it is important to understand how the homology groups can grow in a tower of finite covers. Thus, we are led to the following related invariants of a group Γ:

1. the first Betti number $b_1(\Gamma)$,
2. the torsion part $H_1(\Gamma)_{\text{tor}}$ of first homology,

3. the rank $d_p(\Gamma)$ of mod p homology,

4. the rank of Γ, denoted $d(\Gamma)$, which is the minimal number of generators.

For each of these invariants, it is natural to consider its growth rate in a nested sequence of finite index subgroups Γ_i. For example, one can define the *rank gradient* which is

$$\liminf_i \frac{d(\Gamma_i)}{[\Gamma : \Gamma_i]}.$$

The *mod p homology gradient* and *first Betti number gradient* are defined similarly. The latter is in fact, by a theorem of Lück [47], related to the first L^2 Betti number of Γ (denoted $b_1^{(2)}(\Gamma)$). More precisely, when Γ_i is a nested sequence of finite-index normal subgroups of a finitely presented group Γ, and their intersection is the identity, then their first Betti number gradient is equal to $b_1^{(2)}(\Gamma)$. When Γ is the fundamental group of a finite-volume hyperbolic 3-manifold, $b_1^{(2)}(\Gamma)$ is known to be zero, and hence $b_1(\Gamma_i)$ always grows sub-linearly as a function of the covering degree $[\Gamma : \Gamma_i]$. Interestingly, there is no corresponding theory for mod p homology gradient, and the following question is at present unanswered.

Question. *Let Γ be a finitely presented group, let Γ_i be a nested sequence of finite-index normal subgroups that intersect in the identity. Then does their mod p homology gradient depend only on Γ and possibly p, but not the sequence Γ_i?*

This is unknown, but it seems very likely that the mod p homology gradient is always zero when Γ is the fundamental group of a finite-volume hyperbolic 3-manifold and the subgroups intersect in the identity. However, if we drop the condition that the subgroups intersect in the identity, then there is an interesting situation where positive mod p homology gradient is known to hold, by the following result of the author [30].

Theorem 6.1. *Let Γ be the fundamental group of either an arithmetic hyperbolic 3-manifold or a finite-volume hyperbolic 3-orbifold with non-empty singular locus. Then, for some prime p, Γ has a nested strictly descending sequence of finite-index subgroups with positive mod p homology gradient.*

We will explore this in more detail in Section 9. But we observe here that it rapidly implies Theorem 5.3. The existence of such a sequence of finite-index subgroups seems to be a very strong conclusion. In fact, the following is unknown.

Question. *Suppose that a finitely presented group Γ has a strictly descending sequence of finite-index subgroups with positive mod p homology gradient, for some prime p. Does this imply that Γ is large?*

We will give some affirmative evidence for this in Section 10. Somewhat surprisingly, this question is related to the theory of error-correcting codes (see [32]).

We have mostly focused on the existence of very fast homology growth for certain covers of hyperbolic 3-manifolds. But one can also consider the other end of the spectrum, and ask how slowly the homology groups of a tower of covers can grow. In this context, the following theorem of Boston and Ellenberg [7] is striking (see also [12]).

Theorem 6.2. *There is an example of a closed hyperbolic 3-manifold, with fundamental group that has a sequence of nested finite-index normal subgroups* Γ_i *which intersect in the identity, such that* $b_1(\Gamma_i) = 0$ *and* $d_3(\Gamma_i) = 3$ *for all* i.

This is proved using the theory of pro-p groups. This is a particularly promising set of techniques, which will doubtless have other applications to 3-manifold theory.

7. The Behaviour of Geometric and Topological Invariants in Finite Covers

In addition to the above algebraic invariants, it seems to be important also to understand the behaviour of various geometric and topological invariants in a tower of covers, including the following:

1. the first eigenvalue of the Laplacian,
2. the Cheeger constant,
3. the Heegaard genus.

We will recall the definitions of these terms below.

It is well known that the Laplacian on a closed Riemannian manifold M has a discrete set of eigenvalues, and hence there is a smallest positive eigenvalue, denoted $\lambda_1(M)$. This exerts considerable control over the geometry of the manifold. In particular, it is related to the *Cheeger constant* $h(M)$, which is defined to be

$$\inf_S \frac{\text{Area}(S)}{\min\{\text{Volume}(M_1), \text{Volume}(M_2)\}},$$

as S ranges over all codimension-one submanifolds that divide M into submanifolds M_1 and M_2. It is a famous theorem of Cheeger [14] and Buser [11] that if M is closed Riemannian n-manifold with Ricci curvature at least $-(n-1)a^2$ (for some $a \geq 0$), then

$$h(M)^2/4 \leq \lambda_1(M) \leq 2a(n-1)h(M) + 10(h(M))^2.$$

A consequence is that if M_i is a sequence of finite covers of M, then $\lambda_1(M_i)$ is bounded away from zero if and only if $h(M_i)$ is. In this case, $\pi_1(M)$ is said to have *Property* (τ) with respect to the subgroups $\{\pi_1(M_i)\}$. As the definition implies, this depends only on the fundamental group $\pi_1(M)$ and the subgroups

$\{\pi_1(M_i)\}$, and not on the choice of particular Riemannian metric on M. Also, $\pi_1(M)$ is said to have *Property* (τ) if it has Property (τ) with respect to the collection of all its finite-index subgroups. Since any finitely presented group is the fundamental group of some closed Riemannian manifold, this is therefore a property that is or is not enjoyed by any finitely presented group. In fact, one can extend the definition to finitely generated groups that need not be finitely presented. (See [42, 46] for excellent surveys of this concept.)

The reason for the 'τ' terminology is that Property (τ) is a weak form of Kazhdan's Property (T). In particular, any finitely generated group with Property (T) also has Property (τ) [49]. A harder result is due to Selberg [58], which implies that $\mathrm{SL}(2,\mathbb{Z})$ has Property (τ) with respect to its congruence subgroups.

The simplest example of a group without Property (τ) is $\mathbb{Z}$. Also, if there is a surjective group homomorphism $\Gamma \to \overline{\Gamma}$ and $\overline{\Gamma}$ does not have Property (τ), then nor does Γ. Hence, if a group has a finite-index subgroup with positive first Betti number, then it does not have Property (τ). Strikingly, it remains an open question whether the converse holds in the finitely presented case.

Question. *If a finitely presented group does not have Property (τ), then must it have a finite-index subgroup with positive first Betti number?*

The assumption that the group is finitely presented here is critical. For example, Grigorchuk's group [23] is residually finite and amenable, and hence does not have Property (τ), and yet it is a torsion group, and so no finite-index subgroup has positive first Betti number. Although a positive answer to this question is unlikely, it might have some striking applications. For example, every residually finite group with sub-exponential growth does not have Property (τ). So, a positive answer to the above question might be a step in establishing that such groups are virtually nilpotent, provided they are finitely presented.

One small piece of evidence for an affirmative answer to the question is given by the following theorem of the author [29], which relates the behaviour of λ_1 and h to the existence of a finite-index subgroup with positive first Betti number.

Theorem 7.1. *Let M be a closed Riemannian manifold. Then the following are equivalent:*

- *there exists a tower of finite covers $\{M_i\}$ of M with degree d_i, where each $M_i \to M_1$ is regular, and such that $\lambda_1(M_i)d_i \to 0$;*
- *there exists a tower of finite covers $\{M_i\}$ of M with degree d_i, where each $M_i \to M_1$ is regular, and such that $h(M_i)\sqrt{d_i} \to 0$;*
- *there exists a finite-index subgroup of $\pi_1(M)$ with positive first Betti number.*

However, it remains unlikely that the above question has a positive answer. Hence, the following conjecture [44] about 3-manifolds is *a priori* much weaker than the Positive Virtual b_1 Conjecture.

Conjecture (Lubotzky-Sarnak). *The fundamental group of any closed hyperbolic 3-manifold does not have Property* (τ).

This is a natural question for many reasons. It is known that if Γ is a lattice in a semi-simple Lie group G, then whether or not Γ has Kazhdan's Property (T) depends only on G. It remains an open question whether a similar phenomenon holds for Property (τ), but if it did, then this would of course imply the Lubotzky-Sarnak Conjecture.

An 'infinite' version of the Lubotzky-Sarnak Conjecture is known to hold, according to the following result of the author, Long and Reid [36].

Theorem 7.2. *Any closed hyperbolic 3-manifold has a sequence of infinite-sheeted covers M_i where $\lambda_1(M_i)$ and $h(M_i)$ both tend to zero.*

Of course, $\lambda_1(M_i)$ and $h(M_i)$ need to be defined appropriately, since each M_i has infinite volume. In the case of $\lambda_1(M_i)$, this is just the bottom of the spectrum of the Laplacian on L^2 functions on M_i. To define $h(M_i)$, one considers all compact codimension-zero submanifolds of M_i, one evaluates the ratio of the area of their boundary to their volume, and then one takes the infimum. Just as in the finite-volume case, there is a result of Cheeger [14] which asserts that $\lambda_1(M_i) \geq h(M_i)^2/4$. Also, in the case of hyperbolic 3-manifolds M_i, these quantities are related to another important invariant $\delta(M_i)$, which is the critical exponent. A theorem of Sullivan [61] asserts that

$$\lambda_1(M_i) = \begin{cases} \delta(M_i)(2-\delta(M_i)) & \text{if } \delta(M_i) \geq 1 \\ 1 & \text{if } \delta(M_i) \leq 1. \end{cases}$$

The proof of Theorem 7.2 relies crucially on a recent result of Bowen [9], which asserts that, given any closed hyperbolic 3-manifold M and any finitely generated discrete free convex-cocompact subgroup F of $\mathrm{PSL}(2,\mathbb{C})$, there is an arbitrarily small 'perturbation' of F which places a finite-index subgroup of F as a subgroup of $\pi_1(M)$. Starting with a group F where $\delta(F)$ is very close to 2, the critical exponent of this perturbation remains close to 2. Thus, this produces subgroups of $\pi_1(M)$ with critical exponent arbitrarily close to 2. By Sullivan's theorem, the corresponding covers M_i of M have $\lambda_1(M_i)$ arbitrarily close to zero. By Cheeger's theorem, $h(M_i)$ also tends to zero, proving Theorem 7.2.

Although the infinite version of the Lubotzky-Sarnak Conjecture does not seem to have any immediate consequence for finite covering spaces of closed hyperbolic 3-manifolds, it can be used to produce surface subgroups. Indeed, it is a key step in the proof of Theorem 2.1. We will give more details in Section 9.

In addition to understanding $\lambda_1(M_i)$ and $h(M_i)$ for finite covering spaces M_i, it also seems to be important to understand the growth rate of their Heegaard genus. Recall that any closed orientable 3-manifold M can be obtained

by gluing two handlebodies via a homeomorphism between their boundaries. This is a *Heegaard splitting* for M, and the image of the boundary of each handlebody is a *Heegaard surface*. The minimal genus of a Heegaard surface in M is known as the *Heegaard genus* $g(M)$. A related quantity is the *Heegaard Euler characteristic* $\chi_-^h(M)$, which is $2g(M)-2$. These are widely-studied invariants of 3-manifolds, and there is now a well-developed theory of Heegaard splittings [56]. It is therefore natural to consider the *Heegaard gradient* of a sequence of finite covers $\{M_i\}$, which is

$$\liminf_i \frac{\chi_-^h(M_i)}{\text{degree}(M_i \to M)}.$$

Somewhat surprisingly, the Cheeger constant and the Heegaard genus of a closed hyperbolic 3-manifold are related by the following inequality of the author [29].

Theorem 7.3. *Let M be a closed orientable hyperbolic 3-manifold. Then*

$$h(M) \leq \frac{8\pi(g(M)-1)}{\text{Volume}(M)}.$$

A consequence is that if the Heegaard gradient of a sequence of finite covers of M is zero, then the corresponding subgroups of $\pi_1(M)$ do not have Property (τ). Equivalently, if a sequence of finite covers has Property (τ), then these covers have positive Heegaard gradient.

We now give a sketch of the proof of Theorem 7.3. Any Heegaard splitting for a 3-manifold M determines a 'sweepout' of the manifold by surfaces, as follows. The Heegaard surface divides the manifold into two handlebodies, each of which is a regular neighbourhood of a core graph. Thus, there is a 1-parameter family of copies of the surface, starting with the boundary of a thin regular neighbourhood of one core graph and ending with the boundary of a thin regular neighbourhood of the other graph. Consider sweepouts where the maximum area of the surfaces is as small as possible. Then, using work of Pitts and Rubinstein [53], one can arrange that the surfaces of maximal area tend (in a certain sense) to a minimal surface S, which is obtained from the Heegaard surface possibly by performing some compressions. Since S is a minimal surface in a hyperbolic 3-manifold, Gauss-Bonnet implies that its area is at most $-2\pi\chi(S) \leq 4\pi(g(M)-1)$. Hence, we obtain a sweepout of M by surfaces, each of which has area at most this bound (plus an arbitrarily small $\epsilon > 0$). One of these surfaces divides M into two parts of equal volume. This decomposition gives the required upper bound on the Cheeger constant $h(M)$.

There is an important special case when the Heegaard gradient of a sequence of finite covers is zero. Suppose that M fibres over the circle with fibre F. Then it is easy to construct a Heegaard splitting for M with genus at most $2g(F)+1$, where $g(F)$ is the genus of F. Hence, the finite cyclic covers of M dual to F have uniformly bounded Heegaard genus. In particular, their Heegaard gradient is

zero. This is the only known method of constructing sequences of finite covers of a hyperbolic 3-manifold with zero Heegaard gradient. And so we are led to the following conjecture of the author, called the Heegaard Gradient Conjecture [29].

Conjecture. *A closed orientable hyperbolic 3-manifold has zero Heegaard gradient if and only if it virtually fibres over the circle.*

This remains a difficult open problem. However, a qualitative version of it is known to be true. More specifically, if a closed hyperbolic 3-manifold has a sequence of finite covers with Heegaard genus that grows 'sufficiently slowly', then these covers are eventually fibred, by the following result of the author [27].

Theorem 7.4. *Let M be a closed orientable hyperbolic 3-manifold, and let M_i be a sequence of finite regular covers, with degree d_i. Suppose that $g(M_i)/\sqrt[4]{d_i} \to 0$. Then, for all sufficiently large i, M_i fibres over the circle.*

The proof of this uses several of the results mentioned above. Using Theorem 7.3, the hypothesis that $g(M_i)/\sqrt[4]{d_i} \to 0$ implies that $h(M_i)d_i^{3/4} \to 0$. Hence, by Theorem 7.1, we deduce that some finite-sheeted cover of M has positive first Betti number. In fact, if we go back to the proof of Theorem 7.1, we see that this is true of each M_i sufficiently far down the sequence, and with further work, one can actually prove that these manifolds fibre over the circle.

We will see that the two conjectures introduced in this section, the Lubotzky-Sarnak Conjecture and the Heegaard Gradient Conjecture, may be a route to proving the Virtually Haken Conjecture.

8. Two Approaches to the Virtually Haken Conjecture

The two conjectures introduced in the previous section can be combined to form an approach to the Virtually Haken Conjecture, via the following theorem of the author [29].

Theorem 8.1. *Let M be a closed orientable irreducible 3-manifold, and let M_i be a tower of finite regular covers of M such that*

1. *their Heegaard gradient is positive, and*
2. *they do not have Property (τ).*

Then, for all sufficiently large i, M_i is Haken.

Hence, the Lubotzky-Sarnak Conjecture and the Heegaard Gradient Conjecture together imply the Virtually Haken Conjecture. For, assuming the

Lubotzky-Sarnak Conjecture, a closed orientable hyperbolic 3-manifold M has a tower of finite regular covers without Property (τ). If these have positive Heegaard gradient, then by Theorem 8.1, they are eventually Haken. On the other hand, if they have zero Heegaard gradient, then by the Heegaard Gradient Conjecture, M is virtually fibred.

The proof requires some ideas from the theory of Heegaard splittings. A central concept in this theory is the notion of a *strongly irreducible* Heegaard surface S, which means that any compression disc on one side of S must intersect any compression disc on the other side. A key theorem of Casson and Gordon [13] implies that if a closed orientable irreducible 3-manifold has a minimal genus Heegaard splitting that is *not* strongly irreducible, then the manifold is Haken. A quantified version of this is as follows. Suppose that S is a Heegaard surface for the closed 3-manifold M, and that there are d disjoint non-parallel compression discs on one side of S that are all disjoint from d disjoint non-parallel compression discs on the other side of S. Then, either $g(M) \leq g(S) - (d/6)$ or M is Haken.

Suppose now that M_i is a sequence of covers of M as in Theorem 8.1. Let S be a minimal genus Heegaard surface for M. Its inverse image in each M_i is a Heegaard surface S_i. We are assuming that the Heegaard gradient of these covers is positive. Hence (by replacing M by some M_i if necessary), we may assume that $g(M_i)$ is roughly $g(S_i)$. Now, we are also assuming that these covers do not have Property (τ). Hence, there is a way of decomposing M_i into two pieces A_i and B_i with large volume, and with small intersection. By using compression discs on one side of S_i that lie in A_i and compression discs on the other side of S_i that lie in B_i, we obtain d_i discs on each side of S_i which are all disjoint and non-parallel, and where d_i grows linearly as a function of the covering degree of $M_i \to M$. Hence, by the quantified version of Casson-Gordon, if M_i is not Haken, then $g(M_i)$ is substantially less than $g(S_i)$, which is a contradiction, thereby proving Theorem 8.1.

Of course, it remains unclear whether the hypotheses of Theorem 8.1 always hold, and hence the Virtually Haken Conjecture remains open. However, there is another intriguing approach. By using results of Bourgain and Gamburd [8] which give lower bounds on the first eigenvalue of the Laplacian on certain Cayley graphs of $\mathrm{SL}(2,p)$, Long, Lubotzky and Reid [39] were able to establish the following theorem.

Theorem 8.2. *Let M be a closed orientable hyperbolic 3-manifold. Then M has a sequence of finite covers M_i with Property (τ) and such that the subgroups $\pi_1(M_i)$ of $\pi_1(M)$ intersect in the identity.*

Combining this with Theorem 7.3, we deduce that these covers have positive Heegaard gradient. Now, Theorem 8.2 does not provide a *tower* of finite regular covers, but it is not unreasonable to suppose that this can be achieved. Hence, by replacing M by some M_i if necessary, we may assume that the Heegaard gradient of these covers is very close to $\chi_-^h(M)$. Let S be any minimal genus

Heegaard surface for M. Its inverse image S_i in M_i is a Heegaard splitting of M_i, and it therefore is nearly of minimal genus. It is reasonable to conjecture that there is a minimal genus splitting $\overline{S}_i$ for M_i with geometry 'approximating' that of S_i. Now, S_i inevitably fails to be strongly irreducible when the degree of $M_i \to M$ is large, via a simple argument that counts compression discs and their points of intersection. One might conjecture that this is also true of $\overline{S}_i$, which would therefore imply that M_i is Haken for all sufficiently large i. Of course, this is somewhat speculative, and the conjectural relationship between S_i and $\overline{S}_i$ may not hold. But it highlights the useful interaction between Heegaard splittings, Property (τ) and the Virtually Haken Conjecture.

9. Covering Spaces of Hyperbolic 3-orbifolds and Arithmetic 3-manifolds

The material in the previous section is, without doubt, rather speculative. However, the ideas behind it have been profitably applied in some important special cases. It seems to be easiest to make progress when analysing finite covers of either of the following spaces:

1. hyperbolic 3-orbifolds with non-empty singular locus;
2. arithmetic hyperbolic 3-manifolds.

There is a well-developed theory of orbifolds, their fundamental groups and their covering spaces. We will only give a very brief introduction here, and refer the reader to [57] for more details.

Recall that an orientable hyperbolic 3-orbifold O is the quotient of hyperbolic 3-space $\mathbb{H}^3$ by a discrete group Γ of orientation-preserving isometries. This group may have non-trivial torsion, in which case it does not act freely. The images in O of points in $\mathbb{H}^3$ with non-trivial stabiliser form the *singular locus* $\mathrm{sing}(O)$. This is a collection of 1-manifolds and trivalent graphs. Each 1-manifold and each edge of each graph has an associated positive integer, its *order*, which is the order of the finite stabiliser of corresponding points in $\mathbb{H}^3$. For any positive integer n, $\mathrm{sing}_n(O)$ denotes the closure of the union of singular edges and 1-manifolds that have order a multiple of n. The underlying topological space of a 3-orbifold O is always a 3-manifold, denoted $|O|$.

One can define the fundamental group $\pi_1(O)$ of any orbifold O, which is, in general, different from the usual fundamental group of $|O|$. When O is hyperbolic, and hence of the form $\mathbb{H}^3/\Gamma$, its fundamental group is Γ. One can also define the notion of a covering map between orbifolds. In the hyperbolic case, these maps are of the form $\mathbb{H}^3/\Gamma' \to \mathbb{H}^3/\Gamma$, for some subgroup Γ' of Γ. Note that this need not be a cover in the usual topological sense.

The following result of the author, Long and Reid [35] allows one to apply orbifold technology in the arithmetic case.

Theorem 9.1. *Any arithmetic hyperbolic 3-manifold is commensurable with a 3-orbifold O with non-empty singular locus. Indeed, one may arrange that every curve and arc of the singular locus has order 2 and that there is at least one singular vertex.*

The main reason why 3-orbifolds are often more tractable than 3-manifolds is the following lower bound on the rank of their homology [30]. For an orbifold O and prime p, we let $d_p(O)$ denote $d_p(\pi_1(O))$.

Theorem 9.2. *Let O be a compact orientable 3-orbifold. Then for any prime p, $d_p(O) \geq b_1(\mathrm{sing}_p(O))$.*

The reason is that $\pi_1(O)$ can be computed by starting with the usual fundamental group of the manifold $O - \mathrm{sing}(O)$ and then quotienting out powers of the meridians of the singular locus, where the power is the relevant edge's singularity order. If this order is a multiple of p, then quotienting out this power of this meridian has no effect on d_p. On the other hand, if the order of a singular edge or curve is coprime to p, then we may replace these points by manifold points without changing d_p. Hence, $d_p(O) = d_p(|O| - \mathrm{int}(N(\mathrm{sing}_p(O))))$. Now, the latter space is a compact orientable 3-manifold M with boundary, and it is a well-known consequence of Poincaré duality that $d_p(M)$ is at least $d_p(\partial M)/2$. From this, the required inequality rapidly follows.

So, as far as mod p homology is concerned, orbifolds O where $\mathrm{sing}_p(O)$ is non-empty behave as though they have non-empty boundary. And 3-manifolds with non-empty boundary are often much more tractable than closed ones.

Theorem 9.2 is the basis behind Theorem 6.1. Here, we are given a finite-volume hyperbolic 3-orbifold O with non-empty singular locus. The main case is when O is closed. Let p be a prime that divides the order of some edge or curve in the singular locus. We first show that one can find a finite cover $\tilde{O}$ where $\mathrm{sing}(\tilde{O})$ is a non-empty collection of simple closed curves with singularity order p, and where $d_p(\tilde{O}) \geq 11$, using techniques that are generalisations of those in Section 3. Let λ and μ be a longitude and meridian of some component L of $\mathrm{sing}(\tilde{O})$, viewed as elements of $\pi_1(\tilde{O})$. Then, using the Golod-Shafarevich inequality [40], we can show that $\pi_1(\tilde{O})/\langle\langle \lambda, \mu \rangle\rangle$ is infinite, and in fact has an infinite sequence of finite-index subgroups. These pull back to finite-index subgroups of $\pi_1(\tilde{O})$, which determine a sequence of covering spaces O_i. Because these subgroups contain the normal subgroup $\langle\langle \lambda, \mu \rangle\rangle$, the inverse image of L in each O_i is a disjoint union of copies of L. Hence, there is a linear lower bound on the number of components of $\mathrm{sing}_p(O_i)$ as a function of the covering degree. Therefore, by Theorem 9.2, $d_p(O_i)$ grows linearly, as required.

For any closed orientable 3-manifold M, there are obvious inequalities $g(M) \geq d(\pi_1(M)) \geq d_p(M)$, and the same is true for closed orientable 3-orbifolds (with an appropriate definition of Heegaard genus). Hence, Theorem 6.1 provides a sequence of finite covers of the orbifold O with positive Heegaard gradient. If we also knew that these covers did not have Property (τ), then by (an orbifold version of) Theorem 8.1, we would deduce that they are eventually

Haken. In fact, we would get much more than this. We would be able to deduce that $\pi_1(O)$ is large, via the following theorem of the author [32].

Theorem 9.3. *Let Γ be a finitely presented group, let p be a prime and suppose that $\Gamma \geq \Gamma_1 \triangleright \Gamma_2 \triangleright \dots$ is a sequence of finite-index subgroups, where each Γ_{i+1} is normal in Γ_i and has index a power of p. Suppose that*

1. *the subgroups Γ_i have positive mod p homology gradient, and*
2. *the subgroups Γ_i do not have Property (τ).*

Then Γ is large.

We will explain the proof of this and related results in the next section. Similar reasoning also gives the the following theorem [35].

Theorem 9.4. *The Lubotzky-Sarnak Conjecture implies that any closed hyperbolic 3-orbifold that has at least one singular vertex has large fundamental group. In particular, the Lubotzky-Sarnak Conjecture implies that every arithmetic hyperbolic 3-manifold has large fundamental group.*

It is quite striking that the Lubotzky-Sarnak Conjecture, which is a question solely about the spectrum of the Laplacian, should have such far-reaching consequences for arithmetic hyperbolic 3-manifolds.

The way that this is proved is as follows. One starts with the closed hyperbolic 3-orbifold O with at least one singular vertex. Its fundamental group therefore contains a finite non-cyclic subgroup. For simplicity, suppose that this is $\mathbb{Z}/2 \times \mathbb{Z}/2$ (which is the case considered in [35]). One can then pass to a finite cover $\tilde{O}$ where every arc and circle of the singular locus has order 2 and which has at least one singular vertex. Any finite cover of the underlying manifold $|\tilde{O}|$ induces a finite cover O_i of $\tilde{O}$ where $\mathrm{sing}(O_i)$ is the inverse image of $\mathrm{sing}(\tilde{O})$. Since $\mathrm{sing}(\tilde{O})$ contains a trivalent vertex, $b_1(\mathrm{sing}_2(O_i))$ grows linearly as a function of the covering degree. Hence, $\{\pi_1(O_i)\}$ has positive mod 2 homology gradient. With some further work, and using the solution to the Geometrisation Conjecture, we may arrange that $|\tilde{O}|$ has a hyperbolic structure or has a finite cover with positive b_1. Hence, assuming the Lubotzky-Sarnak Conjecture, one can find finite covers $|O_i|$ with Cheeger constants tending to zero. Thus, $\pi_1(O)$ is large, by Theorem 9.3.

The above arguments are closely related to those behind Theorem 2.1. In fact, we can prove the following stronger version [33].

Theorem 9.5. *Let Γ be the fundamental group of a finite-volume hyperbolic 3-orbifold or 3-manifold. Suppose that Γ has a finite non-cyclic subgroup or is arithmetic. Then Γ contains the fundamental group of a closed orientable surface with positive genus.*

The proof runs as follows. One uses the same finite cover $\tilde{O}$ as above. We do not know that the Lubotzky-Sarnak Conjecture holds, but we have Theorem

7.2, which provides a sequence of infinite-sheeted covers $|O_i|$ of $|\tilde{O}|$ with Cheeger constants tending to zero. These induce covers O_i of $\tilde{O}$. One can use the singular locus of O_i to find a finite cover with more than one end. It then follows quickly that $\pi_1(O_i)$ contains a surface subgroup.

To make further progress with finite covers, it seems to be necessary to establish the Lubotzky-Sarnak Conjecture. But there is an important special case where this holds trivially: when the manifold or orbifold has a finite cover with positive first Betti number. For example, suppose that O is a compact orientable 3-orbifold with singular locus that contains a simple closed curve C. Suppose also that there is a surjective homomorphism $\pi_1(O) \to \mathbb{Z}$ that sends $[C]$ to zero. Then the resulting finite cyclic covers have linear growth of mod p homology (where p divides the order of C) and also their Cheeger constants tend to zero. So, by Theorem 9.3, $\pi_1(O)$ is large. Using this observation, the author, Long and Reid were able to prove the following [35].

Theorem 9.6. *Let Γ be the fundamental group of a finite-volume hyperbolic 3-manifold or 3-orbifold. Suppose that Γ is arithmetic or contains $\mathbb{Z}/2 \times \mathbb{Z}/2$. Suppose also that Γ has a finite-index subgroup $\tilde{\Gamma}$ with $b_1(\tilde{\Gamma}) \geq 4$. Then Γ is large.*

This is significant because such a finite-index subgroup $\tilde{\Gamma}$ is known to exist in many cases. Indeed, arithmetic techniques, due to Clozel [15], Labesse-Schwermer [26], Lubotzky [44] and others, often provide a congruence subgroup with positive first Betti number. Then, using a theorem of Borel [6], one can find congruence subgroups with arbitrarily large first Betti number. The consequence of Theorem 9.6 is that one can in fact strengthen the conclusion to deduce that these groups are large.

10. Group-theoretic Generalisations

We have discussed several topological results, such as Theorem 8.1, which are helpful in tackling the Virtually Haken Conjecture. It is natural to ask whether there are more general group-theoretic versions of these theorems. In many cases, there are. For example, the following is a version of Theorem 8.1, due to the author [28].

Theorem 10.1. *Let Γ be a finitely presented group, and let $\{\Gamma_i\}$ be a nested sequence of finite-index normal subgroups. Suppose that*

1. *their rank gradient is positive, and*
2. *they do not have Property (τ).*

Then, for all sufficiently large i, Γ_i is an amalgamated free product or HNN extension.

We now give an indication of the proof. As is typical with arguments in this area, one starts with a finite cell complex K with fundamental group Γ. Let K_i be the finite covering space corresponding to Γ_i. The hypothesis that Γ does not have Property (τ) with respect to $\{\Gamma_i\}$ implies that one can form a decomposition of K_i into two sets B_i and C_i with large volume but small intersection. Via the Seifert - van Kampen theorem, this then determines a decomposition of Γ_i into a graph of groups. We must show that this is a non-trivial decomposition. In other words, we must ensure that neither $\pi_1(B_i)$ nor $\pi_1(C_i)$ surjects onto Γ_i. This is where the hypothesis that $\{\Gamma_i\}$ has positive rank gradient is used. The number of 1-cells of B_i (or C_i) gives an upper bound to the rank of $\pi_1(B_i)$, and this is a definite fraction of the total number of 1-cells of K_i. Hence if $\pi_1(B_i)$ or $\pi_1(C_i)$ were to surject onto Γ_i, one could use this to deduce that the rank of Γ_i was too small.

We have also seen Theorem 9.3, which starts with the stronger hypothesis of positive mod p homology gradient, and which ends with the strong conclusion of largeness. The proof follows similar lines, but now the goal is to show that neither $H_1(B_i;\mathbb{F}_p)$ nor $H_1(C_i;\mathbb{F}_p)$ surjects onto $H_1(\Gamma_i;\mathbb{F}_p)$. Instead of using the Seifert - van Kampen theorem, the Meyer-Vietoris theorem is used. One deduces that if $H_1(B_i;\mathbb{F}_p)$ or $H_1(C_i;\mathbb{F}_p)$ were to surject onto $H_1(\Gamma_i;\mathbb{F}_p)$, then $H_1(\Gamma_i;\mathbb{F}_p)$ would be too small, contradicting the assumption that the subgroups Γ_i have positive mod p homology gradient. Hence, again we get a graph of groups decomposition for Γ_i. This induces a graph of groups decomposition for $\tilde{\Gamma}_i = [\Gamma_i,\Gamma_i](\Gamma_i)^p$. Its underlying graph has valence at least p at each vertex. And $\tilde{\Gamma}_i$ surjects onto the fundamental group of this graph, which is a non-abelian free group (when $p \neq 2$), as required.

One might wonder whether Theorem 9.3 remains true even if we do not assume that the subgroups $\{\Gamma_i\}$ do not have Property (τ). The above proof breaks down. But is the hypothesis that the subgroups Γ_i have positive mod p homology gradient enough to deduce largeness? As mentioned in Section 6, this question relates to error-correcting codes. More details can be found in [32]. However, there is one interesting and natural situation where the hypothesis of positive mod p homology gradient is enough to deduce largeness, according to the following theorem of the author [34].

Theorem 10.2. *Let Γ be a finitely presented group. Suppose that its derived p-series has positive mod p homology gradient. Then Γ is large.*

The main part of the proof is showing that if the derived p-series of Γ has positive mod p homology gradient, then it does not have Property (τ). Hence, by Theorem 9.3, Γ is large. Once again, let K be a finite 2-complex with fundamental group Γ, and let K_i be the covering space corresponding to the subgroup Γ_i in the derived p-series. One needs to show that the Cheeger constant of K_i is arbitrarily small. This is achieved by finding non-trivial 1-cocycles on K_{i-1} with small support size, compared with the total number of edges of K_{i-1}. If $\tilde{K}_{i-1}$ denotes the cyclic covering space dual to such a cocycle,

then the inverse image of the cocycle determines a decomposition of $\tilde{K}_{i-1}$ into two parts with large volume and small intersection. Since K_i finitely covers $\tilde{K}_{i-1}$, it too has small Cheeger constant. In fact, one keeps track of not just one cocycle on K_{i-1}, but several of them, and one uses these to create cocycles on K_i with slightly smaller relative support size, and so on. This is achieved using the technology explained in Section 4 for constructing cocycles on abelian covers, together with an elementary theorem from coding theory, known as the Plotkin bound.

11. Subgroup Separability, Special Cube Complexes and Virtual Fibering

There are many other interesting directions in the theory of finite covers of 3-manifolds, which we can only briefly discuss here.

The first of these is the notion of subgroup separability. A subgroup H of a group Γ is *separable* if for every element $\gamma \in \Gamma$ that does not lie in H, there is a homomorphism ϕ from Γ onto a finite group such that $\phi(\gamma) \notin \phi(H)$. A group Γ is said to be *LERF* if every finitely generated subgroup is separable. The relevance of this concept to 3-manifolds arises from the following theorem [38].

Theorem 11.1. *Let M be a compact orientable irreducible 3-manifold, and suppose that $\pi_1(M)$ has a separable subgroup that is isomorphic to the fundamental group of a closed orientable surface with positive genus. Then either M is virtually fibred or $\pi_1(M)$ is large. In particular, M is virtually Haken.*

This raises the question of which 3-manifolds have LERF fundamental group. There are examples of certain graph 3-manifolds M for which $\pi_1(M)$ is not LERF [10]. But it is conjectured that the fundamental group of every closed hyperbolic 3-manifold is LERF. A piece of evidence for this conjecture is given by the following important theorem, which is an amalgamation of work by Agol, Long and Reid [4] and Bergeron, Haglund and Wise [5].

Theorem 11.2. *Let M be an arithmetic hyperbolic 3-manifold which contains a closed immersed totally geodesic surface. Then every geometrically finite subgroup of $\pi_1(M)$ is separable.*

There is an important new concept, introduced by Haglund and Wise [24], that relates to subgroup separability. They considered a certain type of cell complex, known as a *special cube complex*. A group is said to be *virtually special* if it has a finite index subgroup which is the fundamental group of a compact special cube complex. One major motivation for introducing this concept is the following theorem of Haglund and Wise [24].

Theorem 11.3. *Let Γ be a word-hyperbolic group that is virtually special. Then every quasi-convex subgroup of Γ is separable.*

In the 3-dimensional case, it is known that this condition is equivalent to having 'enough' surface subgroups that are separable. Indeed Theorem 11.2 is proved in the case when M is closed by using the surface subgroups arising from totally geodesic surfaces to deduce that $\pi_1(M)$ is virtually special.

This is related to work of Agol [1]. He introduced a condition on a group, called RFRS. We will not give the definition of this here, but we note that if a group is virtually special then it is virtually RFRS. Agol was able to show that this condition can be used to prove that a 3-manifold virtually fibres over the circle.

Theorem 11.4. *Let M be a compact orientable irreducible 3-manifold with boundary a (possibly empty) collection of tori. Suppose that $\pi_1(M)$ is virtually RFRS. Then M has a finite cover that fibres over the circle.*

The hypotheses that $\pi_1(M)$ is virtually RFRS or virtually special are strong ones. However, Wise has recently raised the possibility of showing that if a compact orientable hyperbolic 3-manifold M has a properly embedded orientable incompressible surface that is not a sphere or a virtual fibre, then $\pi_1(M)$ is virtually special, by using induction along a hierarchy for M. While this would not say anything about the Virtually Haken Conjecture itself, it would be a very major development, as it would nearly reduce all the other conjectures to it. For example, combined with Theorem 11.4, it would show that every finite-volume orientable hyperbolic Haken 3-manifold is virtually fibred.

References

[1] I. Agol, *Criteria for virtual fibering.* J. Topol. 1 (2008), no. 2, 269–284.

[2] I. Agol, *Virtual betti numbers of symmetric spaces*, arxiv:math.GT/0611828

[3] I. Agol, S. Boyer, X. Zhang, *Virtually fibered Montesinos links.* J. Topol. 1 (2008), no. 4, 993–1018.

[4] I. Agol, D. Long, A. Reid, *The Bianchi groups are separable on geometrically finite subgroups.* Ann. of Math. (2) 153 (2001), no. 3, 599–621

[5] N. Bergeron, F. Haglund, D. Wise, *Hyperbolic sections in arithmetic hyperbolic manifolds*, Preprint.

[6] A. Borel, *Cohomologie de sous-groupes discrets et représentations de groupes semi-simples,* in Colloque "Analyse et Topologie" en l'Honneur de Henri Cartan (Orsay, 1974), 73–112, Astrisque, 32–33, Soc. Math. France, Paris, 1976

[7] N. Boston, J. Ellenberg, *Pro-p groups and towers of rational homology spheres.* Geom. Topol. 10 (2006), 331–334

[8] J. Bourgain, A. Gamburd, *Uniform expansion bounds for Cayley graphs of* $\mathrm{SL}_2(\mathbb{F}_p)$. Ann. of Math. (2) 167 (2008), no. 2, 625–642.

[9] L. Bowen, *Free groups in lattices.* Geom. Topol. 13 (2009), no. 5, 3021–3054.

[10] R. Burns, A. Karrass, D. Solitar, *A note on groups with separable finitely generated subgroups.* Bull. Austral. Math. Soc. 36 (1987), no. 1, 153–160.

[11] P. Buser, *A note on the isoperimetric constant.* Ann. Sci. École Norm. Sup. (4) 15 (1982), no. 2, 213–230.

[12] F. Calegari, N. Dunfield, *Automorphic forms and rational homology 3-spheres*, Geom. Topol. 10 (2006), 295–329.

[13] A. Casson, C. Gordon, *Reducing Heegaard splittings,* Topology Appl. 27 (1987), no. 3, 275–283.

[14] J. Cheeger, *A lower bound for the smallest eigenvalue of the Laplacian.* Problems in analysis (Papers dedicated to Salomon Bochner, 1969), pp. 195–199. Princeton Univ. Press, Princeton, N. J., 1970.

[15] L. Clozel, *On the cuspidal cohomology of arithmetic subgroups of* $\mathrm{SL}(2n)$ *and the first Betti number of arithmetic* 3*-manifolds.* Duke Math. J. 55 (1987), no. 2, 475–486

[16] D. Cooper, D. Long, *Virtually Haken Dehn-filling.* J. Differential Geom. 52 (1999), no. 1, 173–187.

[17] D. Cooper, D. Long, A. Reid, *Essential closed surfaces in bounded* 3*-manifolds.* J. Amer. Math. Soc. 10 (1997), no. 3, 553–563.

[18] D. Cooper, D. Long, A. Reid, *On the virtual Betti numbers of arithmetic hyperbolic 3-manifolds.* Geom. Topol. 11 (2007), 2265–2276.

[19] D. Cooper, G. Walsh, *Virtually Haken fillings and semi-bundles.* Geom. Topol. 10 (2006), 2237–2245

[20] D. Cooper, G. Walsh, *Three-manifolds, virtual homology, and group determinants.* Geom. Topol. 10 (2006), 2247–2269

[21] N. Dunfield, W. Thurston, *The virtual Haken conjecture: experiments and examples.* Geom. Topol. 7 (2003) 399–441.

[22] F. González-Acuña, H. Short, *Cyclic branched coverings of knots and homology spheres*, Revista Math. 4 (1991) 97–120.

[23] R. Grigorchuk, *Degrees of growth of finitely generated groups and the theory of invariant means.* Izv. Akad. Nauk SSSR Ser. Mat. 48 (1984), no. 5, 939–985.

[24] F. Haglund, D. Wise, *Special cube complexes.* Geom. Funct. Anal. 17 (2008), no. 5, 1551–1620.

[25] J. Kahn, V. Markovic, *Immersing almost geodesic surfaces in a closed hyperbolic three manifold*, arXiv:0910.5501

[26] J.-P. Labesse, J. Schwermer, *On liftings and cusp cohomology of arithmetic groups.* Invent. Math. 83 (1986), no. 2, 383–401.

[27] M. Lackenby, *The asymptotic behaviour of Heegaard genus*, Math. Res. Lett. 11 (2004) 139-149

[28] M. Lackenby, *Expanders, rank and graphs of groups*, Israel J. Math. 146 (2005) 357–370.

[29] M. Lackenby, *Heegaard splittings, the virtually Haken conjecture and Property* (τ), Invent. Math. 164 (2006) 317–359.

[30] M. Lackenby, *Covering spaces of 3-orbifolds*, Duke Math J. 136 (2007) 181–203.

[31] M. Lackenby, *New lower bounds on subgroup growth and homology growth*, Proc. London Math. Soc. 98 (2009) 271–297.

[32] M. Lackenby, *Large groups, Property (τ) and the homology growth of subgroups*, Math. Proc. Camb. Phil. Soc. 146 (2009) 625–648.

[33] M. Lackenby, *Surface subgroups of Kleinian groups with torsion*, Invent. Math. 179 (2010) 175–190.

[34] M. Lackenby, *Detecting large groups*, arxiv:math.GR/0702571

[35] M. Lackenby, D. Long, A. Reid, *Covering spaces of arithmetic 3-orbifolds*, Int. Math. Res. Not. (2008)

[36] M. Lackenby, D. Long, A. Reid, *LERF and the Lubotzky-Sarnak conjecture*, Geom. Topol. 12 (2008) 2047–2056.

[37] C. Leininger, *Surgeries on one component of the Whitehead link are virtually fibered.* Topology 41 (2002), no. 2, 307–320

[38] D. Long, *Immersions and embeddings of totally geodesic surfaces.* Bull. London Math. Soc. 19 (1987) 481–484.

[39] D. Long, A. Lubotzky, A. Reid, *Heegaard genus and property τ for hyperbolic 3-manifolds.* J. Topol. 1 (2008), no. 1, 152–158.

[40] A. Lubotzky, *Group presentation, p-adic analytic groups and lattices in* $\mathrm{SL}_2(C)$. Ann. of Math. (2) 118 (1983), no. 1, 115–130

[41] A. Lubotzky, *On finite index subgroups of linear groups.* Bull. London Math. Soc. 19 (1987), no. 4, 325–328.

[42] A. Lubotzky, *Discrete groups, expanding graphs and invariant measures.* With an appendix by Jonathan D. Rogawski. Progress in Mathematics, 125. Birkhuser Verlag, Basel, 1994

[43] A. Lubotzky, *Subgroup growth and congruence subgroups.* Invent. Math. 119 (1995), no. 2, 267–295.

[44] A. Lubotzky, *Eigenvalues of the Laplacian, the first Betti number and the congruence subgroup problem.* Ann. of Math. (2) 144 (1996), no. 2, 441–452.

[45] A. Lubotzky, D. Segal, *Subgroup growth.* Progress in Mathematics, 212. Birkhäuser Verlag, Basel, 2003.

[46] A. Lubotzky, A. Zuk, *On Property* (τ), Preprint.

[47] W. Lück, *Approximating L^2-invariants by their finite-dimensional analogues,* Geom. Funct. Anal. 4 (1994), no. 4, 455–481.

[48] C. Maclachlan, A. Reid, *The arithmetic of hyperbolic 3-manifolds.* Graduate Texts in Mathematics, 219. Springer-Verlag, New York, 2003.

[49] G. Margulis, *Explicit constructions of expanders.* Problemy Peredaci Informacii 9 (1973), no. 4, 71–80.

[50] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications,* arxiv:math.DG/0211159

[51] G. Perelman, *Ricci flow with surgery on three-manifolds,* arxiv:math.DG/0303109

[52] G. Perelman, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds,* arxiv:math.DG/0307245

[53] J. Pitts, J. H. Rubinstein, *Applications of minimax to minimal surfaces and the topology of* 3*-manifolds.* Miniconference on geometry and partial differential equations, 2 (Canberra, 1986), 137–170, Proc. Centre Math. Anal. Austral. Nat. Univ., 12, Austral. Nat. Univ., Canberra, 1987.

[54] A. Reid, *A non-Haken hyperbolic* 3*-manifold covered by a surface bundle.* Pacific J. Math. 167 (1995), no. 1, 163–182.

[55] R. Riley, *Growth of order of homology of cyclic branched covers of knots,* Bull. London Math. Soc. 22 (1990) 287–297.

[56] M. Scharlemann, *Heegaard splittings of compact 3-manifolds.* Handbook of geometric topology, 921–953, North-Holland, Amsterdam, 2002.

[57] P. Scott, *The geometries of* 3*-manifolds.* Bull. London Math. Soc. 15 (1983) 401–487.

[58] A. Selberg, *On the estimation of fourier coefficients of modular forms*, Proc. Symp. Pure Math. VIII (1965) 1–15.

[59] P. Shalen, P. Wagreich, *Growth rates, Z_p-homology, and volumes of hyperbolic* 3*-manifolds*, Trans. Amer. Math. Soc. 331 (1992), no. 2, 895–917.

[60] D. Silver, S. Williams, *Torsion numbers of augmented groups with applications to knots and links.* Enseign. Math. (2) 48 (2002), no. 3–4, 317–343.

[61] D. Sullivan, *Related aspects of positivity in Riemannian geometry.* J. Differential Geom. 25 (1987), no. 3, 327–351.

[62] W. Thurston, *Three-dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. Amer. Math. Soc. (N.S.) 6 (1982), no. 3, 357–381.

[63] T. Venkataramana, *Virtual Betti numbers of compact locally symmetric spaces.* Israel J. Math. 166 (2008), 235–238.

[64] B. Weisfeiler, *Strong approximation for Zariski-dense subgroups of semisimple algebraic groups.* Ann. of Math. (2) 120 (1984), no. 2, 271–315.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

K- and L-theory of Group Rings

Wolfgang Lück*

Abstract

This article will explore the K- and L-theory of group rings and their applications to algebra, geometry and topology. The Farrell-Jones Conjecture characterizes K- and L-theory groups. It has many implications, including the Borel and Novikov Conjectures for topological rigidity. Its current status, and many of its consequences are surveyed.

Mathematics Subject Classification (2010). Primary 18F25; Secondary 57XX.

Keywords. K- and L-theory, group rings, Farrell-Jones Conjecture, topological rigidity.

0. Introduction

The algebraic K- and L-theory of group rings — $K_n(RG)$ and $L_n(RG)$ for a ring R and a group G — are highly significant, but are very hard to compute when G is infinite. The main ingredient for their analysis is the Farrell-Jones Conjecture. It identifies them with certain equivariant homology theories evaluated on the classifying space for the family of virtually cyclic subgroups of G. Roughly speaking, the Farrell-Jones Conjecture predicts that one can compute the values of these K- and L-groups for RG if one understands all of the values for RH, where H runs through the virtually cyclic subgroups of G.

Why is the Farrell-Jones Conjecture so important? One reason is that it plays an important role in the classification and geometry of manifolds. A second reason is that it implies a variety of well-known conjectures, such as the ones due to Bass, Borel, Kaplansky and Novikov. (These conjectures are explained in Section 1.) There are many groups for which these conjectures were

*The work was financially supported by the Leibniz-Preis of the author. The author wishes to thank several members and guests of the topology group in Münster for helpful comments.
Mathematisches Institut der Westfälische Wilhelms-Universität, Einsteinstr. 62, 48149 Münster, Germany. E-mail: lueck@math.uni-muenster.de
URL: http://www.math.uni-muenster.de/u/lueck.

previously unknown but are now consequences of the proof that they satisfy the Farrell-Jones Conjecture. A third reason is that most of the explicit computations of K- and L-theory of group rings for infinite groups are based on the Farrell-Jones Conjecture, since it identifies them with equivariant homology groups which are more accessible via standard tools from algebraic topology and geometry (see Section 5).

The rather complicated general formulation of the Farrell-Jones Conjecture is given in Section 3. The much easier, but already very interesting, special case of a torsionfree group is discussed in Section 2. In this situation the K- and L-groups are identified with certain homology theories applied to the classifying space BG.

The recent proofs of the Farrell-Jones Conjecture for hyperbolic groups and CAT(0)-groups are deep and technically very involved. Nonetheless, we give a glimpse of the key ideas in Section 6. In each of these proofs there is decisive input coming from the geometry of the groups that is reminiscent of non-positive curvature. In order to exploit these geometric properties one needs to employ controlled topology and construct flow spaces that mimic the geodesic flow on a Riemannian manifold.

The class of groups for which the Farrell-Jones Conjecture is known is further extended by the fact that it has certain inheritance properties. For instance, subgroups of direct products of finitely many hyperbolic groups and directed colimits of hyperbolic groups belong to this class. Hence, there are many examples of exotic groups, such as groups with expanders, that satisfy the Farrell-Jones Conjecture because they are constructed as such colimits. There are of course groups for which the Farrell-Jones Conjecture has not been proved, like solvable groups, but there is no example or property of a group known that threatens to produce a counterexample. Nevertheless, there may well be counterexamples and the challenge is to develop new tools to find and construct them.

The status of the Farrell-Jones Conjecture is given in Section 4, and open problems are discussed in Section 7.

1. Some Well-known Conjectures

In this section we briefly recall some well-known conjectures. They address topics from different areas, including topology, algebra and geometric group theory. They have one — at first sight not at all obvious — common feature. Namely, their solution is related to questions about the K- and L-theory of group rings.

1.1. Borel Conjecture. A closed manifold M is said to be *topologically rigid* if every homotopy equivalence from a closed manifold to M is homotopic to a homeomorphism. In particular, if M is topologically rigid, then every manifold homotopy equivalent to M is homeomorphic to M. For example, the

spheres S^n are topologically rigid, as predicted by the *Poincaré Conjecture*. A connected manifold is called *aspherical* if its homotopy groups in degree ≥ 2 are trivial. A sphere S^n for $n \geq 2$ has trivial fundamental group, but its higher homotopy groups are very complicated. Aspherical manifolds, on the other hand, have complicated fundamental groups and trivial higher homotopy groups. Examples of closed aspherical manifolds are closed Riemannian manifolds with non-positive sectional curvature, and double quotients $G\backslash L/K$ for a connected Lie group L with $K \subseteq L$ a maximal compact subgroup and $G \subseteq L$ a torsionfree cocompact discrete subgroup. More information about aspherical manifolds can be found, for instance, in [59].

Conjecture 1.1 (Borel Conjecture). *Closed aspherical manifolds are topologically rigid.*

In particular the Borel Conjecture predicts that two closed aspherical manifolds are homeomorphic if and only if their fundamental groups are isomorphic. Hence the Borel Conjecture may be viewed as the topological version of *Mostow rigidity*. One version of Mostow rigidity says that two hyperbolic closed manifolds of dimension ≥ 3 are isometrically diffeomorphic if and only if their fundamental groups are isomorphic.

It is not true that any homotopy equivalence of aspherical closed smooth manifolds is homotopic to a diffeomorphism. The n-dimensional torus for $n \geq 5$ yields a counterexample (see [88, 15A]). Counterexamples with sectional curvature pinched arbitrarily close to -1 are given in [29, Theorem 1.1].

For more information about topologically rigid manifolds which are not necessarily aspherical, the reader is referred to [48].

1.2. Fundamental groups of closed manifolds. The Borel Conjecture is a uniqueness result. There is also an existence part. The problem is to determine when a given group G is the fundamental group of a closed aspherical manifold. Let us collect some obvious conditions that a group G must satisfy so that $G = \pi_1(M)$ for a closed aspherical manifold M. It must be finitely presented, since the fundamental group of any closed manifold is finitely presented. Since the cellular $\mathbb{Z}G$-chain complex of the universal covering of M yields a finite free $\mathbb{Z}G$-resolution of the trivial $\mathbb{Z}G$-module $\mathbb{Z}$, the group G must be of type FP, i.e., the trivial $\mathbb{Z}G$-module $\mathbb{Z}$ possesses a finite projective $\mathbb{Z}G$-resolution. Since $\widetilde{M}$ is a model for the classifying G-space EG, Poincaré duality implies $H^i(G;\mathbb{Z}G) \cong H_{\dim(M)-i}(\widetilde{M};\mathbb{Z})$, where $H^i(G;\mathbb{Z}G)$ is the cohomology of G with coefficients in the $\mathbb{Z}G$-module $\mathbb{Z}G$ and $H_i(\widetilde{M};\mathbb{Z})$ is the homology of $\widetilde{M}$ with integer coefficients. Since $\widetilde{M}$ is contractible, $H^i(G;\mathbb{Z}G) = 0$ for $i \neq \dim(M)$ and $H^{\dim(M)}(G;\mathbb{Z}G) \cong \mathbb{Z}$. Thus, a group G is called a *Poincaré duality group of dimension* n if G is finitely presented, is of type FP, $H^i(G;\mathbb{Z}G) = 0$ for $i \neq n$, and $H^n(G;\mathbb{Z}G) \cong \mathbb{Z}$.

Conjecture 1.2 (Poincaré duality groups). *A group G is the fundamental group of a closed aspherical manifold of dimension n if and only if G is a Poincaré duality group of dimension n.*

For more information about Poincaré duality groups, see [25, 42, 87].

1.3. Novikov Conjecture. Let G be a group and $u\colon M \to BG$ be a map from a closed oriented smooth manifold M to BG. Let $\mathcal{L}(M) \in \prod_{k\geq 0} H^k(M;\mathbb{Q})$ be the *L-class of M*, which is a certain polynomial in the Pontrjagin classes. Therefore it depends, a priori, on the tangent bundle and hence on the differentiable structure of M. For $x \in \prod_{k\geq 0} H^k(BG;\mathbb{Q})$, define the *higher signature of M associated to x and u* to be the rational number

$$\operatorname{sign}_x(M,u) \quad := \quad \langle \mathcal{L}(M) \cup u^*x, [M]\rangle.$$

We say that sign_x for $x \in \prod_{n\geq 0} H^n(BG;\mathbb{Q})$ is *homotopy invariant* if, for two closed oriented smooth manifolds M and N with reference maps $u\colon M \to BG$ and $v\colon N \to BG$, we have

$$\operatorname{sign}_x(M,u) = \operatorname{sign}_x(N,v)$$

whenever there is an orientation preserving homotopy equivalence $f\colon M \to N$ such that $v \circ f$ and u are homotopic.

Conjecture 1.3 (Novikov Conjecture). *Let G be a group. Then sign_x is homotopy invariant for all $x \in \prod_{k\geq 0} H^k(BG;\mathbb{Q})$.*

The Hirzebruch signature formula says that for $x = 1$ the signature $\operatorname{sign}_1(M,c)$ coincides with the ordinary signature $\operatorname{sign}(M)$ of M if $\dim(M) = 4n$, and is zero if $\dim(M)$ is not divisible by four. Obviously $\operatorname{sign}(M)$ depends only on the oriented homotopy type of M and hence the Novikov Conjecture 1.3 is true for $x = 1$.

A consequence of the Novikov Conjecture 1.3 is that for a homotopy equivalence $f\colon M \to N$ of orientable closed manifolds, we get $f_*\mathcal{L}(M) = \mathcal{L}(N)$ provided M and N are aspherical. This is surprising since it is not true in general. Often the L-classes are used to distinguish the homeomorphism or diffeomorphism types of homotopy equivalent closed manifolds. However, if one believes in the Borel Conjecture 1.1, then the map f above is homotopic to a homeomorphism and a celebrated result of Novikov [69] on the topological invariance of rational Pontrjagin classes says that $f_*\mathcal{L}(M) = \mathcal{L}(N)$ holds for any homeomorphism of closed manifolds.

For more information about the Novikov Conjecture, see, for instance, [37, 47].

1.4. Kaplansky Conjecture. Let F be a field of characteristic zero. Consider a group G. Let $g \in G$ be an element of finite order $|g|$. Set $N_g = \sum_{i=1}^{|g|} g^i$. Then $N_g \cdot N_g = |g| \cdot N_g$. Hence $x = N_g/|g|$ is an idempotent, i.e., $x^2 = x$. There are no other constructions known to produce idempotents different from 0 in FG. If G is torsionfree, this construction yields only the obvious idempotent 1. This motivates:

Conjecture 1.4 (Kaplansky Conjecture). *Let F be a field of characteristic zero and let G be a torsionfree group. Then the group ring FG contains no idempotents except* 0 *and* 1.

1.5. Hyperbolic groups with spheres as boundary. Let G be a hyperbolic group. One can assign to G its boundary ∂G. For information about the boundaries of hyperbolic groups, the reader is referred to [16, 43, 60]. Let M be an n-dimensional closed connected Riemannian manifold with negative sectional curvature. Then its fundamental group $\pi_1(M)$ is a hyperbolic group. The exponential map at a point $x \in M$ yields a diffeomorphism $\exp\colon T_x\mathbb{R}^n \to M$, which sends 0 to x, and a linear ray emanating from 0 in $T_x\mathbb{R}^n \cong \mathbb{R}^n$ is mapped to a geodesic ray in M emanating from x. Hence, it is not surprising that the boundary of $\pi_1(M)$ is $S^{\dim(M)-1}$. This motivates (see Gromov [38, page 192]):

Conjecture 1.5 (Hyperbolic groups with spheres as boundary). *Let G be a hyperbolic group whose boundary ∂G is homeomorphic to S^{n-1}. Then G is the fundamental group of an aspherical closed manifold of dimension n.*

This conjecture has been proved for $n \geq 6$ by Bartels-Lück-Weinberger [9] using the proof of the Farrell-Jones Conjecture for hyperbolic groups (see [4]) and the topology of homology ANR-manifolds (see, for example, [17, 76]).

1.6. Vanishing of the reduced projective class group. Let R be an (associative) ring (with unit). Define its *projective class group* $K_0(R)$ to be the abelian group whose generators are isomorphism classes $[P]$ of finitely generated projective R-modules P, and whose relations are $[P_0] + [P_2] = [P_1]$ for any exact sequence $0 \to P_0 \to P_1 \to P_2 \to 0$ of finitely generated projective R-modules. Define the *reduced projective class group* $\widetilde{K}_0(R)$ to be the quotient of $K_0(R)$ by the abelian subgroup $\{[R^m] - [R^n] \mid n, m \in \mathbb{Z}, m, n \geq 0\}$, which is the same as the abelian subgroup generated by the class $[R]$.

Let P be a finitely generated projective R-module. Then its class $[P] \in \widetilde{K}_0(R)$ is trivial if and only if P is *stably free*, i.e., $P \oplus R^r \cong R^s$ for appropriate integers $r, s \geq 0$. So the reduced projective class group $\widetilde{K}_0(R)$ measures the deviation of a finitely generated projective R-module from being stably free. Notice that stably free does not, in general, imply free.

A ring R is called *regular* if it is Noetherian and every R-module has a finite-dimensional projective resolution. Any principal ideal domain, such as $\mathbb{Z}$ or a field, is regular.

Conjecture 1.6 (Vanishing of the reduced projective class group). *Let R be a regular ring and let G be a torsionfree group. Then the change of rings homomorphism*

$$K_0(R) \to K_0(RG)$$

is an isomorphism.

In particular $\widetilde{K}_0(RG)$ vanishes for every principal ideal domain R and every torsionfree group G.

The vanishing of $\widetilde{K}_0(RG)$ contains valuable information about the finitely generated projective RG-modules over RG. In the case $R = \mathbb{Z}$, it also has the following important geometric interpretation.

Let X be a connected CW-complex. It is called *finite* if it consists of finitely many cells, or, equivalently, if X is compact. It is called *finitely dominated* if there is a finite CW-complex Y, together with maps $i\colon X \to Y$ and $r\colon Y \to X$, such that $r \circ i$ is homotopic to the identity on X. The fundamental group of a finitely dominated CW-complex is always finitely presented. While studying existence problems for spaces with prescribed properties (like group actions, for example), it is occasionally relatively easy to construct a finitely dominated CW-complex within a given homotopy type, whereas it is not at all clear whether one can also find a homotopy equivalent *finite* CW-complex. *Wall's finiteness obstruction*, a certain obstruction element $\widetilde{o}(X) \in \widetilde{K}_0(\mathbb{Z}\pi_1(X))$, decides this question.

The vanishing of $\widetilde{K}_0(\mathbb{Z}G)$, as predicted in Conjecture 1.6 for torsionfree groups, has the following interpretation: For a finitely presented group G, the vanishing of $\widetilde{K}_0(\mathbb{Z}G)$ is equivalent to the statement that any connected finitely dominated CW-complex X with $G \cong \pi_1(X)$ is homotopy equivalent to a finite CW-complex.

For more information about the finiteness obstruction, see [35, 49, 67, 86].

1.7. Vanishing of the Whitehead group. The *first algebraic K-group* $K_1(R)$ of a ring R is defined to be the abelian group whose generators $[f]$ are conjugacy classes of automorphisms $f\colon P \to P$ of finitely generated projective R-modules P and has the following relations. For each exact sequence $0 \to (P_0, f_0) \to (P_1, f_1) \to (P_2, f_2) \to 0$ of automorphisms of finitely generated projective R-modules, there is the relation $[f_0] - [f_1] + [f_2] = 0$; and for every two automorphisms $f, g\colon P \to P$ of the same finitely generated projective R-module, there is the relation $[f \circ g] = [f] + [g]$. Equivalently, $K_1(R)$ is the abelianization of the general linear group $GL(R) = \operatorname{colim}_{n\to\infty} GL_n(R)$.

An invertible matrix A over R represents the trivial element in $K_1(R)$ if it can be transformed by elementary row and column operations and by stabilization, $A \to A \oplus 1$ or the inverse, to the empty matrix.

Let G be a group, and let $\{\pm g \mid g \in G\}$ be the subgroup of $K_1(\mathbb{Z}G)$ given by the classes of $(1,1)$-matrices of the shape $(\pm g)$ for $g \in G$. The *Whitehead group* $\mathrm{Wh}(G)$ *of* G is the quotient $K_1(\mathbb{Z}G)/\{\pm g \mid g \in G\}$.

Conjecture 1.7 (Vanishing of the Whitehead group). *The Whitehead group of a torsionfree group vanishes.*

This conjecture has the following geometric interpretation.

An *n-dimensional cobordism* $(W; M_0, M_1)$ consists of a compact oriented n-dimensional smooth manifold W together with a disjoint decomposition $\partial W = M_0 \coprod M_1$ of the boundary ∂W of W. It is called an *h-cobordism* if the inclusions $M_i \to W$ for $i = 0, 1$ are homotopy equivalences. An h-cobordism $(W; M_0, M_1)$ is trivial if it is diffeomorphic relative M_0 to the trivial h-cobordism $(M_0 \times [0,1], M_0 \times \{0\}, M_0 \times \{1\})$. One can assign to an h-cobordism its *Whitehead torsion* $\tau(W, M_0)$ in $\mathrm{Wh}(\pi_1(M_0))$.

Theorem 1.8 (s-Cobordism Theorem). *Let M_0 be a closed connected oriented smooth manifold of dimension $n \geq 5$ with fundamental group $\pi = \pi_1(M_0)$. Then:*

(i) An h-cobordism $(W; M_0, M_1)$ is trivial if and only if its Whitehead torsion $\tau(W, M_0) \in \mathrm{Wh}(\pi)$ vanishes;

(ii) For any $x \in \mathrm{Wh}(\pi)$ there is an h-cobordism $(W; M_0, M_1)$ with $\tau(W, M_0) = x \in \mathrm{Wh}(\pi)$.

The s-Cobordism Theorem 1.8 is due to Barden, Mazur, Stallings. Its topological version was proved by Kirby and Siebenmann [45, Essay II]. More information about the s-Cobordism Theorem can be found, for instance, in [44], [52, Chapter 1], [66]. The Poincaré Conjecture of dimension ≥ 5 is a consequence of the s-Cobordism Theorem 1.8. The s-Cobordism Theorem 1.8 is an important ingredient in the surgery theory due to Browder, Novikov, Sullivan and Wall, which is the main tool for the classification of manifolds.

The s-Cobordism Theorem tells us that the vanishing of the Whitehead group, as predicted in Conjecture 1.7, has the following geometric interpretation: For a finitely presented group G the vanishing of the Whitehead group $\mathrm{Wh}(G)$ is equivalent to the statement that every h-cobordism W of dimension ≥ 6 with fundamental group $\pi_1(W) \cong G$ is trivial.

1.8. The Bass Conjecture.

For a finite group G there is a well-known fact that the homomorphism from the complexification of the complex representation ring of G to the $\mathbb{C}$-algebra of complex-valued class functions on G, given by taking the character of a finite-dimensional complex representation, is an isomorphism. The Bass Conjecture aims at a generalization of this fact to arbitrary groups.

Let $\mathrm{con}(G)$ be the set of conjugacy classes (g) of elements $g \in G$. Denote by $\mathrm{con}(G)_f$ the subset of $\mathrm{con}(G)$ consisting of those conjugacy classes (g) for which

each representative g has finite order. Let $\mathrm{class}_0(G)$ and $\mathrm{class}_0(G)_f$ respectively be the $\mathbb{C}$-vector spaces with the set $\mathrm{con}(G)$ and $\mathrm{con}(G)_f$ respectively as basis. This is the same as the $\mathbb{C}$-vector space of $\mathbb{C}$-valued functions on $\mathrm{con}(G)$ and $\mathrm{con}(G)_f$ with finite support. Define the *universal* $\mathbb{C}$*-trace* as

$$\mathrm{tr}^u_{\mathbb{C}G}\colon \mathbb{C}G \to \mathrm{class}_0(G), \quad \sum_{g\in G} \lambda_g \cdot g \mapsto \sum_{g\in G} \lambda_g \cdot (g).$$

It extends to a function $\mathrm{tr}^u_{\mathbb{C}G}\colon M_n(\mathbb{C}G) \to \mathrm{class}_0(G)$ on (n,n)-matrices over $\mathbb{C}G$ by taking the sum of the traces of the diagonal entries. Let P be a finitely generated projective $\mathbb{C}G$-module. Choose a matrix $A \in M_n(\mathbb{C}G)$ such that $A^2 = A$ and the image of the $\mathbb{C}G$-map $r_A\colon \mathbb{C}G^n \to \mathbb{C}G^n$ given by right multiplication with A is $\mathbb{C}G$-isomorphic to P. Define the *Hattori-Stallings rank* of P as

$$\mathrm{HS}_{\mathbb{C}G}(P) \quad := \quad \mathrm{tr}^u_{\mathbb{C}G}(A) \in \mathrm{class}_0(G).$$

The Hattori-Stallings rank depends only on the isomorphism class of the $\mathbb{C}G$-module P and induces a homomorphism $\mathrm{HS}_{\mathbb{C}G}\colon K_0(\mathbb{C}G) \to \mathrm{class}_0(G)$.

Conjecture 1.9 ((Strong) Bass Conjecture for $K_0(\mathbb{C}G)$). *The Hattori-Stalling rank yields an isomorphism*

$$\mathrm{HS}_{\mathbb{C}G}\colon K_0(\mathbb{C}G) \otimes_{\mathbb{Z}} \mathbb{C} \to \mathrm{class}_0(G)_f.$$

More information and further references about the Bass Conjecture can be found in [8, 0.5], [13],[54, Subsection 9.5.2], and [63, 3.1.3].

2. The Farrell-Jones Conjecture for Torsionfree Groups

2.1. The K-theoretic Farrell-Jones Conjecture for torsionfree groups and regular coefficient rings. We have already explained $K_0(R)$ and $K_1(R)$ for a ring R. There exist algebraic K-groups $K_n(R)$, for every $n \in \mathbb{Z}$, defined as the homotopy groups of the associated K-theory spectrum $\mathbf{K}(R)$. For the definition of higher algebraic K-theory groups and the (connective) K-theory spectrum see, for instance, [20, 74, 82, 85]. For information about negative K-groups, we refer the reader to [12, 32, 72, 73, 80, 82].

How can one come to a conjecture about the structure of the groups $K_n(RG)$? Let us consider the special situation, where the coefficient ring R is regular. Then one gets isomorphisms

$$\begin{aligned} K_n(R[\mathbb{Z}]) &\cong K_n(R) \oplus K_{n-1}(R); \\ K_n(R[G*H]) \oplus K_n(R) &\cong K_n(RG) \oplus K_n(RH). \end{aligned}$$

Now notice that for any generalized homology theory $\mathcal{H}$, we obtain isomorphisms

$$\begin{aligned} \mathcal{H}_n(B\mathbb{Z}) &\cong \mathcal{H}_n(\{\bullet\}) \oplus \mathcal{H}_{n-1}(\{\bullet\}); \\ \mathcal{H}_n(B(G*H)) \oplus \mathcal{H}_n(\{\bullet\}) &\cong \mathcal{H}_n(BG) \oplus \mathcal{H}(BH). \end{aligned}$$

This and other analogies suggest that $K_n(RG)$ may coincide with $\mathcal{H}_n(BG)$ for an appropriate generalized homology theory. If this is the case, we must have $\mathcal{H}_n(\{\bullet\}) = K_n(R)$. Hence, a natural guess for $\mathcal{H}_n$ is $H_n(-;\mathbf{K}(R))$, the homology theory associated to the algebraic K-theory spectrum $\mathbf{K}(R)$ of R. These considerations lead to:

Conjecture 2.1 (K-theoretic Farrell-Jones Conjecture for torsionfree groups and regular coefficient rings). *Let R be a regular ring and let G be a torsionfree group. Then there is an isomorphism*

$$H_n(BG;\mathbf{K}(R)) \xrightarrow{\cong} K_n(RG).$$

Remark 2.2 (The Farrell-Jones Conjecture and the vanishing of middle K-groups). If R is a regular ring, then $K_q(R) = 0$ for $q \leq -1$. Hence the Atiyah-Hirzebruch spectral sequence converging to $H_n(BG;\mathbf{K}(R))$ is a first quadrant spectral sequence. Its E^2-term is $H_p(BG;K_q(R))$. The edge homomorphism at $(0,0)$ obviously yields an isomorphism $H_0(BG;K_0(R)) \xrightarrow{\cong} H_0(BG;\mathbf{K}(R))$. The Farrell-Jones Conjecture 2.1 predicts, because of $H_0(BG;K_0(R)) \cong K_0(R)$, that there is an isomorphism $K_0(R) \xrightarrow{\cong} K_0(RG)$. We have not specified the isomorphism appearing in the Farrell-Jones Conjecture 2.1 above. However, we remark that it is easy to check that this isomorphism $K_0(R) \xrightarrow{\cong} K_0(RG)$ must be the change of rings map associated to the inclusion $R \to RG$. Thus, we see that the Farrell-Jones Conjecture 2.1 implies Conjecture 1.6.

The Atiyah-Hirzebruch spectral sequence yields an exact sequence $0 \to K_1(R) \to H_1(BG;\mathbf{K}(R)) \to H_1(G,K_0(R)) \to 0$. In the special case $R = \mathbb{Z}$, this reduces to an exact sequence $0 \to \{\pm 1\} \to H_1(BG;\mathbf{K}(R)) \to G/[G,G] \to 0$. This implies that the assembly map sends $H_1(BG;\mathbf{K}(R))$ bijectively onto the subgroup $\{\pm g \mid g \in G\}$ of $K_1(\mathbb{Z}G)$. Hence, the Farrell-Jones Conjecture 2.1 implies Conjecture 1.7.

Remark 2.3 (The Farrell-Jones Conjecture and the Kaplansky Conjecture). The Farrell-Jones Conjecture 2.1 also implies the Kaplansky Conjecture 1.4 (see [8, Theorem 0.12]).

Remark 2.4 (The conditions torsionfree and regular are needed in Conjecture 2.1). The version of the Farrell-Jones Conjecture 2.1 cannot be true without the assumptions that R is regular and G is torsionfree. The Bass-Heller-Swan decomposition yields an isomorphism $K_n(R[\mathbb{Z}]) \cong K_n(R) \oplus K_{n-1}(R) \oplus NK_n(R) \oplus NK_n(R)$, whereas $H_n(B\mathbb{Z};\mathbf{K}(R)) \cong K_n(R) \oplus K_{n-1}(R)$. If R is regular, then $NK_n(R)$ is trivial, but there are rings R with non-trivial $NK_n(R)$.

Suppose that $R = \mathbb{C}$ and G is finite. Then $H_0(BG; \mathbf{K}_{\mathbb{C}}) \cong K_0(\mathbb{C}) \cong \mathbb{Z}$, whereas $K_0(\mathbb{C}G)$ is the complex representation ring of G, which is isomorphic to $\mathbb{Z}$ if and only if G is trivial.

2.2. The *L*-theoretic Farrell-Jones Conjecture for torsion-free groups.

There is also an L-theoretic version of Conjecture 2.1:

Conjecture 2.5 (L-theoretic Farrell-Jones Conjecture for torsionfree groups). *Let R be a ring with involution and let G be a torsionfree group. Then there is an isomorphism*

$$H_n\big(BG; \mathbf{L}(R)^{\langle -\infty \rangle}\big) \xrightarrow{\cong} L_n^{\langle -\infty \rangle}(RG).$$

Here $\mathbf{L}(R)^{\langle -\infty \rangle}$ is the periodic quadratic L-theory spectrum of the ring with involution R with decoration $\langle -\infty \rangle$, and $L_n^{\langle -\infty \rangle}(R)$ is the n-th quadratic L-group with decoration $\langle -\infty \rangle$, which can be identified with the n-th homotopy group of $\mathbf{L}_{RG}^{\langle -\infty \rangle}$. For more information about the various types of L-groups and decorations and L-theory spectra we refer the reader to [18, 19, 75, 78, 79, 80, 81, 88]. Roughly speaking, L-theory deals with quadratic forms. For even n, $L_n(R)$ is related to the Witt group of quadratic forms and for odd n, $L_n(R)$ is related to automorphisms of quadratic forms. Moreover, the L-groups are four-periodic, i.e., $L_n(R) \cong L_{n+4}(R)$.

Theorem 2.6 (The Farrell-Jones Conjecture implies the Borel Conjecture in dimensions ≥ 5). *Suppose that a torsionfree group G satisfies Conjecture 2.1 and Conjecture 2.5 for $R = \mathbb{Z}$. Then the Borel Conjecture 1.1 holds for any closed aspherical manifold of dimension ≥ 5 whose fundamental group is isomorphic to G.*

Sketch of proof. The *topological structure set* $\mathcal{S}^{\mathrm{top}}(M)$ of a closed manifold M is defined to be the set of equivalence classes of homotopy equivalences $f\colon M' \to M$, with a topological closed manifold as its source and M as its target, for which $f_0\colon M_0 \to M$ and $f_1\colon M_1 \to M$ are equivalent if there is a homeomorphism $g\colon M_0 \to M_1$ such that $f_1 \circ g$ and f_0 are homotopic. The Borel Conjecture 1.1 can be reformulated in the language of surgery theory to the statement that $\mathcal{S}^{\mathrm{top}}(M)$ consists of a single point if M is an aspherical closed topological manifold.

The *surgery sequence* of a closed topological manifold M of dimension $n \geq 5$ is the exact sequence

$$\begin{aligned} \ldots \to \mathcal{N}_{n+1}(M \times [0,1], M \times \{0,1\}) \xrightarrow{\sigma} L_{n+1}^s(\mathbb{Z}\pi_1(M)) \xrightarrow{\partial} \mathcal{S}^{\mathrm{top}}(M) \\ \xrightarrow{\eta} \mathcal{N}_n(M) \xrightarrow{\sigma} L_n^s(\mathbb{Z}\pi_1(M)), \end{aligned} \tag{2.7}$$

which extends infinitely to the left. It is the fundamental tool for the classification of topological manifolds. (There is also a smooth version of it.) The map σ appearing in the sequence sends a normal map of degree one

to its surgery obstruction. This map can be identified with the version of the L-theory assembly map, where one works with the 1-connected cover $\mathbf{L}^s(\mathbb{Z})\langle 1\rangle$ of $\mathbf{L}^s(\mathbb{Z})$. The map $H_k(M;\mathbf{L}^s(\mathbb{Z})\langle 1\rangle) \to H_k(M;\mathbf{L}^s(\mathbb{Z}))$ is injective for $k = n$ and an isomorphism for $k > n$. Because of the K-theoretic assumptions (and the so-called Rothenberg sequence), we can replace the s-decoration with the $\langle -\infty\rangle$-decoration. Therefore the Farrell-Jones Conjecture 2.5 implies that the map $\sigma\colon \mathcal{N}_n(M) \to L_n^s(\mathbb{Z}\pi_1(M))$ is injective and the map $\mathcal{N}_{n+1}(M \times [0,1], M \times \{0,1\}) \xrightarrow{\sigma} L_{n+1}^s(\mathbb{Z}\pi_1(M))$ is bijective. Thus, by the surgery sequence, $\mathcal{S}^{\mathrm{top}}(M)$ is a point and hence the Borel Conjecture 1.1 holds for M. More details can be found in [36, pages 17,18,28], [79, Chapter 18]. $\square$

For more information about surgery theory, see [18, 19, 46, 52, 81, 88].

3. The General Formulation of the Farrell-Jones Conjecture

3.1. Classifying spaces for families. Let G be a group. A *family $\mathcal{F}$ of subgroups of G* is a set of subgroups which is closed under conjugation with elements of G and under taking subgroups. A G-CW-complex, all of whose isotropy groups belong to $\mathcal{F}$ and whose H-fixed point sets are contractible for all $H \in \mathcal{F}$, is called a *classifying space for the family $\mathcal{F}$* and will be denoted $E_{\mathcal{F}}(G)$. Such a space is unique up to G-homotopy, because it is characterized by the property that for any G-CW-complex X, all whose isotropy groups belong to $\mathcal{F}$, there is precisely one G-map from X to $E_{\mathcal{F}}(G)$ up to G-homotopy. These spaces were introduced by tom Dieck [84]. A functorial "bar-type" construction is given in [23, section 7].

The space $E_{\mathcal{TR}}(G)$, for $\mathcal{TR}$ the family consisting of the trivial subgroup only, is the same as the space EG, which is by definition the total space of the universal G-principal bundle $G \to EG \to BG$, or, equivalently, the universal covering of BG. A model for $E_{\mathcal{ALL}}(G)$, for the family $\mathcal{ALL}$ of all subgroups, is given by the space $G/G = \{\bullet\}$ consisting of one point.

The space $E_{\mathcal{F}\mathrm{in}}(G)$, for $\mathcal{F}$in the family of finite subgroups, is also known as *the classifying space for proper G-actions*, and is denoted by $\underline{E}G$ in the literature. Recall that a G-CW-complex X is proper if and only if all of its isotropy groups are finite (see for instance [50, Theorem 1.23 on page 18]).

There are often nice models for $\underline{E}G$. If G is word hyperbolic in the sense of Gromov, then the Rips-complex is a finite model [65]. If G is a discrete subgroup of a Lie group L with finitely many path components, then for any maximal compact subgroup $K \subseteq L$, the space L/K with its left G-action is a model for $\underline{E}G$. More information about $\underline{E}G$ can be found in [14, 27, 51, 58, 62].

Let $\mathcal{VC}$yc be the family of virtually cyclic subgroups, i.e., subgroups which are either finite or contain $\mathbb{Z}$ as subgroup of finite index. We often abbreviate $\underline{\underline{E}}G = E_{\mathcal{VC}\mathrm{yc}}(G)$.

3.2. G-homology theories. Fix a group G. A *G-homology theory* $\mathcal{H}_*^G$ is a collection of covariant functors $\mathcal{H}_n^G$ from the category of G-CW-pairs to the category of abelian groups indexed by $n \in \mathbb{Z}$ together with natural transformations

$$\partial_n^G(X,A)\colon \mathcal{H}_n^G(X,A) \to \mathcal{H}_{n-1}^G(A) := \mathcal{H}_{n-1}^G(A,\emptyset)$$

for $n \in \mathbb{Z}$, such that four axioms hold; namely, G-homotopy invariance, long exact sequence of a pair, excision, and the disjoint union axiom. The obvious formulation of these axioms is left to the reader or can be found in [53]. Of course a G-homology theory for the trivial group $G = \{1\}$ is a homology theory (satisfying the disjoint union axiom) in the classical non-equivariant sense.

Remark 3.1 (G-homology theories and spectra over $\mathsf{Or}(G)$). The *orbit category* $\mathsf{Or}(G)$ has as objects the homogeneous spaces G/H and as morphisms G-maps. Given a covariant functor $\mathbf{E}$ from $\mathsf{Or}(G)$ to the category of spectra, there exists a G-homology theory $\mathcal{H}_*^G$ such that $\mathcal{H}_n^G(G/H) = \pi_n(\mathbf{E}(G/H))$ holds for all $n \in \mathbb{Z}$ and subgroups $H \subseteq G$ (see [23], [63, Proposition 6.3 on page 737]). For trivial G, this boils down to the classical fact that a spectrum defines a homology theory.

3.3. The Meta-Isomorphism Conjecture. Now we can formulate the following Meta-Conjecture for a group G, a family of subgroups $\mathcal{F}$, and a G-homology theory $\mathcal{H}_*^G$.

Conjecture 3.2 (Meta-Conjecture). *The so-called assembly map*

$$A_{\mathcal{F}}\colon \mathcal{H}_n^G(E_{\mathcal{F}}(G)) \to \mathcal{H}_n^G(\mathrm{pt}),$$

which is the map induced by the projection $E_{\mathcal{F}}(G) \to \mathrm{pt}$, is an isomorphism for $n \in \mathbb{Z}$.

Notice that the Meta-Conjecture 3.2 is always true if we choose $\mathcal{F} = \mathcal{ALL}$. So given G and $\mathcal{H}_*^G$, the point is to choose $\mathcal{F}$ as small as possible.

3.4. The Farrell-Jones Conjecture. Let R be a ring. Then one can construct for every group G, using Remark 3.1, G-homology theories $H_*^G(-;\mathbf{K}_R)$ and $H_*^G\big(-;\mathbf{L}_R^{\langle -\infty\rangle}\big)$ satisfying $H_n^G(G/H;\mathbf{K}_R) \cong K_n(RH)$ and $H_n^G\big(G/H;\mathbf{L}_R^{\langle -\infty\rangle}\big) \cong L_n^{\langle -\infty\rangle}(RH)$. The Meta-Conjecture 3.2 for $\mathcal{F} = \mathcal{VC}\mathrm{yc}$ is the Farrell-Jones Conjecture:

Conjecture 3.3 (Farrell-Jones Conjecture). *The maps induced by the projection $\underline{\underline{E}}G \to G/G$ are, for every $n \in \mathbb{Z}$, isomorphisms*

$$H_n^G(\underline{\underline{E}}G;\mathbf{K}_R) \to H_n^G(G/G;\mathbf{K}_R) = K_n(RG);$$
$$H_n^G\left(\underline{\underline{E}}G;\mathbf{L}_R^{\langle -\infty\rangle}\right) \to H_n^G\left(G/G;\mathbf{L}_R^{\langle -\infty\rangle}\right) = L_n^{\langle -\infty\rangle}(RG).$$

The version of the Farrell-Jones Conjecture 3.3 is equivalent to the original version due to Farrell-Jones [30, 1.6 on page 257]. The decoration $\langle -\infty \rangle$ cannot be replaced by the decorations h, s or p in general, since there are counterexamples for these decorations (see [34]).

Remark 3.4 (Generalized Induction Theorem). One may interpret the Farrell-Jones Conjecture as a kind of generalized induction theorem. A prototype of an induction theorem is Artin's Theorem, which essentially says that the complex representation ring of a finite group can be computed in terms of the representation rings of the cyclic subgroups. In the Farrell-Jones setting one wants to compute $K_n(RG)$ and $L_n^{\langle -\infty \rangle}(RG)$ in terms of the values of these functors on virtually cyclic subgroups, where one has to take into account all the relations coming from inclusions and conjugations, and the values in degree n depend on all the values in degree $k \leq n$ on virtually cyclic subgroups.

Remark 3.5 (The choice of the family $\mathcal{VC}$yc). One can show that, in general, $\mathcal{VC}$yc is the smallest family of subgroups for which one can hope that the Farrell-Jones Conjecture is true for all G and R. The family $\mathcal{F}$in is definitely too small. Under certain conditions one can use smaller families, for instance, $\mathcal{F}$in is sufficient if R is regular and contains $\mathbb{Q}$, and $\mathcal{TR}$ is sufficient if R is regular and G is torsionfree. This explains that Conjecture 3.3 reduces to Conjecture 2.1 and Conjecture 2.5. More information about reducing the family of subgroups can be found in [3], [22], [24], [57, Lemma 4.2], [63, 2.2], [77].

Remarks 3.4 and 3.5 can be illustrated by the following consequence of the Farrell-Jones Conjecture 3.3: Given a field F of characteristic zero and a group G, the obvious map

$$\bigoplus_{H \subseteq G, |H| < \infty} K_0(FH) \to K_0(FG)$$

coming from the various inclusions $H \subseteq G$ is surjective, and actually induces an isomorphism

$$\operatorname{colim}_{H \subseteq G, |H| < \infty} K_0(FH) \xrightarrow{\cong} K_0(FG).$$

Remark 3.6 (The K-theoretic Farrell-Jones Conjecture and the Bass Conjecture). The K-theoretic Farrell-Jones Conjecture 3.3 implies the Bass Conjecture 3.3 (see [8, Theorem 0.9]).

Remark 3.7 (Coefficients in additive categories). It is sometimes important to consider twisted group rings, where we take a G-action on R into account, or more generally, crossed product rings $R * G$. In the L-theory case we also want to allow orientation characters. All of these generalizations can be uniformly handled if one allows coefficients in an additive category. These more general versions of the Farrell-Jones Conjectures are explained for K-theory in [10] and

for L-theory in [5]. These generalizations also encompass the so-called fibered versions. One of their main features is that they have much better inheritance properties, (e.g., passing to subgroups, direct and free products, directed colimits) than the untwisted version 3.3.

For proofs the coefficients are often dummy variables. In the right setup it does not matter whether one uses coefficients in a ring R or in an additive category.

3.5. The Baum-Connes and the Bost Conjectures. There also exists a G-homology theory $H_*^G\bigl(-;\mathbf{K}_{C_r^*}^{\mathrm{top}}\bigr)$ with the property that $H_n^G\bigl(G/H;\mathbf{K}_{C_r^*}^{\mathrm{top}}\bigr) = K_n(C_r^*(H))$, where $K_n(C_r^*(H))$ is the topological K-theory of the reduced group C^*-algebra. For a proper G-CW-complex X, the equivariant topological K-theory $K_n^G(X)$ agrees with $H_n^G\bigl(X;\mathbf{K}_{C_r^*}^{\mathrm{top}}\bigr)$. The Meta-Conjecture 3.2 for $\mathcal{F} = \mathcal{F}\mathrm{in}$ is:

Conjecture 3.8 (Baum-Connes Conjecture). *The maps induced by the projection $\underline{E}G \to G/G$*

$$K_n^G(\underline{E}G) = H_n^G\bigl(\underline{E}G;\mathbf{K}_{C_r^*}^{\mathrm{top}}\bigr) \to H_n^G\bigl(G/G;\mathbf{K}_{C_r^*}^{\mathrm{top}}\bigr) = K_n(C_r^*(G)).$$

are isomorphisms for every $n \in \mathbb{Z}$.

The original version of the Baum-Connes Conjecture is stated in [14, Conjecture 3.15 on page 254]. For more information about the Baum-Connes Conjecture, see, for instance, [40, 63, 68].

Remark 3.9 (The relation between the conjectures of Novikov, Farrell-Jones and Baum-Connes). Both the L-theoretic Farrell-Jones Conjecture 3.3 and the Baum-Connes Conjecture 3.8 imply the Novikov Conjecture. See [47, Section 23], where the relation between the L-theoretic Farrell-Jones Conjecture 3.3 and the Baum-Connes Conjecture 3.8 is also explained.

4. The Status of the Farrell-Jones Conjecture

4.1. The work of Farrell-Jones and the status in 2004. One of the highlights of the work of Farrell and Jones is their proof of the Borel Conjecture 1.1 for manifolds of dimension ≥ 5 which support a Riemannian metric of non-positive sectional curvature [31]. They were able to extend this result to cover compact complete affine flat manifolds of dimension ≥ 5 [33]. This was done by considering complete non-positively curved manifolds that are not necessarily compact. Further results by Farrell and Jones about their conjecture for K-theory and pseudo-isotopy can be found in [30]. For a detailed report about the status of the Baum-Connes Conjecture and Farrell-Jones Conjecture

in 2004 we refer to [63, Chapter 5], where one can also find further references to relevant papers.

4.2. Hyperbolic groups and CAT(0)-groups. In recent years, the class of groups for which the Farrell-Jones Conjecture, and hence the other conjectures appearing in Section 1, are true has been extended considerably beyond fundamental groups of non-positively curved manifolds. In what follows, a *hyperbolic group* is to be understood in the sense of Gromov. A CAT(0)-group is a group that admits a proper isometric cocompact action on some CAT(0)-space of finite topological dimension.

Theorem 4.1 (Hyperbolic groups). *The Farrell-Jones Conjecture with coefficients in additive categories (see Remark 3.7) holds for both K- and L-theory for every hyperbolic group.*

Proof. The K-theory part is proved in Bartels-Lück-Reich [7], the L-theory part in Bartels-Lück [4]. □

Theorem 4.2 (CAT(0)-groups).

(i) The L-theoretic Farrell-Jones Conjecture with coefficients in additive categories (see Remark 3.7) holds for every CAT(0)*-group;*

(ii) The assembly map for the K-theoretic Farrell-Jones Conjecture with coefficients in additive categories (see Remark 3.7) is bijective in degrees $n \leq 0$ and surjective in degree $n = 1$ for every CAT(0)*-group.*

Proof. This is proved in Bartels-Lück [4]. □

For the proofs that the Farrell-Jones Conjecture implies the conjectures mentioned in Section 1, it suffices to know the statements appearing in Theorem 4.2. For instance Theorem 4.2 implies the Borel Conjecture for every closed aspherical manifold of dimension ≥ 5 whose fundamental group is a CAT(0)-group.

4.3. Inheritance properties. We have already mentioned that the version of the Farrell-Jones Conjecture with coefficients in additive categories (see Remark 3.7) does not only include twisted group rings and allow one to insert orientation homomorphisms, but it also has very valuable inheritance properties.

Theorem 4.3 (Inheritance properties). *Let (A) be one of the following assertions for a group G:*

- *The K-theoretic Farrell-Jones Conjecture with coefficients in additive categories (see Remark 3.7) holds for G;*

- *The K-theoretic Farrell-Jones Conjecture with coefficients in additive categories (see Remark 3.7) holds for G up to degree one, i.e., the assembly map is bijective in dimension $n \leq 0$ and surjective for $n = 1$;*
- *The L-theoretic Farrell-Jones Conjecture with coefficients in additive categories (see Remark 3.7) holds for G.*

Then the following is true:

(i) If G satisfies assertion (A), then also every subgroup $H \subseteq G$ satisfies (A);

(ii) If G_1 and G_2 satisfies assertion (A), then also the free product $G_1 * G_2$ and the direct product $G_1 \times G_2$ satisfy assertion (A);

(iii) Let $\pi\colon G \to Q$ be a group homomorphism. If Q satisfies (A) and for every virtually cyclic subgroup $V \subseteq Q$, its preimage $\pi^{-1}(V)$ satisfies (A), then G satisfies assertion (A);

(iv) Let $\{G_i \mid i \in I\}$ be a directed system of groups (with not necessarily injective structure maps). If each G_i satisfies assertion (A), then the colimit $\operatorname{colim}_{i\in I} G_i$ satisfies assertion (A).

Proof. See [4, Lemma 2.3]. □

Examples. Let $\mathcal{FJ}$ be the class of groups satisfying both the K-theoretic and L-theoretic Farrell-Jones Conjecture with additive categories as coefficients (see Remark 3.7). Let $\mathcal{FJ}_{\leq 1}$ be the class of groups which satisfy the L-theoretic Farrell-Jones Conjecture with additive categories as coefficients and the K-theoretic Farrell-Jones Conjecture with additive categories as coefficients up to degree one.

In view of the results above, these classes contain many groups which lie in the region *Hic Abundant Leones* in Martin Bridson's universe of groups (see [15]). Theorem 4.1 and Theorem 4.3 (iv) imply that directed colimits of hyperbolic groups belong to $\mathcal{FJ}$. This class of groups contains a number of groups with unusual properties. Counterexamples to the Baum-Connes Conjecture with coefficients are groups with expanders [41]. The only known construction of such groups is as a directed colimit of hyperbolic groups (see [2]). Thus the Farrell-Jones Conjecture in K- and L-theory holds for the only presently known counterexamples to the Baum-Connes Conjecture with coefficients. (We remark that the formulation of the Farrell-Jones Conjecture we are considering allows for twisted group rings, so this includes the correct analog of the Baum-Connes Conjecture with coefficients.) The class of directed colimits of hyperbolic groups contains, for instance, a torsionfree non-cyclic group all whose proper subgroups are cyclic, constructed by Ol'shanskii [70]. Further examples are lacunary groups (see [71]).

Davis and Januszkiewicz used Gromov's hyperbolization technique to construct exotic aspherical manifolds. They showed that for every $n \geq 5$ there are

closed aspherical n-dimensional manifolds such that their universal covering is a CAT(0)-space whose fundamental group at infinity is non-trivial [26, Theorem 5b.1]. In particular, these universal coverings are not homeomorphic to Euclidean space. Because these examples are non-positively curved polyhedron, their fundamental groups are CAT(0)-groups and belong to $\mathcal{FJ}_{\leq 1}$. There is a variation of this construction that uses the strict hyperbolization of Charney-Davis [21] and produces closed aspherical manifolds whose universal cover is not homeomorphic to Euclidean space and whose fundamental group is hyperbolic. All of these examples are topologically rigid.

Limit groups in the sense of Zela have been a focus of geometric group theory in recent years. Alibegović-Bestvina [1] have shown that limit groups are CAT(0)-groups.

Let G be a (not necessarily cocompact) lattice in $SO(n,1)$, e.g., the fundamental group of a hyperbolic Riemannian manifold with finite volume. Then G acts properly cocompactly and isometrically on a CAT(0)-space by [16, Corollary 11.28 in Chapter II.11 on page 362], and hence belongs to $\mathcal{FJ}_{\leq 1}$.

5. Computational Aspects

It is very hard to compute $K_n(RG)$ or $L_n^{\langle -\infty \rangle}(RG)$ directly. It is easier to compute the source of the assembly map appearing in the Farrell-Jones Conjecture 3.3, since one can apply standard techniques for the computation of equivariant homology theories and there are often nice models for $\underline{E}G$. Rationally, equivariant Chern characters, as developed in [53, 55, 56] give rather general answers. We illustrate this with the following result taken from [53, Example 8.11].

Theorem 5.1. *Let G be a group for which the Farrell-Jones Conjecture 3.3 holds for $R = \mathbb{C}$. Let T be the set of conjugacy classes (g) of elements $g \in G$ of finite order. For an element $g \in G$, denote by $C_G\langle g \rangle$ the centralizer of g. Then we obtain isomorphisms*

$$\begin{aligned} \bigoplus_{p+q=n} \bigoplus_{(g)\in T} H_p(C_G\langle g\rangle;\mathbb{C}) \otimes_{\mathbb{Z}} K_q(\mathbb{C}) &\to \mathbb{C} \otimes_{\mathbb{Z}} K_n(\mathbb{C}G); \\ \bigoplus_{p+q=n} \bigoplus_{(g)\in T} H_p(C_G\langle g\rangle;\mathbb{C}) \otimes_{\mathbb{Z}} L_q^{\langle -\infty \rangle}(\mathbb{C}) &\to \mathbb{C} \otimes_{\mathbb{Z}} L_n^{\langle -\infty \rangle}(\mathbb{C}G), \end{aligned}$$

where we use the involutions coming from complex conjugation in the definition of $L_q^{\langle -\infty \rangle}(\mathbb{C})$ and $L_n^{\langle -\infty \rangle}(\mathbb{C}G)$.

Integral computations can only be given in special cases. For example, the semi-direct product $\mathbb{Z}^r \rtimes \mathbb{Z}/n$ cannot be handled in general. Not even its ordinary group homology is known, so it is not a surprise that the K- and L-theory of the associated group ring are unknown in general. Sometimes explicit answers can be found in the literature, see for instance [63, 8.3]. As an illustration

we mention the following result which follows from Theorem 4.1 using [11, Theorem 1.3], and [64, Corollary 2.11 and Example 3.6].

Theorem 5.2 (Torsionfree hyperbolic groups). *Let G be a torsionfree hyperbolic group. Let $\mathcal{M}$ be a complete system of representatives of the conjugacy classes of maximal infinite cyclic subgroups of G.*

(i) For every $n \in \mathbb{Z}$, there is an isomorphism

$$H_n\big(BG; \mathbf{K}(R)\big) \oplus \bigoplus_{V \in \mathcal{M}} NK_n(R) \oplus NK_n(R) \quad \xrightarrow{\cong} \quad K_n(RG),$$

where $NK_n(R)$ the Bass-Nil-groups of R*;*

(ii) For every $n \in \mathbb{Z}$, there is an isomorphism

$$H_n\big(BG; \mathbf{L}^{\langle -\infty \rangle}(R)\big) \quad \xrightarrow{\cong} \quad L_n^{\langle -\infty \rangle}(RG).$$

Computations of L-groups of group rings are important in the classification of manifolds since they appear in the surgery sequence (2.7).

6. Methods of Proof

Here is a brief sketch of the strategy of proof which has led to the results about hyperbolic groups and CAT(0)-groups mentioned above. It is influenced by ideas of Farrell and Jones. However, we have to deal with spaces that are not manifolds, and hence new ideas and techniques are required. A more detailed survey about methods of proof can be found in [4, Section 1], [6, Section 1], [7, Section 1], [61] and [63, Chapter 7].

Assembly and forget control. We have defined the assembly map appearing in the Farrell-Jones Conjecture as a map induced by the projection $\underline{\underline{E}}G \to G/G$. A homotopy theoretic interpretation by homotopy colimits and a description in terms of the universal property that it is the best approximation from the left by a homology theory is presented in [23]. This interpretation is good for structural and computational aspects but is not helpful for actual proofs. For this purpose the interpretation of the assembly map as a *forget control map* is the right one. This fundamental idea is due to Quinn.

Roughly speaking, one attaches to a metric space certain categories, to these categories spectra and then takes their homotopy groups, where everything depends on a choice of certain control conditions which in some sense measure sizes of cycles. If one requires certain control conditions, one obtains the source of the assembly map. If one requires no control conditions, one obtains the target of the assembly map. The assembly map itself is forgetting the control condition.

One of the basic features of a homology theory is excision. It often comes from the fact that a representing cycle can be found with arbitrarily good control. An example is the technique of subdivision which allows to make the representing cycles for simplicial homology arbitrarily controlled. That is, the diameter of any simplex appearing with non-zero coefficient is very small. One may say that requiring control conditions amounts to implementing homological properties.

With this interpretation it is clear what the main task in the proof of surjectivity of the assembly map is: *achieve control*, i.e., manipulate cycles without changing their homology class so that they become sufficiently controlled. There is a general principle that a proof of surjectivity also gives injectivity, Namely, proving injectivity means that one must construct a cycle whose boundary is a given cycle, i.e., one has to solve a surjectivity problem in a relative situation.

Contracting maps and open coverings. Contracting maps on suitable control spaces are very useful for gaining control. The idea is that the contraction improves the control of a cycle without changing its homology class if the contracting map is, roughly speaking, homotopic to the identity. Of course one has to choose the contracting maps and control spaces with care. If a G-space X has a fixed point, the projection to this fixed point is a contracting G-equivariant map, but it turns out that this is just enough to prove the trivial version of the Meta Conjecture, where the family $\mathcal{F}$ is not $\mathcal{VC}$yc as desired, but rather consists of all subgroups.

Let $\mathcal{F}$ be a family of subgroups and let X be a metric space with an isometric G-action. An *$\mathcal{F}$-covering* $\mathcal{U}$ is an open covering $\mathcal{U}$ such that $gU \in \mathcal{U}$ holds for $U \in \mathcal{U}, g \in G$, for every $U \in \mathcal{U}$ and $g \in G$ we have $gU \cap U \neq \emptyset \implies gU = U$, and for every $U \in \mathcal{U}$ the subgroup $G_U = \{g \in G \mid gU = U\}$ belongs to $\mathcal{F}$. Associated to these data there is a map $f_{\mathcal{U}} \colon X \to |\mathcal{U}|$ from X to the simplicial nerve of $\mathcal{U}$. The larger the Lebesgue number of $\mathcal{U}$ is, the more contracting the map becomes with respect to the L^1-metric on $|\mathcal{U}|$, provided we are able to fix a uniform bound on its covering dimension (see [7, Proposition 5.3]).

Notice that the simplicial nerve carries a G-CW-complex structure and all its isotropy groups belong to $\mathcal{F}$. We see that $\mathcal{F}$-coverings can yield contracting maps, as long as the covering dimension of the possible $\mathcal{U}$ are uniformly bounded.

An axiomatic description of the properties such an equivariant covering has to fulfill can be found in [7, Section 1] and more generally in [4, Section 1]. The equivariant coverings satisfy conditions that are similar to those for finite asymptotic dimension, but with extra requirements about equivariance. A key technical paper for the construction of such equivariant coverings is [6], where the connection to asymptotic dimension is explained.

Enlarging *G* and transfer. Let us try to find $\mathcal{F}$-coverings for G considered as a metric space with the word metric. If we take $\mathcal{U} = \{G\}$, we obtain a G-

invariant open covering with arbitrarily large Lebesgue number, but the open set G is an $\mathcal{F}$-set only if we take $\mathcal{F}$ to be the family of all subgroups. If we take $\mathcal{U} = \big\{\{g\} \mid g \in G\big\}$ and denote by $\mathcal{TR}$ the family consisting only of the trivial subgroup, we obtain a $\mathcal{TR}$-covering of topological dimension zero, but the Lebesgue number is not very impressive, it's just 1. In order to increase the Lebesgue number, we could take large balls around each element. Since the covering has to be G-invariant, we could start with $\mathcal{U} = \big\{B_R(g) \mid g \in G\big\}$, where $B_R(g)$ is the open ball of radius R around g. This is a G-invariant open covering with Lebesgue number R, but the sets $B_R(g)$ are not $\mathcal{F}$-sets in general and the covering dimension grows with R.

One of the main ideas is not to cover G itself, but to enlarge G to $G \times \overline{X}$ for an appropriate compactification $\overline{X}$ of a certain contractible metric space X that has an isometric proper cocompact G-action. This allows us to spread out the open sets and avoid having too many intersections. This strategy has also been successfully used in measurable group theory, where the role of the topological space $\overline{X}$ is played by a probability space with measure preserving G-action (see Gromov [39, page 300]).

The elements under consideration lie in K- or L-theory spaces associated to the control space G. Using a transfer they can be lifted to $G \times \overline{X}$. (This step corresponds in the proofs of Farrell and Jones to the passage to the sphere tangent bundle.) We gain control there and then push the elements down to G. Since the space $\overline{X}$ is contractible, its Euler characteristic is 1 and hence the composite of the push-down map with the transfer map is the identity on the K-theory level. On the L-theory level one needs something with signature 1. On the algebra level this corresponds to the assignment of a finitely generated projective $\mathbb{Z}$-module P to its *multiplicative hyperbolic form* $H_{\otimes}(P)$. It is given by replacing $\oplus$ by $\otimes$ in the standard definition of a hyperbolic form, i.e., the underlying $\mathbb{Z}$-module is $P^* \otimes P$ and the symmetric form is given by the formula $(\alpha, p) \otimes (\beta, q) \mapsto \alpha(q) \cdot \beta(p)$. Notice that the signature of $H_{\otimes}(\mathbb{Z})$ is 1 and taking the multiplicative hyperbolic form yields an isomorphism of rings $K_0(\mathbb{Z}) \to L^0(\mathbb{Z})$.

We can construct $\mathcal{VC}$yc-coverings that are contracting in the G-direction but will actually expand in the $\overline{X}$-direction. The latter defect can be compensated for because the transfer yields elements over $G \times \overline{X}$ with arbitrarily good control in the $\overline{X}$-direction.

Flows. To find such coverings of $G \times \overline{X}$, it is crucial to construct, for hyperbolic and CAT(0)-spaces, flow spaces $\mathrm{FS}(X)$ which are the analog of the geodesic flow on a simply connected Riemannian manifold with negative or non-positive sectional curvature. One constructs appropriate coverings on $\mathrm{FS}(X)$, often called *long and thin coverings*, and then pulls them back with a certain map $G \times \overline{X} \to \mathrm{FS}(X)$. The flow is used to improve a given covering. The use flow spaces to gain control is one of the fundamental ideas of Farrell and Jones (see for instance [28]).

Let us look at a special example to illustrate the use of a flow. Consider two points with coordinates (x_1, y_1) and (x_2, y_2) in the upper half plane model of two-dimensional hyperbolic space. We want to use the geodesic flow to make their distance smaller in a functorial fashion. This is achieved by letting these points flow towards the boundary at infinity along the geodesic given by the vertical line through these points, i.e., towards infinity in the y-direction. There is a fundamental problem: if $x_1 = x_2$, then the distance of these points is unchanged. Therefore we make the following prearrangement. Suppose that $y_1 < y_2$. Then we first let the point (x_1, y_1) flow so that it reaches a position where $y_1 = y_2$. Inspecting the hyperbolic metric, one sees that the distance between the two points (x_1, τ) and (x_2, τ) goes to zero if τ goes to infinity. This is the basic idea to gain control in the negatively curved case.

Why is the non-positively curved case harder? Again, consider the upper half plane, but this time equip it with the flat Riemannian metric coming from Euclidean space. Then the same construction makes sense, but the distance between two points (x_1, τ) and (x_2, τ) is unchanged if we change τ. The basic first idea is to choose a focal point far away, say $f := \big((x_1 + x_2)/2, \tau + 169356991\big)$, and then let (x_1, τ) and (x_2, τ) flow along the rays emanating from them and passing through the focal point f. In the beginning the effect is indeed that the distance becomes smaller, but as soon as we have passed the focal point the distance grows again. Either one chooses the focal point very far away or uses the idea of moving the focal point towards infinity while the points flow. Roughly speaking, we are suggesting the idea of a *dog and sausage* principle. We have a dog, and attached to it is a long stick pointing in front of it with a delicious sausage on the end. The dog will try to reach the sausage, but the sausage is moving away according to the movement of the dog, so the dog will never reach the sausage. (The dog will become long and thin this way, but this is a different effect). The problem with this idea is obvious, we must describe this process in a functorial way and carefully check all the estimates to guarantee the desired effects.

7. Open Problems

7.1. Virtually poly-cyclic groups, cocompact lattices and 3-manifold groups. It is conceivable that our methods can be used to show that virtually poly-cyclic groups belong to $\mathcal{FJ}$ or $\mathcal{FJ}_{\leq 1}$. This already implies the same conclusion for cocompact lattices in almost connected Lie groups following ideas of Farrell-Jones [30] and for fundamental groups of (not necessarily compact) 3-manifolds (possibly with boundary) following ideas of Roushon [83].

7.2. Solvable groups. Show that solvable groups belong to $\mathcal{FJ}$ or $\mathcal{FJ}_{\leq 1}$. In view of the large class of groups belonging to $\mathcal{FJ}$ or $\mathcal{FJ}_{\leq 1}$, it is very surprising that it is not known whether a semi-direct product $A \rtimes_\varphi \mathbb{Z}$ for a (not

necessarily finitely generated) abelian group A belongs to $\mathcal{FJ}$ or $\mathcal{FJ}_{\leq 1}$. The problem is the possibly complicated dynamics of the automorphism φ of A.

Such groups are easy to handle in the Baum-Connes setting, where one can use the long exact Wang sequence for topological K-theory associated to a semi-direct product. Such a sequence does not exists for algebraic K-theory, and new contributions involving Nil-terms occur.

7.3. Other open cases. Show that mapping class groups, $\mathrm{Out}(F_n)$ and Thompson's groups belong to $\mathcal{FJ}$ or $\mathcal{FJ}_{\leq 1}$. The point here is not that this has striking consequence in and of itself, but rather their proofs will probably give more insight in the Farrell-Jones Conjecture and will require some new input about the geometry of these groups which may be interesting in its own right.

A very interesting open case is $SL_n(\mathbb{Z})$. The main obstacle is that $SL_n(\mathbb{Z})$ does not act cocompactly isometrically properly on a CAT(0)-space; the canonical action on $SL_n(\mathbb{R})/SO(n)$ is proper and isometric and of finite covolume but not cocompact. The Baum-Connes Conjecture is also open for $SL_n(\mathbb{Z})$.

7.4. Searching for counterexamples. There is no group known for which the Farrell-Jones Conjecture is false. There has been some hope that groups with expanders may yield counterexamples, but this hope has been dampened since colimits of hyperbolic groups satisfy it. At the moment one does not know any property of a group which makes it likely to produce a counterexample. The same holds for the Borel Conjecture. Many of the known exotic examples of closed aspherical manifolds are known to satisfy the Borel Conjecture.

In order to find counterexamples one seems to need completely new ideas, maybe from random groups or logic.

7.5. Pseudo-isotopy. Extend our results to pseudo-isotopy spaces. There are already interesting results for these proved by Farrell-Jones [30].

7.6. Transfer of methods. The Baum-Connes Conjecture is unknown for all CAT(0)-groups. Can one use the techniques of the proof of the Farrell-Jones Conjecture for CAT(0)-groups to prove the Baum-Connes Conjecture for them? In particular it is not at all clear how the transfer methods in the Farrell-Jones setting carry over to the Baum-Connes case. In the other direction, the Dirac-Dual Dirac method, which is the main tool for proofs of the Baum-Connes Conjecture, lacks an analog on the Farrell-Jones side.

7.7. Classification of (non-aspherical) manifolds. The Farrell-Jones Conjecture is also very useful when one considers not necessarily aspherical manifolds. Namely, because of the surgery sequence (2.7), it gives an interpretation of the structure set as a relative homology group. So it simplifies the classification of manifolds substantially and opens the door to explicit

answers in favorable interesting cases. Here, a lot of work can and will have to be done.

References

[1] E. Alibegović and M. Bestvina. Limit groups are CAT(0). *J. London Math. Soc. (2)*, 74(1):259–272, 2006.

[2] G. Arzhantseva and T. Delzant. Examples of random groups. Preprint, 2008.

[3] A. Bartels and W. Lück. Induction theorems and isomorphism conjectures for K- and L-theory. *Forum Math.*, 19:379–406, 2007.

[4] A. Bartels and W. Lück. The Borel conjecture for hyperbolic and CAT(0)-groups. Preprintreihe SFB 478 — Geometrische Strukturen in der Mathematik, Heft 506 Münster, arXiv:0901.0442v1 [math.GT], 2009.

[5] A. Bartels and W. Lück. On twisted group rings with twisted involutions, their module categories and L-theory. In *Cohomology of groups and algebraic K-theory*, volume 12 of *Advanced Lectures in Mathematics*, pages 1–55, Somerville, U.S.A., 2009. International Press.

[6] A. Bartels, W. Lück, and H. Reich. Equivariant covers for hyperbolic groups. *Geom. Topol.*, 12(3):1799–1882, 2008.

[7] A. Bartels, W. Lück, and H. Reich. The K-theoretic Farrell-Jones conjecture for hyperbolic groups. *Invent. Math.*, 172(1):29–70, 2008.

[8] A. Bartels, W. Lück, and H. Reich. On the Farrell-Jones Conjecture and its applications. *Journal of Topology*, 1:57–86, 2008.

[9] A. Bartels, W. Lück, and S. Weinberger. On hyperbolic groups with spheres as boundary. arXiv:0911.3725v1 [math.GT], 2009.

[10] A. Bartels and H. Reich. Coefficients for the Farrell-Jones Conjecture. *Adv. Math.*, 209(1):337–362, 2007.

[11] A. C. Bartels. On the domain of the assembly map in algebraic K-theory. *Algebr. Geom. Topol.*, 3:1037–1050 (electronic), 2003.

[12] H. Bass. *Algebraic K-theory.* W. A. Benjamin, Inc., New York-Amsterdam, 1968.

[13] H. Bass. Euler characteristics and characters of discrete groups. *Invent. Math.*, 35:155–196, 1976.

[14] P. Baum, A. Connes, and N. Higson. Classifying space for proper actions and K-theory of group C^*-algebras. In *C^*-algebras: 1943–1993 (San Antonio, TX, 1993)*, pages 240–291. Amer. Math. Soc., Providence, RI, 1994.

[15] M. R. Bridson. Non-positive curvature and complexity for finitely presented groups. In *International Congress of Mathematicians. Vol. II*, pages 961–987. Eur. Math. Soc., Zürich, 2006.

[16] M. R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature.* Springer-Verlag, Berlin, 1999. Die Grundlehren der mathematischen Wissenschaften, Band 319.

[17] J. Bryant, S. Ferry, W. Mio, and S. Weinberger. Topology of homology manifolds. *Ann. of Math. (2)*, 143(3):435–467, 1996.

[18] S. Cappell, A. Ranicki, and J. Rosenberg, editors. *Surveys on surgery theory. Vol. 1.* Princeton University Press, Princeton, NJ, 2000. Papers dedicated to C. T. C. Wall.

[19] S. Cappell, A. Ranicki, and J. Rosenberg, editors. *Surveys on surgery theory. Vol. 2.* Princeton University Press, Princeton, NJ, 2001. Papers dedicated to C. T. C. Wall.

[20] G. Carlsson. Deloopings in algebraic K-theory. In *Handbook of K-theory. Vol. 1, 2*, pages 3–37. Springer, Berlin, 2005.

[21] R. M. Charney and M. W. Davis. Strict hyperbolization. *Topology*, 34(2):329–350, 1995.

[22] J. F. Davis, Q. Khan, and A. A. Ranicki. Algebraic K-theory over the infinite dihedral group. Preprint, arXiv:math.KT/0803.1639, 2008.

[23] J. F. Davis and W. Lück. Spaces over a category and assembly maps in isomorphism conjectures in K- and L-theory. *K-Theory*, 15(3):201–252, 1998.

[24] J. F. Davis, F. Quinn, and H. Reich. Algebraic K-theory over the infinite dihedral group: a controlled topology approach. Preprint, arXiv:math.KT/1002.3702v1, 2010.

[25] M. W. Davis. Poincaré duality groups. In *Surveys on surgery theory, Vol. 1*, volume 145 of *Ann. of Math. Stud.*, pages 167–193. Princeton Univ. Press, Princeton, NJ, 2000.

[26] M. W. Davis and T. Januszkiewicz. Hyperbolization of polyhedra. *J. Differential Geom.*, 34(2):347–388, 1991.

[27] S. Echterhoff, W. Lück, C. Philipps, and S. Walters. The structure of crossed products of irrational rotation algebras by finite subgroups of $sl_2(\digamma)$. *Crelle's Journal für reine und angewandte Mathematik*, 639:141–221, 2010.

[28] F. T. Farrell and L. E. Jones. K-theory and dynamics. I. *Ann. of Math. (2)*, 124(3):531–569, 1986.

[29] F. T. Farrell and L. E. Jones. Negatively curved manifolds with exotic smooth structures. *J. Amer. Math. Soc.*, 2(4):899–908, 1989.

[30] F. T. Farrell and L. E. Jones. Isomorphism conjectures in algebraic K-theory. *J. Amer. Math. Soc.*, 6(2):249–297, 1993.

[31] F. T. Farrell and L. E. Jones. Topological rigidity for compact non-positively curved manifolds. In *Differential geometry: Riemannian geometry (Los Angeles, CA, 1990)*, pages 229–274. Amer. Math. Soc., Providence, RI, 1993.

[32] F. T. Farrell and L. E. Jones. The lower algebraic K-theory of virtually infinite cyclic groups. *K-Theory*, 9(1):13–30, 1995.

[33] F. T. Farrell and L. E. Jones. Rigidity for aspherical manifolds with $\pi_1 \subset GL_m(\mathbb{R})$. *Asian J. Math.*, 2(2):215–262, 1998.

[34] F. T. Farrell, L. E. Jones, and W. Lück. A caveat on the isomorphism conjecture in L-theory. *Forum Math.*, 14(3):413–418, 2002.

[35] S. C. Ferry and A. A. Ranicki. A survey of Wall's finiteness obstruction. In *Surveys on surgery theory, Vol. 2*, volume 149 of *Ann. of Math. Stud.*, pages 63–79. Princeton Univ. Press, Princeton, NJ, 2001.

[36] S. C. Ferry, A. A. Ranicki, and J. Rosenberg. A history and survey of the Novikov conjecture. In *Novikov conjectures, index theorems and rigidity, Vol. 1 (Oberwolfach, 1993)*, pages 7–66. Cambridge Univ. Press, Cambridge, 1995.

[37] S. C. Ferry, A. A. Ranicki, and J. Rosenberg, editors. *Novikov conjectures, index theorems and rigidity. Vol. 1 and 2.* Cambridge University Press, Cambridge, 1995. Including papers from the conference held at the Mathematisches Forschungsinstitut Oberwolfach, Oberwolfach, September 6–10, 1993.

[38] M. Gromov. Asymptotic invariants of infinite groups. In *Geometric group theory, Vol. 2 (Sussex, 1991)*, pages 1–295. Cambridge Univ. Press, Cambridge, 1993.

[39] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces.* Modern Birkhäuser Classics. Birkhäuser Boston Inc., Boston, MA, english edition, 2007. Based on the 1981 French original, With appendices by M. Katz, P. Pansu and S. Semmes, Translated from the French by Sean Michael Bates.

[40] N. Higson. The Baum-Connes conjecture. In *Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998)*, pages 637–646 (electronic), 1998.

[41] N. Higson, V. Lafforgue, and G. Skandalis. Counterexamples to the Baum-Connes conjecture. *Geom. Funct. Anal.*, 12(2):330–354, 2002.

[42] F. E. A. Johnson and C. T. C. Wall. On groups satisfying Poincaré duality. *Ann. of Math. (2)*, 96:592–598, 1972.

[43] I. Kapovich and N. Benakli. Boundaries of hyperbolic groups. In *Combinatorial and geometric group theory (New York, 2000/Hoboken, NJ, 2001)*, volume 296 of *Contemp. Math.*, pages 39–93. Amer. Math. Soc., Providence, RI, 2002.

[44] M. A. Kervaire. Le théorème de Barden-Mazur-Stallings. *Comment. Math. Helv.*, 40:31–42, 1965.

[45] R. C. Kirby and L. C. Siebenmann. *Foundational essays on topological manifolds, smoothings, and triangulations.* Princeton University Press, Princeton, N.J., 1977. With notes by J. Milnor and M. F. Atiyah, Annals of Mathematics Studies, No. 88.

[46] M. Kreck. Surgery and duality. *Ann. of Math. (2)*, 149(3):707–754, 1999.

[47] M. Kreck and W. Lück. *The Novikov conjecture: Geometry and algebra*, volume 33 of *Oberwolfach Seminars.* Birkhäuser Verlag, Basel, 2005.

[48] M. Kreck and W. Lück. Topological rigidity for non-aspherical manifolds. *Pure and Applied Mathematics Quarterly*, 5 (3):873–914, 2009. special issue in honor of Friedrich Hirzebruch.

[49] W. Lück. The geometric finiteness obstruction. *Proc. London Math. Soc. (3)*, 54(2):367–384, 1987.

[50] W. Lück. *Transformation groups and algebraic K-theory*, volume 1408 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 1989.

[51] W. Lück. The type of the classifying space for a family of subgroups. *J. Pure Appl. Algebra*, 149(2):177–203, 2000.

[52] W. Lück. A basic introduction to surgery theory. In F. T. Farrell, L. Göttsche, and W. Lück, editors, *High dimensional manifold theory*, number 9 in ICTP Lecture Notes, pages 1–224. Abdus Salam International Centre for Theoretical Physics, Trieste, 2002. Proceedings of the summer school "High dimensional manifold theory" in Trieste May/June 2001, Number 1. http://www.ictp.trieste.it/˜pub_off/lectures/vol9.html.

[53] W. Lück. Chern characters for proper equivariant homology theories and applications to K- and L-theory. *J. Reine Angew. Math.*, 543:193–234, 2002.

[54] W. Lück. *L^2-Invariants: Theory and Applications to Geometry and K-Theory*, volume 44 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin, 2002.

[55] W. Lück. The relation between the Baum-Connes conjecture and the trace conjecture. *Invent. Math.*, 149(1):123–152, 2002.

[56] W. Lück. Equivariant cohomological Chern characters. *Internat. J. Algebra Comput.*, 15(5-6):1025–1052, 2005.

[57] W. Lück. K- and L-theory of the semi-direct product of the discrete 3-dimensional Heisenberg group by $\mathbb{Z}/4$. *Geom. Topol.*, 9:1639–1676 (electronic), 2005.

[58] W. Lück. Survey on classifying spaces for families of subgroups. In *Infinite groups: geometric, combinatorial and dynamical aspects*, volume 248 of *Progr. Math.*, pages 269–322. Birkhäuser, Basel, 2005.

[59] W. Lück. Survey on aspherical manifolds. Preprintreihe SFB 478 — Geometrische Strukturen in der Mathematik, Heft 511, Münster, arXiv:0902.2480v1 [math.GT], to appear in the Proceedings of the 5-th ECM 2008 in Amsterdam, 2008.

[60] W. Lück. Survey on geometric group theory. *Münster J. of Mathematics*, 1:73–108, 2008.

[61] W. Lück. On the Farrell-Jones Conjecture and related conjectures. In *Cohomology of groups and algebraic K-theory*, volume 12 of *Advanced Lectures in Mathematics*, pages 269–341, Somervile, U.S.A., 2009. International Press.

[62] W. Lück and D. Meintrup. On the universal space for group actions with compact isotropy. In *Geometry and topology: Aarhus (1998)*, pages 293–305. Amer. Math. Soc., Providence, RI, 2000.

[63] W. Lück and H. Reich. The Baum-Connes and the Farrell-Jones conjectures in K- and L-theory. In *Handbook of K-theory. Vol. 1, 2*, pages 703–842. Springer, Berlin, 2005.

[64] W. Lück and M. Weiermann. On the classifying space of the family of virtually cyclic subgroups. Preprintreihe SFB 478 — Geometrische Strukturen in der Mathematik, Heft 453, Münster, arXiv:math.AT/0702646v2, to appear in the Proceedings in honour of Farrell and Jones in Pure and Applied Mathematic Quarterly, 2007.

[65] D. Meintrup and T. Schick. A model for the universal space for proper actions of a hyperbolic group. *New York J. Math.*, 8:1–7 (electronic), 2002.

[66] J. Milnor. *Lectures on the h-cobordism theorem.* Princeton University Press, Princeton, N.J., 1965.

[67] G. Mislin. Wall's finiteness obstruction. In *Handbook of algebraic topology*, pages 1259–1291. North-Holland, Amsterdam, 1995.

[68] G. Mislin and A. Valette. *Proper group actions and the Baum-Connes conjecture.* Advanced Courses in Mathematics. CRM Barcelona. Birkhäuser Verlag, Basel, 2003.

[69] S. P. Novikov. Topological invariance of rational classes of Pontrjagin. *Dokl. Akad. Nauk SSSR*, 163:298–300, 1965.

[70] A. Y. Ol'shanskii. An infinite simple torsion-free Noetherian group. *Izv. Akad. Nauk SSSR Ser. Mat.*, 43(6):1328–1393, 1979.

[71] A. Y. Ol'shanskiĭ, D. V. Osin, and M. V. Sapir. Lacunary hyperbolic groups. *Geom. Topol.*, 13(4):2051–2140, 2009. With an appendix by Michael Kapovich and Bruce Kleiner.

[72] E. K. Pedersen. On the K_{-i}-functors. *J. Algebra*, 90(2):461–475, 1984.

[73] E. K. Pedersen and C. A. Weibel. A non-connective delooping of algebraic K-theory. In *Algebraic and Geometric Topology; proc. conf. Rutgers Uni., New Brunswick 1983*, volume 1126 of *Lecture Notes in Mathematics*, pages 166–181. Springer, 1985.

[74] D. Quillen. Higher algebraic K-theory. I. In *Algebraic K-theory, I: Higher K-theories (Proc. Conf., Battelle Memorial Inst., Seattle, Wash., 1972)*, pages 85–147. Lecture Notes in Math., Vol. 341. Springer-Verlag, Berlin, 1973.

[75] F. Quinn. A geometric formulation of surgery. In *Topology of Manifolds (Proc. Inst., Univ. of Georgia, Athens, Ga., 1969)*, pages 500–511. Markham, Chicago, Ill., 1970.

[76] F. Quinn. An obstruction to the resolution of homology manifolds. *Michigan Math. J.*, 34(2):285–291, 1987.

[77] F. Quinn. Hyperelementary assembly for K-theory of virtually abelian groups. Preprint, arXiv:math.KT/0509294, 2005.

[78] A. A. Ranicki. *Exact sequences in the algebraic theory of surgery.* Princeton University Press, Princeton, N.J., 1981.

[79] A. A. Ranicki. *Algebraic L-theory and topological manifolds.* Cambridge University Press, Cambridge, 1992.

[80] A. A. Ranicki. *Lower K- and L-theory.* Cambridge University Press, Cambridge, 1992.

[81] A. A. Ranicki. *Algebraic and geometric surgery.* Oxford Mathematical Monographs. Clarendon Press, Oxford, 2002.

[82] J. Rosenberg. *Algebraic K-theory and its applications.* Springer-Verlag, New York, 1994.

[83] S. K. Roushon. The Farrell-Jones isomorphism conjecture for 3-manifold groups. *J. K-Theory*, 1(1):49–82, 2008.

[84] T. tom Dieck. Orbittypen und äquivariante Homologie. I. *Arch. Math. (Basel)*, 23:307–317, 1972.

[85] F. Waldhausen. Algebraic K-theory of spaces. In *Algebraic and geometric topology (New Brunswick, N.J., 1983)*, pages 318–419. Springer-Verlag, Berlin, 1985.

[86] C. T. C. Wall. Finiteness conditions for CW-complexes. *Ann. of Math. (2)*, 81:56–69, 1965.

[87] C. T. C. Wall. Poincaré complexes. I. *Ann. of Math. (2)*, 86:213–245, 1967.

[88] C. T. C. Wall. *Surgery on compact manifolds*. American Mathematical Society, Providence, RI, second edition, 1999. Edited and with a foreword by A. A. Ranicki.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Moduli Problems for Ring Spectra

Jacob Lurie*

Abstract

In algebraic geometry, it is common to study a geometric object X (such as a scheme) by means of the functor $R \mapsto \mathrm{Hom}(\mathrm{Spec}\, R, X)$ represented by X. In this paper, we consider functors which are defined on larger classes of rings (such as the class of *ring spectra* which arise in algebraic topology), and sketch some applications to deformation theory.

Mathematics Subject Classification (2010). Primary 55P43, Secondary 14B12.

Keywords. Structured ring spectra, deformation theory, derived algebraic geometry.

Introduction

The following thesis plays a central role in deformation theory:

($*$) If X is a moduli space over a field k of characteristic zero, then a formal neighborhood of any point $x \in X$ is controlled by a differential graded Lie algebra.

This idea was developed in unpublished work of Deligne, Drinfeld, and Feigin, and has powerfully influenced subsequent contributions of Hinich, Kontsevich-Soibelman, Manetti, and many others. The goal of this paper is to give a precise formulation of ($*$) using the language of higher category theory. Our main result is Theorem 6.20, which can be regarded as an analogue of ($*$) in the setting of noncommutative geometry. Our proof uses a method which can be adapted to prove a version of ($*$) itself (Theorem 5.3).

Let us now outline the contents of this paper. Our first step is to define precisely what we mean by a moduli space. We will adopt Grothendieck's "functor of points" philosophy: giving the moduli space X is equivalent to specifying the functor $R \mapsto X(R) = \mathrm{Hom}(\mathrm{Spec}\, R, X)$. We will consider several variations on this theme:

*Harvard University. E-mail: lurie@math.harvard.edu.

(a) Allowing R to range over the category Ring of commutative rings, we obtain the notion of a *classical moduli problem* (Definition 1.3). We will discuss this notion and give several examples in §1.

(b) To understand the deformation theory of a moduli space X, it is often useful to extend the definition of the functor $R \mapsto X(R)$ to a more general class of rings. Algebraic topology provides such a generalization via the theory of *E_∞-ring spectra* (or, as we will call them, *E_∞-rings*). We will review this theory in §3 and use it to formulate the notion of a *derived moduli problem* (Definition 3.3).

(c) Let k be a field. To study the local structure of a moduli space X near a point $x \in X(k)$, it is useful to restrict our attention to the values $X(R)$ where R is a ring which is, in some sense, very similar to k (for example, local Artin algebras having residue field k). In §4, we will make this precise by introducing the notion of a *formal moduli problem* (Definition 4.6).

(d) Another way of enlarging the category of commutative rings is by weakening the requirement of commutativity. In the setting of ring spectra there are several flavors of commutativity available, given by the theory of *E_n-rings* for $0 \leq n < \infty$. We will review the theory of E_n-rings in §6, and use it to formulate the notion of a *formal E_n-moduli problem*.

In order to adequately treat cases (b) through (d), it is important to note that for $0 \leq n \leq \infty$, an E_n-ring is an essentially homotopy-theoretic object, and should therefore be treated using the formalism of higher category theory. In §2 we will give an overview of this formalism; in particular, we introduce the notion of an *∞-category* (Definition 2.9). Most of the basic objects under consideration in this paper form ∞-categories, and the main results announced here can be formulated as equivalences of ∞-categories:

($*'$) If k is a field of characteristic zero, then the ∞-category of formal moduli problems over k is equivalent to the ∞-category of differential graded Lie algebras over k (Theorem 5.3).

($*''$) If k is any field and $0 \leq n < \infty$, the ∞-category of formal E_n moduli problems over k is equivalent to the ∞-category of augmented E_n-algebras over k (Theorem 6.20).

We will formulate these statements more precisely in §5 and §6, respectively.

Remark 0.1. The subject of deformation theory has a voluminous literature, some of which has substantial overlap with the material discussed in this paper. Though we have tried to provide relevant references in the body of the text, there are undoubtedly many sins of omission for which we apologize in advance.

Warning 0.2. The approach to the study of deformation theory described in this paper makes extensive use of higher category theory. We will sketch some

of the central ideas of this theory in §2, and then proceed to use these ideas in an informal way. For a more comprehensive approach, we refer the reader to the author's book [22] and the series of papers [23].

I would like to thank David Ben-Zvi, Vladimir Drinfeld, Pavel Etingof, John Francis, Dennis Gaitsgory, Mike Hopkins, David Nadler, Bertrand Toën, and Gabriele Vezzosi for helpful conversations related to the subject of this paper.

1. Moduli Problems for Commutative Rings

Let Ring denote the category of commutative rings and $\mathcal{S}$et the category of sets. Throughout this paper, we will make extensive use of Grothendieck's "functor of points" philosophy: that is, we will identify a geometric object X (such as a scheme) with the functor $\mathrm{Ring} \to \mathcal{S}\mathrm{et}$ represented by X, given by the formula $R \mapsto \mathrm{Hom}(\mathrm{Spec}\, R, X)$.

Example 1.1. Let $X : \mathrm{Ring} \to \mathcal{S}\mathrm{et}$ be the functor which assigns to each commutative ring R the set $R^\times$ of invertible elements of R. For any commutative ring R, we have a canonical bijection $X(R) = R^\times \simeq \mathrm{Hom}_{\mathrm{Ring}}(\mathbf{Z}[t^{\pm 1}], R)$. In other words, we can identify X with the functor represented by the commutative ring $\mathbf{Z}[t^{\pm 1}]$.

Example 1.2. Fix an integer $n \geq 0$. We define a functor $X : \mathrm{Ring} \to \mathcal{S}\mathrm{et}$ by letting $F(R)$ denote the set of all submodules $M \subseteq R^{n+1}$ such that the quotient R^{n+1}/M is a projective R-module of rank n (from which it follows that M is a projective R-module of rank 1). The functor X is not representable by a commutative ring. However, it is representable in the larger category Sch of *schemes*. That is, for any commutative ring R we have a canonical bijection $X(R) \simeq \mathrm{Hom}_{\mathrm{Sch}}(\mathrm{Spec}\, R, \mathbb{P}^n)$, where $\mathbb{P}^n \simeq \mathrm{Proj}\, \mathbf{Z}[x_0, \ldots, x_n]$ denotes projective space of dimension n.

For some purposes, the notion of a functor $X : \mathrm{Ring} \to \mathcal{S}\mathrm{et}$ is too restrictive. We often want to study moduli problems X which assign to a commutative ring R some class of geometric objects which depend on R. The trouble is that this collection of geometric objects is naturally organized into a category, rather than a set. This motivates the following definition:

Definition 1.3. Let Gpd denote the collection of *groupoids*: that is, categories in which every morphism is an isomorphism. We regard Gpd as a 2-category: morphisms are given by functors between groupoids, and 2-morphisms are given by natural transformations (which are automatically invertible). A *classical moduli problem* is a functor $X : \mathrm{Ring} \to \mathrm{Gpd}$.

Remark 1.4. Every set S can be regarded as a groupoid by setting

$$\mathrm{Hom}_S(x, y) = \begin{cases} \{\mathrm{id}_x\} & \text{if } x = y \\ \emptyset & \text{if } x \neq y. \end{cases}$$

This construction allows us to identify the category $\mathcal{S}\mathrm{et}$ with a full subcategory of the 2-category Gpd. In particular, every functor $X : \mathrm{Ring} \to \mathcal{S}\mathrm{et}$ can be identified with a classical moduli problem in the sense of Definition 1.3.

Example 1.5. For every commutative ring R, let $X(R)$ be the category of elliptic curves $E \to \operatorname{Spec} R$ (morphisms in the category $X(R)$ are given by isomorphisms of elliptic curves). Then F determines a functor $\mathrm{Ring} \to \mathrm{Gpd}$, and can therefore be regarded as a moduli problem in the sense of Definition 1.3. This moduli problem cannot be represented by a commutative ring or even by a scheme: for any scheme Y, $\mathrm{Hom}_{\mathrm{Sch}}(\operatorname{Spec} R, Y)$ is a set. In particular, if we regard $\mathrm{Hom}_{\mathrm{Sch}}(\operatorname{Spec} R, Y)$ as a groupoid, every object has a trivial automorphism group. In contrast, every object of $X(R)$ has a *nontrivial* automorphism group: every elliptic curve admits a nontrivial automorphism, given by multiplication by -1.

Nevertheless, the moduli problem X is representable if we work not in the category of schemes but in the larger 2-category $\mathrm{St}_{\mathrm{DM}}$ of *Deligne-Mumford stacks*. More precisely, there exists a Deligne-Mumford stack $\mathcal{M}_{\mathrm{Ell}}$ (the *moduli stack of elliptic curves*) for which there is a canonical equivalence of categories $X(R) \simeq \mathrm{Hom}_{\mathrm{St}_{\mathrm{DM}}}(\operatorname{Spec} R, \mathcal{M}_{\mathrm{Ell}})$ for every commutative ring R.

Example 1.6. Fix an integer $n \geq 0$. For every commutative ring R, let $X(R)$ denote the category whose objects are projective R-modules of rank n, and whose morphisms are given by isomorphisms of R-modules. Then X can be regarded as a moduli problem $\mathrm{Ring} \to \mathrm{Gpd}$. This moduli problem is not representable in the 2-category $\mathrm{St}_{\mathrm{DM}}$ of Deligne-Mumford stacks, because projective R-modules admit continuous families of automorphisms. However, F is representable in the larger 2-category $\mathrm{St}_{\mathrm{Art}}$ of *Artin stacks*. Namely, there is an Artin stack $\mathrm{BGL}(n) \in \mathrm{St}_{\mathrm{Art}}$ for which there is a canonical bijection $X(R) \simeq \mathrm{Hom}_{\mathrm{St}_{\mathrm{Art}}}(\operatorname{Spec} R, \mathrm{BGL}(n))$ for every commutative ring R.

2. Higher Category Theory

In §1, we discussed the notion of a moduli problem in classical algebraic geometry. Even very simple moduli problems involve the classification of geometric objects which admit nontrivial automorphisms, and should therefore be treated as categories rather than as sets (Examples 1.5 and 1.6). Consequently, moduli problems themselves (and the geometric objects which represent them) are organized not into a category, but into a 2-category. Our discussion in this paper will take us much further into the realm of higher categories. We will devote this section to providing an informal overview of the ideas involved.

Definition 2.1 (Informal). Let $n \geq 0$ be a nonnegative integer. The notion of an *n-category* is defined by induction on n. If $n = 0$, an n-category is simply a set. If $n > 0$, an n-category $\mathcal{C}$ consists of the following:

(1) A collection of objects $X, Y, Z, \ldots$

(2) For every pair of objects $X, Y \in \mathcal{C}$, an $(n-1)$-category $\mathrm{Hom}_{\mathcal{C}}(X, Y)$.

(3) Composition laws $\phi_{X,Y,Z} : \mathrm{Hom}_{\mathcal{C}}(X, Y) \times \mathrm{Hom}_{\mathcal{C}}(Y, Z) \to \mathrm{Hom}_{\mathcal{C}}(X, Z)$ which are required to be unital and associative.

If η is an object of the $(n-1)$-category $\mathrm{Hom}_{\mathcal{C}}(X, Y)$ for some pair of objects $X, Y \in \mathcal{C}$, then we will say that η is a 1-*morphism of* $\mathcal{C}$. More generally, a k-*morphism* in $\mathcal{C}$ is a $(k-1)$-morphism in some $(n-1)$-category $\mathrm{Hom}_{\mathcal{C}}(X, Y)$.

Example 2.2. Every topological space X determines an n-category $\pi_{\leq n} X$, the *fundamental n-groupoid of* X. If $n = 0$, we let $\pi_{\leq n} X = \pi_0 X$ be the set of path components of X. For $n > 0$, we let $\pi_{\leq n} X$ be the n-category whose objects are points of X, where $\mathrm{Hom}_{\pi_{\leq n}}(x, y)$ is the fundamental $(n-1)$-groupoid $\pi_{\leq n-1} P_{x,y}(X)$, where $P_{x,y}(X) = \{p : [0, 1] \to X : p(0) = x, p(1) = y\}$ is the space of paths from x to y in X. Composition in $\pi_{\leq n} X$ is given by concatenation of paths. If $n = 1$, this definition recovers the usual fundamental groupoid of X.

Definition 2.1 is informal because we did not specify precisely what sort of associative law the composition in $\mathcal{C}$ is required to satisfy. If $n = 1$, there is no real ambiguity and Definition 2.1 recovers the usual definition of a category. When $n = 2$, the situation is more subtle: the associative law should posit the commutativity of a diagram having the form

$$\begin{array}{ccc} \mathrm{Hom}_{\mathcal{C}}(W, X) \times \mathrm{Hom}_{\mathcal{C}}(X, Y) \times \mathrm{Hom}_{\mathcal{C}}(Y, Z) & \xrightarrow{\phi_{W,X,Y}} & \mathrm{Hom}_{\mathcal{C}}(W, Y) \times \mathrm{Hom}_{\mathcal{C}}(Y, Z) \\ \downarrow{\scriptstyle \phi_{X,Y,Z}} & & \downarrow{\scriptstyle \phi_{W,Y,Z}} \\ \mathrm{Hom}_{\mathcal{C}}(W, X) \times \mathrm{Hom}_{\mathcal{C}}(X, Z) & \xrightarrow{\phi_{W,X,Z}} & \mathrm{Hom}_{\mathcal{C}}(W, Z). \end{array}$$

Since this is a diagram of categories and functors, rather than sets and functions, we are faced with a question: do we require this diagram to commute "on the nose" or only up to isomorphism? In the former case, we obtain the definition of a *strict* 2-*category*. This generalizes in a straightforward way: we can require strict associativity in Definition 2.1 to obtain a notion of strict n-category for every n. However, this notion turns out to be of limited use. For example, the fundamental n-groupoid of a topological space $\pi_{\leq n} X$ usually cannot be realized as a strict n-category when $n > 2$.

To accommodate Example 2.2, it is necessary to interpret Definition 2.1 differently. In place of equality, we require the existence of *isomorphisms*

$$\gamma_{W,X,Y,Z} : \phi_{W,X,Z} \circ (\mathrm{id}_{\mathrm{Hom}_{\mathcal{C}}(W,X)} \times \phi_{X,Y,Z}) \simeq \phi_{W,Y,Z} \circ (\phi_{W,X,Y} \times \mathrm{id}_{\mathrm{Hom}_{\mathcal{C}}(Y,Z)}).$$

These isomorphisms are themselves part of the structure of $\mathcal{C}$, and are required to satisfy certain coherence conditions. When $n > 2$, these coherence conditions are themselves only required to hold up to isomorphism: *these* isomorphisms must also be specified and required to satisfy further coherences, and so forth.

As n grows, it becomes prohibitively difficult to specify these coherences explicitly.

The situation is dramatically simpler if we wish to study not arbitrary n-categories, but *n-groupoids*. An n-category $\mathcal{C}$ is called an n-groupoid if every k-morphism in $\mathcal{C}$ is invertible. If X is any topological space, then the n-category $\pi_{\leq n}X$ is an example of an n-groupoid: for example, the 1-morphisms in $\pi_{\leq n}X$ are given by paths $p : [0,1] \to X$, and every path p has an inverse q (up to homotopy) given by $q(t) = p(1-t)$. In fact, all n-groupoids arise in this way. To formulate this more precisely, let us recall that a topological space X is an *n-type* if the homotopy groups $\pi_m(X,x)$ are trivial for every $m > n$ and every point $x \in X$. The following idea goes back (at least) to Grothendieck:

Thesis 2.3. *The construction $X \mapsto \pi_{\leq n}X$ establishes a bijective correspondence between n-types (up to weak homotopy equivalence) and n-groupoids (up to equivalence).*

We call Thesis 2.3 a thesis, rather than a theorem, because we have not given a precise definition of n-categories (or n-groupoids) in this paper. Thesis 2.3 should instead be regarded as a requirement that any reasonable definition of n-category must satisfy: when we restrict to n-categories where all morphisms are invertible, we should recover the usual homotopy theory of n-types. On the other hand, it is easy to concoct a definition of n-groupoid which tautologically satisfies this requirement:

Definition 2.4. An *n-groupoid* is an n-type.

Definition 2.4 has an evident extension to the case $n = \infty$:

Definition 2.5. An *∞-groupoid* is a topological space.

It is possible to make sense of Definition 2.1 also in the case where $n = \infty$: that is, we can talk about higher categories which have k-morphisms for every positive integer k. In the case where all of these morphisms turn out to be invertible, this reduces to the classical homotopy theory of topological spaces. We will be interested in the next simplest case:

Definition 2.6 (Informal). An *$(\infty,1)$-category* is an ∞-category in which every k-morphism is invertible for $k > 1$.

In other words, an $(\infty,1)$-category $\mathcal{C}$ consists of a collection of objects together with an ∞-groupoid $\mathrm{Hom}_{\mathcal{C}}(X,Y)$ for every pair of objects $X, Y \in \mathcal{C}$, which are equipped with an associative composition law. We can therefore use Definition 2.5 to formulate a more precise version of Definition 2.6.

Definition 2.7. A *topological category* is a category $\mathcal{C}$ for which each of the sets $\mathrm{Hom}_{\mathcal{C}}(X,Y)$ is equipped with a topology, and each of the compositions maps $\mathrm{Hom}_{\mathcal{C}}(X,Y) \times \mathrm{Hom}_{\mathcal{C}}(Y,Z) \to \mathrm{Hom}_{\mathcal{C}}(X,Z)$ is continuous. If $\mathcal{C}$ and $\mathcal{D}$ are topological categories, we will say that a functor $F : \mathcal{C} \to \mathcal{D}$ is *continuous* if,

for every pair of objects $X, Y \in \mathcal{C}$, the map $\mathrm{Hom}_{\mathcal{C}}(X, Y) \to \mathrm{Hom}_{\mathcal{D}}(FX, FY)$ is continuous. The collection of (small) topological categories and continuous functors forms a category, which we will denote by $\mathcal{C}\mathrm{at}_t$.

Construction 2.8. Let $\mathcal{C}$ be a topological category. We can associate to $\mathcal{C}$ an ordinary category $\mathrm{h}\mathcal{C}$ as follows:

- The objects of $\mathrm{h}\mathcal{C}$ are the objects of $\mathcal{C}$.
- For every pair of objects $X, Y \in \mathcal{C}$, we let $\mathrm{Hom}_{\mathrm{h}\mathcal{C}}(X, Y) = \pi_0 \mathrm{Hom}_{\mathcal{C}}(X, Y)$: that is, maps from X to Y in $\mathrm{h}\mathcal{C}$ are homotopy classes of maps from X to Y in $\mathcal{C}$.

We say that a morphism f in $\mathcal{C}$ is an *equivalence* if the image of f in $\mathrm{h}\mathcal{C}$ is an isomorphism.

Definition 2.9. Let $F : \mathcal{C} \to \mathcal{D}$ be a continuous functor between topological categories. We will say that F is a *weak equivalence* if the following conditions are satisfied:

(1) The functor F induces an equivalence of ordinary categories $\mathrm{h}\mathcal{C} \to \mathrm{h}\mathcal{D}$.

(2) For every pair of objects $X, Y \in \mathcal{C}$, the induced map

$$\mathrm{Hom}_{\mathcal{C}}(X, Y) \to \mathrm{Hom}_{\mathcal{D}}(FX, FY)$$

is a weak homotopy equivalence.

Let $\mathrm{h}\mathcal{C}\mathrm{at}_\infty$ be the category obtained from $\mathcal{C}\mathrm{at}_t$ by formally inverting the collection of weak equivalences. An $(\infty, 1)$-*category* is an object of $\mathrm{h}\mathcal{C}\mathrm{at}_\infty$. We will refer to $\mathrm{h}\mathcal{C}\mathrm{at}_\infty$ as the *homotopy category of* $(\infty, 1)$-*categories*.

Remark 2.10. More precisely, we should say that $\mathrm{h}\mathcal{C}\mathrm{at}_\infty$ is the homotopy category of *small* $(\infty, 1)$-categories. We will also consider $(\infty, 1)$-categories which are not small.

Remark 2.11. There are numerous approaches to the theory of $(\infty, 1)$-categories which are now known to be equivalent, in the sense that they generate categories equivalent to $\mathrm{h}\mathcal{C}\mathrm{at}_\infty$. The approach described above (based on Definitions 2.7 and 2.9) is probably the easiest to grasp psychologically, but is one of the most difficult to actually work with. We refer the reader to [1] for a description of some alternatives to Definition 2.7 and their relationship to one another.

All of the higher categories we consider in this paper will have k-morphisms invertible for $k > 1$. Consequently, it will be convenient for us to adopt the following:

Convention 2.12. The term ∞-*category* will refer to an $(\infty, 1)$-category $\mathcal{C}$ in the sense of Definition 2.9. That is, we will implicitly assume that all k-morphisms in $\mathcal{C}$ are invertible for $k > 1$.

With some effort, one can show that Definition 2.7 gives rise to a rich and powerful theory of ∞-categories, which admits generalizations of most of the important ideas from classical category theory. For example, one can develop ∞-categorical analogues of the theories of limits, colimits, adjoint functors, sheaves, and so forth. Throughout this paper, we will make free use of these ideas; for details, we refer the reader to [22].

Example 2.13. Let $\mathcal{C}$ and $\mathcal{D}$ be ∞-categories. Then there exists another ∞-category $\mathrm{Fun}(\mathcal{C}, \mathcal{D})$ with the following universal property: for every ∞-category $\mathcal{C}'$, there is a canonical bijection

$$\mathrm{Hom}_{\mathrm{h}\mathcal{C}\mathrm{at}_\infty}(\mathcal{C}', \mathrm{Fun}(\mathcal{C}, \mathcal{D})) \simeq \mathrm{Hom}_{\mathrm{h}\mathcal{C}\mathrm{at}_\infty}(\mathcal{C} \times \mathcal{C}', \mathcal{D}).$$

We will refer to objects of $\mathrm{Fun}(\mathcal{C}, \mathcal{D})$ simply as *functors* from $\mathcal{C}$ to $\mathcal{D}$.

Warning 2.14. By definition, an ∞-category $\mathcal{C}$ is simply an object of $\mathrm{h}\mathcal{C}\mathrm{at}_\infty$: that is, a topological category. However, there are generally objects of $\mathrm{Fun}(\mathcal{C}, \mathcal{D})$ which are not given by continuous functors between the underlying topological categories.

Warning 2.15. The process of generalizing from the setting of ordinary categories to the setting of ∞-categories is not always straightforward. For example, if $\mathcal{C}$ is an ordinary category, then a *product* of a pair of objects X and Y is another object Z equipped with a pair of maps $X \leftarrow Z \rightarrow Y$ having the following property: for every object $C \in \mathcal{C}$, the induced map $\theta : \mathrm{Hom}_{\mathcal{C}}(C, Z) \rightarrow \mathrm{Hom}_{\mathcal{C}}(C, X) \times \mathrm{Hom}_{\mathcal{C}}(C, Y)$ is a bijection. In the ∞-categorical context, it is natural to demand not that θ is bijective but instead that it is a weak homotopy equivalence. Consequently, products in $\mathcal{C}$ viewed as an ordinary category (enriched over topological spaces) are not necessarily the same as products in $\mathcal{C}$ when viewed as an ∞-category. To avoid confusion, limits and colimits in the ∞-category $\mathcal{C}$ are sometimes called *homotopy limits* and *homotopy colimits*.

We close this section by describing a method which can be used to construct a large class of examples of ∞-categories.

Construction 2.16. Let $\mathcal{C}$ be an ordinary category and let W be a collection of morphisms in $\mathcal{C}$. Then we let $\mathcal{C}[W^{-1}]$ denote an ∞-category which is equipped with a functor $\alpha : \mathcal{C} \rightarrow \mathcal{C}[W^{-1}]$ having the following universal property: for every ∞-category $\mathcal{D}$, composition with α induces a fully faithful embedding

$$\mathrm{Fun}(\mathcal{C}[W^{-1}], \mathcal{D}) \rightarrow \mathrm{Fun}(\mathcal{C}, \mathcal{D})$$

whose essential image consists of those functors which carry every morphism in W to an equivalence in $\mathcal{D}$. More informally: $\mathcal{C}[W^{-1}]$ is the ∞-category obtained from $\mathcal{C}$ by formally inverting the morphisms in W.

Example 2.17. Let $\mathcal{C}$ be the category of all topological spaces and let W be the collection of weak homotopy equivalences. We will refer to $\mathcal{C}[W^{-1}]$ as the *∞-category of spaces*, and denote it by $\mathcal{S}$.

Example 2.18. Let R be an associative ring and let Chain_R denote the category of chain complexes of R-modules. A morphism $f : M_\bullet \to N_\bullet$ in Chain_R is said to be a *quasi-isomorphism* if the induced map of homology groups $\mathrm{H}_n(M) \to \mathrm{H}_n(N)$ is an isomorphism for every integer n. Let W be the collection of all quasi-isomorphisms in $\mathcal{C}$; then $\mathrm{Chain}_R[W^{-1}]$ is an ∞-category which we will denote by Mod_R. The homotopy category hMod_R can be identified with the classical *derived category of R-modules*.

Example 2.19. Let k be a field of characteristic zero. A *differential graded Lie algebra* over k is a Lie algebra object of the category Chain_k: that is, a chain complex of k-vector spaces $\mathfrak{g}_\bullet$ equipped with a Lie bracket operation $[,] : \mathfrak{g}_\bullet \otimes \mathfrak{g}_\bullet \to \mathfrak{g}_\bullet$ which satisfies the identities

$$[x, y] + (-1)^{d(x)d(y)}[y, x] = 0$$

$$(-1)^{d(z)d(x)}[x, [y, z]] + (-1)^{d(x)d(y)}[y, [z, x]] + (-1)^{d(y)d(z)}[z, [x, y]] = 0$$

for homogeneous elements $x \in \mathfrak{g}_{d(x)}$, $y \in \mathfrak{g}_{d(y)}$, $z \in \mathfrak{g}_{d(z)}$. Let $\mathcal{C}$ be the category of differential graded Lie algebras over k and let W be the collection of morphisms in $\mathcal{C}$ which induce a quasi-isomorphism between the underlying chain complexes. Then $\mathcal{C}[W^{-1}]$ is an ∞-category which we will denote by $\mathrm{Lie}_k^{\mathrm{dg}}$; we will refer to $\mathrm{Lie}_k^{\mathrm{dg}}$ as the *∞-category of differential graded Lie algebras over k*.

Example 2.20. Let $\mathcal{C}\mathrm{at}_t$ be the ordinary category of Definition 2.9, whose objects are topologically enriched categories and whose morphisms are continuous functors. Let W be the collection of all weak equivalences in $\mathcal{C}\mathrm{at}_t$ and set $\mathcal{C}\mathrm{at}_\infty = \mathcal{C}\mathrm{at}_t[W^{-1}]$. We will refer to $\mathcal{C}\mathrm{at}_\infty$ as the *∞-category of (small) ∞-categories*. By construction, the homotopy category of $\mathcal{C}\mathrm{at}_\infty$ is equivalent to the category $\mathrm{h}\mathcal{C}\mathrm{at}_\infty$ of Definition 2.9.

3. Higher Algebra

Arguably the most important example of an ∞-category is the ∞-category $\mathcal{S}$ of *spaces* of Example 2.17. A more explicit description of $\mathcal{S}$ can be given as follows:

(a) The objects of $\mathcal{S}$ are CW complexes.

(b) For every pair of CW complexes X and Y, we let $\mathrm{Hom}_{\mathcal{S}}(X, Y)$ denote the space of continuous maps from X to Y (endowed with the compact-open topology).

The role of $\mathcal{S}$ in the theory of ∞-categories is analogous to that of the ordinary category of sets in classical category theory. For example, for any ∞-category $\mathcal{C}$ one can define a *Yoneda embedding* $j : \mathcal{C} \to \mathrm{Fun}(\mathcal{C}^{op}, \mathcal{S})$, given by $j(C)(D) = \mathrm{Hom}_{\mathcal{C}}(D, C) \in \mathcal{S}$.

In this paper, we will be interested in studying the ∞-categorical analogues of more algebraic structures like commutative rings. As a first step, we recall the following notion from stable homotopy theory:

Definition 3.1. A *spectrum* is a sequence of pointed spaces $X_0, X_1, \ldots \in \mathcal{S}_*$ equipped with weak homotopy equivalences $X_n \simeq \Omega X_{n+1}$; here $\Omega : \mathcal{S}_* \to \mathcal{S}_*$ denotes the based loop space functor $X \mapsto \{p : [0,1] \to X | p(0) = p(1) = *\}$.

To any spectrum X, we can associate abelian groups $\pi_k X$ for every integer k, defined by $\pi_k X = \pi_{k+n} X_n$ for $n \gg 0$. We say that X is *connective* if $\pi_n X \simeq 0$ for $n < 0$.

The collection of spectra is itself organized into an ∞-category which we will denote by Sp. If $X = \{X_n, \alpha_n : X_n \simeq \Omega X_{n+1}\}_{n \geq 0}$ is a spectrum, then we will refer to X_0 as the 0*th space* of X. The construction $X \mapsto X_0$ determines a forgetful functor $\mathrm{Sp} \to \mathcal{S}$, which we will denote by Ω^∞.

We will say a spectrum X is *discrete* if the homotopy groups $\pi_i X$ vanish for $i \neq 0$. The construction $X \mapsto \pi_0 X$ determines an equivalence from the ∞-category of discrete spectra to the ordinary category of abelian groups. In other words, we can regard the ∞-category Sp as an *enlargement* of the ordinary category of abelian groups, just as the ∞-category $\mathcal{S}$ is an enlargment of the ordinary category of sets.

The category Ab of abelian groups is an example of a *symmetric monoidal* category: that is, there is a tensor product operation $\otimes : \mathrm{Ab} \times \mathrm{Ab} \to \mathrm{Ab}$ which is commutative and associative up to isomorphism. This operation has a counterpart in the setting of spectra: namely, the ∞-category Sp admits a symmetric monoidal structure $\wedge : \mathrm{Sp} \times \mathrm{Sp} \to \mathrm{Sp}$. This operation is called the *smash product*, and is compatible with the usual tensor product of abelian groups in the following sense: if X and Y are connective spectra, then there is a canonical isomorphism of abelian groups $\pi_0(X \wedge Y) \simeq \pi_0 X \otimes \pi_0 Y$. The unit object for the the smash product $\wedge$ is called the *sphere spectrum* and denoted by S.

The symmetric monoidal structure on the ∞-category Sp allows us to define an ∞-category CAlg(Sp) of *commutative algebra objects* of Sp. An object of CAlg(Sp) is a spectrum R equipped with a multiplication $R \wedge R \to R$ which is unital, associative, and commutative up to coherent homotopy. We will refer to the objects of CAlg(Sp) as E_∞*-rings*, and to CAlg(Sp) as the ∞*-category of* E_∞*-rings*. The sphere spectrum S can be regarded as an E_∞-ring in an essentially unique way, and is an initial object of the ∞-category CAlg(Sp).

For any E_∞-ring R, the product on R determines a multiplication on the direct sum $\pi_* R = \oplus_n \pi_n R$. This multiplication is unital, associative, and commutative in the graded sense (that is, for $x \in \pi_i R$ and $y \in \pi_j R$ we have $xy = (-1)^{ij} yx \in \pi_{i+j}(R)$). In particular, $\pi_0 R$ is a commutative ring and each

$\pi_i R$ has the structure of a module over $\pi_0 R$. The construction $R \mapsto \pi_0 R$ determines an equivalence between the ∞-category of *discrete* E_∞-rings and the ordinary category of commutative rings. Consequently, we can view $\mathrm{CAlg}(\mathrm{Sp})$ as an enlargement of the ordinary category of commutative rings.

Remark 3.2. To every E_∞-ring R, we can associate an ∞-category $\mathrm{Mod}_R(\mathrm{Sp})$ of *R-module spectra*: that is, modules over R in the ∞-category of spectra. If M and N are R-module spectra, we will denote the space $\mathrm{Hom}_{\mathrm{Mod}_R(\mathrm{Sp})}(M, N)$ simply by $\mathrm{Hom}_R(M, N)$. If M is an R-module spectrum, then $\pi_* M$ is a graded module over the ring $\pi_* R$. In particular, each homotopy group $\pi_n M$ has the structure of a $\pi_0 R$-module. If R is a discrete commutative ring, then $\mathrm{Mod}_R(\mathrm{Sp})$ can be identified with the ∞-category $\mathrm{Mod}_R = \mathrm{Chain}_R[W^{-1}]$ of Example 2.18. In particular, the homotopy category $\mathrm{hMod}_R(\mathrm{Sp})$ is equivalent to the classical derived category of R-modules.

We have the following table of analogies:

Classical Notion	∞-Categorical Analogue
Set	topological space
Category	∞-Category
Abelian group	Spectrum
Commutative Ring	E_∞-Ring
Ring of integers $\mathbf{Z}$	Sphere spectrum S

Motivated by these analogies, we introduce the following variant of Definition 1.3:

Definition 3.3. A *derived moduli problem* is a functor X from the ∞-category $\mathrm{CAlg}(\mathrm{Sp})$ of E_∞-rings to the ∞-category $\mathcal{S}$ of spaces.

Remark 3.4. Suppose that $X_0 : \mathrm{Ring} \to \mathrm{Gpd}$ is a classical moduli problem. We will say that a derived moduli problem $X : \mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$ is an *enhancement of F* if, whenever R is a commutative ring (regarded as a discrete E_∞-ring), we have an equivalence of categories $X_0(R) \simeq \pi_{\leq 1} X(R)$, and the homotopy groups $\pi_i X(R)$ vanish for $i \geq 2$ (and any choice of base point).

Example 3.5. Let A be an E_∞-ring. Then R defines a derived moduli problem, given by the formula $X(R) = \mathrm{Hom}_{\mathrm{CAlg}(\mathrm{Sp})}(A, R)$. Assume that A is connective: that is, the homotopy groups $\pi_i A$ vanish for $i < 0$. Then X can be regarded as an enhancement of the classical moduli problem $\mathrm{Spec}(\pi_0 A) : R \mapsto \mathrm{Hom}_{\mathrm{Ring}}(\pi_0 A, R)$.

Example 3.6. Let R be an E_∞-ring and let M be an R-module spectrum. We say that M is *projective of rank* r if the $\pi_0 R$-module $\pi_0 M$ is projective of rank r, and the map $\pi_k R \otimes_{\pi_0 R} \pi_0 M \to \pi_k M$ is an isomorphism for every integer k. Fix an integer $n \geq 0$. For every E_∞-ring R, let $X(R)$ denote the ∞-category space for maps of R-modules $f : M \to R^{n+1}$ such that the cofiber R^{n+1}/M is a projective R-module of rank n; the maps in $X(R)$ are given by homotopy equivalences of R-modules (compatible with the map to R^{n+1}). The $X(R)$ is an ∞-groupoid, so we can regard X as a functor $\mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$. Then X is a derived moduli problem, which is an enhancement of the classical moduli problem represented by the scheme $\mathbb{P}^n = \operatorname{Proj} \mathbf{Z}[x_0, x_1, \ldots, x_n]$ (Example 1.2). We can think of X as providing a generalization of projective space to the setting of E_∞-rings.

Example 3.7. Let X be the functor which associates to every E_∞-ring R the ∞-groupoid of projective R-modules of rank n. Then $X : \mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$ is a derived moduli problem, which can be regarded as an enhancement of the classical moduli problem represented by the Artin stack $\mathrm{BGL}(n)$ (Example 1.6).

Let us now summarize several motivations for the study of derived moduli problems:

(a) Let $\mathcal{X}_0$ be a scheme (or, more generally, an algebraic stack), and let X_0 be the classical moduli problem given by the formula $F_0(R) = \mathrm{Hom}(\operatorname{Spec} R, \mathcal{X}_0)$. Examples 3.6 and 3.7 illustrate the following general phenomenon: we can often give a conceptual description of $X_0(R)$ which continues to make sense in the case where R is an arbitrary E_∞-ring, and thereby obtain a derived moduli problem $X : \mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$ which enhances X_0. In these cases, one can often think of X as itself being represented by a scheme (or algebraic stack) $\mathcal{X}$ in the setting of E_∞-rings (see, for example, [32], [31], or [23]). A good understanding of the derived moduli problem X (or, equivalently, the geometric object $\mathcal{X}$) is often helpful for analyzing $\mathcal{X}_0$.

For example, let Y be a smooth algebraic variety over the complex numbers, and let $\overline{\mathcal{M}}_g(Y)$ denote the Kontsevich moduli stack of curves of genus g equipped with a stable map to Y (see, for example, [12]). Then $\overline{\mathcal{M}}_g(Y)$ represents a functor $X_0 : \mathrm{Ring} \to \mathrm{Gpd}$ which admits a natural enhancement $X : \mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$. This enhancement contains a great deal of useful information about the original moduli stack $\overline{\mathcal{M}}_g(Y)$: for example, it determines the *virtual fundamental class* of $\overline{\mathcal{M}}_g(Y)$ which plays an important role in Gromov-Witten theory.

(b) Let $G_{\mathbf{C}}$ be a reductive algebraic group over the complex numbers. Then $G_{\mathbf{C}}$ is canonically defined over the ring $\mathbf{Z}$ of integers. More precisely, there exists a split reductive group scheme $G_{\mathbf{Z}}$ over $\operatorname{Spec} \mathbf{Z}$ (well-defined up to isomorphism) such that $G_{\mathbf{Z}} \times \operatorname{Spec} \mathbf{C} \simeq G_{\mathbf{C}}$ ([4]). Since $\mathbf{Z}$ is the initial object in the category of commutative rings, the group scheme $G_{\mathbf{Z}}$

can be regarded as a "universal version" of the reductive algebraic group $G_{\mathbf{C}}$: it determines a reductive group scheme $G_R = G_{\mathbf{Z}} \times \operatorname{Spec} R$ over any commutative ring R. However, there are some suggestions that $G_{\mathbf{Z}}$ might admit an even more primordial description (for example, it has been suggested that we should regard the Weyl group W of $G_{\mathbf{C}}$ as the set of points $G(\mathbf{F}_1)$ of G with values in the "field with 1 element"; see [28]). The language of ring spectra provides one way of testing this hypothesis: the initial object in the ∞-category $\mathrm{CAlg}(\mathrm{Sp})$ is given by the sphere spectrum S, rather than the discrete ring $\mathbf{Z} \simeq \pi_0 S$. Therefore it makes sense to ask if the algebraic group $G_{\mathbf{C}}$ can be defined over the sphere spectrum; it was this question which originally motivated the theory described in this paper.

(c) Let $X_0 : \mathrm{Ring} \to \mathrm{Gpd}$ be the classical moduli problem of Example 1.5, which assigns to each commutative ring R the groupoid $\mathrm{Hom}(\operatorname{Spec} R, \mathcal{M}_{1,1})$ of elliptic curves over R. It is possible to make sense of the notion of an elliptic curve over R when R is an arbitrary E_∞-ring, and thereby obtain an enhancement $X : \mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$ of $\mathcal{M}_{\mathrm{Ell}}$. One can use this enhancement to give a moduli-theoretic reformulation of the Goerss-Hopkins-Miller theory of topological modular forms; we refer the reader to [21] for are more detailed discussion.

(d) The framework of derived moduli problems (or, more precisely, their formal analogues: see Definition 4.6) provides a good setting for the study of deformation theory. We will explain this point in more detail in the next section.

4. Formal Moduli Problems

Let $X : \mathrm{CAlg}_{\mathrm{Sp}} \to \mathcal{S}$ be a derived moduli problem. We define a *point* of X to be a pair $x = (k, \eta)$, where k is a field (regarded as a discrete E_∞-ring) and $\eta \in X(k)$. Our goal in this section is to study the local structure of the moduli problem X "near" the point x. More precisely, we will study the restriction of X to E_∞-rings which are closely related to the field k. To make this idea precise, we need to introduce a bit of terminology.

Definition 4.1. Let k be a field. We let CAlg_k denote the ∞-category whose objects are E_∞-rings A equipped with a map $k \to A$, where morphisms are given by commutative triangles

$$\begin{tikzcd} & k \arrow[dl] \arrow[dr] & \\ A \arrow[rr] & & A'. \end{tikzcd}$$

We will refer to the objects of CAlg_k as E_∞*-algebras over* k.

Remark 4.2. We say that a k-algebra A is *discrete* if it is discrete as an E_∞-ring: that is, if the homotopy groups $\pi_i A$ vanish for $i \neq 0$. The discrete k-algebras determine a full subcategory of CAlg_k, which is equivalent to the ordinary category of commutative rings A with a map $k \to A$.

Remark 4.3. Let k be a field. The category Chain_k of chain complexes over k admits a symmetric monoidal structure, given by the usual tensor product of chain complexes. A commutative algebra in the category Chain_k is called a *commutative differential graded algebra* over k. The functor $\mathrm{Chain}_k \to \mathrm{Mod}_k$ is symmetric monoidal, and determines a functor $\phi : \mathrm{CAlg}(\mathrm{Chain}_k) \to \mathrm{CAlg}(\mathrm{Mod}_k) \simeq \mathrm{CAlg}_k$. We say that a morphism $f : A_\bullet \to B_\bullet$ in $\mathrm{CAlg}_k^{\mathrm{dg}}$ is a *quasi-isomorphism* if it induces a quasi-isomorphism between the underlying chain complexes of $A_\bullet$ and $B_\bullet$. The functor ϕ carries every quasi-isomorphism of commutative differential graded algebras to an equivalence in CAlg_k. If k is a field of characteristic zero, then ϕ induces an equivalence $\mathrm{CAlg}(\mathrm{Chain}_k)[W^{-1}] \simeq \mathrm{CAlg}_k$, where W is the collection of quasi-isomorphisms: in other words, we can think of the ∞-category of E_∞-algebras over k as obtained from the ordinary category of commutative differential graded k-algebras by formally inverting the collection of quasi-isomorphisms.

Definition 4.4. Let k be a field and let $V \in \mathrm{Mod}_k$ be a k-module spectrum. We will say that V is *small* if the following conditions are satisfied:

(1) For every integer n, the homotopy group $\pi_n V$ is finite dimensional as a k-vector space.

(2) The homotopy groups $\pi_n V$ vanish for $n < 0$ and $n \gg 0$.

Let A be an E_∞-algebra over k. We will say that A is *small* if it is small as a k-module spectrum, and satisfies the following additional condition:

(3) The commutative ring $\pi_0 A$ has a unique maximal ideal $\mathfrak{p}$, and the map

$$k \to \pi_0 A \to \pi_0 A/\mathfrak{p}$$

is an isomorphism.

We let $\mathrm{Mod}_{\mathrm{sm}}$ denote the full subcategory of Mod_k spanned by the small k-module spectra, and $\mathrm{CAlg}_{\mathrm{sm}}$ denote the full subcategory of CAlg_k spanned by the small E_∞-algebras over k.

Remark 4.5. Let A be a small E_∞-algebra over k. Then there is a unique morphism $\epsilon : A \to k$ in CAlg_k; we will refer to ϵ as the *augmentation* on A.

Let $X : \mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$ be a derived moduli problem, and let $x = (k, \eta)$ be a point of X. We define a functor $X_x : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ as follows: for every small E_∞-algebra A over k, we let $X_x(A)$ denote the fiber of the map $X(A) \to X(k)$ (induced by the augmentation $\epsilon : A \to k$) over the point η. The intuition is

that X_x encodes the local structure of the derived moduli problem X near the point x.

Let us now axiomatize the expected behavior of the functor X_x:

Definition 4.6. Let k be a field. A *formal moduli problem over k* is a functor $X : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ with the following properties:

(1) The space $X(k)$ is contractible.

(2) Suppose that $\phi : A \to B$ and $\phi' : A' \to B$ are maps between small E_∞-algebras over k which induce surjections $\pi_0 A \to \pi_0 B$, $\pi_0 A' \to \pi_0 B$. Then the canonical map

$$X(A \times_B A') \to X(A) \times_{X(B)} X(A')$$

is a homotopy equivalence.

Remark 4.7. Let X be a derived moduli problem and let $x = (k, \eta)$ be a point of X. Then the functor $X_x : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ automatically satisfies condition (1) of Definition 4.6. Condition (2) is not automatic, but holds whenever the functor X is defined in a sufficiently "geometric" way. To see this, let us imagine that there exists some ∞-category of geometric objects $\mathcal{G}$ with the following properties:

(a) To every small k-algebra A we can assign an object $\operatorname{Spec} A \in \mathcal{G}$, which is contravariantly functorial in A.

(b) There exists an object $\mathcal{X} \in \mathcal{G}$ which represents X, in the sense that $X(A) \simeq \mathrm{Hom}_{\mathcal{G}}(\operatorname{Spec} A, \mathcal{X})$ for every small k-algebra A.

To verify that X_x satisfies condition (2) of Definition 4.6, it suffices to show that when $\phi : A \to B$ and $\phi : A' \to B$ are maps of small E_∞ algebras over k which induce surjections $\pi_0 A \to \pi_0 B \leftarrow \pi_0 A'$, then the diagram

$$\begin{tikzcd} \operatorname{Spec} B \arrow[r] \arrow[d] & \operatorname{Spec} A' \arrow[d] \\ \operatorname{Spec} A \arrow[r] & \operatorname{Spec}(A \times_B A') \end{tikzcd}$$

is a pushout square in $\mathcal{G}$. This assumption expresses the idea that $\operatorname{Spec}(A \times_B A')$ should be obtained by "gluing" $\operatorname{Spec} A$ and $\operatorname{Spec} B$ together along the common closed subobject $\operatorname{Spec} B$.

For examples of ∞-categories $\mathcal{G}$ satisfying the above requirements, we refer the reader to the work of Toën and Vezzosi on derived stacks (see, for example, [32]).

Remark 4.8. Let $X : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ be a formal moduli problem. Then X determines a functor $\overline{X} : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}\mathrm{et}$, given by the formula $\overline{X}(A) = \pi_0 X(A)$.

It follows from condition (2) of Definition 4.6 that if we are given maps of small E_∞-algebras $A \to B \leftarrow A'$ which induce surjections $\pi_0 A \to \pi_0 B \leftarrow \pi_0 A'$, then the induced map

$$\overline{X}(A \times_B A') \to \overline{X}(A) \times_{\overline{X}(B)} \overline{X}(A')$$

is a surjection of sets (in fact, this holds under weaker assumptions: see Remark 6.19). There is a substantial literature on set-valued moduli functors of this type; see, for example, [24] and [18].

5. Tangent Complexes

Let $X : \mathrm{Ring} \to \mathcal{S}\mathrm{et}$ be a classical moduli problem. Let k be a field and let $\eta \in X(k)$, so that the pair $x = (k, \eta)$ can be regarded as a point of X. Following Grothendieck, we define the *tangent space* $T_{X,x}$ to be the fiber of the map $X(k[\epsilon]/(\epsilon^2)) \to X(k)$ over the point η. Under very mild assumptions, one can show that this fiber has the structure of a vector space over k: for example, if $\lambda \in k$ is a scalar, then the action of λ on $T_{X,x}$ is induced by the ring homomorphism $k[\epsilon]/(\epsilon^2) \to k[\epsilon]/(\epsilon^2)$ given by $\epsilon \mapsto \lambda\epsilon$.

Now suppose that $X : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ is a formal moduli problem over a field k. Then $X(k[\epsilon]/(\epsilon^2)) \in \mathcal{S}$ is a topological space, which we will denote by $T_X(0)$. As in the classical case, $T_X(0)$ admits a great deal of algebraic structure. To see this, we need to introduce a bit of notation.

Let k be a field and let V be a k-module spectrum. We let $k \oplus V$ denote the direct sum of k and V (as a k-module spectrum). We will regard $k \oplus V$ as an E_∞-algebra over k, with a "square-zero" multiplication on the submodule V. Note that if V is a small k-module, then $k \oplus V$ is a small k-algebra (Definition 4.4). For each integer $n \geq 0$, we let $k[n]$ denote the n-fold shift of k as a k-module spectrum: it is characterized up to equivalence by the requirement

$$\pi_i k[n] \simeq \begin{cases} k & \text{if } i = n \\ 0 & \mathrm{if} i \neq n \end{cases}$$

If X is a formal moduli problem over k, we set $T_X(n) = X(k \oplus k[n])$ (this agrees with our previous definition in the case $n = 0$). For $n > 0$, we have a pullback diagram of E_∞-algebras

$$\begin{tikzcd} k \oplus k[n-1] \arrow[r] \arrow[d] & k \arrow[d] \\ k \arrow[r] & k \oplus k[n] \end{tikzcd}$$

which, using conditions (1) and (2) of Definition 4.6, gives a pullback diagram

$$\begin{tikzcd} T_X(n-1) \arrow[r] \arrow[d] & * \arrow[d] \\ * \arrow[r] & T_X(n) \end{tikzcd}$$

in the ∞-category of spaces. That is, we can identify $T_X(n-1)$ with the loop space of $T_X(n)$, so that the sequence $\{T_X(n)\}_{n \geq 0}$ can be regarded as a spectrum, which we will denote by T_X. We will refer to T_X as the *tangent complex* to the formal moduli problem X.

In fact, we can say more: the spectrum T_X admits the structure of a module over k. Roughly speaking, this module structure comes from the following construction: for each scalar $\lambda \in k$, multiplication by λ induces a map from $k[n]$ to itself, and therefore a map from $T_X(n)$ to itself; these maps are compatible with one another and give an action of k on the spectrum T_X.

Remark 5.1. Here is a more rigorous construction of the k-module structure on the tangent complex T_X. We say that a functor $U : \mathrm{Mod}_{\mathrm{sm}} \to \mathcal{S}$ is *excisive* if it satisfies the following linear version of the conditions of Definition 4.6:

(1) The space $U(0)$ is contractible.

(2) For every pushout diagram

$$\begin{tikzcd} V \arrow[r] \arrow[d] & V' \arrow[d] \\ W \arrow[r] & W' \end{tikzcd}$$

in the ∞-category $\mathrm{Mod}_{\mathrm{sm}}$, the induced diagram of spaces

$$\begin{tikzcd} U(V) \arrow[r] \arrow[d] & U(V') \arrow[d] \\ U(W) \arrow[r] & U(W') \end{tikzcd}$$

is a pullback square.

If $W \in \mathrm{Mod}_k$ is an arbitrary k-module spectrum, then the construction $V \mapsto \mathrm{Hom}_k(V^\vee, W)$ gives an excisive functor from $\mathrm{Mod}_{\mathrm{sm}}$ to $\mathcal{S}$ (here $V^\vee$ denotes the k-linear dual of V). In fact, every excisive functor arises in this way: the above construction determines a fully faithful embedding $\mathrm{Mod}_k \hookrightarrow \mathrm{Fun}(\mathrm{Mod}_{\mathrm{sm}}, \mathcal{S})$ whose essential image is the collection of excisive functors.

If $X : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ is a formal moduli problem, then one can show that the functor $V \mapsto X(k \oplus V)$ is excisive. It follows that there exists a k-module

spectrum W (which is determined uniquely up to equivalence) for which $X(k \oplus V) \simeq \mathrm{Hom}_k(V^\vee, W)$. This k-module spectrum W can be identified with the tangent complex T_X; for example, we have

$$\Omega^\infty T_X = T_X(0) = X(k \oplus k[0]) \simeq \mathrm{Hom}_k(k[0]^\vee, W) \simeq \Omega^\infty W$$

Remark 5.2. The tangent complex to a formal moduli problem X carries a great deal of information about X. For example, if $\alpha : X \to X'$ is a natural transformation between formal moduli problems, then α is an equivalence if and only if it induces a homotopy equivalence of k-module spectra $T_X \to T_{X'}$. In concrete terms, this means that if α induces a homotopy equivalence $X(k \oplus k[n]) \to X'(k \oplus k[n])$ every integer $n \geq 0$, then α induces a homotopy equivalence $F(A) \to F'(A)$ for every small E_∞-algebra A over k. This follows from the fact that A admits a "composition series"

$$A = A(m) \to A(m-1) \to \cdots \to A(0) = k,$$

where each of the maps $A(j) \to A(j-1)$ fits into a pullback diagram

$$\begin{tikzcd} A(j) \arrow[r] \arrow[d] & A(j-1) \arrow[d] \\ k \arrow[r] & k \oplus k[n_j] \end{tikzcd}$$

for some $n_j > 0$.

Remark 5.2 suggests that it should be possible to reconstruct a formal moduli problem X from its tangent complex T_X. If k is a field of characteristic zero, then mathematical folklore asserts that every formal moduli problem is "controlled" by a differential graded Lie algebra over k. This can be formulated more precisely as follows:

Theorem 5.3. *Let k be a field of characteristic zero, and let* Moduli *denote the full subcategory of* $\mathrm{Fun}(\mathrm{CAlg}_{\mathrm{sm}}, \mathcal{S})$ *spanned by the formal moduli problems over k. Then there is an equivalence of ∞-categories $\Phi : \mathrm{Moduli} \to \mathrm{Lie}_k^{\mathrm{dg}}$, where $\mathrm{Lie}_k^{\mathrm{dg}}$ denotes the ∞-category of differential graded Lie algebras over k (Example 2.19). Moreover, if $U : \mathrm{Lie}_k^{\mathrm{dg}} \to \mathrm{Mod}_k$ denotes the forgetful functor (which assigns to each differential graded Lie algebra its underlying chain complex), then the composition $U \circ \Phi$ can be identified with the functor $X \mapsto T_X[-1]$.*

In other words, if X is a formal moduli problem, then the shifted tangent complex $T_X[-1] \in \mathrm{Mod}_k$ can be realized as a differential graded Lie algebra over k. Conversely, every differential graded Lie algebra over k arises in this way (up to quasi-isomorphism).

Remark 5.4. The functor $\Phi^{-1} : \mathrm{Lie}_k^{\mathrm{dg}} \to \mathrm{Moduli} \subseteq \mathrm{Fun}(\mathrm{CAlg}_{\mathrm{sm}}, \mathcal{S})$ is constructed by Hinich in [14]. Roughly speaking, if $\mathfrak{g}$ is a differential graded Lie

algebra and A is a small E_∞-algebra over k, then $\Phi^{-1}(\mathfrak{g})(A)$ is the space of solutions to the Maurer-Cartan equation $dx = [x, x]$ in the differential graded Lie algebra $\mathfrak{g} \otimes_k \mathfrak{m}_A$.

Remark 5.5. The notion that differential graded Lie algebras should play an important role in the description of moduli spaces goes back to Quillen's work on rational homotopy theory ([33]), and was developed further in unpublished work of Deligne, Drinfeld, and Feigin. Many mathematicians have subsequently taken up these ideas: see, for example, the book of Kontsevich and Soibelman ([18]).

Remark 5.6. For applications of Theorem 5.3 to the classification of deformations of algebraic structures, we refer the reader to [15] and [17].

Remark 5.7. Suppose that R is a commutative k-algebra equipped with an augmentation $\epsilon : R \to k$. Then R defines a formal moduli problem X over k, which carries a small E_∞-algebra A over k to the fiber of the map

$$\mathrm{Hom}_{\mathrm{CAlg}_k}(R, A) \to \mathrm{Hom}_{\mathrm{CAlg}_k}(R, k).$$

When k is of characteristic zero, the tangent complex T_X can be identified with the complex Andre-Quillen cochains taking values in k. In this case, the existence of a natural differential graded Lie algebra structure on $T_X[-1]$ is proven in [26].

Remark 5.8. Here is a heuristic explanation of Theorem 5.3. Let $X : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ be a formal moduli problem. Since every k-algebra A comes equipped with a canonical map $k \to A$, we get an induced map $* \simeq X(k) \to X(A)$: in other words, each of the spaces $X(A)$ comes equipped with a natural base point. We can then define a new functor $\Omega X : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ by the formula $(\Omega X)(A) = \Omega X(A)$ (here Ω denotes the loop space functor from the ∞-category of pointed spaces to itself). Then ΩX is another formal moduli problem, and an elementary calculation gives $T_{\Omega X} \simeq T_X[-1]$. However, ΩX is equipped with additional structure: composition of loops gives a multiplication on ΩX (which is associative up to coherent homotopy), so we can think of ΩX as a *group object* in the ∞-category of formal moduli problems.

In classical algebraic geometry, the tangent space to an algebraic group G at the origin admits a Lie algebra structure. In characteristic zero, this Lie algebra structure permits us to reconstruct the formal completion of G (via the Campbell-Hausdorff formula). Theorem 5.3 can be regarded as an analogous statement in the context of formal moduli problems: the group structure on ΩX determines a Lie algebra structure on its tangent complex $T_{\Omega X} \simeq T_X[-1]$. Since we are working in a formal neighborhood of a fixed point, allows us to reconstruct the group ΩX (and, with a bit more effort, the original formal moduli problem X).

Example 5.9. Let $X : \mathrm{CAlg}(\mathrm{Sp}) \to \mathcal{S}$ be the formal moduli problem of Example 3.7, which assigns to every E_∞-ring A the ∞-groupoid $F(A)$ of projective A-modules of rank n. Giving a point $x = (k, \eta)$ of X is equivalent to giving a field k together with a vector space V_0 of dimension n over k. In this case, the functor $X_x : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ can be described as follows: to every small k-algebra A, the functor X_x assigns the ∞-category of pairs (V, α), where V is a projective A-module of rank n and $\alpha : k \wedge_A V \to V_0$ is an isomorphism of k-vector spaces. It is not difficult to show that X_x is a formal moduli problem in the sense of Definition 4.6. We will denote its tangent complex $T_{X,x}$.

Unwinding the definitions, we see that $T_{X,x}(0) = X_x(k[\epsilon]/(\epsilon^2))$ can be identified with a classifying space for the groupoid of projective $k[\epsilon]/(\epsilon^2)$-modules V which deform V_0. This groupoid has only one object up to isomorphism, given by the tensor product $k[\epsilon]/(\epsilon^2) \otimes_k V_0$. It follows that $T_{X,x}(0)$ can be identified with the classifying space BG for the group G of automorphisms of $k[\epsilon]/(\epsilon^2) \otimes_k V_0$ which reduce to the identity moduli ϵ. Such an automorphism can be written as $1 + \epsilon M$, where $M \in \mathrm{End}(V_0)$. Consequently, $T_{X,x}(0)$ is homotopy equivalent to the classifying space for the k-vector space $\mathrm{End}_k(V_0)$, regarded as a group under addition.

Amplifying this argument, we obtain an equivalence of k-module spectra $T_{X,x} \simeq \mathrm{End}_k(V_0)[1]$. The shifted tangent complex $T_{X,x}[-1] \simeq \mathrm{End}_k(V_0)$ has the structure of a Lie algebra over k (and therefore of a differential graded Lie algebra over k, with trivial grading and differential), given by the usual commutator bracket of endomorphisms.

6. Noncommutative Geometry

Our goal in this paper is to describe an analogue of Theorem 5.3 in the setting of noncommutative geometry. We begin by describing a noncommutative analogue of the theory of E_∞-rings.

Definition 6.1. Let $\mathcal{C}$ be a symmetric monoidal ∞-category. We can associate to $\mathcal{C}$ a new ∞-category $\mathrm{Alg}(\mathcal{C})$ of *associative algebra* objects of $\mathcal{C}$. The ∞-category $\mathrm{Alg}(\mathcal{C})$ inherits the structure of a symmetric monoidal ∞-category. We can therefore define a sequence of ∞-categories $\mathrm{Alg}^{(n)}(\mathcal{C})$ by induction on n:

(a) If $n = 1$, we let $\mathrm{Alg}^{(n)}(\mathcal{C}) = \mathrm{Alg}(\mathcal{C})$.

(b) If $n > 1$, we let $\mathrm{Alg}^{(n)}(\mathcal{C}) = \mathrm{Alg}(\mathrm{Alg}^{(n-1)}(\mathcal{C}))$.

We will refer to $\mathrm{Alg}^{(n)}(\mathcal{C})$ as the *∞-category of E_n-algebras in $\mathcal{C}$*.

Remark 6.2. We can summarize Definition 6.1 informally as follows: an E_n-algebra object of a symmetric monoidal ∞-category $\mathcal{C}$ is an object $A \in \mathcal{C}$ which is equipped with n multiplication operations $\{m_i : A \otimes A \to A\}_{1 \leq i \leq n}$;

these multiplications are required to be associative and unital (up to coherent homotopy) and to be compatible with one another in a suitable sense.

Example 6.3. Let $\mathcal{C} = \mathcal{S}$ be the ∞-category of spaces, endowed with the symmetric monoidal structure given by Cartesian products of spaces. For every pointed space X, the loop space ΩX has the structure of an algebra object of $\mathcal{S}$: the multiplication on ΩX is given by concatenation of loops. In fact, we can say a bit more: the algebra object $\Omega X \in \mathrm{Alg}(\mathcal{S})$ is *grouplike*, in the sense that the multiplication on $\Omega(X)$ determines a group structure on the set $\pi_0 \Omega(X) \simeq \pi_1 X$. This construction determines an equivalence from the ∞-category of *connected* pointed spaces to the full subcategory of $\mathrm{Alg}(\mathcal{C})$ spanned by the *grouplike* associative algebras.

More generally, the construction $X \mapsto \Omega^n X$ establishes an equivalence between the ∞-category of $(n-1)$-connected pointed spaces and the full subcategory of grouplike E_n-algebras of $\mathcal{S}$. See [25] for further details.

Example 6.4. Fix an E_∞-ring k, and let $\mathrm{Mod}_k = \mathrm{Mod}_k(\mathrm{Sp})$ denote the ∞-category of k-module spectra. Then Mod_k admits a symmetric monoidal structure, given by the relative smash product $(M, N) \mapsto M \wedge_k N$. We will refer to E_n-algebra objects of Mod_k as *E_n-algebras over k*. We let $\mathrm{Alg}_k^{(n)} = \mathrm{Alg}^{(n)}(\mathrm{Mod}_k)$ denote the ∞-category of E_n-algebras over k. When k is the sphere spectrum S, we will refer to an E_n-algebra over k simply as an *E_n-ring*.

Remark 6.5. For any symmetric monoidal ∞-category $\mathcal{C}$, there is a forgetful functor $\mathrm{Alg}(\mathcal{C}) \to \mathcal{C}$, which assigns to an associative algebra its underlying object of $\mathcal{C}$. These forgetful functors determine rise to a tower of ∞-categories

$$\cdots \to \mathrm{Alg}^{(3)}(\mathcal{C}) \to \mathrm{Alg}^{(2)}(\mathcal{C}) \to \mathrm{Alg}^{(1)}(\mathcal{C}).$$

The inverse limit of this tower can be identified with the ∞-category $\mathrm{CAlg}(\mathcal{C})$ of commutative algebra objects of $\mathcal{C}$.

Remark 6.6. There is a non-inductive description of the ∞-category $\mathrm{Alg}^{(n)}(\mathcal{C})$ of E_n-algebra objects in $\mathcal{C}$: it can be obtained as the ∞-category of representations in $\mathcal{C}$ of the *little n-cubes* operad introduced by Boardman and Vogt; see [2].

Remark 6.7. It is convenient to extend Definition 6.1 to the case $n = 0$: an E_0-algebra object of $\mathcal{C}$ is an object $A \in \mathcal{C}$ which is equipped with a distinguished map $\mathbf{1} \to A$, where $\mathbf{1}$ denotes the unit with respect to the tensor product on $\mathcal{C}$.

Remark 6.8. When $\mathcal{C}$ is an ordinary category, Definition 6.1 is somewhat degenerate: the categories $\mathrm{Alg}^{(n)}(\mathcal{C})$ coincide with $\mathrm{CAlg}(\mathcal{C})$ for $n \geq 2$. This is a consequence of the classical Eckmann-Hilton argument: if $A \in \mathcal{C}$ is equipped with two commuting unital multiplication operations m_1 and m_2, then m_1 and m_2 are commutative and coincide with one another. If $\mathcal{C}$ is the category of sets, the proof can be given as follows. Since the unit map $\mathbf{1} \to A$ for the

multipication m_1 is a homomorphism with multiplication m_2, we see that the unit elements of A for the multiplications m_1 and m_2 coincide with a single element $u \in A$. Then

$$m_1(a,b) = m_1(m_2(a,u), m_2(u,b)) = m_2(m_1(a,u), m_1(u,b)) = m_2(a,b).$$

A similar calculation gives $m_1(a,b) = m_2(b,a)$, so that $m_1 = m_2$ is commutative.

Remark 6.9. Let k be a field, and let Chain_k be the ordinary category of chain complexes over k. The functor $\mathrm{Chain}_k \to \mathrm{Mod}_k$ of Remark 3.2 is symmetric monoidal: in other words, the relative smash product $\wedge_k$ is compatible with the usual tensor product of chain complexes. In particular, we get a functor $\theta : \mathrm{Alg}(\mathrm{Chain}_k) \to \mathrm{Alg}(\mathrm{Mod}_k) = \mathrm{Alg}_k^{(1)}$. The category $\mathrm{Alg}(\mathrm{Chain}_k)$ can be identified with the category of *differential graded algebras over* k. We say that a map of differential graded algebras $f : A_\bullet \to B_\bullet$ is a quasi-isomorphism if it induces a quasi-isomorphism between the underlying chain complexes of $A_\bullet$ and $B_\bullet$; in this case, the morphism $\theta(f)$ is an equivalence in $\mathrm{Alg}_k^{(1)}$. Let W be the collection of quasi-isomorphisms between differential graded algebras. One can show that θ induces an equivalence $\mathrm{Alg}(\mathrm{Chain}_k)[W^{-1}] \to \mathrm{Alg}_k^{(1)}$: that is, E_1-algebras over a field k (of any characteristic) can be identified with differential graded algebras over k.

Remark 6.10. Let k be a field and let A be an E_n-algebra over k. If $n \geq 1$, then A has an underlying associative multiplication. This multiplication endows $\pi_* A$ with the structure of a graded algebra over k. In particular, $\pi_0 A$ is an associative k-algebra.

Definition 6.11. Let k be a field and let A be an E_n-algebra over k, where $n \geq 1$. We will say that A is *small* if the following conditions are satisfied:

(1) The algebra A is small when regarded as a k-module spectrum: that is, the homotopy groups $\pi_i A$ are finite dimensional, and vanish if $i < 0$ or $i \gg 0$.

(2) Let $\mathfrak{p}$ be the radical of the (finite-dimensional) associative k-algebra $\pi_0 A$. Then the composite map $k \to \pi_0 A \to \pi_0 A/\mathfrak{p}$ is an isomorphism.

We let $\mathrm{Alg}_{\mathrm{sm}}^{(n)}$ denote the full subcategory of $\mathrm{Alg}_k^{(n)}$ spanned by the small E_n-algebras over k.

Remark 6.12. Let A be an E_n-algebra over a field k. An *augmentation* on A is a map of E_n-algebras $A \to k$. The collection of augmented E_n-algebras over k can be organized into an ∞-category, which we will denote by $\mathrm{Alg}_{\mathrm{aug}}^{(n)}$. If $n \geq 1$ and A is a small E_n-algebra over k, then A admits a unique augmentation $A \to k$ (up to a contractible space of choices). Consequently, we can view $\mathrm{Alg}_{\mathrm{sm}}^{(n)}$ as a full subcategory of $\mathrm{Alg}_{\mathrm{aug}}^{(n)}$.

If $\eta : A \to k$ is an augmented E_n-algebra over k, we let $\mathfrak{m}_A$ denote the fiber of the map η. We will refer to $\mathfrak{m}_A$ as the *augmentation ideal* of A.

Remark 6.13. If $n = 0$, then an augmentation on an E_0-algebra $A \in \mathrm{Alg}_k^{(0)}$ is a map of k-module spectra $\eta : A \to k$ which is left inverse to the unit map $k \to A$. The construction $(\eta : A \to k) \mapsto \mathfrak{m}_A$ determines an equivalence of ∞-categories $\mathrm{Alg}_{\mathrm{aug}}^{(0)} \simeq \mathrm{Mod}_k$.

It is convenient to extend Definition 6.11 to the case $n = 0$. We say that an augmented E_0-algebra A is *small* if A (or, equivalently, the augmentation ideal $\mathfrak{m}_A$) is small when regarded as a k-module spectrum. We let $\mathrm{Alg}_{\mathrm{sm}}^{(0)} \subseteq \mathrm{Alg}_{\mathrm{aug}}^{(0)} \simeq \mathrm{Mod}_k$ denote the full subcategory spanned by the small E_0-algebras over k.

The following elementary observation will be used several times in this paper:

Claim 6.14. *Let $f : A \to B$ be a map of small E_n-algebras over k which induces a surjection $\pi_0 A \to \pi_0 B$. Then there exists a sequence of maps*

$$A = A(0) \to A(1) \to \cdots \to A(m) = B$$

with the following property: for each integer $0 \leq i < m$, there is a pullback diagram of small E_n-algebras

$$\begin{array}{ccc} A(i) & \longrightarrow & A(i+1) \\ \downarrow & & \downarrow \\ k & \longrightarrow & k \oplus k[j] \end{array}$$

for some $j > 0$ (in other words, $A(i)$ can be identified with the fiber of some map $A(i+1) \to k \oplus k[j]$).

Remark 6.15. Claim 6.14 is most useful in the case where f is the augmentation map $A \to k$. We will refer to a sequence of maps

$$A = A(0) \to A(1) \to \cdots \to A(m) \simeq k$$

satisfying the requirements of Claim 6.14 as a *composition series* for A.

Definition 6.16. Let k be a field and let $n \geq 0$ be an integer. A *formal E_n moduli problem* over k is a functor $X : \mathrm{Alg}_{\mathrm{sm}}^{(n)} \to \mathcal{S}$ with the following properties:

(1) The space $X(k)$ is contractible.

(2) Suppose we are given a pullback diagram of small E_n-algebras

$$\begin{array}{ccc} A' & \longrightarrow & A \\ \downarrow & & \downarrow \\ B' & \longrightarrow & B \end{array}$$

such that the maps $\pi_0 A \to \pi_0 B$ and $\pi_0 B' \to \pi_0 B$ are surjective. Then the diagram

$$\begin{array}{ccc} X(A') & \longrightarrow & X(A) \\ \downarrow & & \downarrow \\ X(B') & \longrightarrow & X(B) \end{array}$$

is a pullback diagram in $\mathcal{S}$.

Remark 6.17. Every formal E_n moduli problem $X : \mathrm{Alg}_{\mathrm{sm}}^{(n)} \to \mathcal{S}$ determines a formal moduli problem X' in the sense of Definition 4.6, where X' is given by the composition

$$\mathrm{Alg}_{\mathrm{sm}} \to \mathrm{Alg}_{\mathrm{sm}}^{(n)} \xrightarrow{X} \mathcal{S}\,.$$

We define the *tangent complex of* X to be the tangent complex of X', as defined in §5. We will denote the tangent complex of X by $T_X \in \mathrm{Mod}_k$.

Remark 6.18. Let X be as in Definition 6.16. By virtue of Claim 6.14, it suffices to check condition (2) in the special case where $A = k$ and $B = k \oplus k[j]$, for some $j > 0$. In other words, condition (2) is equivalent to the requirement that for every map $B' \to k \oplus k[j]$, we have a fiber sequence

$$X(B \times_{k \oplus k[j]} k) \to X(B) \to X(k \oplus k[j]).$$

The final term in this sequence can be identified with $T_X(j) = \Omega^\infty(T_X[j])$.

Remark 6.19. The argument of Remark 6.18 shows that condition (2) of Definition 6.16 is equivalent to the following apparently stronger condition:

(2′) Suppose we are given a pullback diagram of small E_n-algebras

$$\begin{array}{ccc} A' & \longrightarrow & A \\ \downarrow & & \downarrow \\ B' & \longrightarrow & B \end{array}$$

such that the maps $\pi_0 A \to \pi_0 B$ is surjective. Then the diagram

$$\begin{array}{ccc} X(A') & \longrightarrow & X(A) \\ \downarrow & & \downarrow \\ X(B') & \longrightarrow & X(B) \end{array}$$

is a pullback diagram in $\mathcal{S}$.

Let V_0 be a finite dimensional vector space over k, and let $X_x : \mathrm{CAlg}_{\mathrm{sm}} \to \mathcal{S}$ be the formal moduli problem of Example 5.9, so that X_x assigns to every small

E_∞-algebra A over k the ∞-groupoid of pairs (V, α), where V is an A-module and $\alpha : k \wedge_A V \simeq V_0$ is an equivalence. The definition of X_x does not make any use of the commutativity of A. Consequently, X_x extends naturally to a functor $\widehat{X}_x : \mathrm{Alg}^{(1),}_{\mathrm{sm}} \to \mathcal{S}$, By definition, the shifted tangent complex of $\widehat{X}_x[-1]$ is given by the Lie algebra $T_{X_x}[-1] \simeq \mathrm{End}(V_0)$. If k is of characteristic zero, then Theorem 5.3 implies that the formal moduli problem X_x can be canonically reconstructed from the vector space $\mathrm{End}(V_0)$ together with its Lie algebra structure. However, the formal E_1 moduli problem $\widehat{X}_x$ is additional data, since we can evaluate $\widehat{X}_x$ on algebras which are not necessarily commutative. Consequently, it is natural to expect the existence of $\widehat{X}_x$ to be reflected in some additional structure on the Lie algebra $\mathrm{End}(V_0)$. We observe that $\mathrm{End}(V_0)$ is not merely a Lie algebra: there is an associative product (given by composition) whose commutator gives the Lie bracket on $\mathrm{End}(V_0)$. In fact, this is a general phenomenon:

Theorem 6.20. *Let k be a field, let $n \geq 0$, and let Moduli_n be the full subcategory of $\mathrm{Fun}(\mathrm{Alg}^{(n)}_{\mathrm{sm}}, \mathcal{S})$ spanned by the formal E_n moduli problems. Then there exists an equivalence of ∞-categories $\Phi : \mathrm{Moduli}_n \to \mathrm{Alg}^{(n)}_{\mathrm{aug}}$. Moreover, if $U : \mathrm{Alg}^{(n)}_{\mathrm{aug}} \to \mathrm{Mod}_k$ denotes the forgetful functor $A \mapsto \mathfrak{m}_A$ which assigns to each augmented E_n-algebra its augmentation ideal, then the composition $U \circ \Phi$ can be identified with the functor $X \mapsto T_X[-n]$.*

In other words, if X is a formal E_n-module problem, then the shifted tangent complex $T_X[-n]$ can be identified with the augmentation ideal in an augmented E_n-algebra A: that is, $T_X[-n]$ admits a nonunital E_n-algebra structure. Moreover, this structure determines the formal E_n moduli problem up to equivalence.

Example 6.21. Suppose that $n = 0$. The construction $V \mapsto k \oplus V$ determines an equivalence $\mathrm{Mod}_{\mathrm{sm}} \simeq \mathrm{Alg}^{(0)}_{\mathrm{sm}}$. Under this equivalence, we can identify the ∞-category Moduli_0 of formal E_0 moduli problems with the full subcategory of $\mathrm{Fun}(\mathrm{Mod}_{\mathrm{sm}}, \mathcal{S})$ spanned by the excisive functors (see Remark 5.1). In this case, Theorem 6.20 reduces to the claim of Remark 5.1: every excisive functor $U : \mathrm{Mod}_{\mathrm{sm}} \to \mathcal{S}$ has the form $V \mapsto \mathrm{Hom}_k(V^\vee, W) \simeq \Omega^\infty(V \wedge_k W)$ for some object $W \in \mathrm{Mod}_k$, which is determined up to equivalence. Note that we can identify W with the tangent complex to the formal E_0 moduli problem $A \mapsto U(\mathfrak{m}_A)$.

Remark 6.22. Unlike Theorem 5.3, Theorem 6.20 does not require any assumption on the characteristic of the ground field k.

Remark 6.23. Theorem 6.20 is a consequence of the Koszul self-duality of the little cubes operad E_n (see [11]). More precisely, for every field k one can define a Koszul duality functor $\mathbb{D} : \mathrm{Alg}^{(n)}_{\mathrm{aug}} \to (\mathrm{Alg}^{(n)}_{\mathrm{aug}})^{op}$. The construction $\Phi^{-1} : \mathrm{Alg}^{(n)}_{\mathrm{aug}} \to \mathrm{Moduli}_n$ is then given by the formula

$$\Phi^{-1}(A)(B) = \mathrm{Hom}_{\mathrm{Alg}^{(n)}_{\mathrm{aug}}}(\mathbb{D}B, A).$$

References

[1] Bergner, J. *Three models for the homotopy theory of homotopy theories.* Topology 46 (2007), no. 4, 397–436.

[2] Boardman, J. and R. Vogt. *Homotopy Invariant Algebraic Structures on Topological Spaces.* Lecture Notes in Mathematics, 347, Springer-Verlag (1973).

[3] Crane, L. and D.N. Yetter. *Deformations of (bi)tensor categories.* Cahiers Topologie Geom. Differentielle Categ. 39 (1998), no. 3, 163–180.

[4] Demazure, M. and Grothendieck, A., eds. *Schémas en groupes. III: Structure des schémas en groupes réductifs.* Séminaire de Géométrie Algbrique du Bois Marie 1962/64 (SGA 3). Lecture Notes in Mathematics, Vol. 153 Springer-Verlag, Berlin-New York 1962/1964 viii+529 pp.

[5] Efimov, A., Lunts, V., and D. Orlov. *Deformation theory of objects in homotopy and derived categories. I: General Theory.* Adv. Math. 222 (2009), no. 2, 359–401.

[6] Efimov, A., Lunts, V., and D. Orlov. *Deformation theory of objects in homotopy and derived categories. II: Pro-representability of the deformation functor.* Available at arXiv:math/0702839v3 .

[7] Efimov, A., Lunts, V., and D. Orlov. *Deformation theory of objects in homotopy and derived categories. III: Abelian categories.* Available as arXiv:math/0702840v3 .

[8] Etingof, P., Nikshych, D., and V. Ostrik. *On fusion categories.* Ann. of Math. (2) 162 (2005), no. 2, 581–642.

[9] Getzler, E. *Lie theory for L_∞-algebras.* Ann. of Math. (2) 170 (2009), no. 1, 271–301.

[10] Frenkel, E., Gaitsgory, D., and K. Vilonen. *Whittaker patterns in the geometry of moduli spaces of bundles on curves.* Ann. of Math. (2) 153 (2001), no. 3, 699–748.

[11] Fresse, B. *Koszul duality of E_n-operads.* Available as arXiv:0904.3123v6 .

[12] Fulton, W. and R. Pandharipande. *Notes on stable maps and quantum cohomology.* Algebraic geometry—Santa Cruz 1995, 45–96, Proc. Sympos. Pure Math., 62, Part 2, Amer. Math. Soc., Providence, RI, 1997.

[13] Gaitsgory, D. *Twisted Whittaker model and factorizable sheaves.* Selecta Math. (N.S.) 13 (2008), no. 4, 617–659.

[14] Hinich, V. *DG coalgebras as formal stacks.* J. Pure Appl. Algebra, 162 (2001), 209-250.

[15] Hinich, V. *Deformations of homotopy algebras.* Communication in Algebra, 32 (2004), 473-494.

[16] Keller, B. and W. Lowen. *On Hochschild cohomology and Morita deformations.* Int. Math. Res. Not. IMRN 2009, no. 17, 3221–3235.

[17] Kontsevich, M. and Y. Soibelman. *Deformations of algebras over operads and the Deligne conjecture.* Conference Moshe Flato 1999, Vol. I (Dijon), 255–307, Math. Phys. Stud., 21, Kluwer Acad. Publ., Dordrecht, 2000.

[18] Kontsevich, M. and Y. Soibelman. *Deformation Theory.* Unpublished book available at http://www.math.ksu.edu/s̃oibel/Book-vol1.ps .

[19] Lowen, W. *Obstruction theory for objects in abelian and derived categories.* Comm. Algebra 33 (2005), no. 9, 3195–3223.

[20] Lowen, W. *Hochschild cohomology, the characteristic morphism, and derived deformations.* Compos. Math. 144 (2008), no. 6, 1557–1580.

[21] Lurie, J. *A Survey of Elliptic Cohomology.* Algebraic Topology: The Abel Symposium 2007. Springer, Berlin, 2009, 219–277.

[22] Lurie, J. *Higher Topos Theory.* Annals of Mathematics Studies, 170. Princeton University Press, Princeton, NJ, 2009. xviii+925 pp.

[23] Lurie, J. *Derived Algebraic Geometry I - VI.* Available for download at www.math.harvard.edu/~lurie/.

[24] Manetti, M. *Extended deformation functors.* Int. Math. Res. Not. 2002, no. 14, 719–756.

[25] May, P. *The Geometry of Iterated Loop Spaces.* Lectures Notes in Mathematics, Vol. 271. Springer-Verlag, Berlin-New York, 1972. viii+175 pp.

[26] Schlessinger, M. and J. Stasheff. *The Lie algebra structure of tangent cohomology and deformation theory.* Journal of Pure and Applied Algebra 38 (1985), 313-322.

[27] Schlessinger, M. *Functors of Artin Rings.* Trans. Amer. Math. Soc. 130 1968 208–222.

[28] Tits, J. *Sur les analogues algbriques des groupes semi-simples complexes.* Colloque dalgebre superieure, tenu a Bruxelles du 19 au 22 decembre 1956, Centre Belge de Recherches Mathematiques Etablissements Ceuterick, Louvain, Paris: Librairie Gauthier-Villars, pp. 261289.

[29] Toën, B. *The homotopy theory of dg-categories and derived Morita theory.* Invent. Math. 167 (2007), no. 3, 615–667.

[30] Toën, B., and M. Vaquié. *Moduli of objects in dg-categories.* Ann. Sci. cole Norm. Sup. (4) 40 (2007), no. 3, 387–444.

[31] Toën, B., and G. Vezzosi. *From HAG to DAG: derived moduli stacks.* Axiomatic, enriched and motivic homotopy theory, 173–216, NATO Sci. Ser. II Math. Phys. Chem., 131, Kluwer Acad. Publ., Dordrecht, 2004.

[32] Toën, B. and G. Vezzosi. *Brave new algebraic geometry and global derived moduli spaces of ring spectra.* Elliptic cohomology, 325–359, London Math. Soc. Lecture Note Ser., 342, Cambridge Univ. Press, Cambridge, 2007.

[33] Quillen, D. *Rational homotopy theory.* Ann. of Math. (2) 90 1969 205–295.

[34] Yetter, D. *Braided deformations of monoidal categories and Vassiliev invariants.* Higher category theory (Evanston, IL, 1997), 117–134, Contemp. Math., 230, Amer. Math. Soc., Providence, RI, 1998.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

On Weil-Petersson Volumes and Geometry of Random Hyperbolic Surfaces

Maryam Mirzakhani*

Abstract

This paper investigates the geometric properties of random hyperbolic surfaces with respect to the Weil-Petersson measure. We describe the relationship between the behavior of lengths of simple closed geodesics on a hyperbolic surface and properties of the moduli space of such surfaces. First, we study the asymptotic behavior of Weil-Petersson volumes of the moduli spaces of hyperbolic surfaces of genus g as $g \to \infty$. Then we apply these asymptotic estimates to study the geometric properties of random hyperbolic surfaces, such as the length of the shortest simple closed geodesic of a given combinatorial type.

Mathematics Subject Classification (2010). Primary 32G15; Secondary 57M50

Keywords. Moduli space, Weil-Petersson volume form, simple closed geodesic, hyperbolic surface

1. Introduction

The space of hyperbolic surfaces of a given genus is equipped with a natural notion of measure, which is induced by the *Weil-Petersson* symplectic form. We are interested in geometric properties of a random hyperbolic surface with respect to this measure. In particular, we are interested in the behavior of the length of the shortest separating/non-separating simple closed geodesic on a random surface of genus g as $g \to \infty$.

*The author has been supported by a Clay Fellowship (2004-08) and an NSF Research Grant.

Stanford University, Dept. of Mathematics, Building 380, Stanford, CA 94305, USA.
E-mail: mmirzakh@math.stanford.edu.

Notation. Let $\mathcal{M}_{g,n}$ be the moduli space of complete hyperbolic surfaces of genus g with n punctures. The universal cover of $\mathcal{M}_{g,n}$ is the Teichmüller space $\mathcal{T}_{g,n}$; every $X \in \mathcal{T}_{g,n}$ represents a *marked* hyperbolic structure on a surface of genus g with n punctures. The space $\mathcal{M}_{g,n}$ is a connected orbifold of dimension $6g-6+2n$, while $\mathcal{T}_{g,n}$ is homeomorphic to $\mathbb{R}^{3g-3+n} \times \mathbb{R}_+^{3g-3+n}$.

Every isotopy class of a closed curve on a hyperbolic surface contains a unique closed geodesic. Given a homotopy class of a closed curve α on a topological surface $S_{g,n}$ of genus g with n marked points and $X \in \mathcal{T}_{g,n}$, let $\ell_\alpha(X)$ be the length of the unique geodesic in the homotopy class of α on X. This defines a length function ℓ_α on the Teichmüller space $\mathcal{T}_{g,n}$.

When studying the behavior of these length functions, it proves fruitful to consider more generally bordered hyperbolic surfaces with geodesic boundary components. Given $L = (L_1, \ldots, L_n) \in \mathbb{R}_+^n$, we consider the Teichmüller space $\mathcal{T}_{g,n}(L)$ of hyperbolic structures with geodesic boundary components of length $L_1, \ldots, L_n$. Note that a geodesic of length zero is the same as a puncture. The space $\mathcal{T}_{g,n}(L)$ is naturally equipped with a symplectic structure; this symplectic form $\omega = \omega_{wp}$ is called the Weil-Petersson symplectic form. When $L = 0$, this form is the symplectic form of a Kähler noncomplete metric on the moduli space $\mathcal{M}_{g,n}$ introduced by Weil [IT]. Wolpert showed that the Weil-Petersson symplectic form has a simple expression in terms of the Fenchel-Nielsen twist-length coordinates on the Teichmüller space (§2). As a result, there is a close relationship between the Weil-Petersson geometry and the lengths of simple closed geodesics on surfaces in $\mathcal{M}_g$.

Our results. In this paper, we present the following results, old and new:

1. In §2, following [M2] and [M1], we discuss a method to integrate *geometric* functions given in terms of the hyperbolic length functions over $\mathcal{M}_{g,n}$. This implies that the Weil-Petersson volume $V_{g,n}(L)$ of $\mathcal{M}_{g,n}(L_1, \ldots, L_n)$ is a polynomial in $L_1^2, \ldots, L_n^2$. The constant term of this polynomial, $V_{g,n} = V_{g,n}(0, \ldots, 0)$, is the Weil-Petersson volume of the moduli space of complete hyperbolic surfaces of genus g with n punctures. More generally, the coefficients of $V_{g,n}(L)$ can be written in terms of the intersection pairings of tautological line bundles over Deligne-Mumford compactification $\overline{\mathcal{M}}_{g,n}$ of the moduli space.

2. Next, in §3, we study the asymptotic behavior of Weil-Petersson volumes, and the other coefficients of volume polynomials. In particular, we show that for $n \geq 0$
$$\lim_{g\to\infty} \frac{V_{g,n}}{V_{g-1,n+2}} = 1,$$
and
$$\lim_{g\to\infty} \frac{V_{g,n+1}}{2gV_{g,n}} = 4\pi^2.$$

These results were predicted by Zograf [Z2]. We obtain several related estimates for the growth of the volumes of moduli spaces.

3. Finally, in §4, we describe the relationship between the asymptotic behavior of the Weil-Petersson volumes and the geometry of a random hyperbolic surface. In particular, we will see that in a typical hyperbolic surface of large genus, the shortest non-separating simple closed geodesic tends to be shorter than any separating simple closed geodesic. Further, we get lower bounds on the expected length of the shortest closed geodesic of a given type. For example, the shortest simple closed geodesic separating the surface into two roughly equal areas has expected length at least linear in g.

Notation. In this paper, $A \asymp B$ means that $A/C < B < AC$ for some universal constant C. Also, $A = O(B)$ means that $A < BC$, for some universal constant C.

Notes and remarks.

1. In [BM] Brooks and Makover developed a method for the study of *typical* Riemann surfaces with large genus by using trivalent graphs. In this model the expected value of the systole of a random Riemann surface turns out to be bounded (independent of the genus) [MM]. See also [Ga]. It seems that a random Riemann surface with respect to the Weil-Petersson volume form has some similar features. However, it is not clear how the measure induced by their model is related to the measure induced by the Weil-Petersson volume.

2. The distribution of hyperbolic surfaces of genus g produced randomly by gluing Riemann surfaces with long geodesic boundary components is closely related to the measure induced by ω on $\mathcal{M}_{g,n}$. See [M4] for details.

3. The following exact asymptotic formula was proved in [MZ]. There exists $C > 0$ such that for any fixed $g \geq 0$

$$V_{g,n} = n! C^n n^{(5g-7)/2}(a_g + O(1/n)), \tag{1}$$

as $n \to \infty$.

Moreover, Zograf developed a fast algorithm for calculating the volume polynomials, and made the following conjecture on the basis of the numerical data obtained by his algorithm [Z2]:

Conjecture 1.1 (Zograf). *For any fixed $n \geq 0$*

$$V_{g,n} = (4\pi^2)^{2g+n-3}(2g-3+n)!\frac{1}{\sqrt{g\pi}}\left(1 + \frac{c_n}{g} + O\left(\frac{1}{g^2}\right)\right)$$

as $g \to \infty$.

4. We warn the reader that there are some small differences in the normalization of the Weil-Petersson volume form in the literature; in this paper,

$$V_{g,n} = V_{g,n}(0,\dots,0) = \frac{1}{(3g-3+n)!}\int_{\mathcal{M}_{g,n}} \omega^{3g-3+n}$$

which is slightly different from the notation used in [Z2] and [ST]. Also, in [Z1] the Weil-Petersson Kähler form is 1/2 the imaginary part of the Weil-Petersson pairing, while here the factor 1/2 does not appear. So our answers are different by a power of 2.

Acknowledgement. I would like to thank Peter Zograf for many discussions regarding the growth of Weil-Petersson volumes. I am grateful to Curt McMullen for his guidance which initiated this work. I would also like to thank all my teachers and friends from Sharif University of Technology for showing me the beauty of mathematics. Finally, I am indebted to my family for their unceasing love, emotional support and encouragement.

2. Weil-Petersson Measure on $\mathcal{M}_{g,n}$

First, we briefly recall some background material and constructions in Teichmüller theory of Reimann surfaces with geodesic boundary components. For further details see [IT], [M2] and [Bu].

Teichmüller Space. A point in the *Teichmüller space* $\mathcal{T}(S)$ is a complete hyperbolic surface X equipped with a diffeomorphism $f : S \to X$. The map f provides a *marking* on X by S. Two marked surfaces $f : S \to X$ and $g : S \to Y$ define the same point in $\mathcal{T}(S)$ if and only if $f \circ g^{-1} : Y \to X$ is isotopic to a conformal map. When ∂S is nonempty, consider hyperbolic Riemann surfaces homeomorphic to S with geodesic boundary components of fixed length. Let $A = \partial S$ and $L = (L_\alpha)_{\alpha\in A} \in \mathbb{R}_+^{|A|}$. A point $X \in \mathcal{T}(S, L)$ is a marked hyperbolic surface with geodesic boundary components such that for each boundary component $\beta \in \partial S$, we have

$$\ell_\beta(X) = L_\beta.$$

Let $S_{g,n}$ be an oriented connected surface of genus g with n boundary components $(\beta_1,\dots,\beta_n)$. Then

$$\mathcal{T}_{g,n}(L_1,\dots,L_n) = \mathcal{T}(S_{g,n}, L_1,\dots,L_n),$$

denote the Teichmüller space of hyperbolic structures on $S_{g,n}$ with geodesic boundary components of length $L_1,\dots,L_n$. By convention, a geodesic of length zero is a cusp and we have

$$\mathcal{T}_{g,n} = \mathcal{T}_{g,n}(0,\dots,0).$$

Let $\mathrm{Mod}(S)$ denote the mapping class group of S, or the group of isotopy classes of orientation preserving self homeomorphisms of S leaving each boundary component setwise fixed. The mapping class group $\mathrm{Mod}_{g,n} = \mathrm{Mod}(S_{g,n})$ acts on $\mathcal{T}_{g,n}(L)$ by changing the marking. The quotient space

$$\mathcal{M}_{g,n}(L) = \mathcal{M}(S_{g,n}, \ell_{\beta_i} = L_i) = \mathcal{T}_{g,n}(L_1, \ldots, L_n)/\,\mathrm{Mod}_{g,n}$$

is the moduli space of Riemann surfaces homeomorphic to $S_{g,n}$ with n boundary components of length $\ell_{\beta_i} = L_i$. Also, we have

$$\mathcal{M}_{g,n} = \mathcal{M}_{g,n}(0, \ldots, 0).$$

For a disconnected surface $S = \bigcup_{i=1}^{k} S_i$ such that $A_i = \partial S_i \subset \partial S$, we have

$$\mathcal{M}(S, L) = \prod_{i=1}^{k} \mathcal{M}(S_i, L_{A_i}),$$

where $L_{A_i} = (L_s)_{s \in A_i}$.

The Weil-Petersson symplectic form. By work of Goldman [Go], the space $\mathcal{T}_{g,n}(L_1, \ldots, L_n)$ carries a natural symplectic form invariant under the action of the mapping class group. This symplectic form is called the *Weil-Petersson symplectic form,* and denoted by ω or ω_{wp}. We investigate the volume of the moduli space with respect to the volume form induced by the Weil-Petersson symplectic form. Also, when S is disconnected, we have

$$\mathrm{Vol}(\mathcal{M}(S, L)) = \prod_{i=1}^{k} \mathrm{Vol}(\mathcal{M}(S_i, L_{A_i})).$$

When $L = 0$, there is a natural complex structure on $\mathcal{T}_{g,n}$, and this symplectic form is in fact the Kähler form of a Kähler metric [IT].

The Fenchel-Nielsen coordinates. A *pants decomposition* of S is a set of disjoint simple closed curves which decompose the surface into pairs of pants. Fix a system of pants decomposition of $S_{g,n}$, $\mathcal{P} = \{\alpha_i\}_{i=1}^{k}$, where $k = 6g-6+2n$. For a marked hyperbolic surface $X \in \mathcal{T}_{g,n}(L)$, the *Fenchel-Nielsen coordinates* associated with $\mathcal{P}$, $\{\ell_{\alpha_1}(X), \ldots, \ell_{\alpha_k}(X), \tau_{\alpha_1}(X), \ldots, \tau_{\alpha_k}(X)\}$, consists of the set of lengths of all geodesics used in the decomposition and the set of the *twisting* parameters used to glue the pieces. We have an isomorphism

$$\mathcal{T}_{g,n}(L) \cong \mathbb{R}_+^{\mathcal{P}} \times \mathbb{R}^{\mathcal{P}}$$

by the map

$$X \to (\ell_{\alpha_i}(X), \tau_{\alpha_i}(X)).$$

See [Bu] for more details.

By work of Wolpert, over Teichmüller space the Weil-Petersson symplectic structure has a simple form in Fenchel-Nielsen coordinates [W1].

Theorem 2.1 (Wolpert). *The Weil-Petersson symplectic form is given by*

$$\omega_{wp} = \sum_{i=1}^{k} d\ell_{\alpha_i} \wedge d\tau_{\alpha_i}.$$

Given a simple closed geodesic α on X and $t \in \mathbb{R}$, we can deform the hyperbolic structure of X by a right twist along α as follows. First, cut X along α and then reglue back after twisting distance t to the right. We observe that the hyperbolic structure of the complement of the cut extends to a new hyperbolic structure on S. The resulting continuous path in Teichmüller space is the Fenchel-Nielsen deformation of X along α. By Theorem 2.1 the vector field generated by twisting around is symplectically dual to the exact one-form $d\ell_\alpha$. In other words, the natural twisting around α is the Hamiltonian flow of the length function of α.

Integrating geometric functions over moduli spaces. Here, we develop a method for integrating certain geometric functions over $\mathcal{M}_{g,n}(L)$. Working with bordered Riemann surfaces allows us to exploit the existence of commuting Hamiltonian S^1-actions on certain coverings of the moduli space in order to integrate certain geometric functions over the moduli space of curves.

Let $S_{g,n}$ be a closed surface of genus g with n boundary components and let $Y \in \mathcal{T}_{g,n}$. For a simple closed curve γ on $S_{g,n}$, let $[\gamma]$ denote the homotopy class of γ and let $\ell_\gamma(Y)$ denote the hyperbolic length of the geodesic representative of $[\gamma]$ on Y. To each simple closed curve γ on $S_{g,n}$, we associate the set

$$\mathcal{O}_\gamma = \{[\alpha] \,|\alpha \in \mathrm{Mod}_{g,n} \cdot \gamma\}$$

of homotopy classes of simple closed curves in the $\mathrm{Mod}_{g,n}$-orbit of γ on $X \in \mathcal{M}_{g,n}$. Given a function $f : \mathbb{R}_+ \to \mathbb{R}_+$, and a multicurve γ on $S_{g,n}$ define

$$f_\gamma : \mathcal{M}_{g,n} \to \mathbb{R}$$

by

$$f_\gamma(X) = \sum_{[\alpha]\in\mathcal{O}_\gamma} f(\ell_\alpha(X)). \tag{2}$$

The main idea for integrating over $\mathcal{M}^\gamma_{g,n}$ is that the decomposition of the surface along γ gives rise to a description of $\mathcal{M}^\gamma_{g,n}$ in terms of moduli spaces corresponding to simpler surfaces. This leads to formulas for the integral of f_γ in terms of the Weil-Petersson volumes of moduli spaces of bordered Riemann surfaces and the function f.

We sketch the proof for the case where γ is a connected simple closed curve. See Theorem 2.2 for the general case.

First, consider the covering space of $\mathcal{M}_{g,n}$

$$\pi^\gamma : \mathcal{M}^\gamma_{g,n} = \{(X,\alpha) \mid X \in \mathcal{M}_{g,n}, \text{and } \alpha \in \mathcal{O}_\gamma \text{ is a geodesic on } X\} \to \mathcal{M}_{g,n},$$

where $\pi^\gamma(X,\alpha) = X$. The hyperbolic length function descends to the function,

$$\ell : \mathcal{M}^\gamma_{g,n} \to \mathbb{R}$$

defined by $\ell(X,\eta) = \ell_\eta(X)$. Therefore, we have

$$\int\limits_{\mathcal{M}_{g,n}} f_\gamma(X)\, dX = \int\limits_{\mathcal{M}^\gamma_{g,n}} f \circ \ell(Y)\, dY.$$

On the other hand, the function f is constant on each level set of ℓ and we have

$$\int\limits_{\mathcal{M}^\gamma_{g,n}} f \circ \ell(Y)\, dY = \int\limits_0^\infty f(t)\, \mathrm{Vol}(\ell^{-1}(t))\, dt,$$

where the volume is taken with respect to the volume form $- * d\ell$ on $\ell^{-1}(t)$.

Let $S_{g,n}(\gamma)$ be the result of cutting the surface $S_{g,n}$ along γ; that is $S_{g,n}(\gamma) \cong S_{g,n} - U_\gamma$, where U_γ is an open neighborhood of γ homeomorphic to $\gamma \times (0,1)$. Thus $S_{g,n}(\gamma)$ is a possibly disconnected compact surface with $n+2$ boundary components. We define $\mathcal{M}(S_{g,n}(\gamma), \ell_\gamma = t)$ to be the moduli space of Riemann surfaces homeomorphic to $S_{g,n}(\gamma)$ such that the lengths of the 2 boundary components corresponding to γ are equal to t. We have a natural circle bundle

$$\begin{array}{ccc} S^1 & \longrightarrow & \ell^{-1}(t) \subset \mathcal{M}^\gamma_{g,n} \\ & & \Big\downarrow \\ & & \mathcal{M}(S_{g,n}(\gamma), \ell_\gamma = t) \end{array}$$

We will study the S^1-action on the level set $\ell^{-1}(t) \subset \mathcal{M}^\gamma_{g,n}$ induced by twisting the surface along γ. The quotient space $\ell^{-1}(t)/S^1$ inherits a symplectic form from the Weil-Petersson symplectic form. On the other hand, $\mathcal{M}(S_{g,n}(\gamma), \ell_\gamma = t)$ is equipped with the Weil-Petersson symplectic form. By investigating these S^1-actions in more detail, one can show that

$$\ell^{-1}(t)/S^1 \cong \mathcal{M}(S_{g,n}(\gamma), \ell_\gamma = t)$$

as symplectic manifolds. Therefore, we have

$$\mathrm{Vol}(\ell^{-1}(t)) = t\ \mathrm{Vol}(\mathcal{M}(S_{g,n}(\gamma), \ell_\gamma = t)).$$

For any connected simple closed curve γ on $S_{g,n}$, we have

$$\int_{\mathcal{M}_{g,n}} f_\gamma(X)\, dX = \int\limits_0^\infty f(t)\ t\ \mathrm{Vol}(\mathcal{M}(S_{g,n}(\gamma), \ell_\gamma = t))\, dt. \tag{3}$$

In general, we have ([M2]):

Theorem 2.2. *For any multicurve* $\gamma = \sum_{i=1}^{k} c_i\gamma_i$*, the integral of* f_γ *over* $\mathcal{M}_{g,n}(L)$ *with respect to the Weil-Petersson volume form is given by*

$$\int_{\mathcal{M}_{g,n}(L)} f_\gamma(X)\, dX = \frac{2^{-M(\gamma)}}{|\operatorname{Sym}(\gamma)|} \int_{\mathbf{x}\in\mathbb{R}_+^k} f(|\mathbf{x}|)\, V_{g,n}(\Gamma, \mathbf{x}, \beta, L)\, \mathbf{x}\cdot d\mathbf{x},$$

where $\Gamma = (\gamma_1, \ldots, \gamma_k)$*,* $|\mathbf{x}| = \sum_{i=1}^{k} c_i\, x_i$*,* $\mathbf{x}\cdot d\mathbf{x} = x_1 \cdots x_k \cdot dx_1 \wedge \cdots \wedge dx_k$*, and*

$$M(\gamma) = |\{i | \gamma_i \quad \textit{separates off a one-handle from} \;\; S_{g,n}\}|.$$

Given a multicurve $\gamma = \sum_{i=1}^{k} c_i\gamma_i$, the symmetry group of γ, $\operatorname{Sym}(\gamma)$, is defined by

$$\operatorname{Sym}(\gamma) = \operatorname{Stab}(\gamma)/\cap_{i=1}^{k} \operatorname{Stab}(\gamma_i).$$

Recall that given $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}_+^k$, $V_{g,n}(\Gamma, \mathbf{x}, \beta, L)$ is defined by

$$V_{g,n}(\Gamma, \mathbf{x}, \beta, L) = \operatorname{Vol}(\mathcal{M}(S_{g,n}(\gamma), \ell_\Gamma = \mathbf{x}, \ell_\beta = L)).$$

Also,

$$V_{g,n}(\Gamma, \mathbf{x}, \beta, L) = \prod_{i=1}^{s} V_{g_i,n_i}(\ell_{A_i}),$$

where

$$S_{g,n}(\gamma) = \bigcup_{i=1}^{s} S_i\ , \tag{4}$$

$S_i \cong S_{g_i,n_i}$, and $A_i = \partial S_i$.

By Theorem 2.2 integrating f_γ, even for a compact Riemann surface, reduces to the calculation of volumes of moduli spaces of bordered Riemann surfaces. This formula can be used to relate the growth of the number of simple closed geodesics on $X \in \mathcal{M}_g$ to the volume polynomials [M3].

Remark. Let $g \in \operatorname{Sym}(\gamma)$, where $\gamma = \sum_{i=1}^{k} c_i\gamma_i$. Then $g(\gamma_i) = \gamma_j$ implies that $c_i = c_j$.

Connection with the intersection pairings of tautological line bundles. The moduli space $\mathcal{M}_{g,n}$ is endowed with natural cohomology classes. An example of such a class is the Chern class of a vector bundle on the moduli space. When $n > 0$, there are n tautological line bundles defined on $\overline{\mathcal{M}}_{g,n}$ as follows. For each marked point i, there exists a canonical line bundle $\mathcal{L}_i$ in the orbifold sense whose fiber at the point $(C, x_1, \ldots, x_n) \in \overline{\mathcal{M}}_{g,n}$ is the cotangent space of C at x_i. The first Chern class of this bundle is denoted by $\psi_i = c_1(\mathcal{L}_i)$. Note that although the complex curve C may have nodes, x_i never coincides

with the singular points. For any set $\{d_1, \ldots, d_n\}$ of integers, define the top intersection number of ψ classes by

$$\langle \tau_{d_1}, \ldots, \tau_{d_n} \rangle_g = \int\limits_{\overline{\mathcal{M}}_{g,n}} \prod_{i=1}^{n} \psi_i^{d_i} \quad .$$

Such products are well defined when the d_i's are non-negative integers and $\sum\limits_{i=1}^{n} d_i = 3g - 3 + n$. In other cases $\langle \tau_{d_1}, \ldots, \tau_{d_n} \rangle_g$ is defined to be zero. Since we are in the orbifold setting, these intersection numbers are rational numbers. See [HM] and [AC] for more details. In [M1], we use the symplectic geometry of moduli spaces of bordered Riemann surfaces to relate these intersection pairings to the volume polynomials. This method allows us to read off the intersection numbers of tautological line bundles from the volume polynomials:

Theorem 2.3. *In terms of the above notation,*

$$\mathrm{Vol}(\mathcal{M}_{g,n}(L_1, \ldots, L_n)) = \sum_{|\mathbf{d}| \leq 3g-3+n} C_g(\mathbf{d}) \;\; L_1^{2d_1} \ldots L_n^{2d_n},$$

where $\mathbf{d} = (d_1, \ldots, d_n)$, *and* $C_g(\mathbf{d})$ *is equal to*

$$\frac{2^{m(g,n)|\mathbf{d}|}}{2^{|\mathbf{d}|} \; |\mathbf{d}|! \; (3g-3+n-|\mathbf{d}|)!} \int\limits_{\overline{\mathcal{M}}_{g,n}} \psi_1^{d_1} \cdots \psi_n^{d_n} \cdot \omega^{3g-3+n-|\mathbf{d}|}.$$

Here $m(g,n) = \delta(g-1) \times \delta(n-1)$, $\mathbf{d}! = \prod_{i=1}^{n} d_i!$, *and* $|\mathbf{d}| = \sum_{i=1}^{n} d_i$.

Recursive formulas for volume polynomials. We approach the study of the volumes of $\mathcal{M}_{g,n}(L)$ via the length functions of simple closed geodesics on a hyperbolic surface in $\mathcal{T}_{g,n}(L)$. Our point of departure for calculating these volume polynomials is a result due to McShane [Mc] which gives an identity for the lengths of certain types of simple closed geodesics on a surface $X \in \mathcal{M}_{g,n}$ when $n > 0$. Here we just cite the simplest case of this identity for $g = n = 1$.

Let $X \in \mathcal{T}_{1,1}$ be a hyperbolic once-punctured torus. Then we have

$$\sum_{\gamma} (1 + e^{\ell_\gamma(X)})^{-1} = \frac{1}{2},$$

where the sum is over all simple closed geodesics γ on X. Note that the left hand side of this identity is a *geometric function* for $f(t) = 1/(1+e^t)$ in the sense of (2), and the right hand side is independent of X.

This identity can be generalized to hyperbolic surfaces with finitely many geodesic boundary components or cone singularities [TWZ2]. In [LM], Labourie and McShane generalize the length identities to arbitrary cross ratios; as a result they obtain new identities for the Hitchin representations of surface groups in $SL(n, \mathbb{R})$. For further generalization of these identities see [TWZ1] and [Bo].

Remark. A recursive formula for the Weil-Petersson volume of the moduli space of punctured spheres was obtained by Zograf [Z1]. Moreover, Zograf and Manin have obtained generating functions for the Weil-Petersson volume of $\mathcal{M}_{g,n}$[MZ]. See also [KMZ]. Penner has developed a different method for calculating the Weil-Petersson volume of the moduli spaces of curves with marked points by using decorated Teichmüller theory [Pe].

3. Asymptotic Behavior of Weil-Petersson Volumes and Tautological Intersection Pairings

In this section, we study the asymptotics behavior of the Weil-Petersson volume of $\mathcal{M}_{g,n}$ as $g \to \infty$.

It is known [Gr] that for a fixed $n > 0$ there are $c_1, c_2 > 0$ such that

$$c_2^g (2g)! < \mathrm{Vol}(\mathcal{M}_{g,n}) < c_1^g (2g)!.$$

This result was extended to the case of $n = 0$ in [ST]. However, these estimates do not give much information about the growth of

$$B_{g,n} = V_{g,n}/V_{g-1,n+2}$$

and

$$C_{g,n} = V_{g,n+1}/(2gV_{g,n})$$

when $g \to \infty$.

Notation. For $\mathbf{d} = (d_1, \ldots, d_n)$ with $d_i \in \mathbb{N} \cup \{0\}$ and $|\mathbf{d}| = d_1 + \ldots + d_n \leq 3g - 3 + n$, let $d_0 = 3g - 3 - |\mathbf{d}|$ and define

$$\left[\prod_{i=1}^{n} \tau_{d_i}\right]_{g,n} = \frac{\prod_{i=1}^{n}(2d_i+1)!2^{|\mathbf{d}|}}{\prod_{i=0}^{n} d_i!} \int_{\overline{\mathcal{M}}_{g,n}} \psi_1^{d_1} \cdots \psi_n^{d_n} \omega^{d_0} =$$

$$= \frac{(2\pi^2)^{d_0} \prod_{i=1}^{n}(2d_i+1)!!2^{2|d|}}{d_0!} \int_{\overline{\mathcal{M}}_{g,n}} \psi_1^{d_1} \cdots \psi_n^{d_n} \kappa_1^{d_0},$$

where $\kappa_1 = \frac{\omega}{2\pi^2}$ is the first Mumford class on $\overline{\mathcal{M}}_{g,n}$ [AC]. By Theorem 2.3 for $L = (L_1, \ldots, L_n)$ we have:

$$V_{g,n}(2L) = \sum_{|\mathbf{d}| \leq 3g-3+n} [\tau_{d_1}, \ldots \tau_{d_n}]_{g,n} \frac{L_1^{2d_1}}{(2d_1+1)!} \cdots \frac{L_n^{2d_n}}{(2d_n+1)!}. \tag{5}$$

Some useful recursive formulas for the intersection pairings. Given $\mathbf{d} = (d_1, \ldots, d_n)$ with $|\mathbf{d}| \leq 3g - 3 + n$, the following recursive formulas hold:

I.

$$\left[\tau_0\tau_1 \prod_{i=1}^{n} \tau_{d_i}\right]_{g,n+2} = \left[\tau_0^4 \prod_{i=1}^{n} \tau_{d_i}\right]_{g-1,n+4} +$$

$$+\frac{1}{2} \sum_{\substack{g_1+g_2=g \\ \{1,\ldots,n\}=I\amalg J}} \left[\tau_0^2 \prod_{i\in I} \tau_{d_i}\right]_{g_1,|I|+2} \cdot \left[\tau_0^2 \prod_{i\in J} \tau_{d_i}\right]_{g_2,|J|+2},$$

II.

$$(2g-2+n)\left[\prod_{i=1}^{n} \tau_{d_i}\right]_{g,n} = \frac{1}{2} \sum_{L=0}^{3g-3+n} (-1)^L (L+1) \frac{\pi^{2L}}{(2L+3)!} \left[\tau_{L+1} \prod_{i=1}^{n} \tau_{d_i}\right]_{g,n+1}.$$

III. Let $a_0 = 1/2$, and for $n \geq 1$,

$$a_n = \zeta(2n)(1 - 2^{1-2n}).$$

Then we have

$$[\tau_{d_1}, \ldots, \tau_{d_n}]_{g,n} = \sum_{j=2}^{n} \mathcal{A}_{\mathbf{d}}^j + \frac{1}{2}\mathcal{B}_{\mathbf{d}} + \frac{1}{2}\mathcal{C}_{\mathbf{d}},$$

where

$$\mathcal{A}_{\mathbf{d}}^j = \sum_{L=0}^{d_0} (2d_j + 1)\, a_L \left[\tau_{d_1+d_j+L-1}, \prod_{i\neq 1,j} \tau_{d_i}\right]_{g,n-1},$$

$$\mathcal{B}_{\mathbf{d}} = \sum_{L=0}^{d_0} \sum_{k_1+k_2=L+d_1-2} a_L \left[\tau_{k_1}\tau_{k_2} \prod_{i\neq 1} \tau_{d_i}\right]_{g-1,n+1},$$

and

$$\mathcal{C}_{\mathbf{d}} = \sum_{\substack{I\amalg J=\{2,\ldots,n\} \\ 0\leq g'\leq g}} \sum_{L=0}^{d_0} \sum_{k_1+k_2=L+d_1-2} a_L \left[\tau_{k_1} \prod_{i\in I} \tau_{d_i}\right]_{g',|I|+1} \times [\tau_{k_2} \prod_{i\in J} \tau_{d_i}]_{g-g',|J|+1}.$$

Remarks.

- Formula (**I**) is a special case of Proposition 3.3 in [LX1]. For different proofs of (**II**) see [DN] and [LX1]. The proof presented in [DN] uses the properties of moduli spaces of of hyperbolic surfaces with cone points. See also [KMZ] and [AC].

- In terms of the volume polynomials (**II**) can be written as ([DN]):

$$\frac{\partial V_{g,n+1}}{\partial L}(L, 2\pi i) = 2\pi i(2g-2+n)V_{g,n}(L).$$

When $n = 0$,

$$V_{g,1}(2\pi i) = 0,$$

and

$$\frac{\partial V_{g,1}}{\partial L}(2\pi i) = 2\pi i(2g-2)V_g. \tag{6}$$

Note that (**III**) applies only when $n > 0$. In the case of $n = 0$, (6) allows us to prove necessary estimates for the growth of $V_{g,0}$.

- Although (**III**) has been described in purely combinatorial terms, it is closely related to the topology of different types of pairs of pants in a surface. In fact, this formula gives us the volume of $\mathcal{M}_{g,n}(L)$ in terms of volumes of moduli spaces of Riemann surfaces that we get by removing a pair of pants containing at least one boundary component of $S_{g,n}$.

- If $d_1 + \ldots + d_n = 3g - 3 + n$, (**III**) gives rise to a recursive formula for the intersection pairings of ψ_i classes which is the same as the Virasoro constraints for a point. This result is equivalent to the Witten-Kontsevich formula [M2], [LX2]. See also [MS]. For different proofs and discussions related to these relations see [Wi], [Ko], [OP], [M1], [KL], and [EO]. In this paper, we are mainly interested in the intersection parings only containing κ_1 and ψ_i classes. For generalizations of (**III**) to the case of higher Mumford's κ classes see [LX1] and [E].

Basic general estimates. The main advantage of using (**III**) is that all the coefficients are positive. Moreover, it is easy to check that

$$\zeta(2n)(1-2^{1-2n}) = \frac{1}{(2n-1)!}\int_0^\infty \frac{t^{2n-1}}{1+e^t}\,dt.$$

Hence,

$$a_{n+1} - a_n = \int_0^\infty \frac{1}{(1+e^t)^2}\left(\frac{t^{2n+1}}{(2n+1)!} + \frac{t^{2n}}{2n!}\right)dt.$$

As a result, we have:

1. $\{a_n\}_{n=1}^\infty$ is an increasing sequence, and $\lim_{n\to\infty} a_n = 1$;

2. $$a_{n+1} - a_n \asymp 1/2^{2n}. \tag{7}$$

Using this observation one can prove the following general estimates:

- For any $\mathbf{d} = (d_1, \ldots, d_n)$
$$[\tau_{d_1}, \ldots, \tau_{d_n}]_{g,n} \leq [\tau_0, \ldots, \tau_0]_{g,n} = V_{g,n}.$$
- Then (5) implies that
$$V_{g,n}(2L_1, \ldots, 2L_n) \leq e^L V_{g,n}, \tag{8}$$
where $L = L_1 + \ldots + L_n$.
- Moreover, since
$$[\tau_1, \tau_0, \ldots, \tau_0] \leq V_{g,n}$$
(**I**) and (**II**) for $\mathbf{d} = 0$ imply that for any $g, n \geq 0$,
$$V_{g,n+2} \geq V_{g-1,n+4}, \quad \text{and} \quad b \cdot V_{g,n+1} > (2g-2+n)V_{g,n}, \tag{9}$$
where $b = \sum_{L=0}^{\infty} \pi^{2L}(L+1)/(2L+3)!$.

Remark. We will show that as $g \to \infty$ the first inequality of (9) is asymptotically sharp. However, (1) implies that when g is fixed and n is large this inequality is far from being sharp; in fact, given $g \geq 1$ as $n \to \infty$
$$V_{g,n+2} \asymp \sqrt{n}\, V_{g-1,n+4}.$$

Asymptotic behavior of the coefficients of volume polynomials. Let $n \geq 0$. The following estimates hold:

- Combining (9) and (7) with a more careful analysis of (**III**) implies that for any $k \in \mathbb{N}$
$$\frac{[\tau_k, \tau_0, \ldots, \tau_0]_{g,n}}{V_{g,n}} = 1 + O(1/g), \tag{10}$$
as $g \to \infty$.

  This is a special case of Conjecture 2 in [Z2]. However, (10) fails if k is not very small compare to g.
- Also, (**III**) implies that:
$$\sum_{i=1}^{g-1} i(g-i)V_{i,1+n}V_{g-i,1+n} = O(V_{g,1+n}),$$
and hence by (9), we get
$$\sum_{i=1}^{g-1} V_{i,2+n}V_{g-i,2+n} = O(V_{g,2+n}/g), \tag{11}$$
as $g \to \infty$.

In fact, one can prove a stronger version of (11) by replacing the right-hand side with $O(V_{g,2+n}/g^2)$.

Asymptotic behavior of ratios $B_{g,n}$ and $C_{g,n}$. We use (**II**) to show that when n is fixed

$$V_{g,n} \asymp V_{g-1,n+2}, \quad \text{and} \quad V_{g,n+1} \asymp gV_{g,n}. \tag{12}$$

More generally, we show:

Theorem 3.1. *Let $n \geq 0$. As $g \to \infty$*

a):

$$\frac{V_{g,n+1}}{2gV_{g,n}} \to 4\pi^2,$$

b):

$$\frac{V_{g,n}}{V_{g-1,n+2}} \to 1.$$

Sketch of Proof. We use the following elementary observation to prove (a):

Elementary fact. *Let $\{r_i\}_{i=1}^{\infty}$ be a sequence of real numbers and $\{k_g\}_{g=1}^{\infty}$ be an increasing sequence of positive integers. Assume that for $g \geq 1$, and $i \in \mathbb{N}$, $0 \leq c_{g,i} \leq c_i$, and $\lim_{g\to\infty} c_{g,i} = c_i$. If $\sum_{i=1}^{\infty} |c_i r_i| < \infty$, then*

$$\lim_{g\to\infty} \sum_{i=1}^{k_g} r_i c_{g,i} = \sum_{i=1}^{\infty} r_i c_i. \tag{13}$$

Now, let

$$r_i = (-1)^i \frac{\pi^{2i}(i+1)}{(2i+3)!}, \; k_g = 3g-3+n \;, c_i = 1 \;\text{ and }\; c_{g,i} = \frac{[\tau_{i+1}\tau_0 \ldots \tau_0]_{g,n}}{V_{g,n+1}}.$$

By (13), and (**II**) for $\mathbf{d} = 0$ we get

$$\lim_{g\to\infty} \frac{2(2g-2+n)V_{g,n}}{V_{g,n+1}} = \frac{1}{3!} - \frac{2\pi^2}{5!} + \ldots + (-1)^L (L+1)\frac{\pi^{2L}}{(2L+3)!} + \ldots = \frac{1}{2\pi^2}.$$

On the other hand, from (**I**) and (11) we get that for $n \geq 2$:

$$\lim_{g\to\infty} \frac{V_{g,n}}{V_{g-1,n+2}} = 1.$$

Finally, we can use part (a) to get the same result for $n = 0, 1$. □

In fact, we can prove that as $g \to \infty$:

$$\frac{V_{g,n+2}}{2gV_{g,n+1}} = 4\pi^2 + O(1/g), \quad \text{and} \quad \frac{V_{g,n}}{V_{g-1,n+2}} = 1 + O(1/g). \tag{14}$$

These stronger results imply that:

$$\sum_{i=1}^{g-1} V_{i,1} \times V_{g-i,1} \asymp V_{g,1}/g^2 \asymp V_g/g. \tag{15}$$

Remark. These estimates are all consistent with the conjectures of Zograf [Z2] on the growth of Weil-Petersson volumes as $g \to \infty$.

4. Random Riemann Surfaces of High Genus

In this section, we will discuss the typical behavior of a Riemann surfaces of large genus with respect to the Weil-Petersson measure.

Notation. The mapping class group $\mathrm{Mod}_{g,n}$ acts naturally on the set of isotopy classes of simple closed curves on $S_{g,n}$: Two simple closed curves α_1 and α_2 are of the same *type* if and only if there exists $g \in \mathrm{Mod}_{g,n}$ such that $g \cdot \alpha_1 = \alpha_2$. The type of a simple closed curve is determined by the topology of $S_{g,n} - \alpha$, the surface that we get by cutting $S_{g,n}$ along α.

To simplify the notation, let γ_0 be a non-separating simple closed curve on S_g, and γ_i be a separating simple closed curve on S_g such that

$$S_g - \gamma_i = S_{i,1} \cup S_{g-i,1}.$$

Thin part of $\mathcal{M}_g$. First, we discuss the probability of appearance of a short closed geodesic in a random surface. Recall that every hyperbolic surface has a thick-thin decomposition; the thin part is the region of injectivity radius is less than a fixed small number. The thin components of a hyperbolic surface are neighborhoods of cusps or tubular neighborhoods of short geodesics.

The set of hyperbolic surfaces with lengths of closed geodesics bounded below by a constant $\epsilon > 0$ is a compact subset $\mathcal{C}^{\epsilon}_{g,n}$ of the moduli space $\mathcal{M}_{g,n}$. Some geometric properties of the moduli space can be controlled more easily in $\mathcal{C}^{\epsilon}_{g,n}$. See [Hu] and [Te]. Let $\mathcal{M}^{\epsilon}_{g,n} = \mathcal{M}_{g,n} - \mathcal{C}^{\epsilon}_{g,n}$.

Theorem 4.1. *Given $c > 0$, and $n \geq 0$, there exists $\epsilon > 0$ such that for any $g \geq 2$*

$$\mathrm{Vol}_{wp}(\mathcal{M}^{\epsilon}_{g,n}) < c\,\mathrm{Vol}_{wp}(\mathcal{M}_{g,n}).$$

Here we sketch the proof for the case of $n = 0$. Consider the function

$$F^{\epsilon} : \mathcal{M}_g \to \mathbb{R}_+$$

defined by

$$F^{\epsilon}(X) = |\{\gamma | \ell_{\gamma}(X) \leq \epsilon\}| = F^{\epsilon}_0(X) + \ldots F^{\epsilon}_{g/2}(X),$$

where $F_i^\epsilon(X) = |\{\gamma | \gamma \in \mathcal{O}_{\gamma_i}, \ell_\gamma(X) \leq \epsilon\}|$. Then by Theorem 2.2, we have

$$\mathrm{Vol}_{wp}(\mathcal{M}_g^\epsilon) \leq \int_{\mathcal{M}_g} F^\epsilon(X)\, dX \leq$$

$$\leq \sum_{i=1}^{g/2} \int_0^\epsilon t\, \mathrm{Vol}_{wp}(\mathcal{M}(S_g - \gamma_i, t, t))\, dt \; + \int_0^\epsilon t\, \mathrm{Vol}_{wp}(\mathcal{M}_{g-1,2}(t,t))\, dt$$

On the other hand, by (8) we know that if t is small enough for $i \geq 1$,

$$\mathrm{Vol}_{wp}(\mathcal{M}(S_g - \gamma_i, t, t)) \leq 2V_{i,1} \times V_{g-i,1},$$

and

$$\mathrm{Vol}_{wp}(\mathcal{M}_{g-1,2}(t,t)) \leq 2V_{g-1,2}.$$

Hence, when ϵ is small (independent of g), from (12) and (15) we get

$$\mathrm{Vol}_{wp}(\mathcal{M}_g^\epsilon) = O\left(\epsilon^2 \left(\sum_{i=1}^{g/2} V_{i,1}V_{g-i,1} + V_{g-1,2}\right)\right) = O(\epsilon^2 V_g).$$

Remark. Even though we can make the ratio $T_{g,n}^\epsilon = \mathrm{Vol}(\mathcal{M}_{g,n}^\epsilon)/\mathrm{Vol}(\mathcal{M}_{g,n})$ small, for any fixed $\epsilon > 0$, $T_{g,n}^\epsilon$ does not tend to zero as $g \to \infty$.

Behavior of the systoles. Next, we would like to know how the length of the shortest closed geodesic on a random Riemann surface grows with the genus. In general, the systole of a compact metric space X is defined to be the least length of a noncontractible loop in X. It is known that there are Riemann surfaces of large genus whose systole behaves logarithmically in the the genus [BP]. In fact, by [KSV] there is a principal congruence tower of Hurwitz surfaces (PCH), such that

$$\ell_{syst}(X_{PCH}) \geq \frac{4}{3}\log(g(X_{PCH})),$$

where $\ell_{syst}(X)$ is the length of a shortest simple closed geodesic on X. However, such a closed geodesic could be separating or non-separating. For more on properties of the function $\ell_{syst} : \mathcal{T}_{g,n} \to \mathbb{R}_+$ see [S1] and [S2].

First, consider the set of separating simple closed geodesics of typer γ_1. Since

$$\frac{V_g}{V_{1,1} \times V_{g-1,1}} \asymp g,$$

Theorem 2.2 and (8) imply that we can not cover $\mathcal{M}_g$ with surfaces which have a short separating curve $\gamma \in \mathcal{O}_{\gamma_1}$. More precisely, let

$$\mathcal{C}_g(L) = \{X \in \mathcal{M}_g | \; \exists \text{ separating curve } \alpha,\; \ell_\alpha(X) \leq L\} \subset \mathcal{M}_g.$$

Then

$$\mathcal{C}_g(L) = \bigcup_{i=1}^{g/2} \mathcal{C}_g(\gamma_i, L),$$

where

$$\mathcal{C}_g(\gamma, L) = \{X \in \mathcal{M}_g | \ \exists \alpha \in \mathcal{O}_\gamma \ , \ell_\alpha(X) \leq L\} \subset \mathcal{M}_g.$$

Then by Theorem 2.2 and (8), for $1 \leq i \leq g/2$

$$\mathrm{Vol}_{wp}(\mathcal{C}_g(\gamma_i, L)) \leq V_{i,1} \times V_{g-i,1} e^L L^2.$$

Hence (15) implies that

$$\frac{\mathrm{Vol}_{wp}(\mathcal{C}_g(L))}{V_g} \leq \frac{\sum_{i=1}^{g/2} \mathrm{Vol}_{wp}(\mathcal{C}_g(\gamma_i, L))}{V_g} \leq L^2 e^L \sum_{i=1}^{g/2} \frac{V_{i,1} \times V_{g-i,1}}{V_g} = O\left(\frac{L^2 e^L}{g}\right).$$

This implies:

Theorem 4.2. *The probability that a Riemann surface in $\mathcal{M}_g$ has a separating simple closed geodesic of length $\leq \frac{1}{3}\log(g)$ tends to zero as $g \to \infty$.*

Remark. On the other hand, because $\frac{V_g}{V_{g-1,2}}$ is bounded, the situation is very different for a non-separating simple closed curve. In fact, the probability that a random Riemann surface has a short non-separating simple closed geodesic is asymptotically positive.

Finally, we consider the following quantity similar to the Cheeger constant [Bu] of a Riemann surface. Given $X \in \mathcal{T}_g$, let

$$L(X) = \inf_C \frac{\ell_C(X)}{\min[\mathrm{area}(A), \mathrm{area}(B)]},$$

where C runs over (possibly disconnected) simple closed *geodesics* on X, with $X - C = A \cup B$. Then there exists $c > 0$ such that

$$\frac{\mathrm{Vol}_{wp}\{X \ | X \in \mathcal{M}_g, L(X) < c\}}{V_g} \to 0$$

as $g \to \infty$.

References

[AC] E. Arbarello and M. Cornalba. *Combinatorial and algebro-geometric cohomology classes on the Moduli Spaces of Curves.* J. Algebraic Geometry 5 (1996), 705–709.

[Bo] B. Bowditch. *Markoff triples and quasifuchsian groups.* Proceedings of the London Mathematical Society, Vol. 77 (1998), 697–736.

[BM] R. Brooks and E. Makover. *Random Construction of Riemann Surfaces.* J. Differential Geom. 68:1 (2004), 121–157.

[Bu] P. Buser. Geometry and spectra of compact Riemann surfaces, Birkhauser Boston, 1992.

[BP] P. Buser and P. Sarnak. *On the period matrix of a Riemann surface of large genus.* Invent. Math. 117:1 (1994), 27–56.

[DN] N. Do and P. Norbury. *Weil-Petersson volumes and cone surfaces.* Geom. Dedicata 141 (2009), 93–107

[E] B. Eynard. *Recursion between Mumford volumes of moduli spaces.* Preprint.

[EO] B. Eynard and N. Orantin. *Invariants of algebraic curves and topological expansion.* Commun. Number Theory Phys. 1:2 (2007), 347–452.

[Ga] A. Gamburd. *Poisson-Dirichlet distribution for random Belyi surfaces.* Ann. Probab. 34:5 (2006), 1827–1848.

[Go] W. Goldman. *The symplectic nature of fundamental groups of surfaces.* Adv. Math. 54 (1984), 200–225.

[Gr] S. Grushevsky. *An explicit upper bound for Weil-Petersson volumes of the moduli spaces of punctured Riemann surfaces.* Mathematische Annalen 321 (2001) 1, 1–13.

[HM] J. Harris and I. Morrison. Moduli of Curves. Graduate Texts in Mathematics, vol. 187, Springer-Verlag, 1998.

[Hu] Z. Huang. *The Weil-Petersson geometry on the thick part of the moduli space of Riemann surfaces.* Proc. Amer. Math. Soc. 135 (2007)

[IT] Y. Imayoshi and M. Taniguchi. *An introduction to Teichmüller spaces* Springer-Verlag, 1992.

[KSV] M. Katz, M. Schaps and U. Vishne. *Logarithmic growth of systole of arithmetic Riemann surfaces along congruence subgroups.* J. Differential Geom. 76:3 (2007), 399–422.

[KMZ] R. Kaufmann, Y. Manin, and D. Zagier. *Higher Weil-Petersson volumes of moduli spaces of stable n-pointed curves.* Comm. Math. Phys. 181 (1996), 736–787.

[KL] M. E. Kazarian and S. K. Lando. *An algebro-geometric proof of Witten's conjecture.* J. Amer. Math. Soc. 20 (2007), 1079–1089.

[Ko] M. Kontsevich. *Intersection on the moduli space of curves and the matrix Airy function.* Comm. Math. Phys. 147 (1992).

[LM] F. Labourie and G. McShane. *Cross ratios and identities for higher Teichmller-Thurston theory.* Duke Math. J. 149:2 (2009), 279–345.

[LX1] K. Liu and H. Xu. *Recursion formulae of higher Weil-Petersson volumes* Int. Math. Res. Not. IMRN 2009, no. 5, 835–859.

[LX2] K. Liu, and H. Xu. *Mirzakharni's recursion formula is equivalent to the Witten-Kontsevich theorem.* Preprint.

[MM] E. Makover and J. McGowan. *The length of closed geodesics on random Riemann Surfaces.* Preprint.

[MZ] Yu. Manin and P. Zograf. *Invertible cohomological field theories and Weil-Petersson volumes.* Ann. Inst. Fourier 50:2 (2000), 519–535.

[Mc] G. McShane. *Simple geodesics and a series constant over Teichmüller space.* Invent. Math. 132 (1998), 607–632.

[M1] M. Mirzakhani. *Weil-Petersson volumes and intersection theory on the moduli space of curves.* J. Amer. Math. Soc. 20:1 (2007), 1–23.

[M2] M. Mirzakhani. *Simple geodesics and Weil-Petersson volumes of moduli spaces of bordered Riemann surfaces.* Invent. Math. 167 (2007), 179–222.

[M3] M. Mirzakhani. *Growth of the number of simple closed geodesics on hyperbolic surfaces.* Annals of Math.168:1 (2008), 97–125.

[M4] M. Mirzakhani. Random hyperbolic surfaces and measured laminations. In the tradition of Ahlfors-Bers. IV, 179–198, Contemp. Math., 432, Amer. Math. Soc., Providence, RI, 2007.

[MS] Y. Mulase and P. Safnuk. *Mirzakhani's recursion relations, Virasoro constraints and the KdV hierarchy.* Indian Journal of Mathematics 50 (2008), 189–228.

[OP] A. Okounkov and R. Pandharipande. *Gromov-Witten theory, Hurwitz theory, and matrix models, I.* Preprint.

[Pe] R. Penner. *Weil-Petersson volumes.* J. Differential Geom. 35 (1992), 559–608.

[S1] P. Schmutz. *Geometry of Riemann surfaces based on closed geodesics.* Bulletin (New Series) of the American Mathematical Society 35:3 (1998), 193–214.

[S2] P. Schmutz. *Systoles on Riemann surfaces.* Manuscripta Math., 85:(3-4), 429–447, 1994.

[ST] G. Schumacher and S. Trapani. *Estimates of Weil-Petersson volumes via effective divisors* Comm. Math. Phys. 222, No.1 (2001), 1–7.

[TWZ1] S. Tan, Y. Wong andY Zhang. *Generalized Markoff maps and McShane's identity.* Adv. Math. 217 (2008), no. 2, 761–813.

[TWZ2] S. Tan, Y. Wong and Y. Zhang. *Generalizations of McShane's identity to hyperbolic cone-surfaces.* J. Differ. Geom. 72 (2006), 73–111.

[Te] L. Teo. *The Weil-Petersson geometry of the moduli space of Riemann surfaces.* Proc. Amer. Math. Soc. 137 (2009), 541–552.

[Wi] E. Witten. *Two-dimensional gravity and intersection theory on moduli spaces.* Surveys in Differential Geometry 1 (1991), 243–269.

[W1] S. Wolpert. *An elementary formula for the Fenchel-Nielsen twist.* Comment. Math. Helv. 56 (1981), 132–135.

[W2] S. Wolpert. *On the symplectic geometry of deformations of a hyperbolic surface.* Ann. of Math. 117:2 (1983), 207–234.

[W3] S. Wolpert. *Behavior of geodesic-length functions on Teichmüller space.* J. Differential Geom. 79:2 (2008), 277–334.

[W3] S. Wolpert. *The Weil-Petersson metric geometry.* In Handbook of Teichmüller theory. Vol. II, volume 13 of IRMA Lect. Math. Theor. Phys., 47–64. Eur. Math. Soc., Zurich, 2009.

[Z1] P. Zograf. The Weil-Petersson volume of the moduli space of punctured spheres, Mapping class groups and moduli spaces of Riemann surfaces. Contemp. Math., vol. 150, Amer. Math. Soc., 1993, 367–372.

[Z2] P. Zograf. *On the large genus asymptotics of Weil-Petersson volumes.* Preprint.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

A New Family of Complex Surfaces of General Type with $p_g = 0$

Jongil Park*

Abstract

In this article we review how to construct new families of simply connected complex surfaces of general type with $p_g = 0$ and $2 \leq K^2 \leq 4$ using a rational blow-down surgery and $\mathbb{Q}$-Gorenstein smoothing theory. Furthermore, we also explain that this technique is a very powerful tool to construct many other interesting families of complex surfaces.

Mathematics Subject Classification (2010). Primary 14J29; Secondary 14J17, 53D03.

Keywords. $\mathbb{Q}$-Gorenstein smoothing, rational blow-down, surface of general type

1. Introduction

One of the fundamental problems in the classification of complex surfaces is to find a new family of simply connected surfaces of general type with $p_g = 0$. Surfaces with $p_g = 0$ are interesting in view of Castelnuovo's criterion: An irrational surface with $q = 0$ must have $P_2 \geq 1$. This class of surfaces has been studied extensively by algebraic geometers and topologists for a long time. Nonetheless, simply connected surfaces of general type with $p_g = 0$ are little known. Although a large number of non-simply connected complex surfaces of general type with $p_g = 0$ have been known due to Godeaux, Campedelli and so on ([BHPV], Chapter VII), it was only in 1983 that the first example of a *simply connected* surface of general type with $p_g = 0$ appeared, the so-called Barlow surface [B]. Barlow surface has $K^2 = 1$. Therefore it has been a very important problem to find a new family of simply connected surfaces with $p_g = 0$. In

*This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (2009-0093866).

Department of Mathematical Sciences, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-747, Korea & Korea Institute for Advanced Study, Seoul 130-722, Korea. E-mail: jipark@snu.ac.kr.

particular, it is very intriguing whether there is a simply connected surface of general type with $p_g = 0$ and $K^2 \geq 2$.

In 2004 the author constructed a new simply connected symplectic 4-manifold with $b_2^+ = 1$ and $K^2 = 2$ using a rational blow-down surgery ([P2]). After this construction, it has been an interesting question whether such a symplectic 4-manifold admits a complex structure. In 2006 Y. Lee and the author successfully constructed a simply connected, minimal, complex surface of general type with $p_g = 0$ and $K^2 = 2$ by modifying the symplectic 4-manifold constructed in [P2] ([LP1]). Our main techniques involved in the construction are a rational blow-down surgery and $\mathbb{Q}$-Gorenstein smoothing theory, which are very different techniques from other classical constructions such as a finite group quotient and a multiple covering, due to Godeaux, Campedelli, Burniat and others. In the following year H. Park, D. Shin and the author also found proper configurations to produce simply connected surfaces of general type with $p_g = 0$ and $3 \leq K^2 \leq 4$ using the same technique as above ([PPS1, PPS2]). Furthermore, we notice that many other families of complex surfaces of general type such as surfaces with $p_g = 0$ and small homology group, Horikawa surfaces, and simply connected surfaces with $p_g = 1$ and $q = 0$ can also be constructed using a rational blow-down surgery and $\mathbb{Q}$-Gorenstein smoothing theory ([LP2, LP3, PPS3, PPS4]).

The aim of this article is to survey these constructions above. It is organized as follows: In Section 2 we briefly review two main techniques, a rational blow-down surgery and $\mathbb{Q}$-Gorenstein smoothing theory, and we sketch in Section 3 how to construct new families of simply connected surfaces of general type with $p_g = 0$ and $2 \leq K^2 \leq 4$. And then we mention in Section 4 that many other families of complex surfaces can also be constructed via a rational blow-down surgery and $\mathbb{Q}$-Gorenstein smoothing theory.

2. Preliminaries

In this section we first briefly review a rational blow-down surgery initially introduced by R. Fintushel and R. Stern and extended by the author ([FS, P1] for details): For any relatively prime integers p and q with $p > q > 0$, we define a configuration $C_{p,q}$ as a smooth 4-manifold obtained by plumbing disk bundles over the 2-spheres instructed by the following linear diagram

$$\underset{u_k}{\overset{-b_k}{\circ}} - \underset{u_{k-1}}{\overset{-b_{k-1}}{\circ}} - \cdots - \underset{u_2}{\overset{-b_2}{\circ}} - \underset{u_1}{\overset{-b_1}{\circ}}$$

where $\frac{p^2}{pq-1} = [b_k, b_{k-1}, \ldots, b_1]$ is the unique continued fraction with all $b_i \geq 2$, and each vertex u_i represents a disk bundle over the 2-sphere whose Euler number is $-b_i$. Orient the 2-spheres in $C_{p,q}$ so that $u_i \cdot u_{i+1} = +1$. Then the configuration $C_{p,q}$ is a negative definite simply connected smooth 4-manifold whose boundary is the lens space $L(p^2, 1-pq)$. Note that the lens space $L(p^2, 1-pq)$ also bounds a rational ball $B_{p,q}$ with $\pi_1(B_{p,q}) \cong \mathbb{Z}_p$.

Definition. Suppose M is a smooth 4-manifold containing a configuration $C_{p,q}$. Then we construct a new smooth 4-manifold M_p, called a (*generalized*) *rational blow-down* of M, by replacing $C_{p,q}$ with the rational ball $B_{p,q}$. Note that this process is well-defined, that is, a new smooth 4-manifold M_p is uniquely determined (up to diffeomorphism) from M because each diffeomorphism of $\partial B_{p,q}$ extends over the rational ball $B_{p,q}$. We call the procedure replacing $C_{p,q}$ with the rational ball $B_{p,q}$ a *rational blow-down surgery*. Furthermore, M. Symington proved that a rational blow-down manifold M_p admits a symplectic structure in some cases. For example, if M is a symplectic 4-manifold containing a configuration $C_{p,q}$ such that all 2-spheres u_i in $C_{p,q}$ are symplectically embedded and intersect positively, then the rational blow-down manifold M_p also admits a symplectic structure [Sy1, Sy2].

Next, we briefly review $\mathbb{Q}$-Gorenstein smoothing theory for projective surfaces with special quotient singularities and we quote some basic facts developed in [LP1].

Definition. Let X be a normal projective surface with quotient singularities. Let $\mathcal{X} \to \Delta$ (or $\mathcal{X}/\Delta$) be a flat family of projective surfaces over a small disk Δ. The one-parameter family of surfaces $\mathcal{X} \to \Delta$ is called a $\mathbb{Q}$-*Gorenstein smoothing* of X if it satisfies the following three conditions;
(i) the general fiber X_t is a smooth projective surface,
(ii) the central fiber X_0 is X,
(iii) the relative canonical divisor $K_{\mathcal{X}/\Delta}$ is $\mathbb{Q}$-Cartier.

A $\mathbb{Q}$-Gorenstein smoothing for a germ of a quotient singularity $(X_0, 0)$ is defined similarly. A quotient singularity which admits a $\mathbb{Q}$-Gorenstein smoothing is called a *singularity of class T*.

Proposition 2.1 ([KSB, Ma]). *Let $(X_0, 0)$ be a germ of two dimensional quotient singularity. If $(X_0, 0)$ admits a $\mathbb{Q}$-Gorenstein smoothing over the disk, then $(X_0, 0)$ is either a rational double point or a cyclic quotient singularity of type $\frac{1}{dn^2}(1, dna - 1)$ for some integers a, n, d with a and n relatively prime.*

Proposition 2.2 ([KSB, Ma]).

1. *The singularities $\overset{-4}{\circ}$ and $\overset{-3}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \cdots - \overset{-2}{\circ} - \overset{-3}{\circ}$ are of class T.*
2. *If the singularity $\overset{-b_1}{\circ} - \cdots - \overset{-b_r}{\circ}$ is of class T, then so are*
$$\overset{-2}{\circ} - \overset{-b_1}{\circ} - \cdots - \overset{-b_{r-1}}{\circ} - \overset{-b_r-1}{\circ} \quad \text{and} \quad \overset{-b_1-1}{\circ} - \overset{-b_2}{\circ} - \cdots - \overset{-b_r}{\circ} - \overset{-2}{\circ}.$$
3. *Every singularity of class T that is not a rational double point can be obtained by starting with one of the singularities described in* (1) *and iterating the steps described in* (2).

Let X be a normal projective surface with singularities of class T. Then the natural question arises whether this local $\mathbb{Q}$-Gorenstein smoothing can be

extended over the global surface X or not. Roughly geometric interpretation is the following: Let $\cup V_\alpha$ be an open covering of X such that each V_α has at most one singularity of class T. By the existence of a local $\mathbb{Q}$-Gorenstein smoothing, there is a $\mathbb{Q}$-Gorenstein smoothing $\mathcal{V}_\alpha/\Delta$ of V_α. The question is if these families glue to a global one. The answer can be obtained by figuring out the obstruction map of the sheaves of deformation $T_X^i = Ext_X^i(\Omega_X, \mathcal{O}_X)$ for $i = 0, 1, 2$. For example, if X is a smooth surface, then T_X^0 is the usual holomorphic tangent sheaf T_X and $T_X^1 = T_X^2 = 0$. By applying the standard result of deformations [LS, Pa] to a normal projective surface with quotient singularities, we get the following

Proposition 2.3 ([Wa]). *Let X be a normal projective surface with quotient singularities. Then*

1. *The first order deformation space of X is represented by the global Ext 1-group $\mathbb{T}_X^1 = \mathrm{Ext}_X^1(\Omega_X, \mathcal{O}_X)$.*
2. *The obstruction lies in the global Ext 2-group $\mathbb{T}_X^2 = \mathrm{Ext}_X^2(\Omega_X, \mathcal{O}_X)$.*

Furthermore, by applying a general result of the local-global spectral sequence of ext sheaves ([Pa]) to deformation theory of surfaces with quotient singularities (i.e. $E_2^{p,q} = H^p(T_X^q) \Rightarrow \mathbb{T}_X^{p+q}$) and by the fact that $H^j(T_X^i) = 0$ for $i, j \geq 1$, we also get

Proposition 2.4 ([Ma, Wa]). *Let X be a normal projective surface with quotient singularities. Then*

1. *We have the exact sequence*

$$0 \to H^1(T_X^0) \to \mathbb{T}_X^1 \to \ker[H^0(T_X^1) \to H^2(T_X^0)] \to 0$$

   where $H^1(T_X^0)$ represents the first order deformations of X for which the singularities remain locally a product.

2. *If $H^2(T_X^0) = 0$, every local deformation of the singularities may be globalized.*

Theorem 2.5 ([LP1]). *Let X be a normal projective surface with singularities of class T. Let $\pi : V \to X$ be the minimal resolution and let E be the reduced exceptional divisors. Suppose that $H^2(T_V(-\log\ E)) = 0$. Then $H^2(T_X^0) = 0$ and there is a $\mathbb{Q}$-Gorenstein smoothing of X.*

3. Simply Connected Surfaces of General Type with $p_g = 0$

In this section we explain how to construct a new family of simply connected complex surfaces of general type with $p_g = 0$ using a rational blow-down surgery

and $\mathbb{Q}$-Gorenstein smoothing theory. The following is an overall scheme to construct such surfaces:

STEP 1: We first choose a special rational elliptic surface $Y \to \mathbb{P}^1$ with desired singular fibers by blowing up 9 times from a cubic pencil in $\mathbb{P}^2$. And then we blow up many times again to get a projective normal surface $Z = Y \sharp k\overline{\mathbb{P}}^2$ which contains several disjoint configurations $\{C_{p_1,q_1}, \dots, C_{p_n,q_n}\}$ representing the resolution graphs of singularities of class T.

STEP 2: We perform a rational blow-down surgery along the disjoint configurations $\{C_{p_1,q_1}, \dots, C_{p_n,q_n}\}$ to get a symplectic 4-manifold $Z_{p_1,\dots,p_n}$. On the other hand, we also contract these chains of curves lying in $\cup_{i=1}^n C_{p_i,q_i}$ from Z to produce a projective surface X with n number of singularities of class T.

STEP 3: We apply $\mathbb{Q}$-Gorenstein smoothing theory to prove that the normal projective X has a $\mathbb{Q}$-Gorenstein smoothing, i.e. there exists one-parameter family of surfaces $\mathcal{X} = \cup X_t \to \Delta$ with $X_0 = X$. Note that the existence of a global $\mathbb{Q}$-Gorenstein smoothing of X depends on the configurations $\{C_{p_1,q_1}, \dots, C_{p_n,q_n}\}$ lying in $Z = Y \sharp k\overline{\mathbb{P}}^2$. For example, we have the following proposition which constrains a choice of configurations $\{C_{p_1,q_1}, \dots, C_{p_n,q_n}\}$ lying in Z:

Proposition 3.1 ([LP1]). *Let Y be a rational elliptic surface and let $g : Y \to \mathbb{P}^1$ be a relatively minimal elliptic fibration without multiple fibers. Assume that Z is obtained from Y by blowing-up at the singular points $p_1, \dots, p_j$ on nodal fibers with $j \leq 2$. Let $F_1, \dots, F_j$ be the proper transforms of nodal fibers. Then $H^2(Z, T_Z(-\log(F_1 + \cdots + F_j))) = 0$.*

STEP 4: We show that the rational blow-down symplectic 4-manifold $Z_{p_1,\dots,p_n}$ is simply connected, and we also show that the general fiber X_t of $\mathbb{Q}$-Gorenstein smoothing of X is a minimal surface of general type with $p_g = 0$ and $K^2 > 0$. Note that both the simple connectivity of $Z_{p_1,\dots,p_n}$ and the minimality of X_t also depend on the configurations $\{C_{p_1,q_1}, \dots, C_{p_n,q_n}\}$ lying in $Z = Y \sharp k\overline{\mathbb{P}}^2$. I.e. we can choose appropriate configurations $\{C_{p_1,q_1}, \dots, C_{p_n,q_n}\}$ in a right rational surface $Z = Y \sharp k\overline{\mathbb{P}}^2$ so that they guarantee simple connectivity and minimality.

STEP 5: By using the fact coming from the standard argument about Milnor fibers that the general fiber X_t is diffeomorphic to the rational blow-down 4-manifold $Z_{p_1,p_2,\dots,p_n}$, we conclude that the general fiber X_t is a simply connected, minimal, surface of general type with $p_g = 0$ and $K^2 > 0$.

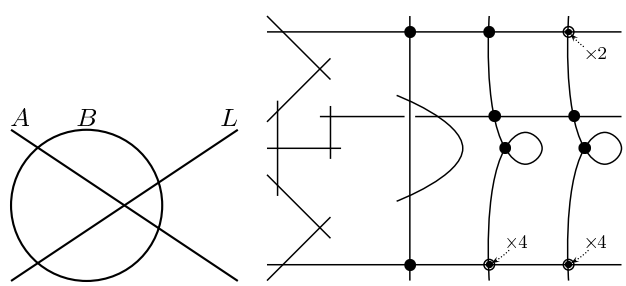
In this way, we were able to construct a series of simply connected complex surfaces of general type with $p_g = 0$ and $1 \leq K^2 \leq 4$:

Theorem 3.2 ([LP1, PPS1, PPS2]). *There exist a family of simply connected, minimal, complex surfaces of general type with $p_g = 0$ and $1 \leq K^2 \leq 4$.*

In the remaining of this section, we explicitly present appropriate configurations $C_{p,q}$'s in a right rational surface $Z = Y\sharp k\overline{\mathbb{P}}^2$ to produce desired simply connected surfaces of general type with $p_g = 0$ and $2 \leq K^2 \leq 4$.

Remark: Although an example with $p_g = 0$ and $K^2 = 1$ can also be constructed similarly, we omit the case in this article because we do not know whether it is diffeomorphic to Barlow surface or not.

3.1. An example with $p_g = 0$ and $K^2 = 2$. We begin with a special elliptic fibration $g : E(1) \to \mathbb{P}^1$ constructed as follows: Let A be a line and B be a smooth conic in $\mathbb{P}^2$. Choose another line L in $\mathbb{P}^2$ which meets B at two distinct points p, q, and which also meets A at a different point r. We may assume that the conic B and the line A meet at two different points which are not p, q, r. We now consider a cubic pencil in $\mathbb{P}^2$ induced by $A + B$ and $3L$, i.e. $\lambda(A + B) + \mu(3L)$, for $[\lambda : \mu] \in \mathbb{P}^1$ (Figure 1-(a)). After we blow up first at three points p, q, r, blow up again three times at the intersection points of the proper transforms of B, A with the three exceptional curves e_1, e_2, e_3. Finally, blowing up again three times at the intersection points of the proper transforms of B and A with the three new exceptional curves e'_1, e'_2, e'_3, we get an elliptic fibration $E(1) = \mathbb{P}^2\sharp 9\overline{\mathbb{P}}^2$ over $\mathbb{P}^1$. Let us denote this elliptic fibration by $g : Y = E(1) \to \mathbb{P}^1$. Note that there is an $\tilde{E}_6$-singular fiber on the fibration $g : Y \to \mathbb{P}^1$ which consists of the proper transforms of $L, e_1, e'_1, e_2, e'_2, e_3, e'_3$. We also note that there is one I_2-singular fiber on $g : Y \to \mathbb{P}^1$ which consists of the proper transforms of the line A and the conic B. Furthermore, by the proper choice of curves A, B and L guarantees two more nodal singular fibers on $g : Y \to \mathbb{P}^1$. Hence the fibration $g : Y \to \mathbb{P}^1$ has one $\tilde{E}_6$-singular fiber, one reducible I_2-singular fiber, and two nodal singular fibers (Figure 1-(b)).



(a) A pencil (b) $Y = E(1)$

Figure 1: A pencil and an elliptic surface Y for $K^2 = 2$

Now we blow up the surface Y totally 17 times at the marked points in Figure 1-(b) above. Then we get a rational surface $Z' := Y\sharp 17\overline{\mathbb{P}}^2$ (Figure 2).

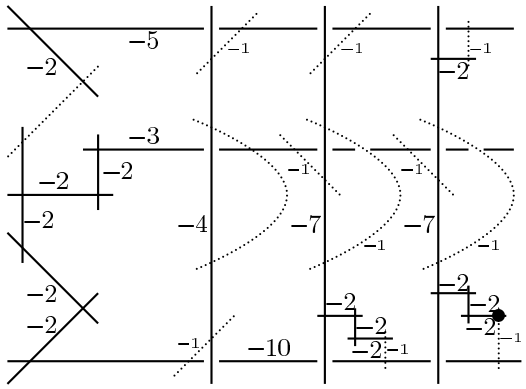


Figure 2: A rational surface $Z' = Y\sharp 17\overline{\mathbb{P}}^2$ for $K^2 = 2$

Then, by blowing up the surface Z' once again at the marked point in Figure 2, we finally get a rational surface $Z := Y\sharp 18\overline{\mathbb{P}}^2$ which contains five disjoint linear chains of $\mathbb{P}^1$'s (Figure 3):
$C_{15,7} = \overset{-2}{\circ} - \overset{-10}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-3}{\circ}$, $C_{5,1} = \overset{-7}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ}$, $C_{9,4} = \overset{-2}{\circ} - \overset{-7}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-3}{\circ}$, $C_{3,1} = \overset{-5}{\circ} - \overset{-2}{\circ}$ and $C_{2,1} = \overset{-4}{\circ}$

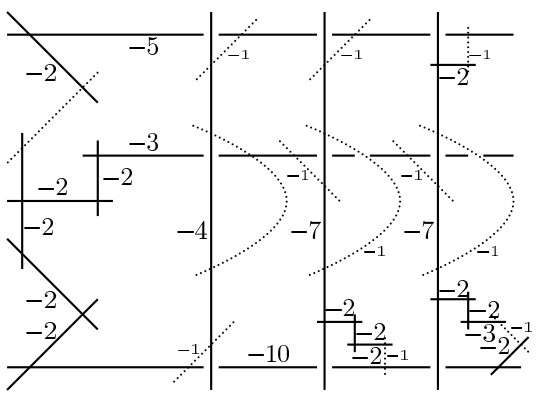
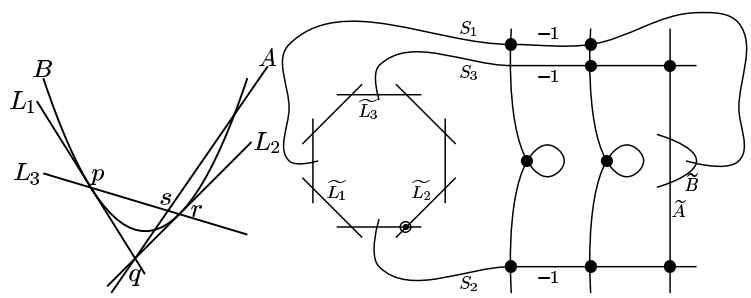


Figure 3: A rational surface $Z = Y\sharp 18\overline{\mathbb{P}}^2$ for $K^2 = 2$

Next, we contract these five disjoint chains $\{C_{15,7}, C_{9,4}, C_{5,1}, C_{3,1}, C_{2,1}\}$ from $Z = Y\sharp 18\overline{\mathbb{P}}^2$ so that it produces a normal projective surface X with five quotient singular points of class T. And then, by proving $H^2(Z, T_Z(-\log D_Z)) = 0$ (so that $H^2(X, T_X^0) = 0$), we can prove that X has a global $\mathbb{Q}$-Gorenstein smoothing due to Theorem 2.5 above. Finally, it is easy to check that the general fiber X_t of the $\mathbb{Q}$-Gorenstein smoothing of X is a simply

connected, minimal, complex surface of general type with $p_g = 0$ and $K^2 = 2$ (see [LP1] for details).

3.2. An example with $p_g = 0$ and $K^2 = 3$. Let A be a line and B be a smooth conic in $\mathbb{P}^2$ such that A and B meet at two different points. Choose a tangent line L_1 to B at a point $p \in B$ so that L_1 intersects with A at a different point $q \in A$, and draw a tangent line L_2 from q to B which tangents at the point $r \in B$. Let L_3 be the line connecting p and r which meets A at s (Figure 4-(a)).



(a) A pencil (b) $Y = E(1)$

Figure 4: A pencil and an elliptic surface Y for $K^2 = 3$

We consider a cubic pencil in $\mathbb{P}^2$ induced by $A + B$ and $L_1 + L_2 + L_3$, and blow up first at p and blow up at the intersection point of the proper transform of B with the exceptional curve e_1. And then blow up again at the intersection point of the proper transform of B with the exceptional curve e_2. Similarly, after blowing up at r, blow up two more times at the intersection point of the proper transform of B with the exceptional curves e_4 and e_5. Next, blow up at q, and then blow up again at the intersection point of the proper transform of A with the exceptional curve e_7. Let e_8 be the exceptional curve induced by the blowing up. Finally, blowing up once at s, which induces the exceptional divisor e_9, we get an elliptic fibration $E(1) = \mathbb{P}^2 \sharp 9 \overline{\mathbb{P}}^2$ over $\mathbb{P}^1$. Let us denote this elliptic fibration by $g : Y \to \mathbb{P}^1$. Note that there is an I_8-singular fiber on $g : Y \to \mathbb{P}^1$ which consists of the proper transforms of L_1, L_2, L_3, e_1, e_2, e_4, e_5, e_7. There is also one I_2-singular fiber on $g : Y \to \mathbb{P}^1$ which consists of the proper transforms of A and B, denoted by $\tilde{A}$ and $\tilde{B}$ respectively. According to the list of Persson [Pe], there exist only two more nodal singular fibers on $g : Y \to \mathbb{P}^1$ (Figure 4-(b)).

Now we blow up the surface Y totally 10 times at the marked points in Figure 4-(b) above. Then we get a rational surface $Z' := Y \sharp 10 \overline{\mathbb{P}}^2$ (Figure 5). And we blow up the surface Z' totally 11 times again at the marked points as in Figure 5.

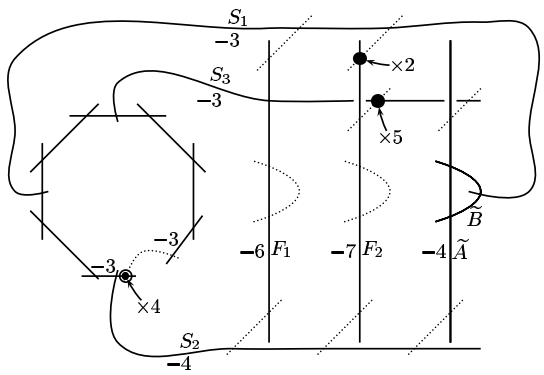


Figure 5: A rational surface $Z' = Y\sharp 10\overline{\mathbb{P}}^2$ for $K^2 = 3$

Then we finally get a rational surface $Z := Y\sharp 21\overline{\mathbb{P}}^2$ which contains four disjoint linear chains of $\mathbb{P}^1$'s, $C_{2,1} = \overset{-4}{\circ}\ (\tilde{A})$, $C_{7,1} = \overset{-9}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ}$ (which contains the proper transform of F_2), $C_{19,5} = \overset{-4}{\circ} - \overset{-7}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-3}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ}$ (which contains the proper transforms of S_1, S_2, and a part of proper transforms of I_8-singular fibers) and $C_{35,6} = \overset{-6}{\circ} - \overset{-8}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-3}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ} - \overset{-2}{\circ}$ (which contains the proper transforms of S_3, F_1, and a part of proper transforms of I_8-singular fibers) (Figure 6).

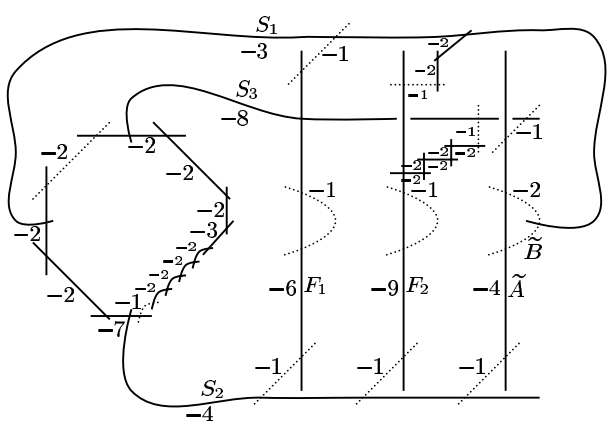
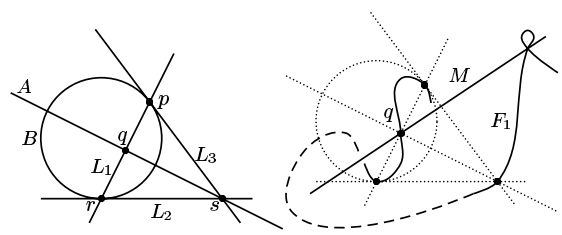


Figure 6: A rational surface $Z = Y\sharp 21\overline{\mathbb{P}}^2$ for $K^2 = 3$

Finally, by applying $\mathbb{Q}$-Gorenstein smoothing theory to Z as above, we construct a simply connected, minimal, complex surface with $p_g = 0$ and $K^2 = 3$. That is, we first contract four disjoint linear chains $C_{2,1}, C_{7,1}, C_{19,5}$ and $C_{35,6}$ of $\mathbb{P}^1$'s from Z so that it produces a normal projective surface X with four

permissible singular points. And then,by proving $H^2(Z, T_Z(-\log D_Z)) = 0$, we conclude that X has a global $\mathbb{Q}$-Gorenstein smoothing. Furthermore, the remaining argument is the same as $K^2 = 2$ case (see [PPS1] for details).

3.3. An example with $p_g = 0$ and $K^2 = 4$. Let L_1, L_2, L_3 and A be lines in $\mathbb{P}^2$ and let B be a smooth conic in $\mathbb{P}^2$ intersecting as in Figure 7(a). We consider a pencil of cubics generated by two cubic curves $L_1 + L_2 + L_3$ and $A + B$, which has 4 base points, say, p, q, r and s. In order to obtain an elliptic fibration over $\mathbb{P}^1$ from the pencil, we blow up three times at p and r, respectively, and twice at s, including infinitely near base-points at each point, and one further blowing-up at the base point q. Then, by blowing-up totally nine times, we resolve all base points of the pencil and we get an elliptic fibration $Y = \mathbb{P}^2 \sharp 9 \overline{\mathbb{P}}^2$ over $\mathbb{P}^1$ (Figure 8). Note that the elliptic fibration Y has an I_8-singular fiber consisting of the proper transforms $\widetilde{L_i}$ of L_i $(i = 1, 2, 3)$. Also Y has an I_2-singular fiber consisting of the proper transforms $\widetilde{A}$ and $\widetilde{B}$ of A and B, respectively. According to the list of Persson [Pe], we may assume that Y has only two more nodal singular fibers F_1 and F_2 by choosing generally L_i's, A and B (Figure 8).



(a) Two generators (b) A bisection

Figure 7: A pencil of cubics for $K^2 = 4$

Let M be the line in $\mathbb{P}^2$ passing through the point q and the node of the nodal cubic curve F_1. The node of F_1 does not lie on any L_i's, A, and B. Hence it satisfies that $M \neq L_1$, $M \neq A$, and $\widetilde{M} \cdot \widetilde{M} = 0$, where $\widetilde{M}$ is the proper transform of M in Y (Figure 7(b)). We may assume further that M does not pass through the node of the other nodal cubic curve F_2 by choosing generally L_i's, A, and B. Since M meets every member in the pencil at three points, $\widetilde{M}$ is a bisection of the elliptic fibration $Y \to \mathbb{P}^1$. Furthermore, since $q \in M$, the section S_2 meets $\widetilde{M}$ at one point (Figure 8).

Next, by blowing-up 9 times at the marked points on Y as in Figure 8, we get a rational surface $Z := Y \sharp 9 \overline{\mathbb{P}}^2$ which contains a special linear chain of $\mathbb{P}^1$'s,

$$C_{252,145} = \underset{u_{13}}{\overset{-2}{\circ}} - \underset{u_{12}}{\overset{-4}{\circ}} - \underset{u_{11}}{\overset{-6}{\circ}} - \underset{u_{10}}{\overset{-2}{\circ}} - \underset{u_{9}}{\overset{-6}{\circ}} - \underset{u_{8}}{\overset{-2}{\circ}} - \underset{u_{7}}{\overset{-4}{\circ}} - \underset{u_{6}}{\overset{-2}{\circ}} - \underset{u_{5}}{\overset{-2}{\circ}} - \underset{u_{4}}{\overset{-2}{\circ}} - \underset{u_{3}}{\overset{-3}{\circ}} - \underset{u_{2}}{\overset{-2}{\circ}} - \underset{u_{1}}{\overset{-3}{\circ}},$$

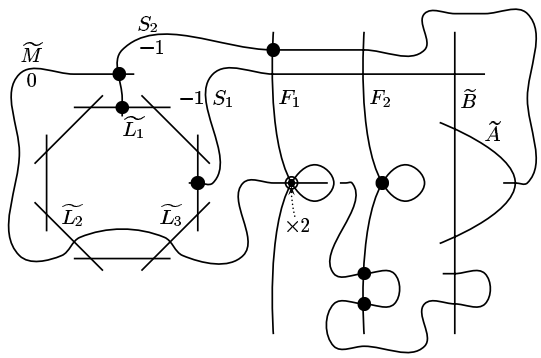


Figure 8: A rational surface Y for $K^2 = 4$

which contains $\widetilde{A}$, S_2, $\widetilde{F_2}$, S_1, $\widetilde{F_1}$, $\widetilde{M}$, $\widetilde{L_2}$, $\widetilde{L_1}$, and $\widetilde{L_3}$, where u_i represents an embedded rational curve (Figure 9).

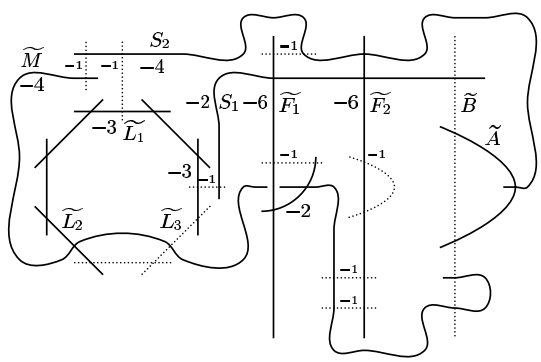


Figure 9: A rational surface $Z = Y \sharp 9\overline{\mathbb{P}}^2$ for $K^2 = 4$

Finally, the remaining argument is the same as above (see [PPS2] for details).

4. Other Surfaces Via $\mathbb{Q}$-Gorenstein Smoothings

As we notice in the previous section, $\mathbb{Q}$-Gorenstein smoothing theory together with a rational blow-down surgery is a very powerful tool to construct a new family of simply connected surfaces of general type with $p_g = 0$. In fact, this technique also produces many other interesting families of complex surfaces. For example, the following results have been obtained via $\mathbb{Q}$-Gorenstein smoothings in last several years.

Theorem 4.1 ([LP2]). *There exist minimal complex surfaces of general type with $p_g = 0$, $K^2 = 2$ and $H_1 = \mathbb{Z}/2\mathbb{Z}, \mathbb{Z}/3\mathbb{Z}$.*

Theorem 4.2 ([PPS3]). *There exists a minimal complex surface of general type with $p_g = 0$, $K^2 = 3$ and $H_1 = \mathbb{Z}/2\mathbb{Z}$.*

Theorem 4.3 ([KL]). *There exist minimal complex surfaces of general type with $p_g = 0$, $1 \leq K^2 \leq 3$ and $\pi_1 = \mathbb{Z}/2\mathbb{Z}$.*

Theorem 4.4 ([Ph]). *There exists a minimal complex surface of general type with $p_g = 0$, $K^2 = 4$ and $\pi_1 = \mathbb{Z}/2\mathbb{Z}$.*

Theorem 4.5 ([PPS4]). *There exist a family of simply connected, minimal, complex surfaces of general type with $p_g = 1, q = 0$ and $K^2 = 1, 2, \ldots, 6, 8$.*

Theorem 4.6 ([LP3]). *The projective surface X_n obtained by contracting two disjoint configurations $C_{n-2,1}$ on an elliptic surface $E(n)$ admits a $\mathbb{Q}$-Gorenstein smoothing of two quotient singularities simultaneously, and a general fiber of the $\mathbb{Q}$-Gorenstein smoothing is a Horikawa surface $H(n)$.*

References

[B] R. Barlow, *A simply connected surface of general type with $p_g = 0$*, Invent. Math. **79** (1984), 293–301.

[BHPV] W. Barth, K. Hulek, C. Peters, A. Van de Ven, *Compact complex surfaces*, 2nd ed. Springer-Verlag, Berlin, 2004.

[FS] R. Fintushel and R. Stern, *Rational blowdowns of smooth 4-manifolds*, J. Differential Geom. **46** (1997), 181–235.

[KL] J. Keum and Y. Lee, *A personal communication*, 2009

[KSB] J. Kollár and N.I. Shepherd-Barron, *Threefolds and deformations of surface singularities*, Invent. Math. **91** (1988), 299–338.

[LP1] Y. Lee and J. Park, *A surface of general type with $p_g = 0$ and $K^2 = 2$*, Invent. Math. **170** (2007), 483–505

[LP2] Y. Lee and J. Park, *A complex surface of general type with $p_g = 0, K^2 = 2$ and $H_1 = \mathbb{Z}/2\mathbb{Z}$*, Math. Res. Lett. **16** (2009), 323–330.

[LP3] Y. Lee and J. Park, *A construction of Horikawa surface via $\mathbb{Q}$-Gorenstein smoothings*, to appear in Math. Zeit., arXiv:0708.3319.

[LS] S. Lichtenbaum and M. Schlessinger, *The cotangent complex of a morphism*, Trans. Amer. Math. Soc. **128** (1967), 41–70.

[Ma] M. Manetti, *Normal degenerations of the complex projective plane*, J. Reine Angew. Math. **419** (1991), 89–118.

[Pa] V. P. Palamodov, *Deformations of complex spaces*, Russian Math. Surveys **31:3** (1976), 129–197.

[Ph] H. Park, *A complex surface of general type with $p_g = 0$, $K^2 = 4$ and $\pi_1 = \mathbb{Z}/2\mathbb{Z}$*, arXiv:0910.3476.

[PPS1] H. Park, J. Park and D. Shin, *A simply connected surface of general type with $p_g = 0$ and $K^2 = 3$*, Geom. Topol. **13** (2009), 743–767.

[PPS2] H. Park, J. Park and D. Shin, *A simply connected surface of general type with $p_g = 0$ and $K^2 = 4$*, Geom. Topol. **13** (2009), 1483–1494.

[PPS3] H. Park, J. Park and D. Shin, *A complex surface of general type with $p_g = 0$, $K^2 = 3$ and $H_1 = \mathbb{Z}/2\mathbb{Z}$*, to appear in Bull. Korean Math. Soc., arXiv:0803.1322.

[PPS4] H. Park, J. Park and D. Shin, *A construction of surfaces of general type with $p_g = 1$ and $q = 0$*, preprint (2009), arXiv:0906.5195.

[P1] J. Park, *Seiberg-Witten invariants of generalized rational blow-downs*, Bull. Austral. Math. Soc. **56** (1997), 363–384.

[P2] J. Park, *Simply connected symplectic 4-manifolds with $b_2^+ = 1$ and $c_1^2 = 2$*, Invent. Math. **159** (2005), 657–667.

[Pe] U. Persson, *Configuration of Kodaira fibers on rational elliptic surfaces*, Math. Zeit. **205** (1990), 1–47.

[Sy1] M. Symington, *Symplectic rational blowdowns*, J. Differential Geom. **50** (1998), 505–518.

[Sy2] M. Symington, *Generalized symplectic rational blowdowns*, Algebr. Geom. Topol. **1** (2001), 503–518.

[Wa] J. Wahl, *Smoothing of normal surface singularities*, Topology **20** (1981), 219–246.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Ozsváth-Szabó Invariants and 3-dimensional Contact Topology

András I. Stipsicz*

Abstract

We review applications of Ozsváth–Szabó homologies (and in particular, the contact Ozsváth–Szabó invariant) in 3-dimensional contact topology.

Mathematics Subject Classification (2010). 57R17; 57R57

Keywords. Contact 3-manifolds, tight contact structures, Heegaard Floer theory, Ozsváth–Szabó invariants, Legendrian and transverse knots

1. Contact 3-manifolds

We start by reviewing basic definitions of 3-dimensional contact topology. (For a more complete treatment the reader is advised to turn to [13].) Let Y be a given closed, oriented, smooth 3-manifold. A 1-form α is a (positive) *contact form* if $\alpha \wedge d\alpha > 0$ (with respect to the given orientation). A 2-plane field $\xi \subset TY$ is a positive, coorientable *contact structure* on Y if there is a contact 1-form $\alpha \in \Omega^1(Y)$ such that $\xi = \ker \alpha$. By fixing α up to multiplication by smooth functions $f\colon Y \to \mathbb{R}^+$, we also fix an orientation for the 2-plane field ξ: the basis $\{v_1, v_2\} \subset \xi_p$ is positive if $\{v_1, v_2, n\}$ with normal vector n satisfying $\alpha(n) > 0$ provides an oriented basis for T_pY.

The 1-form $\alpha = dz + xdy$ induces a contact structure on the 3-dimensional Euclidean space $\mathbb{R}^3$. It turns out that this contact structure extends to the 3-sphere S^3. In addition, the resulting 2-plane field is isotopic (through contact structures) to the 2-plane field of complex tangencies on S^3 when viewed as the boundary of the unit 4-ball in the complex vector space $\mathbb{C}^2$. The above structures are the *standard* contact structures on $\mathbb{R}^3$ and S^3, and we will denote

*The author was supported by OTKA T67928. He wants to thank Paolo Lisca, Peter Ozsváth and Zoltán Szabó for many helpful discussions.

MTA Rényi Institute of Mathematics, Reáltanoda utca 13–15. Budapest, HUNGARY, H-1053. E-mail: stipsicz@renyi.hu.

them by ξ_{st}. According to Darboux's Theorem, locally any contact structure is like the standard one; more precisely, for any $p \in Y$ and any contact form α on Y there are coordinates near p in which α is the standard contact form on the chart. In short, contact structures are locally the same. According to Gray's Stability Theorem, contact structures do not admit deformations, since if ξ_t ($t \in [0,1]$) is a smooth family of contact structures on the closed 3-manifold Y then there is an isotopy $(\phi_t)_{t\in[0,1]}$ such that $(\phi_t)_*(\xi_0) = \xi_t$ for all $t \in [0,1]$.

Fillings. Any closed, oriented 3-manifold is the boundary of a compact 4-manifold (i.e. the third cobordism group Ω_3 is zero). Contact 3-manifolds which are boundaries (in an appropriate sense, to be described below) of symplectic or complex 4-manifolds admit special features. Let (X,ω) be a given compact, symplectic 4-manifold, that is, X is a smooth, compact, oriented 4-manifold with possibly non-empty boundary and ω is a closed 2-form with $\omega \wedge \omega > 0$ (with respect to the given orientation). (X,ω) is a *weak symplectic filling* of the contact 3-manifold (Y,ξ) if $\partial X = Y$ as oriented manifolds and $\omega|_\xi \neq 0$. A symplectic filling (X,ω) is a *strong filling* of (Y,ξ) if there is a 1-form α near ∂X with $\omega = d\alpha$, $d\alpha|_\xi \neq 0$ and $\xi = \{\alpha|_Y = 0\}$, i.e. $\alpha|_Y$ is a contact form for ξ. The compact complex manifold (X,J) with complex structure J is a *Stein filling* of (Y,ξ) if $\partial X = Y$, ξ is given as the oriented 2-plane field of complex tangencies on Y and (X,J) is a Stein domain, that is, it admits a proper, *plurisubharmonic* function $\varphi\colon X \to [0,\infty)$ (with $\partial X = \varphi^{-1}(a)$ for some regular value $a \in \mathbb{R}$), i.e., the 2-form $\omega_\varphi = -d^{\mathbb{C}}d\varphi$ is a Kähler form with associated Kähler metric g_φ. It is not hard to see that a Stein filling is always a strong filling and a strong filling is automatically a weak filling. The converse of any of these inclusions fail to hold. The contact 3-manifold (Y,ξ) is (weakly, strongly or Stein) fillable if it admits a corresponding filling. Once again, weak (and similarly, strong) implies strong (respectively, Stein) fillability, while the converse of these implications do not hold. For more about fillings see [7].

Knots in contact topology. As knot theory plays a special role in the study of 3-manifolds, knots compatible with contact structures are extremally important in contact topology. A knot $K \subset (Y,\xi)$ is called *Legendrian* if it is tangent to ξ, i.e., if for $\xi = \ker\alpha$ we have $\alpha(TK) = 0$. Every knot can be smoothly isotoped to a Legendrian knot, in fact, for every knot there is a C^0-close Legendrian knot smoothly isotopic to it. Legendrian knots in $(\mathbb{R}^3,\xi_{st})$ (and so in (S^3,ξ_{st})) can be depicted by their *front projections* to the yz-plane, since according to the equation $x = -\frac{dz}{dy}$ the slope of the tangent of the front projection determines the x-coordinate. After possibly isotoping, every Legendrian knot admits a front projection with no triple points, only transverse double points and $(2,3)$-cusps instead of vertical tangencies. Conversely, any front projection having cusps instead of vertical tangencies and not admitting crossings with higher slope in front uniquely specifies a Legendrian knot. For this reason we will symbolize Legendrian knots in $(\mathbb{R}^3,\xi_{st})$ (and so in (S^3,ξ_{st}))

by their front projections, see also [19, 38]. Notice that if $L \subset (Y, \xi)$ is Legendrian, it admits a canonical framing: consider the unit orthogonal of the tangent vector of L in ξ. The resulting framing is called the *contact framing* of the Legendrian knot L. If L is null-homologous in Y then it admits another framing, induced by pushing off L along its existing Seifert surface; this latter framing is called the *Seifert framing*. When measuring the contact framing with respect to this Seifert framing we get an integer invariant of the Legendrian knot L called the *Thurston–Bennequin invariant* $\mathrm{tb}(L)$. If $L \subset (Y, \xi)$ is null-homologous then there is another numerical invariant we can associate to it: consider a Seifert surface $\Sigma \subset (Y, \xi)$ and take the relative Euler class of ξ (as an oriented 2-plane bundle) over Σ. For this to make sense we need to trivialize ξ over $\partial\Sigma = L$: choose the trivialization provided by the tangents of L together with their oriented normals in ξ. Note that in order to specify the tangents we need to fix an orientation on L, which provides a compatible orientation on the Seifert surface Σ. The resulting quantity, called the *rotation number* $\mathrm{rot}_\Sigma(L)$, will in general depend on the chosen Seifert surface and the orientation fixed on the knot. It is easy to see that the two 'classical' invariants $\mathrm{tb}(L)$ and $\mathrm{rot}_\Sigma(L)$ of an oriented Legendrian knot remain unchanged under Legendrian isotopy. Notice that for knots in S^3 the rotation number is independent of the chosen surface (since $H_2(S^3; \mathbb{Z}) = 0$), and both the Thurston–Bennequin and the rotation numbers can be easily read off from a front projection, cf. [38, Section 4.2].

Stabilization changes a Legendrian knot in a simple way: in a Darboux chart we replace a segment (depicted by the left of Figure 1) with one of the right diagrams of the same figure. After fixing an orientation on L, we can speak of positive and negative stabilizations, resulting in $L^\pm$. It follows that $\mathrm{tb}(L^\pm) = \mathrm{tb}(L) - 1$ and $\mathrm{rot}_\Sigma(L^\pm) = \mathrm{rot}_\Sigma(L) \pm 1$.

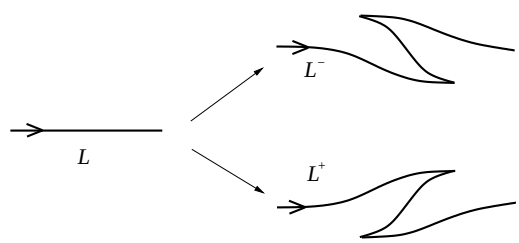


Figure 1. The two stabilizations of an oriented Legendrian knot.

A knot $T \subset (Y, \xi)$ is called *transverse* if the tangent vectors of T are transverse to ξ. A transverse knot comes with a natural orientation by declaring a tangent vector positive if the contact 1-form α evaluates positively on it. A Legendrian knot can be perturbed to become transverse, and in fact every transverse knot occurs in this way. Assume that T is null-homologous and fix a Seifert surface Σ for T. The *self-linking number* $s\ell_\Sigma(T)$ can be conveniently defined using the fact that T can be approximated by Legendrian knots: take an oriented Legendrian knot L which can be perturbed to T and define $s\ell_\Sigma(T)$

to be equal to $tb_\Sigma(L) - rot_\Sigma(L)$. Since by [8] the Legendrian approximation of T is unique up to negative stabilization and Legendrian isotopy, the above quantity is clearly independent of the chosen approximation.

Overtwisted versus tight dichotomy. A contact 3-manifold (Y, ξ) is *overtwisted* if there is an embedded 2-disk $D \subset Y$ which is tangent to ξ along its boundary. Such a disk D is called and *overtwisted disk*. If (Y, ξ) contains no overtwisted disk, we say that it is *tight*.

Theorem 1.1 (Eliashberg–Gromov). *If the contact 3-manifold (Y, ξ) is fillable then it is tight.* □

The above theorem is a major tool in proving tightness of contact structures. For a while, actually, it was unclear whether the reverse implication of the theorem is true or false. As we will show in Section 4, it is now known to be false. Regarding overtwisted contact structures we have Eliashberg's classification:

Theorem 1.2 (Eliashberg, [6]). *Two overtwisted contact structures on a closed 3-manifold Y are isotopic if and only if they are homotopic as oriented 2-plane fields. Moreover, for any oriented 2-plane field there is an overtwisted contact structure homotopic to it.* □

In short, the classification of overtwisted contact structures on a closed 3-manifold Y up to isotopy coincides with the classification of oriented 2-plane fields up to homotopy. There is a more intimate relationship between the geometry of Y and tight structures on it.

Theorem 1.3 (Eliashberg). *For any closed, oriented surface $\Sigma \subset Y$ with Euler characteristic $\chi(\Sigma)$ and tight contact structure ξ on Y either $\Sigma = S^2$ and then $\langle c_1(\xi), [\Sigma]\rangle = 0$ or we have $\chi(\Sigma) \leq 0$ and*

$$\langle c_1(\xi), [\Sigma]\rangle \leq -\chi(\Sigma).$$

For a tight contact structure ξ on Y and a Legendrian knot $L \subset (Y, \xi)$ with Seifert surface Σ the inequality

$$tb_\Sigma(L) + |\mathrm{rot}_\Sigma(L)| \leq -\chi(\Sigma)$$

is satisfied. □

According to a result of Bennequin, the contact structure ξ_{st} on $\mathbb{R}^3$ (and on S^3) is tight. On the other hand, for example, the contact structure $\xi_1 = \ker \alpha_1$ for $\alpha_1 = \cos r dz + r \sin r d\theta$ in cylindircal coordinates $(z, (r, \theta))$ on $\mathbb{R}^3$ can be easily shown to be overtwisted.

The *Giroux torsion* $\mathrm{Tor}(Y, \xi)$ of a contact 3-manifold (Y, ξ) is defined as the supremum of the integers $n \geq 1$ for which there is a contact embedding of

$$\mathbb{T}_n := (T^2 \times [0, 1], \ker(\cos(2\pi n z)dx - \sin(2\pi n z)dy))$$

(with x, y coordinates on T^2 and z on $[0,1]$) into (Y,ξ). We say that $\mathrm{Tor}(Y,\xi) = 0$ if no such embedding exists. (Notice that $\mathrm{Tor}(Y,\xi)$ might be equal to ∞; for example, the Giroux torsion is ∞ for all overtwisted contact structures.) The importance of this invariant stems from the following result.

Theorem 1.4 ([1], Theorem 1.4). *Let Y be a closed 3-manifold. For every natural number n the 3-manifold Y carries at most finitely many isomorphism classes of tight contact structures with Giroux torsion bounded above by n.* $\square$

The central problem of 3-dimensional contact topology is to classify contact structures on 3-manifolds. Since overtwisted structures are classified by their homotopy type, the question reduces to understanding tight contact structures. Tight structures are much harder to find, and seem to carry important information about the geometry of the underlying 3-manifold, as is demonstrated by the successful application of contact topological arguments in the solution of several low-dimensional problems, see for example [23, 24], cf. also [44]. Great advances have been made in the recent past in classifying tight contact structures on some simple 3-manifolds, and this question is still in the focus of active research. In this note we would like to recall an application of contact Ozsváth–Szabó invariants to solve the classification problem on certain classes of 3-manifolds. Closely related to this problem, in the theory of Legendrian and transverse knots, it is a natural question, to which extent do the 'classical' invariants tb and rot (and $s\ell$ in the transverse case) determine the Legendrian (resp. transverse) knot in a given knot type. If these numerical invariants determine the Legendrian (transverse) knot, the knot type is called *Legendrian (transverse) simple*. As we will show, Ozsváth–Szabó invariants can be used to provide examples of non-simple knot types.

Contact surgery. Suppose that $L \subset (Y,\xi)$ is a Legendrian knot in the given contact 3-manifold. Consider the contact framing on L and perform r-surgery with respect to this framing. The resulting 3-manifold is denoted by $Y_r(L)$. According to the classification of tight contact structures on solid tori [20], the contact structure ξ admits an extension from $Y - \nu(L)$ to $Y_r(L)$ as a tight structure on the new glued-up torus provided $r \neq 0$. (The extension might not be tight on the entire closed 3-manifold $Y_r(L)$ but it is required to be tight on the solid torus of the surgery.) Such a tight extension is not unique in general; the different extensions can be determined from the continued fraction coefficients of r. Nevertheless, the extension is unique if $r \in \mathbb{Q}$ is of the form $\frac{1}{k}$ for some integer $k \in \mathbb{Z}$. In particular, according to the above, we have that if $\mathbb{L} = \mathbb{L}^+ \cup \mathbb{L}^- \subset (S^3, \xi_{st})$ is a given Legendrian link, then the result of contact $(+1)$-surgery along components of $\mathbb{L}^+$ and contact (-1)-surgery along components of $\mathbb{L}^-$ uniquely specifies a contact 3-manifold $(Y_{\mathbb{L}}, \xi_{\mathbb{L}})$. In fact, the converse of this statement also holds, namely we have

Theorem 1.5 (Ding–Geiges, [2], cf. also [4]). *For a given contact 3-manifold (Y,ξ) there exists a Legendrian link $\mathbb{L}=\mathbb{L}^+\cup\mathbb{L}^-\subset(S^3,\xi_{st})$ such that $(Y_{\mathbb{L}},\xi_{\mathbb{L}})=(Y,\xi)$. In fact, we can assume that $|\mathbb{L}^+|\le 1$.* □

According to a result of Eliashberg, the contact 3-manifold $(Y_{\mathbb{L}},\xi_{\mathbb{L}})$ is Stein fillable once $\mathbb{L}=\mathbb{L}^-$. The tightness of $(Y_{\mathbb{L}},\xi_{\mathbb{L}})$ in general is, however, a rather delicate question. As we will see, contact Ozsváth–Szabó invariants (and their appropriate variants) provide convenient tools to study such questions.

Open book decompositions and Giroux's theorem. The definition of the contact Ozsváth–Szabó invariant rests on a seminal result of Giroux [18], providing a close connection between open book decompositions and contact structures on a given 3-manifold Y. Here we restrict ourselves to an outline of this beautiful theory; for a more complete treatment the reader is advised to turn to [9, 18].

Suppose that $L\subset Y$ is a fibered link in Y, that is, the complement $Y-L$ fibers as $f\colon Y-L\to S^1$ over the circle S^1, and the fibers of f provide (interiors of) Seifert surfaces for L. A fiber F of f is a *page*, while L is the binding of the open book decomposition. The monodromy φ of the fibration $f\colon Y-L\to S^1$ is called the *monodromy* of the open book decomposition. A contact structure ξ on Y is said to be *compatible* with an open book decomposition (F,φ) on Y if there is a contact 1-form α defining ξ such that the binding L is transverse with respect to ξ and the 2-form $d\alpha$ is a volume form on each page. In addition, we assume that the orientation of the binding as a transverse knot coincides with its orientation as the boundary of a page (which is oriented by $d\alpha$).

According to a classical theorem of Thurston and Winkelnkemper, for any open book decomposition there exists a contact structure compatible with it, and a simple argument shows that if two contact structures are compatible with the same open book decomposition then they are isotopic. Giroux [18] proved that the converse of this statement is also true, namely for any contact structure there is an open book decomposition compatible with it. Let F' denote the surface we get by adding a 1-handle to F. The open book decomposition with page F' and monodromy $\varphi\circ t_a$ is called a *positive stabilization* of (F,φ) if t_a is the right-handed Dehn twist along a simple closed curve $a\subset F'$ intersecting the cocore of the new 1-handle in a single (transverse) point. With this notion in place, we can formulate the central result clarifying the relation between open book decompositions and compatible contact structures.

Theorem 1.6 (Giroux, [18]). **(a)** *For a given open book decomposition of Y there is a compatible contact structure ξ on Y. Contact structures compatible with a fixed open book decomposition are isotopic.*

(b) *For a contact structure ξ on Y there is a compatible open book decomposition of Y. Two open book decompositions compatible with a fixed contact structure admit common positive stabilization.* □

2. Heegaard Floer Theory

In this section we outline the basics of Heegaard Floer theory; we restrict ourselves to a short introduction, highlighting the aspects crucial for contact topological considerations. For a more detailed treatment see [41, 42].

Ozsváth–Szabó homologies of 3-manifolds. Elementary Morse theory shows that a closed, oriented 3-manifold Y admits a *Heegaard decomposition* $Y = U_1 \cup_{\Sigma_g} U_2$ into two solid genus-g handlebodies U_1 and U_2, glued together along a surface Σ_g of genus g. A solid genus-g handlebody with boundary Σ_g can be specified by g disjoint, simple closed curves $\alpha_1, \ldots, \alpha_g \subset \Sigma_g$ which are linearly independent in homology: attaching handles along α_i (together with a 3-ball) we recover the given handlebody. Therefore Y can be described by a *Heegaard diagram*

$$(\Sigma_g, \boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^g, \boldsymbol{\beta} = \{\beta_j\}_{j=1}^g).$$

Consider the gth symmetric power $\mathrm{Sym}^g(\Sigma_g)$ and the g-dimensional tori $\mathbb{T}_\alpha = \alpha_1 \times \ldots \times \alpha_g$ and $\mathbb{T}_\beta = \beta_1 \times \ldots \times \beta_g$ in it. A symplectic structure on Σ_g gives rise to a symplectic structure on $\mathrm{Sym}^g(\Sigma_g)$; let J be an appropriate compatible almost-complex structure. Furthermore, fix a point $w \in \Sigma_g$ (the *basepoint*) distinct from all the $\boldsymbol{\alpha}$- and $\boldsymbol{\beta}$-curves and consider the hypersurface $V_w = \{w\} \times \mathrm{Sym}^{g-1}(\Sigma_g)$, which is disjoint from the tori $\mathbb{T}_\alpha$ and $\mathbb{T}_\beta$. For $\mathbf{x}, \mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ let $\mathfrak{M}_{\mathbf{x},\mathbf{y}}$ denote the moduli space of holomorphic maps $u\colon \Delta \to \mathrm{Sym}^g(\Sigma_g) - V_w$ from the unit disk $\Delta \subset \mathbb{C}$ with the properties that $u(i) = \mathbf{x}, u(-i) = \mathbf{y}$ and the arc connecting i and $-i$ on $\partial\Delta$ is mapped into $\mathbb{T}_\alpha$ (resp. into $\mathbb{T}_\beta$) if the points on the arc have negative (resp. positive) real parts. The space $\mathfrak{M}_{\mathbf{x},\mathbf{y}}$ admits an $\mathbb{R}$-action, let $\mathfrak{N}_{\mathbf{x},\mathbf{y}}$ denote the quotient by this action.

Consider $\widehat{CF}(Y) = \oplus_{\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta} \mathbb{Z}_2\langle x \rangle$ and define the map $\widehat{\partial}\colon \widehat{CF}(Y) \to \widehat{CF}(Y)$ by the matrix element $\langle \partial \mathbf{x}, \mathbf{y} \rangle$ (for $\mathbf{x}, \mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$) as

$$\langle \widehat{\partial} \mathbf{x}, \mathbf{y} \rangle = \#\mathfrak{N}_{\mathbf{x},\mathbf{y}} \qquad (\mathrm{mod}\ 2),$$

where $\#\mathfrak{N}_{\mathbf{x},\mathbf{y}}$ is the number of 0-dimensional components of $\mathfrak{N}_{\mathbf{x},\mathbf{y}}$. (For the sake of simplicity above we used $\mathbb{Z}_2$-coefficients. The theory can be set up using $\mathbb{Z}$-coefficients, in which case a coherent choice of orientations of the various moduli spaces must be made.) When $b_1(Y) \neq 0$, not every Heegaard diagram gives rise to a well-defined theory, and one must assume that the Heegaard diagram is *(weakly) admissible*, cf. [41].

Standard theory of Floer homologies shows that $\widehat{\partial} \circ \widehat{\partial} = 0$, hence $(\widehat{CF}(Y), \widehat{\partial})$ is a chain complex. We define the Ozsváth–Szabó homology $\widehat{HF}(Y)$ of the 3-manifold Y as the homology of this chain complex.

Theorem 2.1 (Ozsváth–Szabó, [41]). *The Abelian group $\widehat{HF}(Y)$ is an invariant of the 3-manifold Y and is independent of the choices made throughout its definition.* □

It can be shown directly that by fixing the basepoint w, any intersection point $\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ determines a *spin^c structure* $\mathfrak{t}_{\mathbf{x}}$, and $\widehat{\partial}\mathbf{x}$ can have components only with the same induced spin^c structure. Consequently, the chain complex $(\widehat{CF}(Y), \widehat{\partial})$ naturally splits as a direct sum $\oplus_{\mathfrak{t} \in Spin^c(Y)}(\widehat{CF}(Y, \mathfrak{t}), \widehat{\partial})$, defining a splitting

$$\widehat{HF}(Y) = \oplus_{\mathfrak{t} \in Spin^c(Y)} \widehat{HF}(Y, \mathfrak{t})$$

of the Ozsváth–Szabó homology groups. It has been proved [41] that the group $\widehat{HF}(Y, \mathfrak{t})$ is an invariant of the spin^c 3-manifold $(Y, \mathfrak{t})$. In addition, for a spin^c structure $\mathfrak{t}$ with $c_1(\mathfrak{t})$ torsion, a relative $\mathbb{Z}$-grading can be defined on $\widehat{HF}(Y, \mathfrak{t})$, which lifts to an absolute $\mathbb{Q}$-grading, called the *homological* (or *Maslov*) grading. In conclusion, the Ozsváth–Szabó homology group $\widehat{HF}(Y, \mathfrak{t})$ with $c_1(\mathfrak{t})$ torsion splits as $\widehat{HF}(Y, \mathfrak{t}) = \oplus_{d \in \mathbb{Q}} \widehat{HF}_d(Y, \mathfrak{t})$.

Suppose now that W is an oriented cobordism between the 3-manifolds Y_1 and Y_2. Using Heegaard triples and counting holomorphic triangles, a map

$$F_W \colon \widehat{HF}(Y_1) \to \widehat{HF}(Y_2)$$

is defined in [42]. As in the 3-dimensional case, the map splits according to spin^c structures on the cobordisms; in the following F_W denotes the sum of the induced maps for all spin^c structures.

In addition, a spin^c cobordism $(W, \mathfrak{s})$ from $(Y_1, \mathfrak{t}_1)$ to $(Y_2, \mathfrak{t}_2)$, with $\mathfrak{t}_1, \mathfrak{t}_2$ torsion spin^c structures shifts the absolute $\mathbb{Q}$-grading by the rational number

$$\tfrac{1}{4}(c_1^2(\mathfrak{s}) - 3\sigma(W) - 2\chi(W)).$$

(Notice that the fact that $\mathfrak{t}_1, \mathfrak{t}_2$ are torsion spin^c structures implies that the square $c_1^2(\mathfrak{s})$ is well-defined as an element of $\mathbb{Q}$.) Although the determination of the set $\mathbb{T}_\alpha \cap \mathbb{T}_\beta$ and hence the generators of $\widehat{CF}(Y)$ is a purely combinatorial question (based on the combinatorics of the Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$), the boundary map $\widehat{\partial}$ requires the study of moduli spaces of certain J-holomorphic maps. This step is typically far from being algorithmic and makes the computation of the homology groups a challenge in general. According to a recent result of Sarkar-Wang [48], however, for specific diagrams (which they called *nice*) the boundary operator $\widehat{\partial}$ can be also combinatorially computed from the Heegaard diagram. In addition, it was also shown in [48] that every 3-manifold admits nice Heegaard diagrams. In [40] it was shown that (by extending the theory to admit more basepoints) an appropriately stabilized version of $\widehat{HF}(Y)$ can be actually *defined* by purely combinatorial means.

An important feature of Heegaard Floer theory is that it admits a *surgery exact triangle*. To describe it, suppose that a 3-manifold Y and a knot $K \subset Y$ are given. Perform integer surgery along K, resulting in a 3-manifold Y_K and a cobordism X_1 from Y to Y_K. Consider a normal circle N to K and attach a 4-dimensional 2-handle to $Y_K \times [0, 1]$ along N with framing (-1). The resulting 3-manifold will be denoted by Y', while the cobordism is X_2. Repeat this last

step, i.e., attach a 2-handle to Y' along a normal circle U of N with framing (-1). It is not hard to see that the resulting 3-manifold is diffeomorphic to Y; denote the last cobordism by X_3. This geometric situation induces a triangle on Ozsváth–Szabó homologies. The central result for computing Ozsváth–Szabó homologies is the following

Theorem 2.2 (Surgery exact triangle, [42]; cf. [46]). *The triangle defined above for Ozsváth–Szabó homologies is exact.* □

Knot Floer homologies. By choosing two points $w, z \in \Sigma - \boldsymbol{\alpha} - \boldsymbol{\beta}$, an oriented knot is specified in the 3-manifold Y: connect z to w in the complement of the $\boldsymbol{\alpha}$-curves and w to z in the complement of the $\boldsymbol{\beta}$-curves (and push the resulting arcs into the corresponding handlebodies). In fact, any pair (Y, K) of a closed 3-manifold and a knot in it can be presented by such a doubly pointed Heegaard diagram. Taking the same group $\widehat{CF}(Y)$, but modifying the boundary map $\widehat{\partial}$ to $\widehat{\partial}_K$ by considering only those maps which have their image in $\mathrm{Sym}^g(\Sigma) - V_z - V_w$, a new chain complex, and therefore a new homology theory $\widehat{HFK}(Y, K)$ can be defined [43], which provides an interesting and powerful invariant for knots. As an application of the theory of nice diagrams, in [34] it was shown that these invariants can be computed combinatorially for all knots (and links) in S^3, and in fact, the theory admits a combinatorial definition for knots and links in S^3 [35]. It is not hard to see that for knots in S^3 the resulting theory (together with the relative spin^c and homological gradings) provides a categorification of the Alexander polynomial. Roughly the same idea (together with the introduction of more basepoints) provides invariants of links in 3-manifolds. The combinatorial definition of (a suitably stabilized version of) these invariants follows the same pattern as the corresponding definition of the homologies of 3-manifolds given in [40].

3. Contact Ozsváth–Szabó Invariants

The most spectacular success of Ozsváth–Szabó homologies stems from its applications to knot theory and to contact topology. In the following we will focus on the contact topological applications. First we discuss the definition and basic properties of the contact invariant defined in [45], and (some) applications will be given in the next chapter.

Contact Ozsváth–Szabó invariants. According to the result of Giroux on open book decompositions, we get that an invariant of an open book decomposition which is invariant under positive stabilization is a contact invariant.

In the definition of the contact Ozsváth–Szabó invariant we will follow the reformulation found by Honda-Kazez-Matić [22]; for the original definition see [45]. Suppose therefore that (Y, ξ) is a given contact 3-manifold, and choose

an open book (F, φ) on Y compatible with ξ. Fix arcs $a_1, \dots, a_n$ properly embedded into the page F which provide a basis of $H_1(F, \partial F; \mathbb{Z})$. Let $b_1, \dots, b_n$ be displacements of the a_i in such a way that a_i intersects b_i in a single (positive) point. A Heegaard surface Σ for the 3-manifold Y can be given by taking the union of F with another page of the open book decomposition (with the reversed orientation). The images of the a_i under the identity (together with the a_i arcs) define a set of simple closed curves $\{\alpha_i\}$, while the same construction applied for the b_j and their images under the monodromy map φ provide $\{\beta_j\}$. It is easy to see that the resulting triple $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is a Heegaard diagram for Y, and if we place the basepoint into the region of F which is not between any a_i and b_i, the intersection point $\mathbf{x}$ contained by F defines an element $c(Y, \xi)$ in $\widehat{HF}(-Y)$.

Remark 3.1. For $\mathbf{x}$ to be a cycle we need to view it in the chain complex given by $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha})$ rather than by $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, explaining the fact that $c(Y, \xi)$ is an element of the Ozsváth–Szabó homology group of $-Y$.

Theorem 3.2.

- *The homology class $c(Y, \xi)$ is an invariant of the (isotopy class of the) contact structure ξ and is independent of the chosen compatible open book decomposition and basis $\{a_1, \dots, a_n\}$.*
- *For a contact 3-manifold (Y, ξ) the contact Ozsváth–Szabó invariant $c(Y, \xi)$ is an element of $\widehat{HF}_{-d_3(\xi)}(-Y, \mathfrak{t}_\xi)$, where $\mathfrak{t}_\xi$ is the spin^c structure induced by ξ (as a 2-plane field) and (if $c_1(\mathfrak{t}_\xi)$ is torsion) $d_3(\xi)$ is its Hopf invariant.*
- *If (Y, ξ) is Stein fillable then $c(Y, \xi) \neq 0$.*
- *It $Tor(Y, \xi) > 0$ then the contact invariant vanishes. In particular, if (Y, ξ) is overtwisted then $c(Y, \xi) = 0$.* □

The next property provides a way for computing the invariant for contact structures given by contact surgery diagrams.

Theorem 3.3. *Suppose that (Y_2, ξ_2) is given as contact $(+1)$-surgery along the Legendrian knot $L \subset (Y_1, \xi_1)$; let X denote the corresponding cobordism. Then*

$$F_{-X}(c(Y_1, \xi_1)) = c(Y_2, \xi_2). \qquad \square$$

Suppose that $(Y, \xi) = (Y_{\mathbb{L}}, \xi_{\mathbb{L}})$ with $\mathbb{L} = \mathbb{L}^+ \cup \mathbb{L}^-$ and $|\mathbb{L}^+| \leq 1$. Since $\mathbb{L}^-$ defines a Stein fillable contact structure, it has $c(Y_{\mathbb{L}^-}, \xi_{\mathbb{L}^-}) \neq 0$. The invariant of (Y, ξ) is given by $F_{-X}(c(Y_{\mathbb{L}^-}, \xi_{\mathbb{L}^-}))$, where X is the 4-dimensional cobordism induced by the single contact $(+1)$-surgery on $\mathbb{L}^+$. The nonvanishing of this invariant therefore depends on the relation between $c(Y_{\mathbb{L}^-}, \xi_{\mathbb{L}^-})$ and $\ker F_{-X} \leq \widehat{\mathrm{HF}}(-Y_{\mathbb{L}^-})$.

Legendrian and transverse invariants. The reformulation of knot Floer homology for knots in S^3 through grid diagrams provided (as a byproduct) Legendrian and transverse invariants for knots in the standard contact 3-sphere (S^3, ξ_{st}) [47]. The definition of these invariants motivated the extension of the invariant of contact structures to Legendrian and transverse knots [27], which we outline below. Suppose that (Y, ξ) is a contact 3-manifold, and (F, φ) is a compatible open book decomposition. Suppose that $\{a_i\}$ is a chosen basis (with $\{b_i\}$ being the pertured arcs of a_i in F). Recall that we have chosen the basepoint w in the domain of F which is not between the arcs a_i and b_i. Now choosing the other basepoint z from one of these regions we get a knot in Y, which can be shown to determine an oriented Legendrian knot type. In turn, any oriented Legendrian knot L in a contact 3-manifold (Y, ξ) can be presented in this way by choosing appropriate open book decomposition and basis $\{a_i\}$.

Suppose therefore that $L \subset (Y, \xi)$ is a given oriented Legendrian knot, and consider an open book decomposition (F, φ) and a basis $\{a_i\}$ compatible with it, together with the two basepoints w and z. The intersection point on the page F providing the contact invariant now defines an element $\widehat{\mathcal{L}}(L) \in \widehat{HFK}(-Y, L)$ in the knot Floer homology of L, which we call the *Legendrian invariant* of L.

Remark 3.4. Recall that in the definition of the contact invariant $c(Y, \xi)$ the role of the $\boldsymbol{\alpha}$- and the $\boldsymbol{\beta}$-curves has been reversed in order to get a cycle $\mathbf{x}$. For the Legendrian invariant $\widehat{\mathcal{L}}$ the same switch has to be performed. In order to restore the orientation of L, we also switch the two basepoints z and w. In fact, the definition extends to provide an invariant $\mathcal{L}(L)$, which is an element of a further version $\mathrm{HFK}^-(-Y, L)$ of Ozsváth–Szabó homologies.

Theorem 3.5.
- *The class $\widehat{\mathcal{L}}(L) \in \widehat{HFK}(-Y, L)$ for the oriented Legendrian knot $L \subset (Y, \xi)$ is an invariant of its (oriented) Legendrian isotopy class, in the sense that if L_1 and L_2 are (oriented) Legendrian isotopic then there is a map $f_* \colon \widehat{HFK}(-Y, L_1) \to \widehat{HFK}(-Y, L_2)$ induced by a map $f \colon Y - L_1 \to Y - L_2$ mapping $\widehat{\mathcal{L}}(L_1)$ into $\widehat{\mathcal{L}}(L_2)$.*
- *The invariant $\widehat{\mathcal{L}}(L)$ vanishes if $(Y - L, \xi|_{Y-L})$ has positive Giroux torsion.*
- *$\widehat{\mathcal{L}}(L^+) = 0$ for the positive stabilization L^+ of L and $\widehat{\mathcal{L}}(L) = \widehat{\mathcal{L}}(L^-)$ for the negative stabilization. Since the oriented Legendrian approximation of a transverse knot is unique up to negative stabilization and isotopy, this property allows us to define the invariant $\widehat{\mathcal{T}}(T)$ of a transverse knot $T \subset (Y, \xi)$ by taking $\widehat{\mathcal{T}}(T) = \widehat{\mathcal{L}}(L)$ for a Legendrian approximation of T.* □

The transformation of the Legendrian invariant under contact (+1)-surgery follows the same pattern as that of the contact Ozsváth–Szabó invariant:

Theorem 3.6 (Ozsváth-Stipsicz, [39]). *Supppose that $S, L \subset (Y, \xi)$ are (disjoint) Legendrian knots. Let (Y_S, ξ_S) denote the result of contact (+1)-surgery*

along S, and let L_S denotes the image of L in (Y_S, ξ_S). Then the surgery gives rise to a map $F\colon \widehat{\mathrm{HFK}}(-Y, L) \to \widehat{\mathrm{HFK}}(-Y_S, L_S)$ which maps the Legendrian invariant $\widehat{\mathcal{L}}(L)$ to $\widehat{\mathcal{L}}(L_S)$. □

4. Results

Surgery along knots in S^3. First we examine the problem of the existence of tight contact structures on 3-manifolds of the form $Y = S^3_r(K)$, i.e., Y can be given by a single Dehn surgery on S^3. (Here the surgery coefficient is measured with respect to the Seifert framing of $K \subset S^3$.) Let us recall that the *maximal Thurston–Bennequin number* $TB(K)$ of a knot $K \subset S^3$ is defined by

$$\max\{\mathrm{tb}(L) \mid L \text{ is smoothly isotopic to } K \text{ and Legendrian in } (S^3, \xi_{st})\}.$$

The *slice-genus* (or 4-ball genus) $g_s(K)$ of $K \subset S^3$ is by definition the minimum of the genera of surfaces smoothly embedded in D^4 with boundary equal to K, that is,

$$\min\{g(F) \mid F \subset D^4, \partial F = K \subset S^3\}.$$

Using gauge theory it has been proved that $TB(K) \leq 2g_s(K) - 1$.

Theorem 4.1 (Lisca-Stipsicz, [30]). *If $TB(K) = 2g_s(K) - 1 > 0$ is satisfied for a knot K then $S^3_r(K)$ admits a positive tight contact structure for any $r \neq TB(K)$.* □

Notice that if K is the (p, q) torus knot $T_{p,q}$ (for $p, q \geq 2$ and relative prime), then it has $TB(T_{p,q}) = pq - p - q$, which is equal to $2g_s(T_{p,q}) - 1$, hence those knots satisfy the assumptions of the theorem. For example, the right-handed trefoil knot $T = T_{2,3}$ depicted in Figure 2 satisfies the assumptions. We sketch

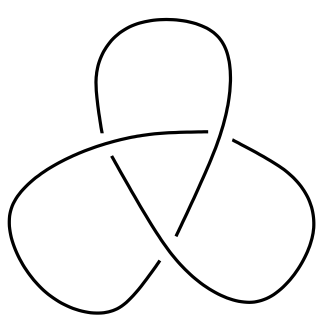


Figure 2. The right-handed trefoil knot.

the argument in the special case of $K = T_{2,3}$ and $r = TB(K) + 1 = 2$.

Proposition 4.2. *The contact structure given by the diagram of Figure 3 on $S^3_2(T_{2,3})$ is tight.*

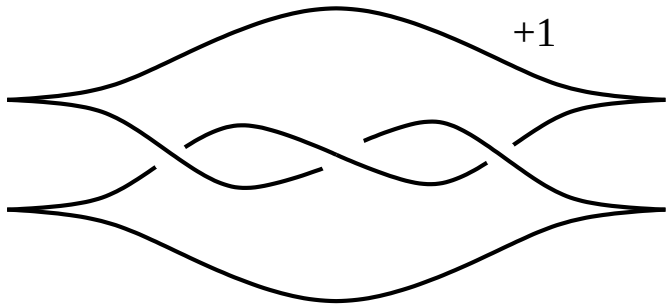


Figure 3. Contact structure on the manifold $Y = S^3_2(T_{(2,3)})$.

Proof. Let ξ denote the contact structure given by the contact surgery diagram of Figure 3. According to Theorem 3.3, the invariant $c(Y,\xi)$ is the image of the unique nontrivial element $c(S^3,\xi_{st}) \in \widehat{HF}(-S^3) \cong \mathbb{Z}_2$ under the map F induced by the cobordism given by the surgery. The surgery defines an exact triangle, where $\widehat{HF}(-S^3) \cong \mathbb{Z}_2$, $\widehat{HF}(-S^3_2(T_{2,3})) = \mathbb{Z}_2^2$ and the third term (which is equal to the invariant of the 3-manifold $-S^3_1(T_{2,3})$) is isomorphic to $\mathbb{Z}_2$. Exactness of the triangle then implies that F is injective, hence $c(S^3_2(T_{2,3}),\xi) = F(c(S^3,\xi_{st})) \in \widehat{\mathrm{HF}}(-S^3_2(T_{2,3}))$ is nontrivial. The nonvanishing of the class $c(S^3_2(T_{2,3}),\xi)$ implies then that ξ is tight. □

Proposition 4.3 (Lisca, [25]). *The 3-manifold $S^3_2(T)$ admits no fillable contact structure.*

Proof (sketch). The argument consists of two steps. In the first step, using gauge theory one shows that any hypothesized filling of any contact structure on $S^3_2(T)$ must be a 4-manifold with negative definite intersection form. In this step the property that $S^3_2(T)$ admits a positive scalar curvature (or more generally, it is an L-space) is used.

For the next step we observe that $-S^3_2(T)$ can be given as the boundary of the plumbing 4-manifold along the negative definite E_7 diagram. Therefore, the existence of a filling would provide (after gluing it to the above plumbing 4-manifold) a closed 4-manifold X with negative definite intersection form, which contains the negative definite E_7 lattice as a sublattice. By Donaldson's Theorem A the intersection form of X (as a closed 4-manifold with negative definite intersection form) is diagonalizable over the integers, while a simple computation shows that the negative definite E_7 lattice does not embed into any negative definite diagonal lattice, providing the desired contradiction with the existence of a filling, hence completing the proof. □

Corollary 4.4. *The contact structure ξ given by the surgery diagram of Figure 3 is a tight, nonfillable contact structure.* □

The first examples of tight, nonfillable structures were found by Etnyre-Honda [11], for more similar examples, see [28, 29].

Notice that Theorem 4.1 deals only with surgeries satisfying $r \neq TB(K)$. This assumption plays an important role in defining the candidate tight contact structure; for $r = TB(K)$ the previous strategy (requiring contact 0-surgery) would provide an overtwisted structure. (For a discussion on contact 0-surgery, see [5].) It seems to be a very subtle question to understand what happens on the 3-manifold $S^3_{TB(K)}(K)$. As it turns out, for the trefoil knot T the surgery coefficient $r = TB(T) = 1$ provides a 3-manifold $-P$, the Poincaré homology sphere with its reversed orientation, which by a theorem of Etnyre-Honda admits no tight contact structures [10]. If $T_{2,2n+1} \subset S^3$ denotes the $(2, 2n+1)$-torus knot and Y_n $(n \geq 1)$ stands for the 3-manifold obtained by performing $TB(T_{2,2n+1}) = (2n-1)$-surgery along $T_{2,2n+1}$, then we get a family of 3-manifolds with the same property: Y_n carries no tight contact structure [31]. (These 3-manifolds are all small Seifert fibered 3-manifolds; we will return to the discussion of contact structures on such 3-manifolds.) As it turns out, for all other torus knots $T_{p,q}$ $(p > q > 2)$ the 3-manifold $S^3_{TB(T_{p,q})}(T_{p,q})$ does admit tight contact structures, cf. Theorem 4.6.

Regarding the general question of existence of tight structures on 3-manifolds of the form $S^3_r(K)$, we restrict ourselves to a conjecture:

Conjecture 4.5. *(High surgery conjecture) Suppose that $K \subset S^3$ is a knot. Then there is an integer n_K such that for all $r \geq n_K$ the surgered 3-manifold $S^3_r(K)$ admits tight contact structure.*

For Seifert fibered 3-manifolds the existence question of tight contact structures is now settled.

Theorem 4.6 (Lisca-Stipsicz, [33]). *Let Y be a closed, oriented Seifert fibered 3-manifold. Then, either Y is orientation-preserving diffeomorphic to $Y_n = S^3_{2n-1}(T_{2,2n+1})$ for some $n \geq 1$, or Y carries a positive, tight contact structure.* □

Recall that by [31, Corollary 1.2] each 3-manifold Y_n carries no positive tight contact structure, hence Theorem 4.6 yields a complete solution to the existence problem for positive tight contact structures on Seifert fibered 3-manifolds.

According to the Milnor-Kneser Theorem any 3-manifold uniquely decomposes as the connected sum of *prime* 3-manifolds (i.e., of those for which a connected sume decomposition necessarily applies S^3 as one of the factors). Tight contact structures obey the same connected sum decomposition theorem, in particular, the set of tight structures on $Y_1 \# Y_2$ is simply the cartesian product of the corresponding sets on Y_1 and Y_2 [3]. In the light of the Geometrization Conjecture (stating that a closed, prime 3-manifold is either Seifert fibered,

or admits a hyperbolic metric, or contains a *essential* torus, that is, one for which the embedding induces an injective map on the fundamental groups), and the fact that the existence of an essential torus guarantees the existence of (infinitely many different) tight contact structures on a 3-manifold, as far as existence goes we are left with the study of hyperbolic 3-manifolds.

Contact structures on small Seifert fibered 3-manifolds. Above we showed a way to produce tight contact structures (and, in particular, verify tightness) in certain suitable cases using contact surgery and Heegaard Floer theory. It is much harder to find *all* the tight structures on a given 3-manifold. In general this question is still open, but for some families of 3-manifolds we have complete classification of tight structures. For lens spaces the problem was answered by Giroux and Honda. In this case all contact structures were Stein fillable, and could be distinguished by their homotopy theoretic invariants (namely the induced spin^c structure). The examples of the contact structures (and their tightness) were fairly straightforward for these 3-manifolds, the difficult part of the classification scheme was in showing that the particular manifolds do not carry any further (tight) contact structures. In this step convex surface theory (initiated by Giroux [17]) was applied. An extension of these methods by Honda [21] led to a complete classification of tight contact structures on torus bundles and circle bundles. By a further delicate application of the convex surface theory techniques, Massot [36] provided a complete classification of tight contact structures on Seifert fibered 3-manifolds with base of positive genus, with the additional assumption for the structures to have *negative maximal twisting*. Recall that a Legendrian knot in a Seifert fibered 3-manifold Y smoothly isotopic to a regular fiber admits two framings: one coming from the fibration and another one coming from the contact structure ξ. The difference between the contact framing and the fibration framing is the *twisting number* of the Legendrian curve. We say that the contact structure ξ on a Seifert fibered 3-manifold has *maximal twisting equal to zero* if there is a Legendrian knot L isotopic to a regular fiber such that L has twisting number zero. If there is no such Legendrian curve in (Y, ξ), then it has negative maximal twisting.

A Seifert fibered 3-manifold M is *small* if it fibers over S^2 with three singular fibers. Equivalently, a 3-manifold is small Seifert fibered, if it can be given by the surgery diagram of Figure 4. Let $M = M(e_0; r_1, r_2, r_3)$ denote the 3-manifold given by the surgery presentation of Figure 4. The classification of tight structures on M with the extra hypothesis $e_0 \neq -1, -2$ was given in [14, 50]. The main feature of the classification was the same as for lens spaces: all tight structures turned out be Stein fillable, and one could distinguish them by their induced spin^c structures.

The classification is slightly more complicated for the case $e_0 = -1$ and $r_1, r_2 \geq \frac{1}{2}$: in this case the 3-manifold M carries a number of nonfillable structures, and the proof of their tightness requires the application of the contact

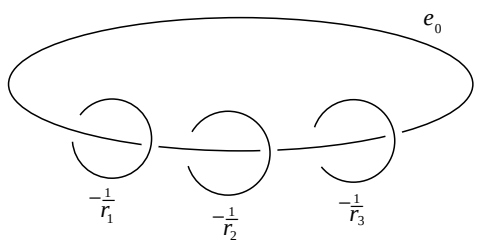


Figure 4. Surgery diagram for the Seifert fibered 3-manifold $M(e_0; r_1, r_2, r_3)$

Ozsváth–Szabó invariants [15]. All tight structures on these small Seifert fibred 3-manifolds admit a straightforward surgery presentation (shown by Figure 5), and this result leads us to the a conjectured classification of tight contact structures on small Seifert 3-manifolds with zero maximal twisting. Suppose that the small Seifert fibered 3-manifold M admits no transverse contact structures, which property, according to [26], can be conveniently phrased in terms of the Seifert invariants. By [32] this property implies that any tight contact structure ξ on M can be given by a surgery diagram of Figure 5. The contact structure ξ

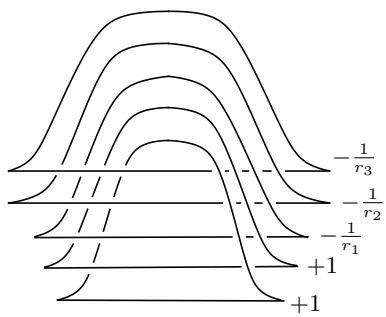


Figure 5. Contact structures on $M(-1; r_1, r_2, r_3)$.

defines a 2-plane field with Hopf invariant $d_3(\xi)$ and a spin^c structure $\mathfrak{t}_\xi$ which has Ozsváth–Szabó d-invariant $d(\xi)$.

Conjecture 4.7. *The contact structure ξ given by Figure 5 is tight if and only if $d_3(\xi) = d(\xi)$. Two contact structures ξ_1, ξ_2 given by Figure 5 are isotopic if and only if they induce isomorphic spin^c structures $\mathfrak{t}_{\xi_1} = \mathfrak{t}_{\xi_2}$.*

The resolution of this conjecture would reduce the classification problem of tight structures for many small Seifert fibered 3-manifold with $e_0 = -1$ to an algebraic computation, but the general classification question still remains open. Further classification results (also relying on the use of the contact Ozsváth–Szabó invariant in twisted Heegaard Floer theory) were given by Ghiggini-Van Horn-Morris in [16] for the Brieskorn spheres $-\Sigma(2, 3, 6n - 1)$ (with

reversed orientation), diffeomorphic to the small Seifert fibered 3-manifolds $M(-1; \frac{1}{2}, \frac{1}{3}, \frac{n}{6n-1})$.

Legendrian and transverse knots. The application of the invariant of [47] given by Ng-Ozsváth-Thurston [37] verified the existence of many transversely non-simple knot types in the standard contact 3-sphere (S^3, ξ_{st}). The connected sum formula of Vértesi [49] extended the applicability of these invariants even further. The similar invariant of [27], however, can also be applied for knots in other 3-manifolds, or in S^3 equipped with some other (necessarily overtwisted) contact structure. With the aid of explicite computations of the Legendrian invariant $\widehat{\mathcal{L}}$ for a number of examples, the application of a connected sum formula for these invariants gives the following:

Theorem 4.8 (Lisca-Ozsváth-Stipsicz-Szabó, [27]). *Let (Y, ξ) be a contact 3-manifold with $c(Y, \xi) \neq 0$. Let ζ be an overtwisted contact structure on Y with induced spin^c structure satisfying $\mathfrak{t}_\zeta = \mathfrak{t}_\xi$. Then, in (Y, ζ) there are null-homologous knot types which admit two non-loose, transversely non-isotopic transverse representatives with the same self-linking number.* □

Consequently, in many overtwisted contact 3-manifolds there are transversely non-simple knots. The relation of the Legendrian invariant to contact surgery (described in Theorem 3.6) provided further computational tools leading us to

Theorem 4.9 (Ozsváth-Stipsicz, [39]). *The Eliashberg–Chekanov twist knot E_n (which is the two-bridge knot of type $\frac{2n+1}{2}$) depicted by Figure 6 is not transversely simple for n odd and $n > 3$. In fact, for n odd there are at least $\lceil \frac{n}{4} \rceil$ transverse knots in the standard contact 3-sphere (S^3, ξ_{st}) with self-linking number equal to 1, all topologically isotopic to E_n, yet not pairwise transverse isotopic.* □

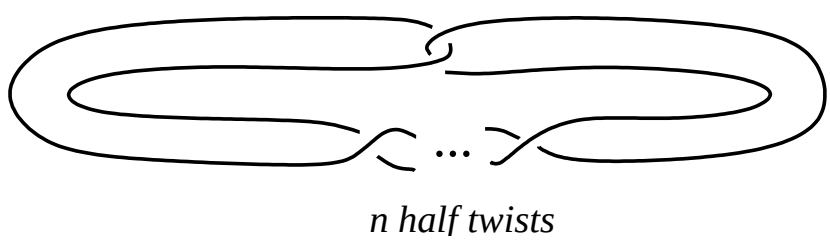


Figure 6. The Eliashberg–Chekanov knot E_n.

As a special case, we have the following:

Corollary 4.10. *The twist knot 7_2 in Rolfsen's table is not transversely simple.*

Proof. The knot 7_2 is the two-bridge knot $\frac{11}{2}$, hence Theorem 4.9 applies with $n = 5$. □

The above result, when coupled with convex surface theory and contact homology, led Etnyre-Ng-Vértesi [12] to a complete classification of Legendrian and transverse knots in the knot types given by Figure 6. In particular, it has been verified that (for n odd) the $\lceil \frac{n}{4} \rceil$ distinct transverse knots in the standard contact 3-sphere (S^3, ξ_{st}) with self-linking number equal to 1, all topologically isotopic to E_n, used in Theorem 4.9 comprise a complete list of transverse knots with the given classical invariants. We hope that similar methods can be applied to wider classes of knots, for example, to 2-bridge knots to settle the Legendrian/transverse classification problem.

References

[1] V. Colin, E. Giroux and K. Honda, *On the coarse classification of tight contact structures*, Topology and geometry of manifolds (Athens, GA, 2001), 109–120, Proc. Sympos. Pure Math., 71, Amer. Math. Soc., Providence, RI, 2003.

[2] F. Ding and H. Geiges, *A Legendrian surgery presentation of contact 3-manifolds*, Proc. Cambridge Philos. Soc. **136** (2004), 583–598.

[3] F. Ding and H. Geiges, *A unique decomposition theorem for tight contact 3-manifolds*, Enseign. Math. (2) **53** (2007), 333–345.

[4] F. Ding, H. Geiges and A. Stipsicz, *Surgery diagrams for contact 3-manifolds*, Turkish J. Math. **28** (2004), 41–74.

[5] F. Ding, H. Geiges and A. Stipsicz, *Lutz twist and contact surgery*, Asian J. Math. **9** (2005), 57–64.

[6] Y. Eliashberg, *Classification of overtwisted contact structures on 3-manifolds*, Invent. Math. **98** (1989), 623–637.

[7] Y. Eliashberg, *A few remarks about symplectic filling*, Geom. Topol. **8** (2004), 277–293.

[8] J. Epstein, D. Fuchs and M. Meyer, *Chekanov-Eliashberg invariants and transverse approximate of Legendrian knots*, Pacific J. Math. **201** (2001), 89-106.

[9] J. Etnyre, *Lectures on open book decompositions and contact structures*, Floer homology, gauge theory, and low-dimensional topology, 103–141, Clay Math. Proc., **5**, Amer. Math. Soc., Providence, RI, 2006.

[10] J. Etnyre and K. Honda, *On the non-existence of tight structures*, Ann. of Math. **153** (2001), 749–766.

[11] J. Etnyre and K. Honda, *Tight contact structures with no symplectic fillings*, Invent. Math. **148** (2002), 609–626.

[12] J. Etnyre, L. Ng and V. Vértesi, *Legendrian and transverse twist knots*, arXiv:1002.2400

[13] H. Geiges, *An introduction to contact topology*, Cambridge Studies in Advanced Mathematics, **109**. Cambridge University Press, Cambridge, 2008.

[14] P. Ghiggini, P. Lisca and A. Stipsicz, *Classification of tight contact structures on small Seifert 3-manifolds with* $e_0 \geq 0$, Proc. Amer. Math. Soc. **134** (2006), 909–916.

[15] P. Ghiggini, P. Lisca and A. Stipsicz, *Tight contact structures on some small Seifert fibered 3-manifolds*, Amer. J. Math. **129** (2007), 1403–1447.

[16] P. Ghiggini and J. Van Horn-Morris, *Tight contact structures on the Brieskorn spheres* $-\Sigma(2,3,6n-1)$ *and contact invariants*, arXiv:0910.2752

[17] E. Giroux, *Convexité en topologie de contact*, Comment. Math. Helv. **66** (1991), 637–677.

[18] E. Giroux, *Contact geometry: from dimension three to higher dimensions*, Proceedings of the International Congress of Mathematicians (Beijing 2002), 405–414.

[19] R. Gompf and A. Stipsicz, *4-manifolds and Kirby calculus*, Graduate Studies in Mathematics, vol. **20**, American Math. Society, Providence 1999.

[20] K. Honda, *On the classification of tight contact structures, I.*, Geom. Topol. **4** (2000), 309–368.

[21] K. Honda, *On the classification of tight contact structures, II.*, J. Differential Geom. **55** (2000), 83–143.

[22] K. Honda, W. Kazez and G. Matić, *On the contact class in Heegaard Floer, homology*, J. Diff. Geom. **83** (2009), 289–311.

[23] P. Kronheimer and T. Mrowka, *Witten's conjecture and Property P*, Geom. Topol. **8** (2004), 295–310.

[24] P. Kronheimer, T. Mrowka, P. Ozsváth and Z. Szabó, *Monopoles and lens space surgeries*, Ann. of Math. (2) **165** (2007), 457–546.

[25] P. Lisca, *Symplectic fillings and positive scalar curvature*, Geom. Topol. **2** (1998), 103–116.

[26] P. Lisca and G. Matić, *Transverse contact structures on Seifert 3-manifolds*, Algebr. Geom. Topol. **4** (2004), 1125–1144.

[27] P. Lisca, P. Ozsváth, A. Stipsicz and Z. Szabó, *Heegaard Floer invariants of Legendrian knots in contact three-manifolds*, Journal of the EMS **11** (2009), 1307–1363.

[28] P. Lisca and A. Stipsicz, *An infinite family of tight, not semi-fillable contact three-manifolds*, Geom. Topol. **7** (2003), 1055–1073.

[29] P. Lisca and A. Stipsicz, *Tight, not semi-fillable contact circle bundles*, Math. Ann. **328** (2004), 285–298.

[30] P. Lisca and A. Stipsicz, *Ozsváth–Szabó invariants and tight contact three-manifolds I.*, Geom. Topol. **8** (2004), 925–945.

[31] P. Lisca and A. Stipsicz, *Ozsváth–Szabó invariants and tight contact three-manifolds II.*, J. Differential Geom. **75** (2007), 109–141.

[32] P. Lisca and A. Stipsicz, *Ozsváth-Szabó invariants and tight contact 3-manifolds III.*, J. Symplectic Geom. **5** (2007), 357–384.

[33] P. Lisca and A. Stipsicz, *On the existence of tight contact structures on Seifert fibered 3-manifolds*, Duke Math. J. **148** (2009), 175–209.

[34] C. Manolescu, P. Ozsváth and S. Sarkar, *A combinatorial description of knot Floer homology*, Ann. of Math. **169** (2009), 633–660.

[35] C. Manolescu, P. Ozsváth, Z. Szabó and D. Thurston, *On combinatorial link Floer homology*, Geom. Topol. **11** (2007), 2339–2412.

[36] P. Massot, *Geodesible contact structures on 3-manifolds*, Geometry and Topology **12** (2008), 1729–1776.

[37] L. Ng, P. Ozsváth and D. Thurston, *Transverse Knots Distinguished by Knot Floer Homology*, J. Symplectic Geom. **6** (2008), 461–490.

[38] B. Ozbagci and A. Stipsicz, *Surgery on contact 3-manifolds and Stein surfaces*, Bolyai Society Mathematical Studies, **13** Springer-Verlag, Berlin; János Bolyai Mathematical Society, Budapest, 2004.

[39] P. Ozsváth and A. Stipsicz, *Contact surgeries and the transverse invariant in knot Floer homology*, Journal of the Institute of Mathematics of Jussiue, to appear, arXiv:0803.1252.

[40] P. Ozsváth, A. Stipsicz and Z. Szabó, *Combinatorial Heegaard Floer homology and nice Heegaard diagrams*, arXiv:0912.0830.

[41] P. Ozsváth and Z. Szabó, *Holomorphic disks and topological invariants for closed three–manifolds*, Ann. of Math. **159** (2004), 1027–1158.

[42] P. Ozsváth and Z. Szabó, *Holomorphic disks and three-manifold invariants: properties and applications*, Ann. of Math. **159** (2004), 1159–1245.

[43] P. Ozsváth and Z. Szabó, *Holomorphic disks and knot invariants*, Adv. Math. **186** (2004), 58–116.

[44] P. Ozsváth and Z. Szabó, *Holomorphic disks and genus bounds*, Geom. Topol. **8** (2004), 311–334.

[45] P. Ozsváth and Z. Szabó, *Heegaard Floer homologies and contact structures*, Duke Math. J. **129** (2005), 39–61.

[46] P. Ozsváth and Z. Szabó, *Lectures on Heegaard Floer homology*, Floer homology, gauge theory, and low-dimensional topology, 29–70, Clay Math. Proc., **5**, Amer. Math. Soc., Providence, RI, 2006.

[47] P. Ozsváth, Z Szabó and D. Thurston, *Legendrian knots, transverse knots and combinatorial Floer homology*, Geom. Topol. **12** (2008), 941–980.

[48] S. Sarkar and J. Wang, *An algorithm for computing some Heegaard Floer homologies*, Ann. of Math., to appear, arXiv:math/0607777.

[49] V. Vértesi, *Transversely nonsimple knots*, Algebr. Geom. Topol. **8** (2008), 1481–1498.

[50] H. Wu, *Legendrian vertical circles in small Seifert spaces*, Commun. Contemp. Math. **8** (2006), 219–246.

Author Index*

(Volumes II, III, and IV)

Adler, Jill, **IV** 3213
Anantharaman, Nalini, **III** 1839
Arnaud, Marie-Claude, **III** 1653
Auroux, Denis, **II** 917

Baake, Ellen, **IV** 3037
Balmer, Paul, **II** 85
Belkale, Prakash, **II** 405
Benjamini, Itai, **IV** 2177
Benson, David J., **II** 113
Bernard, Patrick, **III** 1680
Billera, Louis J., **IV** 2389
Borodin, Alexei, **IV** 2188
Bose, Arup, **IV** 2203
Breuil, Christophe, **II** 203
Brydges, David, **IV** 2232
Buff, Xavier, **III** 1701
Bürgisser, Peter, **IV** 2609
Burq, Nicolas, **III** 1862

Chen, Shuxing, **III** 1884
Cheng, Chong-Qing, **III** 1714
Chéritat, Arnaud, **III** 1701
Cockburn, Bernardo, **IV** 2749
Cohn, Henry, **IV** 2416
Contreras, Gonzalo, **III** 1729
Costello, Kevin, **II** 942
Csörnyei, Marianna, **III** 1379

Dancer, E. N., **III** 1901
De Lellis, Camillo, **III** 1910
del Pino, Manuel, **III** 1934
Delbaen, Freddy, **IV** 3054
den Hollander, Frank, **IV** 2258
Dencker, Nils, **III** 1958
Dwork, Cynthia, **IV** 2634

Einsiedler, Manfred, **III** 1740
Erschler, Anna, **II** 681
Eskin, Alex, **III** 1185
Evans, Steven N., **IV** 2275

Fernández, Isabel, **II** 830
Fomin, Sergey, **II** 125
Frankowska, Hélène, **IV** 2915
Fu, Jixiang, **II** 705
Fusco, Nicola, **III** 1985

Gabai, David, **II** 960
Gaboriau, Damien, **III** 1501
Goldman, William M., **II** 717
Gordon, Iain G., **III** 1209
Greenberg, Ralph, **II** 231
Grodal, Jesper, **II** 973
Guruswami, Venkatesan, **IV** 2648
Guth, Larry, **II** 745

Hacon, Christopher D., **II** 427, 513
Hamenstädt, Ursula, **II** 1002
Heath-Brown, D.R., **II** 249
Hertz, Federico Rodriguez, **III** 1760
Hutchings, Michael, **II** 1022
Huybrechts, Daniel, **II** 450

Its, Alexander R., **III** 1395
Ivanov, Sergei, **II** 769
Iwata, Satoru, **IV** 2943
Izumi, Masaki, **III** 1528

*Names of invited speakers only are shown in the Index.

Kaledin, D., **II** 461
Kapustin, Anton, **III** 2021
Karpenko, Nikita A., **II** 146
Kedlaya, Kiran Sridhara, **II** 258
Khare, Chandrashekhar, **II** 280
Khot, Subhash, **IV** 2676
Kisin, Mark, **II** 294
Kjeldsen, Tinne Hoff, **IV** 3233
Koskela, Pekka, **III** 1411
Kuijlaars, Arno B.J., **III** 1417
Kumar, Shrawan, **III** 1226
Kunisch, Karl, **IV** 3061
Kupiainen, Antti, **III** 2044

Lackenby, Marc, **II** 1042
Lando, Sergei K., **IV** 2444
Lapid, Erez M., **III** 1262
Leclerc, Bernard, **IV** 2471
Liu, Chiu-Chu Melissa, **II** 497
Losev, Ivan, **III** 1281
Lück, Wolfgang, **II** 1071
Lurie, Jacob, **II** 1099

Ma, Xiaonan, **II** 785
Maini, Philip K., **IV** 3091
Marcolli, Matilde, **III** 2057
Markowich, Peter A., **IV** 2776
Marques, Fernando Codá, **II** 811
Martin, Gaven J., **III** 1433
Mastropietro, Vieri, **III** 2078
McKay, Brendan D., **IV** 2489
McKernan, James, **II** 427, 513
Mira, Pablo, **II** 830
Mirzakhani, Maryam, **II** 1126
Moore, Justin Tatch, **II** 3
Morel, Sophie, **II** 312

Nabutovsky, Alexander, **II** 862
Nadirashvili, Nikolai, **III** 2001
Naor, Assaf, **III** 1549
Nazarov, Fedor, **III** 1450
Nešetřil, J., **IV** 2502
Nesterov, Yurii, **IV** 2964
Neuhauser, Claudia, **IV** 2297
Nies, André, **II** 30
Nochetto, Ricardo H., **IV**, 2805

Oh, Hee, **III** 1308

Pacard, Frank, **II** 882
Park, Jongil, **II** 1146
Păun, Mihai, **II** 540
Peterzil, Ya'acov, **II** 58

Quastel, Jeremy, **IV** 2310

Rains, Eric M., **IV** 2530
Reichstein, Zinovy, **II** 162
Riordan, Oliver, **IV** 2555
Rudelson, Mark, **III** 1576

Saito, Shuji, **II** 558
Saito, Takeshi, **II** 335
Sarig, Omri M., **III** 1777
Schappacher, Norbert, **IV** 3258
Schreyer, Frank-Olaf, **II** 586
Schütte, Christof, **IV** 3105
Seregin, Gregory A., **III** 2105
Shah, Nimish A., **III** 1332
Shao, Qi-Man, **IV** 2325
Shapiro, Alexander, **IV** 2979
Shen, Zuowei, **IV** 2834
Shlyakhtenko, Dimitri, **III** 1603
Sodin, Mikhail, **III** 1450
Soundararajan, K., **II** 357
Spielman, Daniel A., **IV** 2698
Spohn, Herbert, **III** 2128
Srinivas, Vasudevan, **II** 603
Starchenko, Sergei, **II** 58
Stipsicz, András I., **II** 1159
Stroppel, Catharina, **III** 1344
Sudakov, Benny, **IV** 2579
Suresh, V., **II** 189

Thomas, Richard P., **II** 624
Toro, Tatiana, **III** 1485
Touzi, Nizar, **IV** 3132
Turaev, Dmitry, **III** 1804

Vadhan, Salil, **IV** 2723
Vaes, Stefaan, **III** 1624

van de Geer, Sara, **IV** 2351
van der Vaart, Aad, **IV** 2370
Venkataramana, T. N., **III** 1366
Venkatesh, Akshay, **II** 383
Vershynin, Roman, **III** 1576

Weismantel, Robert, **IV** 2996
Welschinger, Jean-Yves, **II** 652
Wendland, Katrin, **III** 2144
Wheeler, Mary F., **IV** 2864
Wilkinson, Amie, **III** 1816
Wintenberger, Jean-Pierre, **II** 280

Xu, Jinchao, **IV** 2886
Xu, Zongben, **IV** 3151

Yamaguchi, Takao, **II** 899

Zhang, Xu, **IV** 3008
Zhou, Xun Yu, **IV** 3185