Progress in Mathematics
Volume 235

Series Editors
Hyman Bass
Joseph Oesterlé
Alan Weinstein

Geometric Methods in Algebra and Number Theory

Fedor Bogomolov
Yuri Tschinkel
Editors

Birkhäuser
Boston • Basel • Berlin

Fedor Bogomolov
New York University
Department of Mathematics
Courant Institute of Mathematical Sciences
New York, NY 10012
U.S.A.

Yuri Tschinkel
Princeton University
Department of Mathematics
Princeton, NJ 08544
U.S.A.

AMS Subject Classifications: 11G18, 11G35, 11G50, 11F85, 14G05, 14G20, 14G35, 14G40, 14L30, 14M15, 14M17, 20G05, 20G35

Library of Congress Cataloging-in-Publication Data
Geometric methods in algebra and number theory / Fedor Bogomolov, Yuri Tschinkel, editors.
p. cm. – (Progress in mathematics ; v. 235)
Includes bibliographical references.
ISBN 0-8176-4349-4 (acid-free paper)
1. Algebra. 2. Geometry, Algebraic. 3. Number theory. I. Bogomolov, Fedor, 1946- II. Tschinkel, Yuri. III. Progress in mathematics (Boston, Mass.); v. 235.

QA155.G47 2004
512–dc22 2004059470

ISBN 0-8176-4349-4 Printed on acid-free paper.

©2005 Birkhäuser Boston



All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media Inc., Rights and Permissions, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America. (HP)

9 8 7 6 5 4 3 2 1 SPIN 10987025

www.birkhauser.com

Preface

The transparency and power of geometric constructions has been a source of inspiration for generations of mathematicians. Their applications to problems in algebra and number theory go back to Diophantus, if not earlier. Naturally, the Greek techniques of intersecting lines and conics have given way to much more sophisticated and subtle constructions. What remains unchallenged is the beauty and persuasion of pictures, communicated in words or drawings.

This volume contains a selection of articles exploring geometric approaches to problems in algebra, algebraic geometry and number theory. All papers are strongly influenced by geometric ideas and intuition. Several papers focus on algebraic curves: the themes range from the study of unramified curve covers (Bogomolov–Tschinkel), Jacobians of curves (Zarhin), moduli spaces of curves (Hassett) to modern problems inspired by physics (Hausel). The paper by Bogomolov–Tschinkel explores certain special aspects of the geometry of curves over number fields: there exist many more nontrivial correspondences between such curves than between curves defined over larger fields. Zarhin studies the structure of Jacobians of cyclic covers of the projective line and provides an effective criterion for this Jacobian to be sufficiently generic. Hassett applies the logarithmic minimal model program to moduli spaces of curves and describes it in complete detail in genus two. Hausel studies Hodge-type polynomials for mixed Hodge structure on moduli spaces of representations of the fundamental group of a complex projective curve into a reductive algebraic group. Explicit formulas are obtained by counting points over finite fields on these moduli spaces. Two contributions deal with surfaces: applying the structure theory of finite groups to the construction of interesting surfaces (Bauer–Catanese–Grunewald), and developing a conjecture about rational points of bounded height on cubic surfaces (Swinnerton-Dyer). Representation-theoretic and combinatorial aspects of higher-dimensional geometry are discussed in the papers by de Concini–Procesi and Tamvakis. The papers by Chai and Pink report on current active research exploring special points and special loci on Shimura varieties. Budur studies invariants

of higher-dimensional singular varieties. Spitzweck considers families of motives and describes an analog of limit mixed Hodge structures in the motivic setup. Cluckers–Loeser continue their foundational work on motivic integration. One of the immediate applications is the reduction of a central problem from the theory of automorphic forms (the Fundamental Lemma) from p-adic fields to function fields of positive characteristic, for large p. A different reduction to function fields of positive characteristic is shown in the paper by Ellenberg–Venkatesh: they find a geometric interpretation, via Hurwitz schemes, of Malle's conjectures about the asymptotic of number fields of bounded discriminant and fixed Galois group and establish several upper bounds in this direction. Finally, Pineiro–Szpiro–Tucker relate algebraic dynamical systems on $\mathbb{P}^1$ to Arakelov theory on an arithmetic surface. They define heights associated to such dynamical systems and formulate an equidistribution conjecture in this context.

The authors have been charged with the task of making the ideas and constructions in their papers accessible to a broad audience, by placing their results into a wider mathematical context. The collection as a whole offers a representative sample of modern problems in algebraic and arithmetic geometry. It can serve as an intense introduction for graduate students and others wishing to pursue research in these areas.

Most results discussed in this volume have been presented at the conference "Geometric methods in algebra and number theory" in Miami, December 2003. We thank the Department of Mathematics at the University of Miami for help in organizing this conference.

New York,
August 2004

Fedor Bogomolov
Yuri Tschinkel

Contents

Beauville surfaces without real structures
Ingrid Bauer, Fabrizio Catanese, Fritz Grunewald 1

Couniformization of curves over number fields
Fedor Bogomolov, Yuri Tschinkel 43

On the V-filtration of $\mathcal{D}$-modules
Nero Budur 59

Hecke orbits on Siegel modular varieties
Ching-Li Chai 71

Ax–Kochen–Eršov Theorems for p-adic integrals and motivic integration
Raf Cluckers, François Loeser 109

Nested sets and Jeffrey–Kirwan residues
Corrado De Concini, Claudio Procesi 139

Counting extensions of function fields with bounded discriminant and specified Galois group
Jordan S. Ellenberg, Akshay Venkatesh 151

Classical and minimal models of the moduli space of curves of genus two
Brendan Hassett 169

Mirror symmetry and Langlands duality in the non-Abelian Hodge theory of a curve
Tamás Hausel 193

Mahler measure for dynamical systems on $\mathbb{P}^1$ and intersection theory on a singular arithmetic surface
Jorge Pineiro, Lucien Szpiro, Thomas J. Tucker 219

A Combination of the Conjectures of Mordell–Lang and André–Oort
Richard Pink 251

Motivic approach to limit sheaves
Markus Spitzweck 283

Counting points on cubic surfaces, II
Sir Peter Swinnerton-Dyer 303

Quantum cohomology of isotropic Grassmannians
Harry Tamvakis 311

Endomorphism algebras of superelliptic jacobians
Yuri G. Zarhin 339

Beauville surfaces without real structures

Ingrid Bauer[1], Fabrizio Catanese[1], and Fritz Grunewald[2]

[1] Department of Mathematics, University of Bayreuth, D-95440 Bayreuth, Germany
Ingrid.Bauer@uni-bayreuth.de
Fabrizio.Catanese@uni-bayreuth.de
[2] Mathematisches Institut, Universitätsstrasse 1, D-40225, Germany
grunewald@math.uni-duesseldorf.de

Summary. Inspired by a construction by Arnaud Beauville of a surface of general type with $K^2 = 8,\ p_g = 0$, the second author defined Beauville surfaces as the surfaces which are rigid, i.e., without nontrivial deformations, and which admit an unramified covering which is isomorphic to a product of curves of genus at least 2.

In this case the moduli space of surfaces homeomorphic to the given surface consists either of a unique real point, or of a pair of complex conjugate points corresponding to complex conjugate surfaces. It may also happen that a Beauville surface is biholomorphic to its complex conjugate surface, while failing to admit a real structure.

The first aim of this note is to provide series of concrete examples of the second situation, respectively of the third.

The second aim is to introduce a wider audience, in particular group theorists, to the problem of classification of such surfaces, especially with regard to the problem of existence of real structures on them.

1 Introduction

In [2] (see p. 159) A. Beauville constructed a new surface of general type with $K^2 = 8,\ p_g = 0$ as a quotient of the product of two Fermat curves of degree 5 by the action of the group $(\mathbb{Z}/5\mathbb{Z})^2$. Inspired by this construction, in the article [4], dedicated to the geometrical properties of varieties which admit an unramified covering biholomorphic to a product of curves, the following definition was given

Definition 1.1. *A Beauville surface is a compact complex surface S which*

1) is rigid , i.e., it has no nontrivial deformation,
2) is isogenous to a higher product, i.e., it admits an unramified covering which is isomorphic (i.e., biholomorphic) to a product of two curves C_1, C_2 of genera ≥ 2.

It was proven in [4] (cf. also [5]) that any surface S isogenous to a higher product has a unique minimal realization as a quotient $S = (C_1 \times C_2)/G$, where G is a finite group acting freely and with the property that no element acts trivially on one of the factors C_i. Moreover, any other smooth surface X with the same topological Euler number as S and with isomorphic fundamental group is diffeomorphic to S; and either X or its conjugate surface $\bar{X}$ belongs to an irreducible family of surfaces containing S as an element.

Therefore, if S is a Beauville surface, either X is isomorphic to S, or X is isomorphic to $\bar{S}$.

In order to reduce the description of Beauville surfaces to some group-theoretic statement, we need to recall that surfaces isogenous to a higher product belong to two types:

- **unmixed type**: the action of G does not mix the two factors, i.e., it is the product action of respective actions of G on C_1, resp. C_2. We set $G^0 := G$.
- **mixed type**: C_1 is isomorphic to C_2, and the subgroup $G^0 \subset G$ of transformations which do not mix the factors has index precisely 2 in G.

It is obvious from the above definition that every Beauville surface of mixed type has an unramified double covering which is a Beauville surface of unmixed type.

The rigidity property of the Beauville surface is equivalent to the fact that $C_i/G \cong \mathbb{P}^1$ and that the projection $C_i \to C_i/G \cong \mathbb{P}^1$ is branched in three points. Therefore the datum of a Beauville surface of unmixed type is determined, once we look at the monodromy of each covering of $\mathbb{P}^1$, by the datum of a finite group $G = G^0$ together with two respective systems of generators, (a, c) and (a', c'), which satisfy a further property (*), ensuring that the product action of G on $C \times C'$ is free, where $C := C_1$, $C' := C_2$ are the corresponding curves with an action of G associated to the monodromies determined by (a, c), resp. (a', c').

Define b, b' by the properties $abc = a'b'c' = 1$, let Σ be the union of the conjugates of the cyclic subgroups generated by a, b, c respectively, and define Σ' analogously: then property (*) is the following:

$$(*)\ \Sigma \cap \Sigma' = \{1_G\}.$$

In the mixed case, one requires instead that the two systems of generators be related by an automorphism ϕ of G^0 which should satisfy the further conditions:

- ϕ^2 is an inner automorphism, i.e., there is an element $\tau \in G^0$ such that $\phi^2 = \mathrm{Int}_\tau$,
- $\Sigma \cap \phi(\Sigma) = \{1_{G^0}\}$,
- there is no $g \in G^0$ such that $\phi(g)\tau g \in \Sigma$,
- moreover $\phi(\tau) = \tau$ and indeed the elements in the trivial coset of G^0 are transformations of $C \times C$ of the form

$$g(x,y) = (g(x), \phi(g)(y))$$

while transformations in the nontrivial coset are transformations of the form

$$\tau' g(x,y) = (\phi(g)(y), \tau g(x)).$$

Remark 1.2. The choice of τ is not unique, we can always replace τ by $\phi(g)\tau g$, where $g \in G^0$ is arbitrary, and accordingly replace ϕ by $\phi \circ \mathrm{Int}_g$.

Observe that if G^0 has only inner automorphisms, there can certainly be no Beauville surface of mixed type, since the second of the above properties will be violated.

In this paper we use the definition of Beauville surfaces of unmixed and mixed type to formulate group-theoretic conditions which will allow us to treat the following problems:

1. The biholomorphism problem for Beauville surfaces

For every finite group G we introduce sets of structures $\mathbb{U}(G)$ and $\mathbb{M}(G)$ and groups $\mathrm{A}_{\mathbb{U}}(G)$, $\mathrm{A}_{\mathbb{M}}(G)$ acting on them. We call $\mathbb{U}(G)$ the set of unmixed Beauville structures and $\mathbb{M}(G)$ the set of mixed Beauville structures on G. Using constructions from [4] and [5] we associate an unmixed Beauville surface $S(v)$ to every $v \in \mathbb{U}(G)$ and a mixed Beauville surface $S(u)$ to every $u \in \mathbb{M}(G)$. The minimal Galois representation of every Beauville surface yields a surface $S(v)$ in the unmixed case, respectively a surface $S(u)$ in the mixed case. We show that $S(v)$ is biholomorphic to $S(v')$ $(v, v' \in \mathbb{U}(G))$ if and only if v lies in the $\mathrm{A}_{\mathbb{U}}(G)$-orbit of v'. An analogous result holds in the mixed case.

2. Existence and classification problem for Beauville surfaces

The existence problem asks for finite groups G such that $\mathbb{U}(G)$ or $\mathbb{M}(G)$ is not empty. Previously, only abelian groups G were known with $\mathbb{U}(G) \neq \emptyset$. Here we give many other examples. In the mixed case it is not immediately clear that the requirements for the corresponding structures can be met, and no examples were known. We give a group-theoretic construction which produces finite groups G with $\mathbb{M}(G) \neq \emptyset$.

The classification problem has two meanings. First of all, we wish to find all G with $\mathbb{U}(G) \neq \emptyset$ or $\mathbb{M}(G) \neq \emptyset$. In [5] all finite abelian groups G are found with $\mathbb{U}(G) \neq \emptyset$ (we give a proof of this fact in Section 3, Theorem 3.4). One of our results is that a group G with $\mathbb{U}(G) \neq \emptyset$ cannot be a nontrivial quotient of one of the nonhyperbolic triangle groups.

Our examples show that the classification problem may be hopeless. In fact, in Section 3 (cf. (3.10)) we prove that every finite group G of exponent n with $\gcd(n,6) = 1$, which is generated by two elements and which has

$\mathbb{Z}/n\mathbb{Z}\times\mathbb{Z}/n\mathbb{Z}$ as abelianization has $\mathbb{U}(G)\neq\emptyset$. Even if, by Zelmanov's solution of the restricted Burnside problem, there is for every n a maximal such group, the number of groups involved is very large. We also show

Theorem 1.3. *Let G be one of the groups $\mathbf{SL}(2,\mathbb{F}_p)$ or $\mathbf{PSL}(2,\mathbb{F}_p)$ where $\mathbb{F}_p$ is the prime field with p elements and p is distinct from $2,3,5$. Then there is an unmixed Beauville surface with group G.*

A finer classification entails the determination of all orbits of Beauville structures for a fixed group G or for an interesting series of groups. We do not address this problem here.

3. Is S biholomorphic to $\bar{S}$?

We give examples of Beauville surfaces S such that the complex conjugate surface $\bar{S}$ is not biholomorphic to S. Note that $\bar{S}$ is the same differentiable manifold as S, but with complex structure $-J$ instead of J. To do this we introduce involutions $\iota:\mathbb{U}(G)\to\mathbb{U}(G)$ and $\iota:\mathbb{M}(G)\to\mathbb{M}(G)$. We prove that $S(v)$ $(v\in\mathbb{U}(G))$ is biholomorphic to $\overline{S(v)}$ if and only if v is in the $\mathrm{A}_{\mathbb{U}}(G)$ orbit of $\iota(v)$. We also show the analogous result in the mixed case. We use this to produce the following explicit example:

Theorem 1.4. *Let G be the symmetric group $\mathcal{S}_n$ in $n\geq 8$ letters, let $S(n)$ be the unmixed Beauville surface corresponding to the choice of*

$$a:=(5,4,1)(2,6),\ c:=(1,2,3)(4,5,\dots,n),\ \text{ and of }\ a':=\sigma^{-1},\ c':=\tau\sigma^2,$$

where $\tau:=(1,2)$ and $\sigma:=(1,2,\dots,n)$.

Then $S(n)$ is not biholomorphic to $\overline{S(n)}$ provided that $n\equiv 2 \bmod 3$.

We now give the construction of a mixed Beauville surface with the same property. We first describe the group G and its subgroup G^0.

Let H be a nontrivial group and $\Theta: H\times H\to H\times H$ the automorphism defined by $\Theta(g,h):=(h,g)$ $(g,h\in H)$. Consider the semidirect product

$$G:=H_{[4]}:=(H\times H)\rtimes\mathbb{Z}/4\mathbb{Z}, \tag{1}$$

where the generator 1 of $\mathbb{Z}/4\mathbb{Z}$ acts through Θ on $H\times H$. Since Θ^2 is the identity we find

$$G^0:=H_{[2]}:=H\times H\times 2\mathbb{Z}/4\mathbb{Z}\cong H\times H\times\mathbb{Z}/2\mathbb{Z} \tag{2}$$

as a subgroup of index 2 in $H_{[4]}$.

Theorem 1.5. *Let p be a prime with $p\equiv 3$ mod 4 and $p\equiv 1$ mod 5 and $H:=\mathbf{SL}(2,\mathbb{F}_p)$. Let S be the mixed Beauville surface corresponding to the data $G:=H_{[4]}$, $G^0:=H_{[2]}$ and to a certain system of generators (a,c) of $H_{[2]}$ with $\mathrm{ord}(a)=20$, $\mathrm{ord}(c)=30$, $\mathrm{ord}(a^{-1}c^{-1})=5p$.*

Then S is not biholomorphic to $\overline{S}$.

Different examples of rigid surfaces S not isomorphic to $\bar{S}$ have been constructed in [12], using Hirzebruch type examples of ball quotients.

4. Is S real?

A surface S is called real if there exists a biholomorphism $\sigma : S \to \bar{S}$ such that $\sigma^2 = \mathrm{Id}$. In this case we say that S has a real structure. We translate this problem into group theory and obtain the following examples.

Theorem 1.6. *Let $p > 5$ be a prime with $p \equiv 1$ mod 4, $p \not\equiv 2, 4$ mod 5, $p \not\equiv 5$ mod 13 and $p \not\equiv 4$ mod 11. Set $n := 3p + 1$. Then there is an unmixed Beauville surface S with group $\mathcal{A}_n$ which is biholomorphic to the complex conjugate surface $\bar{S}$ but is not real.*

Further examples of real and nonreal Beauville surfaces will be given in the sequel to this paper.

Acknowledgments. We thank Benjamin Klopsch for help with alternating groups.

2 Triangular curves and group actions

In this section we recall the construction of triangular curves as given in [4], [5]. They are the building blocks for Beauville surfaces of both unmixed and mixed type. We add some group-theoretic observations which will help with the classification problems of Beauville surfaces mentioned above and which will be studied later.

We need the following group-theoretic notation. Let G be a group and M, N two sets equipped with a left-action of G. We call a map $\sigma : M \to N$ *G-twisted-equivariant* if there is an automorphism $\psi : G \to G$ of G with

$$\sigma(gP) = \psi(g)\sigma(P) \quad \text{for all } g \in G, P \in M. \tag{3}$$

Let G be a finite group and (a, c) a pair of elements of G. Define

$$\Sigma(a, c) := \bigcup_{g \in G} \bigcup_{i=0}^{\infty} \{ga^i g^{-1}, gc^i g^{-1}, g(ac)^i g^{-1}\} \tag{4}$$

to be the union of the G-conjugates of the cyclic groups generated by a, c and ac. Set

$$\mu(a, c) := \frac{1}{\mathrm{ord}(a)} + \frac{1}{\mathrm{ord}(c)} + \frac{1}{\mathrm{ord}(ac)}, \tag{5}$$

where $\mathrm{ord}(a)$ stands for the order of the element $a \in G$. Furthermore, call

$$(\mathrm{ord}(a), \mathrm{ord}(c), \mathrm{ord}(ac)) \tag{6}$$

the *type* of the pair (a, c) and define

$$\nu(a, c) := \mathrm{ord}(a)\mathrm{ord}(c)\mathrm{ord}(ac). \tag{7}$$

We consider here finite groups G having a pair (a, c) of generators. Setting $(r, s, t) := (\mathrm{ord}(a), \mathrm{ord}(c), \mathrm{ord}(ac))$, such a group is a quotient of the triangle group

$$T(r, s, t) := \langle x, y \mid x^r = y^s = (xy)^t = 1 \rangle. \tag{8}$$

We define

$$\mathbb{T}(G) := \{(a, c) \in G \times G \mid \langle a, c \rangle = G\}. \tag{9}$$

For $(a, c) \in \mathbb{T}(G)$ we consider its *triangular triple* $(a, b, c) := (a, a^{-1}c^{-1}, c)$. Clearly, the automorphism group $\mathrm{Aut}(G)$ of G acts diagonally on $\mathbb{T}(G)$. If $\mathbb{T}(G) \neq \emptyset$, then this action is faithful. We define additionally the following permutations of $\mathbb{T}(G)$:

$$\sigma_0 : (a, c) \mapsto (a, c),\ \sigma_1 : (a, c) \mapsto (a^{-1}c^{-1}, a),\ \sigma_2 : (a, c) \mapsto (c, a^{-1}c^{-1}), \tag{10}$$

$$\sigma_3 : (a, c) \mapsto (c, a),\ \sigma_4 : (a, c) \mapsto (c^{-1}a^{-1}, c),\ \sigma_5 : (a, c) \mapsto (a, c^{-1}a^{-1}). \tag{11}$$

The set $\mathbb{T}(G)$ is in bijection with the set $\mathbb{T}_{\mathrm{tr}}(G) := \{(a, b, c) | abc = 1\}$. Examining these triples we see that σ_0 is the identity, σ_1 is the 3-cycle $(a, b, c) \to (b, c, a)$, σ_3 is the permutation $(a, b, c) \to (c, c^{-1}bc, a)$, while $\sigma_2 = \sigma_1^2$ and $\sigma_1\sigma_3 = \sigma_4,\ \sigma_1^2\sigma_3 = \sigma_5$. This gives the relations

$$\sigma_1^3 = \sigma_3^2 = \sigma_0,\ \sigma_2 = \sigma_1^2,\ \sigma_1\sigma_3 = \sigma_4,\ \sigma_1^2\sigma_3 = \sigma_5, \tag{12}$$

$$(\sigma_1\sigma_3)^2 = \sigma_4^2 = \mathrm{Int}_{c^{-1}} \circ \sigma_0. \tag{13}$$

Let us write

$$\mathrm{A}_{\mathbb{T}}(G) := \langle \mathrm{Aut}(G), \sigma_1, \ldots, \sigma_5 \rangle \tag{14}$$

for the permutation group generated by these operations. The above equations show that we have a homomorphism of the symmetric group $\mathcal{S}_3$ into $\mathrm{A}_{\mathbb{T}}(G)/\mathrm{Int(G)}$ and that $\mathrm{Aut}(G)$ is a normal subgroup of index ≤ 6 in $\mathrm{A}_{\mathbb{T}}(G)$, with quotient a subgroup of $\mathcal{S}_3$. In particular, every element $\rho \in \mathrm{A}_{\mathbb{T}}(G)$ can be written as

$$\rho = \psi \circ \sigma_i \tag{15}$$

for an automorphism ψ of G and an element σ_i from the above list. Define

$$\mathrm{I}_{\mathbb{T}}(G) := \langle \mathrm{Int}(G), \sigma_1, \ldots, \sigma_5 \rangle, \tag{16}$$

where $\mathrm{Int}(G)$ the (normal) subgroup of $\mathrm{A}_{\mathbb{T}}(G)$ consisting of the inner automorphisms.

By an operation from $\mathrm{A}_{\mathbb{T}}(G)$ we may ensure that a pair $(a, c) \in \mathbb{T}(G)$ satisfies

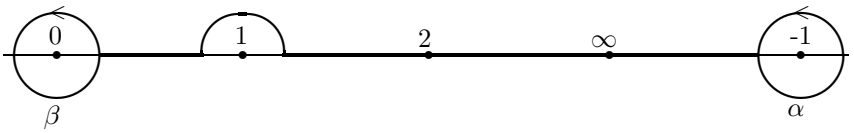
$$\mathrm{ord}(a) \leq \mathrm{ord}(b) = \mathrm{ord}(a^{-1}c^{-1}) \leq \mathrm{ord}(c),$$

in which case we call the pair *normalized*. We call (a, c) *strict*, if all inequalities are strict, *critical* if all the three orders are equal, and *subcritical* otherwise.

To every pair $(a, c) \in \mathbb{T}(G)$ we attach a ramified covering $C(a, c) \to \mathbb{P}^1_{\mathbb{C}}$ as follows. Consider the set $B \subset \mathbb{P}^1_{\mathbb{C}}$ consisting of three real points $B := \{-1, 0, 1\}$. Choose ∞ as a base point in $\mathbb{P}^1_{\mathbb{C}} \setminus B$, and take the following generators α, β, γ of $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty)$:

- α goes from ∞ to $-1 - \epsilon$ along the real line, passing through -2, then makes a full turn counterclockwise around the circumference with centre -1 and radius ϵ, then goes back to 2 along the same way on the real line.
- γ goes from ∞ to $1 + \epsilon$ along the real line, then makes a full turn counterclockwise around the circumference with centre $+1$ and radius ϵ, then goes back to ∞ along the same way on the real line.
- β goes from ∞ to $1 + \epsilon$ along the real line, makes a half turn counterclockwise around the circumference with centre $+1$ and radius ϵ, reaching $1 - \epsilon$, then proceeds along the real line reaching $+\epsilon$, makes a full turn counterclockwise around the circumference with centre 0 and radius ϵ, goes back to $1 - \epsilon$ along the same way on the real line, makes again a half turn clockwise around the circumference with centre $+1$ and radius ϵ, reaching $1 + \epsilon$, finally it proceeds along the real line returning to ∞.

A graphical picture of α, β, is:



Writing α, β, γ for the corresponding elements of $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty)$ we find

$$\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty) = \langle \alpha, \beta, \gamma \mid \alpha\beta\gamma = 1 \rangle$$

and α, γ are free generators of $\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty)$.

Let G be a finite group and $(a, c) \in \mathbb{T}(G)$. By Riemann's existence theorem, the elements a, $b = a^{-1}c^{-1}$, c, once we fix a basis of the fundamental group of $\mathbb{P}^1_{\mathbb{C}} \setminus \{-1, 0, 1\}$ as above, give rise to a surjective homomorphism

$$\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty) \to G, \qquad \alpha \mapsto a,\ \gamma \mapsto c \tag{17}$$

and to a Galois covering $\lambda : C \to \mathbb{P}^1_{\mathbb{C}}$ ramified only in $\{-1, 0, 1\}$ with ramification indices equal to the orders of a, b, c and with group G (beware, this means that these data yield a well-determined action of G on $C(a, c)$).

We call this covering a *triangular covering*. We embed G into $\mathcal{S}_G$ as the transitive subgroup of left translations. The monodromy homomorphism [3]

$$m_\lambda : \pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty) \to \mathcal{S}_G$$

maps onto the embedded subgroup G and equals the homomorphism (17).

By Hurwitz's formula, the genus $g(C(a,c))$ of the curve $C(a,c)$ is given by

$$g(C(a,c)) = 1 + \frac{1-\mu(a,c)}{2}\,|G|. \tag{18}$$

Let (a,c), $(a',c') \in \mathbb{T}(G)$. A *twisted covering isomorphism* from the Galois covering $\lambda : C(a,c) \to \mathbb{P}^1_{\mathbb{C}}$ to the Galois covering $\lambda' : C(a',c') \to \mathbb{P}^1_{\mathbb{C}}$ is a pair (σ,δ) of biholomorphic maps $\sigma : C(a,c) \to C(a',c')$ and $\delta : \mathbb{P}^1_{\mathbb{C}} \to \mathbb{P}^1_{\mathbb{C}}$ with $\delta(B) = B$ such that the diagram

$$\begin{array}{ccc} C(a,c) & \xrightarrow{\sigma} & C(a',c') \\ \downarrow{\scriptstyle \lambda} & & \downarrow{\scriptstyle \lambda'} \\ \mathbb{P}^1_{\mathbb{C}} & \xrightarrow{\delta} & \mathbb{P}^1_{\mathbb{C}}. \end{array} \tag{19}$$

is commutative. We say that we have a *strict covering isomorphism* if δ is the identity.

Consider G as acting on $C(a,c)$ by covering transformations over λ, and conjugate a transformation $g \in G$ by σ: since $\sigma \circ g \circ \sigma^{-1}$ is a covering transformation of $C(a',c')$, we obtain in this way an automorphism ψ of G (attached to the biholomorphic equivalence (σ,δ)) such that

$$\sigma(gP) = \psi(g)\sigma(P) \quad \text{for all } g \in G, P \in C(a,c).$$

That is, the map $\sigma : C(a,c) \to C(a',c')$ is G-twisted-equivariant.

Remark 2.1. We claim that ψ is the identity if we have a strict covering isomorphism. The converse does not necessarily hold, as shown by the example of $G = \mathbb{Z}/3\mathbb{Z}$ as a quotient of $T(3,3,3)$, where all three elements α, β, γ have the same image $= 1 \bmod 3$ (see the following considerations).

In order to understand the equivalence relation induced by the covering isomorphisms on the set $\mathbb{T}(G)$ of triangle structures, recall the following well-known facts from the theory of ramified coverings (see [13]):

Facts 2.2.

A) The monodromy homomorphism is only determined by the choice of a base point ∞' lying over ∞; a different choice alters the monodromy up to composition with an inner automorphism (corresponding to a transformation carrying one base point to the other).

[3] Actually, with the usual conventions the monodromy is an antihomomorphism; there are two ways to remedy this problem, here we shall do it by considering the composition of paths $\gamma \circ \delta$ as the path obtained by following first δ and then γ .

B) The map δ induces isomorphisms

$$\delta_* : \pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty) \to \pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \delta(\infty)) \to \pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty),$$

the second being induced by the choice of a path from ∞ to $\delta(\infty)$. Since the stabilizer of a chosen base point lying over ∞ under the monodromy action equals the kernel of the monodromy homomorphism m_λ, the class of monodromy homomorphisms corresponding to the covering $C(a', c')$ is obtained from the one of the given μ (corresponding to $C(a, c)$) by composing with $(\delta_*)^{-1}$. In particular, we may set $a'' := \mu(\delta_*)^{-1}(\alpha)$, and $c'' := \mu(\delta_*)^{-1}(\gamma)$. It follows that ψ is obtained from the natural isomorphism

$$\pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty)/\ker(\mu) \to \pi_1(\mathbb{P}^1_{\mathbb{C}} \setminus B, \infty)/\ker(\mu \circ (\delta_*)^{-1})$$

induced by (δ_*), and the obvious identifications of these quotient groups with G. (In more concrete terms, ψ sends $a'' \to a'$, $c'' \to c'$.)

C) The above shows that if the isomorphism is strict, then ψ is the identity. The converse does not hold since ψ can be the identity, without δ_* being the identity.

Proposition 2.3. *Let G be a finite group and $(a, c), (a', c') \in \mathbb{T}(G)$. The following are equivalent:*

(i-t) there is a twisted covering isomorphism from $\lambda : C(a, c) \to \mathbb{P}^1_{\mathbb{C}}$ to the Galois covering $\lambda' : C(a', c') \to \mathbb{P}^1_{\mathbb{C}}$,
(ii-t) there is a G-twisted-equivariant biholomorphic map

$$\sigma : C(a, c) \to C(a', c'),$$

(iii-t) (a, c) is in the $\mathrm{A}_{\mathbb{T}}(G)$-orbit of (a', c').

Respectively, the following are equivalent:

(i-s) there is a strict covering isomorphism from $\lambda : C(a, c) \to \mathbb{P}^1_{\mathbb{C}}$ to the Galois covering $\lambda' : C(a', c') \to \mathbb{P}^1_{\mathbb{C}}$,
(ii-s) there is a G-equivariant biholomorphic map $\sigma : C(a, c) \to C(a', c')$,
(iii-s) (a, c) is in the $\mathrm{I}_{\mathbb{T}}(G)$-orbit of (a', c').

Proof. The equivalence of (i) and (ii) follows directly from the definition. In view of A) we only consider triangle structures up to action of $\mathrm{Int}(G)$. We have seen that two triangle structures yield coverings which are twisted covering isomorphic if and only if there is an automorphism δ of $(\mathbb{P}^1_{\mathbb{C}} \setminus B)$ and an automorphism $\psi \in \mathrm{Aut}(G)$ such that $\psi \circ \mu = \mu' \circ \delta_*$.

In particular, $a'' := \mu(\delta_*)^{-1}(\alpha)$, $c'' := \mu(\delta_*)^{-1}(\gamma)$ are ψ equivalent to a', c', and it suffices to show that they are obtained from (a, c) by one of the transformations σ_i. Note however that the group of projectivities $\mathrm{Aut}(\mathbb{P}^1_{\mathbb{C}} \setminus B)$ is isomorphic to the group of permutations of B, by the fundamental theorem on projectivities.

We see immediately the action of an element of order 2: namely, consider the projectivity $z \to -z$: this leaves the base point ∞ fixed, as well as the point 0, and acts by sending $\alpha \to \gamma$, $\gamma \to \alpha$: we obtain in this way the transformation σ_3 on the set of triangle structures.

In order to obtain the transformation σ_1 of order 3, it is more convenient, after a projectivity, to assume that B consists of the three cubic roots of unity. Setting $\omega = \exp(2\pi i/3)$ and $B = \{1, \omega, \omega^2\}$, one sees immediately that σ_1 is induced by the automorphism $z \to \omega z$, which leaves again the base point ∞ fixed and cyclically permutes α, β, γ. □

To be able to treat questions of reality we define

$$\iota(a, c) := (a^{-1}, c^{-1}) \tag{20}$$

for $(a, c) \in \mathbb{T}(G)$ and call it the *conjugate* of (a, c). Note that $\iota(a, c) \in \mathbb{T}(G)$ and also $\Sigma(\iota(a, c)) = \Sigma(a, c)$, $\mu(\iota(a, c)) = \mu(a, c)$. A feature built into our construction is:

Proposition 2.4. *Let G be a finite group and $(a, c) \in \mathbb{T}(G)$. Then*

$$C(\iota(a, c)) = \overline{C(a, c)}. \tag{21}$$

Proof. For the proof note that by construction the complex conjugates of the paths α, γ used in the construction of the triangular curves $C(a, c)$ satisfy $\bar{\alpha} = \alpha^{-1}$, $\bar{\gamma} = \gamma^{-1}$. □

We now remind the reader of the operations $\sigma_0, \ldots, \sigma_5$ defined in (10), (11). For later use we observe:

Lemma 2.5. *Let G be a finite group, $(a, c) \in \mathbb{T}(G)$ and $\rho = \psi \circ \sigma_i \in \mathrm{A}_{\mathbb{T}}(G)$.*

(i) In case $i = 0$, $\rho(a, c) = \iota(a, c)$ if and only if $\psi(a) = a^{-1}$ and $\psi(c) = c^{-1}$.
(ii) In case $i = 1$, $\rho(a, c) = \iota(a, c)$ if and only if $\psi(a) = c^{-1}$ and $\psi(c) = ac$.
(iii) In case $i = 2$, $\rho(a, c) = \iota(a, c)$ if and only if $\psi(a) = ac$ and $\psi(c) = a^{-1}$.
(iv) In case $i = 3$, $\rho(a, c) = \iota(a, c)$ if and only if $\psi(a) = c^{-1}$ and $\psi(c) = a^{-1}$.
(v) In case $i = 4$, $\rho(a, c) = \iota(a, c)$ if and only if $\psi(a) = ac$ and $\psi(c) = c^{-1}$.
(vi) In case $i = 5$, $\rho(a, c) = \iota(a, c)$ if and only if $\psi(a) = a^{-1}$ and $\psi(c) = ac$.

Using this notation we assume that $\rho(a, c) = \psi \circ \sigma_i(a, c) = \iota(a, c)$ and get

$$(\psi\circ\sigma_0)^2(a, c) = (a, c),\ (\psi\circ\sigma_1)^2(a, c) = (c^{-1}ac, c),\ (\psi\circ\sigma_2)^2(a, c) = (a, aca^{-1}),$$

$$(\psi\circ\sigma_3)^2(a, c) = (a, c),\ (\psi\circ\sigma_4)^2(a, c) = (c^{-1}ac, c),\ (\psi\circ\sigma_5)^2(a, c) = (a, aca^{-1}),$$

for the square of ρ on (a, c).

3 The unmixed case

In this section we translate the problem of existence and classification of Beauville surfaces S of unmixed type to purely group-theoretic problems.

Unmixed Beauville surfaces and group actions

To have the group-theoretic background for the construction of Beauville surfaces from [4] we give the following definition.

Definition 3.1. *Let G be a finite group. A quadruple $v = (a_1, c_1; a_2, c_2)$ of elements of G is called an unmixed Beauville structure for G if and only if*

(i) the pairs a_1, c_1, and a_2, c_2 both generate G,
(ii) $\Sigma(a_1, c_1) \cap \Sigma(a_2, c_2) = \{1_G\}$.

The group G admits an unmixed Beauville structure if such a quadruple v exists. We write $\mathbb{U}(G)$ for the set of unmixed Beauville structures on G.

We also need an appropriate notion of equivalence of unmixed Beauville structures. To clarify it, let us observe that a Beauville surface has a unique minimal realization ([4], [5]), and that the Galois group of this covering is isomorphic to G. This yields an action of G on the product $C_1 \times C_2$ (whence, two actions of G^0 on both factors) only after we fix an isomorphism of the Galois group with G. In turn, these two actions of G determine a triangular covering up to strict covering isomorphism, and we can apply Proposition 2.3.

Notice that for $\psi_1, \psi_2 \in \mathrm{I}_{\mathbb{T}}(G)$ and $(a_1, c_1; a_2, c_2) \in \mathbb{U}(G)$ we have $(\psi_1(a_1, c_1); \psi_2(a_2, c_2)) \in \mathbb{U}(G)$. This gives a faithful action of $\mathrm{I}_{\mathbb{T}}(G) \times \mathrm{I}_{\mathbb{T}}(G)$ on $\mathbb{U}(G)$. Consider the group $\mathrm{B}_{\mathbb{U}}(G)$ generated by the action of $\mathrm{I}_{\mathbb{T}}(G) \times \mathrm{I}_{\mathbb{T}}(G)$ and by the diagonal action of $\mathrm{Aut}(G)$ (such that ($\psi \in \mathrm{Aut}(G)$ carries $(a_1, c_1; a_2, c_2) \in \mathbb{U}(G)$ to $(\psi(a_1, c_1); \psi(a_2, c_2)) \in \mathbb{U}(G)$). Define an operation

$$\tau((a_1, c_1; a_2, c_2)) := (a_2, c_2; a_1, c_1) \tag{22}$$

on the elements of $\mathbb{U}(G)$ and let

$$\mathrm{A}_{\mathbb{U}}(G) := \langle \mathrm{B}_{\mathbb{U}}(G),\, \tau \rangle \tag{23}$$

be the group generated by these permutations. Note that $\mathrm{B}_{\mathbb{U}}(G)$ is a normal subgroup of index ≤ 2 in $\mathrm{A}_{\mathbb{U}}(G)$. Given a $v := (a_1, c_1; a_2, c_2) \in \mathbb{U}(G)$ define

$$S(v) := C(a_1, c_1) \times C(a_2, c_2)/G. \tag{24}$$

The second condition in the definition of $\mathbb{U}(G)$ ensures that the action of G on the product has no fixed points, hence the covering $C(a_1, c_1) \times C(a_2, c_2) \to S(v)$ is unramified. We call the surface $S(v)$ an *unmixed Beauville surface*. It is obvious that (24) is a *minimal Galois realization* (see [4], [5]) of $S(v)$. Our next result shows that the unmixed Beauville surface $S(v)$ is isogenous to a *higher product* in the terminology of [4].

Proposition 3.2. *Let G be a finite, nontrivial group with an unmixed Beauville structure $(a_1, c_1; a_2, c_2) \in \mathbb{U}(G)$. Then $\mu(a_1, c_1) < 1$ and $\mu(a_2, c_2) < 1$. Whence we have: $g(C(a_1, c_1)) \geq 2$ and $g(C(a_2, c_2)) \geq 2$.*

Proof. Without loss of generality we may assume that G is not cyclic. Suppose (a_1, c_1) satisfies $\mu(a_1, c_1) > 1$: then the type of (a_1, c_1) is up to permutation among the

$$(2, 2, n)\ (n \in \mathbb{N}),\ (2, 3, 3),\ (2, 3, 4),\ (2, 3, 5).$$

In the first case G is a quotient group of the infinite dihedral group and G cannot admit an unmixed Beauville structure by Lemma 3.7. There are the following isomorphisms of triangular groups

$$T(2, 3, 3) = \mathcal{A}_4,\ T(2, 3, 4) = \mathcal{S}_4,\ T(2, 3, 5) = \mathcal{A}_5,$$

see [6], Chapter 4. These groups do not admit an unmixed Beauville structure by Proposition 3.6.

If $\mu(a_1, c_1) = 1$, then the type of (a_1, c_1) is up to permutation among the

$$(3, 3, 3),\ (2, 4, 4),\ (2, 3, 6)$$

and G is a finite quotient of one of the wallpaper groups and cannot admit an unmixed Beauville structure by the results of Section 6.

The second statement follows now from formula (18) since then $g(C(a_i, c_i))$, for $i = 1, 2$, is an integer strictly greater than 1. □

We may now apply results from [4], [5] to prove:

Proposition 3.3. *Let G be a finite group and $v, v' \in \mathbb{U}(G)$. Then $S(v)$ is biholomorphic to $S(v')$ if and only if v is in the $\mathrm{A}_{\mathbb{U}}(G)$-orbit of v'.*

Proof. Let $v = (a_1, c_1; a_2, c_2)$, $v' = (a_1', c_1'; a_2', c_2')$. Assume that there is a biholomorphism between two unmixed Beauville surfaces $S(v)$ and $S(v')$. This happens, by Proposition 3.2 of [5], if and only if there is a product biholomorphism (up to a possible interchange of the factors)

$$\sigma : C(a_1, c_1) \times C(a_2, c_2) \to C(a_1', c_1') \times C(a_2', c_2')$$

of the product surfaces appearing in the minimal Galois realization (24) which normalizes the G-action.

In the notation introduced previously, this means that σ is twisted G-equivariant. That is, there is an automorphism $\psi : G \to G$ with $\sigma(g(x, y)) = \psi(g)(\sigma(x, y))$ for all $g \in G$ and $(x, y) \in C(a_1, c_1) \times C(a_2, c_2)$. Up to replacing one of the two unmixed Beauville structures by an $\mathrm{Aut}(G)$-equivalent one, we may assume without loss of generality that the map σ is strict G-equivariant.

Note that our surfaces are both isogenous to a higher product by Proposition 3.2. Since σ is of product type it can interchange the factors or not. If it does not, there are biholomorphic maps

$$\sigma_1 : C(a_1, c_1) \to C(a_1', c_1'), \quad \sigma_2 : C(a_2, c_2) \to C(a_2', c_2')$$

such that $\sigma = (\sigma_1, \sigma_2)$. If σ does interchange the factors there are biholomorphic maps

$$\sigma_1 : C(a_1, c_1) \to C(a_2', c_2'), \quad \sigma_2 : C(a_2, c_2) \to C(a_1', c_1')$$

such that $\sigma = (\sigma_1, \sigma_2)$. In both cases we may now use Proposition 2.3 which characterizes strict G-equivariant isomorphisms of triangle coverings. □

Unmixed Beauville structures on finite groups

The question arises: which groups admit Beauville structures? The unmixed case with G abelian is easy to classify, and all examples were essentially given in [4], page 24.

Theorem 3.4. *If $G = G^0$ is abelian, nontrivial and admits an unmixed Beauville structure, then $G \cong (\mathbb{Z}/n\mathbb{Z})^2$, where the integer n is relatively prime to 6. Moreover, the structure is critical for both factors. Conversely, any group $G \cong (\mathbb{Z}/n\mathbb{Z})^2$ admits such a structure.*

Proof. Let $(a, c; a, c')$ be an unmixed Beauville structure on G, set

$$\Sigma := \Sigma(a, c), \ \Sigma' := \Sigma(a', c') \ \text{ and } \ b := a^{-1}c^{-1}, \ b' := a'^{-1}c'^{-1}.$$

Our basic strategy will be to observe that if H is a nontrivial characteristic subgroup of G, and if we show that for each choice of Σ we must have $\Sigma \supset H$, then we obtain a contradiction to $\Sigma \cap \Sigma' = \{1\}$.

Consider the primary decomposition of G,

$$G = \bigoplus_{p \in \{\text{Primes}\}} G_p$$

and observe that since G is 2-generated, then any G_p (which is a characteristic subgroup), is also 2-generated.

Step 1. Let $a = (a_p) \in \bigoplus_{p \in \{\text{Primes}\}} G_p$, and let Σ_p be the set of multiples of a_p, b_p, c_p: then $\Sigma \supset \Sigma_p$. This follows since a_p is a multiple of a.

Step 2. $G_p \cong (\mathbb{Z}/p^m\mathbb{Z})^2$. Since G_p is 2-generated, G_p is either cyclic $G_p \cong \mathbb{Z}/p^m\mathbb{Z}$ or $G_p \cong \mathbb{Z}/p^n\mathbb{Z} \oplus \mathbb{Z}/p^m\mathbb{Z}$ with $n < m$. In both cases the subgroup $H_p := p^{m-1}G_p$ is characteristic in G and isomorphic to $\mathbb{Z}/p\mathbb{Z}$. But $\Sigma \supset \Sigma_p$, and Σ_p contains generators of G_p, whence it contains a nontrivial element in H_p, thus $\Sigma_p \supset H_p$, a contradiction.

Step 3. $G_2 = 0$. Else, by Step 2, $G_2 \cong (\mathbb{Z}/2^m\mathbb{Z})^2$, and $H_2 = \cong (\mathbb{Z}/2\mathbb{Z})^2$. But since $\Sigma \supset \Sigma_2$, and Σ_2 contains a basis of G_2, $\Sigma \supset H_2$, a contradiction.

Step 4. $G_3 = 0$. In this case we have that Σ_3 contains a basis of G_3, whence $\Sigma \cap H_3$ contains at least six nonzero elements, likewise for $\Sigma' \cap H_3$, a contradiction since H_3 has only eight nonzero elements.

Step 5. Whence, $G \cong (\mathbb{Z}/n\mathbb{Z})^2$, and since a, b are generators of G, they are a basis, and without loss of generality a, b are the standard basis e_1, e_2. It follows that all the elements a, b, c, a', b', c' have order exactly n. Write now the elements of G as row vectors, $a' := (x, y), b' := (z, t)$. Then the condition that $\Sigma \cap \Sigma' = \{1\}$ means that any pair of the six vectors yield a basis of G. By using the primary decomposition, we can read out this condition on each primary component: thus it suffices to show that there are solutions in the case where $n = p^m$ is primary.

Step 6. We write up the conditions explicitly, namely, if $n = p^m$ and $U := \mathbb{Z}/n\mathbb{Z}^*$, we want

$$x, y, z, t \in U,\ x - y, x + z, z - t, y + t \in U, x + z - y - t \in U, xt - yz \in U.$$

Again, these conditions only bear on the residue class modulo p, thus we have p^{4m-4} times the number of solutions that we get for $n = p$.

Step 7. Simple counting yields at least $(p-1)(p-2)^2(p-4)$ solutions. In this case we get $p-1$ times the number of solutions that we get for $x = 1$, and for each choice of $y \neq 0, 1$ $z \neq 0, -1$, $d \neq 0$, $d \neq -yz$ we set $t := yz + d$: the other inequalities are then satisfied if d is different from $z - yz, -yz - y, (1 + z)(1 - y)$ so that the number of solutions equals at least $(p-1)(p-2)^2(p-5)$. □

Remark 3.5. The computation above shows that the number of biholomorphism classes of unmixed Beauville surfaces with abelian group $(\mathbb{Z}/n\mathbb{Z})^2$ is asymptotic to at least $(1/36)\ n^4$ (cf. [1] where it is calculated that, for $n = 5$ there are exactly two isomorphism classes).

Proposition 3.6. *No nonabelian group of order* ≤ 128 *admits an unmixed Beauville structure.*

Proof. This result can be obtained by a straightforward computation using MAGMA or by direct considerations. In fact, using the Smallgroups-routine of MAGMA we may list all groups of order ≤ 128 as explicit permutation-groups or given by a polycyclic presentation. Loops which are easily designed can be used to search for appropriate systems of generators. □

Another simple result is:

Lemma 3.7. *If G is a nontrivial finite quotient of the infinite dihedral group $D := \langle x, y \mid x^2, y^2 \rangle$, then G does not admit an unmixed Beauville structure.*

Proof. The infinite cyclic subgroup $N_0 := \langle xy \rangle$ is normal in D of index 2; actually D is thus the semidirect product of $N_0 \cong \mathbb{Z}$ through the subgroup of order 2 generated by x. Let $t \in D$ be not contained in N_0. Then there is an integer n such that $t = x(xy)^n$. Since $yty = x(xy)^{n-2}$, the normal subgroup generated by t then contains $(xy)^2$. Hence every normal subgroup N of D not contained in N_0 has index ≤ 4 and thus the quotient D/N cannot admit an unmixed Beauville structure. Let now $N \leq N_0$ be a normal subgroup of D. The quotient D/N is a finite dihedral group. Let (a, c) be a pair of generators for D/N. It is easy to see that one of the elements a, c, ac lies in the (cyclic) image of N_0 in D/N and generates it. Thus condition (*) is contradicted. □

Proposition 3.8. *The following groups admit an unmixed Beauville structure:*

1. *the alternating groups $\mathcal{A}_n$ for large n,*
2. *the symmetric groups $\mathcal{S}_n$ for $n \in \mathbb{N}$ with $n \geq 8$ and $n \equiv 2$ mod 3,*
3. *the groups $\mathbf{SL}(2, \mathbb{F}_p)$ and $\mathbf{PSL}(2, \mathbb{F}_p)$ for every prime $p \neq 2, 3, 5$.*

Proof. 1. Fix two triples (n_1, n_2, n_3), $(m_1, m_2, m_3) \in \mathbb{N}^3$ such that neither $T(n_1, n_2, n_3)$ nor $T(m_1, m_2, m_3)$ is one of the nonhyperbolic triangle groups. From [8] we infer that, for large enough $n \in \mathbb{N}$, the group $\mathcal{A}_n$ has systems of generators (a_1, c_1) of type (n_1, n_2, n_3) and (a_2, c_2) of type (m_1, m_2, m_3). Adding the property that $\gcd(n_1 n_2 n_3, m_1 m_2 m_3) = 1$ we find that $(a_1, c_1; a_2, c_2)$ is an unmixed Beauville structure on $\mathcal{A}_n$.

By going through the proofs of [8] the minimal choice of such an $n \in \mathbb{N}$ can be made effective.

2. This follows directly from the first Proposition of Section 5.1.

3. Let p be a prime with the property that no prime $q \geq 5$ divides $p^2 - 1$. Then $(p, 1)$ is a primitive solution of the equation

$$y^2 - x^3 = \pm 2^n 3^m \tag{25}$$

with $n, m \in \mathbb{N}$ chosen appropriately. It is known that the collection of these equations has 98 primitive solutions (as n, m vary). A table of them is contained in [3] Table 4, page 125. From this we see that $p = 2, 3, 5, 7, 17$ are the only primes with the property that no prime $q \geq 5$ divides $p^2 - 1$. Notice that a theorem of C.L. Siegel implies directly that there are only finitely many such primes. A special case of this theorem says that any of the equations (25) has only finitely many solutions in $\mathbb{Z}[1/2, 1/3]$.

If p is a prime with $p \neq 2, 3, 5, 7, 17$ we use the system of generators from (40) which is of type $(4, 6, p)$ together with one of the system of generators from (42) or (44) to conclude the result for $\mathbf{SL}(2, \mathbb{F}_p)$. The groups $\mathbf{PSL}(2, \mathbb{F}_p)$ can be treated by reduction of these systems of generators, observing that the two generators belonging to different systems have coprime orders.

For the primes $p = 7, 17$ appropriate systems of generators can be easily found by a computer calculation. □

From the third item of the above proposition we immediately obtain a *proof of Theorem 1.3.*

Remark 3.9. The various systems of generators given in Section 5 for the alternating groups $\mathcal{A}_n$ and for $\mathbf{SL}(2,\mathbb{F}_p)$, $\mathbf{PSL}(2,\mathbb{F}_p)$ can be grouped together in many ways to construct unmixed Beauville structures on these groups. We in turn obtain Beauville surfaces of unmixed type for which the two curves appearing in the minimal Galois realization have different genus.

We now describe groups of a different nature admitting an unmixed Beauville structure. For $n \in \mathbb{N}$, put

$$C[n] := \langle x, y \mid x^n = y^n = (xy)^n = (xy^{-1})^n = (xy^{-2})^n = (xy^{-1}xy^{-2})^n = 1\rangle. \tag{26}$$

Proposition 3.10. *Fix $n \in \mathbb{N}$ with $\gcd(n,6) = 1$ and let $N \leq C[n]'$ be a normal subgroup of finite index in the commutator $C[n]'$ of $C[n]$. Then $C[n]/N$ admits an unmixed Beauville structure.*

Proof. An unmixed Beauville structure for $C[n]/N$ is given by

$$(a_1, c_1; a_2, c_2) = (x, y; xy^{-1}, xy^{-2}).$$

Let G^{ab} be the abelianization of G. Then Σ and Σ' map injectively into G^{ab}, and their images do not meet inside G^{ab}, as verified in [4], lemma 3.21. □

Among the quotients $C[n]/N$ ($N \leq C[n]'$) are all finite groups G of exponent n having $\mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ as abelianization. The proposition can hence be used to construct finite p-groups ($p \geq 5$) admitting an unmixed Beauville structure.

Questions of reality

We now translate to a group-theoretic conditions the two questions concerning an unmixed Beauville surface mentioned in the introduction:

- Is S biholomorphic to the complex conjugate surface $\bar{S}$?
- Is S real, i.e., does there exist such a biholomorphism σ with the property that $\sigma^2 = \mathrm{Id}$?

Let G be a finite group and $v = (a_1, c_1; a_2, c_2) \in \mathbb{U}(G)$. In analogy with (20) we define

$$\iota(v) := (a_1^{-1}, c_1^{-1}; a_2^{-1}, c_2^{-1}) \tag{27}$$

and infer from Proposition 2.4:

$$S(\iota(v)) = \overline{S(v)}. \tag{28}$$

From Proposition 3.3 we get

Proposition 3.11. *Let G be a finite group with an unmixed Beauville structure $v = (a_1, c_1; a_2, c_2) \in \mathbb{U}(G)$. Then*

1. *$S(v)$ is biholomorphic to $\overline{S(v)}$ if and only if $\iota(v)$ is in the $\mathrm{A}_{\mathbb{U}}(G)$-orbit of v,*
2. *$S(v)$ is real if and only if there is a $\rho \in \mathrm{A}_{\mathbb{U}}(G)$ with $\rho(v) = \iota(v)$ and moreover $\rho(\iota(v)) = v$.*

Remark 3.12. The above observations immediately imply that unmixed Beauville surfaces S with abelian group G always have a real structure, since $g \to -g$ is an automorphism (of order 2).

We observe the following:

Corollary 3.13. *Let G be a finite group with an unmixed Beauville structure $v = (a_1, c_1; a_2, c_2) \in \mathbb{U}(G)$. Assume that*

$$\{\mathrm{ord}(a_1), \mathrm{ord}(c_1), \mathrm{ord}(a_1c_1)\} \neq \{\mathrm{ord}(a_2), \mathrm{ord}(c_2), \mathrm{ord}(a_2c_2)\}$$

and that both (a_1, c_1) and (a_2, c_2) are strict. Then $S(v)$ is biholomorphic to $\overline{S(v)}$ if and only if the following holds:

There are inner automorphisms ϕ_1, ϕ_2 of G and an automorphism $\psi \in Aut(G)$ such that, setting $\psi_j := \psi \circ \phi_j$, we have

$$\psi_1(a_1) = a_1^{-1},\ \psi_1(c_1) = c_1^{-1}, \quad \text{and} \quad \psi_2(a_2) = {a_2}^{-1},\ \psi_2(c_2) = {c_2}^{-1}.$$

Thus $S(v)$ is isomorphic to $\overline{S(v)}$ if and only if $S(v)$ has a real structure.

Proof. The first statement follows from our definition of $\mathrm{A}_{\mathbb{U}}(G)$ and Proposition 3.3. In fact let $\rho \in \mathrm{A}_{\mathbb{U}}(G)$ be such that $\rho(v) = \iota(v)$. We have

$$\rho = (\psi_1 \circ \sigma_i, \psi_2 \circ \sigma_j) \circ \tau^e$$

with ψ_1, ψ_2 as above, $i, j \in \{0, \ldots, 5\}$ and $e \in \{0, 1\}$. Our incompatibility conditions on the orders imply that $e = 0$ and $i = j = 0$ (see Lemma 2.5).

For the second statement note that the conclusion implies that both ψ_1 and ψ_2 have order 2. □

Remark 3.14. If the unmixed Beauville structure v does not have the strong incompatibility properties of the corollary, then Lemma 2.5 gives the appropriate conditions.

From our corollary we immediately get:

Proof (of Theorem 1.4). Let $v := (a, c; a', c')$ with $a, c, a', c' \in \mathcal{S}_n$ as in Proposition 5.1. Then v is an unmixed Beauville structure on $\mathcal{S}_n$. The type of (a, c) is $(6, 3(n-3), 3(n-4))$, while the type of (a', c') is $(n, n-1, n)$ or $(n, n-1, (n^2-1)/4)$. Suppose that $S(v)$ is biholomorphic to $\overline{S(v)}$. By Proposition 3.11 (statement 1) $\iota(v)$ is in the $\mathrm{A}_{\mathbb{U}}(\mathcal{S}_n)$ orbit of $S(v)$. The incompatibility

of the types of (a, c) and (a', c') makes Corollary 3.13 applicable. Thus there is a $\psi \in \mathrm{Aut}(\mathcal{S}_n)$ with $\psi(a) = a^{-1}$ and $\psi(c) = c^{-1}$. Since all automorphisms on $\mathcal{S}_n$ $(n \geq 8)$ are inner we obtain a contradiction to Proposition 5.1. □

As noted above the unmixed Beauville surfaces S coming from an abelian group G always have a real structure. It is also possible to construct examples from nonabelian groups:

Proposition 3.15. *Let $p \geq 5$ be a prime with $p \equiv 1$ mod 4. Set $n := 3p + 1$. Then there is an unmixed Beauville structure v for the group $\mathcal{A}_n$ such that $S(v)$ is biholomorphic to $\overline{S(v)}$.*

Proof. We use the first and the second system of generators for $\mathcal{A}_n$ from Proposition 5.9. The first is a system of generators (a_1, c_1) of type $(2, 3, 84)$, the second gives (a_2, c_2) of type $(p, 5p, 2p + 3)$. Since the orders in the two types are coprime, $(a_1, c_1; a_2, c_2)$ is an unmixed Beauville structure on $\mathcal{A}_n$. The existence of the respective elements γ in Proposition 5.9 implies the last assertion. Note that both the elements γ can be chosen to be in $\mathcal{S}_n \setminus \mathcal{A}_n$. □

In further arguments we shall often use the fact that every automorphism of $\mathcal{A}_n$ $(n \neq 6)$ is induced by conjugation by an element of $\mathcal{S}_n$ (see [14], p. 299).

Proposition 3.16. *The following groups admit unmixed Beauville structures v such that $S(v)$ is not biholomorphic to $\overline{S(v)}$:*

1. *the symmetric group $\mathcal{S}_n$ for $n \geq 8$ and $n \equiv 2$ mod 3,*
2. *the alternating group $\mathcal{A}_n$ for $n \geq 16$ and $n \equiv 0$ mod 4, $n \equiv 1$ mod 3, $n \not\equiv 3, 4$ mod 7.*

Proof. 1. This is just the example of Section 5.1.

2. We use the system of generators (a_1, c_1) from Proposition 5.9, 1. It has type $(2, 3, 84)$. We choose $p = 5$ and $q = 11$ and get from Proposition 5.8 a system of generators (a_2, c_2) of type $(11, 5(n - 11), n - 3)$. Both systems are strict. We set $v := (a_1, c_1; a_2, c_2)$. The congruence conditions $n \not\equiv 3, 4$ mod 7 insure that $\nu(a_1, , c_1)$ is coprime to $\nu(a_2, , c_2)$, hence this is an unmixed Beauville structure. It also satisfies the hypotheses of Corollary 3.13. If $S(v)$ is biholomorphic to $\overline{S(v)}$ we obtain an element $\gamma \in \mathcal{S}_n$ with $\gamma a_2 \gamma^{-1} = a_2^{-1}$ and $\gamma c_2 \gamma^{-1} = c_2^{-1}$. This contradicts Proposition 5.8. □

Proposition 3.17. *Let $p > 5$ be a prime with $p \equiv 1$ mod 4, $p \not\equiv 2, 4$ mod 5, $p \not\equiv 5$ mod 13 and $p \not\equiv 4$ mod 11. Set $n := 3p + 1$. Then the alternating group $G := \mathcal{A}_n$ admits an unmixed Beauville structure v such that there is an element $\alpha \in \mathrm{A}_{\mathbb{U}}(G)$ with $\alpha(v) = \iota(v)$ but such that there is no element $\beta \in \mathrm{A}_{\mathbb{U}}(G)$ with $\beta(v) = \iota(v)$ and $\beta(\iota(v)) = v$.*

Proof. We construct v using the system of generators (a_1, c_1) from Proposition 5.10. It has type $(3p-2, 3p-1, 3p-1)$. We then use the system of generators (a_2, c_2) from Proposition 5.9, 2. It has type $(p, 5p, 2p+3)$. The second system is strict. We set $v := (a_1, c_1; a_2, c_2)$. The congruence conditions $p \not\equiv 2, 4 \bmod 5$, $p \not\equiv 5 \bmod 13$ and $p \not\equiv 4 \bmod 11$ ensure that $\nu(a_1, , c_1)$ is coprime to $\nu(a_2, , c_2)$, hence v is an unmixed Beauville structure.

We first show that there exists an $\alpha \in \mathrm{A}_{\mathbb{U}}(\mathcal{A}_n)$ with

$$\alpha(v) = \iota(v) = (a_1^{-1}, c_1^{-1}; a_2^{-1}, c_2^{-1}).$$

We choose γ_1 as in Proposition 5.10 and γ_2 as in Proposition 5.9, 2. Let $w \in \mathcal{S}_n$ be a representative of the nontrivial coset of $\mathcal{A}_n$ in $\mathcal{S}_n$. By Propositions 5.10, 5.9 these choices can be made so that $\gamma_1 = \delta_1 w$, $\gamma_2 = \delta_2 w$ with $\delta_1, \delta_2 \in \mathcal{A}_n$. We have now

$$\begin{aligned}(w^{-1}\delta_1^{-1}\ a_1\ \delta_1 w, w^{-1}\delta_1^{-1}\ c_1^{-1}a_1^{-1}\ \delta_1 w) &= (a_1^{-1}, c_1^{-1}),\\ (w^{-1}\delta_2^{-1}\ a_2\ \delta_2 w, w^{-1}\delta_2^{-1}\ c_2\ \delta_2 w) &= (a_2^{-1}, c_2^{-1}).\end{aligned}$$

Recalling the formula for σ_5 (see (11)) the existence of ρ follows from our definition of $\mathrm{A}_{\mathbb{U}}(\mathcal{A}_n)$.

Suppose now that there is a $\beta \in \mathrm{A}_{\mathbb{U}}(G)$ as indicated, then $\beta^2(v) = v$. By construction of v the transformation β cannot interchange (a_1, c_1) and (a_2, c_2). Hence we find $\beta_1, \beta_2 \in \mathrm{A}_{\mathbb{T}}(\mathcal{A}_n)$ with

$$\beta_1(a_1, c_1) = (a_1^{-1}, c_1^{-1}), \quad \beta_1(a_2, c_2) = (a_2^{-1}, c_2^{-1})$$

from $\beta_1^2(a_1, c_1) = (a_1, c_1)$, and the formulae given immediately after Lemma 2.5 imply that either $\beta_1 = \psi \circ \sigma_0$ or $\beta_1 = \psi \circ \sigma_3$ for a suitable automorphism ψ of $\mathcal{A}_n$. (Note that a_1 and c_1 cannot commute.) Going back to Lemma 2.5 (i), (iv) we find a contradiction with the statement of Proposition 5.10. □

Theorem 1.6 follows immediately from the above proposition and from Proposition 3.11.

4 The mixed case

In this section we fix the algebraic data needed for the construction of Beauville surfaces of mixed type and use this description to give several examples.

Mixed Beauville surfaces and group actions

This subsection contains the translation between the geometrical data of a mixed Beauville surface and the corresponding algebraic data: finite groups endowed with a mixed Beauville structure. This concept is contained in the following:

Definition 4.1. *Let G be a nontrivial finite group. A mixed Beauville quadruple for G is a quadruple $M = (G^0; a, c; g)$ consisting of a subgroup G^0 of index 2 in G, of elements $a, c \in G^0$ and of an element $g \in G$ such that*

1. *G^0 is generated by a, c,*
2. *$g \notin G^0$,*
3. *for every $\gamma \in G^0$ we have $g\gamma g\gamma \notin \Sigma(a, c)$,*
4. *$\Sigma(a, c) \cap \Sigma(gag^{-1}, gcg^{-1}) = \{1_G\}$.*

Forgetting the choice of g, we obtain from a mixed Beauville quadruple a mixed Beauville triple for G, $u = (G^0; a, c)$. The group G is said to admit a mixed Beauville structure if such a quadruple M exists. We let then $\mathbb{M}_4(G)$ be the set of mixed Beauville quadruples on the group G, $\mathbb{M}_3(G)$ be the set of mixed Beauville triples on the group G. These last will also be called mixed Beauville structures.

We describe the correspondence between the data for an unmixed Beauville structure given above and those given in [4] (also described in the introduction). Let $M = (G^0; a, c; g)$ be a mixed Beauville quadruple on a finite group G. Then G^0 is normal in G. By condition (3) the exact sequence

$$1 \to G^0 \to G \to \mathbb{Z}/2\mathbb{Z} \to 1, \tag{29}$$

does not split. Define $\varphi_g : G^0 \to G^0$ to be the automorphism of G^0 induced by conjugation with g, that is $\varphi_g(\gamma) = g\gamma g^{-1}$ for all $\gamma \in G^0$. Suppose that φ_g is an inner automorphism. Then we can find $\delta \in G^0$ with $\varphi_g(\gamma) = \delta\gamma\delta^{-1}$ for all $\gamma \in G^0$. This implies that $\Sigma(gag^{-1}, gcg^{-1}) = \Sigma(a, c)$. Since G is nontrivial condition (4) cannot hold.

Let $\tau := \tau_g := g^2 \in G^0$. We have $\varphi_g(\tau) = \tau$ and $\varphi_g^2 = \mathrm{Int}_\tau$ where Int_τ is the inner automorphism induced by τ. This shows that φ_g is of order 2 in the group of outer automorphisms $\mathrm{Out}(G^0)$ of G^0. Conversely given a nontrivial finite group G^0 together with an an automorphism $\varphi : G^0 \to G^0$ of order 2 in the outer automorphism group allows us to find a group G together with an exact sequence (29).

It is important to observe that the conditions (3), (4) are the ones which guarantee the freeness of the action of G.

Now we describe the appropriate notion of equivalence for mixed Beauville structures. Let $M = (G^0; a, c; g)$ be a mixed Beauville quadruple for G and $\psi : G \to G$ an automorphism; then $\psi(M) := (\psi(G^0); \psi(a), \psi(c); \psi(g))$ is again a mixed Beauville structure on G. Thus we obtain respective actions of $\mathrm{Aut}(G)$ on $\mathbb{M}_4(G)$, $\mathbb{M}_3(G)$. If $\gamma \in G^0$ and $M = (G^0; a, c; g)$ is a mixed Beauville quadruple on G, then so is $M_\gamma = (G^0; a, c; \gamma g)$.

We can therefore, without loss of generality, only consider mixed Beauville triples (beware, such a triple is obtained from a quadruple satisfying conditions (1)–(4) of the previous definition).

The set $\mathbb{M}(G) := \mathbb{M}_3(G)$ of mixed Beauville structures carries the action of the group

$$\mathrm{A}_{\mathbb{M}}(G) :=< \mathrm{Aut}(G),\ \sigma_3, \sigma_4 >, \tag{30}$$

with the understanding that the operations σ_3, σ_4 from (10), (11) are applied to the pair (a, c) of generators of G^0. Note that the operations $\sigma_1, \sigma_2, \sigma_5$ are also in $\mathrm{A}_{\mathbb{M}}(G)$ because of (12).

Recall how the above algebraic data give rise to a Beauville surface of mixed type. Let $u := (G^0; a, c; g)$ be a mixed Beauville quadruple on G. Set $\tau_g := g^2$ and $\varphi_g(\gamma) := g\gamma g^{-1}$ for $\gamma \in G^0$. By Riemann's existence theorem, as in the previous section the elements a, $b = a^{-1}c^{-1}$, c give rise to a Galois covering $\lambda : C(a, c) \to \mathbb{P}^1_{\mathbb{C}}$ ramified only in $\{-1, 0, 1\}$ with ramification indices equal to the respective orders of a, $b = a^{-1}c^{-1}$, c and with group G^0. The group G acts on $C(a, c) \times C(a, c)$ by

$$\gamma(x, y) = (\gamma x, \varphi_g(\gamma)y), \quad g(x, y) = (y, \tau_g x), \tag{31}$$

for all $\gamma \in G^0$ and $(x, y) \in C(a, c) \times C(a, c)$. These formulae determine an action of G uniquely. By our conditions (3), (4) in the definition of a mixed Beauville quadruple on G the above action of G is fixed-point free, yielding a Beauville surface of mixed type

$$S(u) := C(a, c) \times C(a, c)/G. \tag{32}$$

It is obvious that (32) is a *minimal Galois representation* (see [4], [5]) of $S(u)$. From Proposition 3.2 we infer that the mixed Beauville surface $S(u)$ is isogenous to a *higher product* in the terminology of [5].

Observe that a Beauville surface of mixed type $S(u) = C(a, c) \times C(a, c)/G$ has a natural unramified double cover $S^0(u) = (C(a, c) \times C(a, c))/G^0$ which is of unmixed type.

Proposition 4.2. *Let G be a finite group and $u_1, u_2 \in \mathbb{M}(G)$. Then $S(u_1)$ is biholomorphic to $S(u_2)$ if and only if u_1 is in the $\mathrm{A}_{\mathbb{M}}(G)$-orbit of u_2.*

Proof. Follows (as Proposition 3.3) from [4], [5] (Proposition 3.2).

Let $M = (G_1^0, a_1, c_1; g_1)$, $M' = (G_2^0; a_2, c_2; g_2)$. Assume that the unmixed Beauville surfaces $S(u)$ and $S(u')$ are biholomorphic. This happens, by Proposition 3.2 of [5], if and only if there is a product biholomorphism (up to an interchange of the factors)

$$\sigma : C(a_1, c_1) \times C(a_1, c_1) \to C(a_2, c_2) \times C(a_2, c_2)$$

of the product surfaces. Since σ is of product type it can interchange the factors or not. Hence there are biholomorphic maps

$$\sigma_1,\ \sigma_2 : C(a_1, c_1) \to C(a_2, c_2)$$

with

$$\sigma(x, y) = (\sigma_1(x), \sigma_2(y)), \qquad \text{for all } x, y \in C(a_1, c_1)$$

in case σ does not interchange the factors. Otherwise,

$$\sigma(x,y) = (\sigma_1(y), \sigma_2(x)), \qquad \text{for all } x, y \in C(a_1, c_1).$$

The map σ normalizes the G-action if there is an automorphism $\psi : G \to G$ with $\sigma(g(x,y)) = \psi(g)(\sigma(x,y))$ for all $g \in G$ and $(x,y) \in C(a_1,c_1) \times C(a_1,c_1)$.

In both cases we use Proposition 2.3 combined with direct computations to complete the only if statement of our proposition. The converse statement follows from Proposition 2.3. □

Mixed Beauville structures on finite groups

To find a group G with a mixed Beauville structure is rather difficult. For instance the subgroup G^0 cannot be abelian:

Theorem 4.3. *If G admits a mixed Beauville structure, then the subgroup G^0 is nonabelian.*

Proof. By Theorem 3.4 we know that G^0 is isomorphic to $(\mathbb{Z}/n\mathbb{Z})^2$, where n is an odd number not divisible by 3.

In particular, multiplication by 2 is an isomorphism of G^0, thus there is a unique element γ such that $-2\gamma = \tau$. Since $-2\varphi(\gamma) = \varphi(\tau) = \tau = -2\gamma$, it follows that $\varphi(\gamma) = \gamma$, and we have found a solution to the prohibited equation $\varphi(\gamma) + \tau + \gamma \in \Sigma$, since $0 \in \Sigma$. Whence the desired contradiction. □

The following fact was obtained by computer calculations using MAGMA.

Proposition 4.4. *No group of order ≤ 512 admits a mixed Beauville structure.*

Here is a general construction giving finite groups G with a mixed Beauville structure. Let H be a nontrivial group and $\Theta : H \times H \to H \times H$ the involution defined by $\Theta(g,h) := (h,g)$ $(g,h \in H)$. Consider the semidirect product

$$H_{[4]} := (H \times H) \rtimes \mathbb{Z}/4\mathbb{Z}, \tag{33}$$

where the generator 1 of $\mathbb{Z}/4\mathbb{Z}$ acts through Θ on $H \times H$. We find

$$H_{[2]} := H \times H \times 2\mathbb{Z}/4\mathbb{Z} \cong H \times H \times \mathbb{Z}/2\mathbb{Z} \tag{34}$$

as a subgroup of index 2 in $H_{[4]}$.

The exact sequence

$$1 \to H_{[2]} \to H_{[4]} \to \mathbb{Z}/2\mathbb{Z} \to 1$$

does not split because there is no element of order 2 in $H_{[4]}$ which is not already contained in $H_{[2]}$. We have

Lemma 4.5. *Let H be a nontrivial group and let a_1, c_1, a_2, c_2 be elements of H. Assume that*

1. *the orders of a_1, c_1 are even,*
2. *a_1^2, a_1c_1, c_1^2 generate H,*
3. *a_2, c_2 also generate H,*
4. *$\nu(a_1, c_1)$ is coprime to $\nu(a_2, c_2)$.*

Set $G := H_{[4]}$, $G^0 := H_{[2]}$ as above and $a := (a_1, a_2, 2)$, $c := (c_1, c_2, 2)$. Then $(G^0; a, c)$ is a mixed Beauville structure on G.

If H is a perfect group, then the conclusion holds with Condition 2. replaced by Condition

2′. *a_1, c_1 generate H.*

Proof. We first show that a, c generate $G^0 := H_{[2]}$. Let $L := \langle a, c\rangle$. We view $H \times H$ as the subgroup $H \times H \times \{0\}$ of $H_{[2]}$. The elements a^2, ac, c^2 are in this subgroup. Condition 2 implies that $L \cap (H \times H)$ surjects onto the first factor of $H \times H$. Conditions 1, 3, 4 imply that a_2, c_2 have odd order, and that there is an even number $2m$ such that a_2^{2m}, c_2^{2m} generate H, while $a_1^{2m} = c_1^{2m} = 1$. It follows that $H \times H \leq L$, and it is clear that $L = G^0$.

Observe next that

$$(1_H, 1_H, 2) \notin \Sigma(a, c). \tag{35}$$

It would have to be a conjugate of a power of a, c or b. Since the orders of a_1, b_1, c_1 are even, we obtain a contradiction. Note that the third component of ac is 0 by construction.

We now verify the third condition of the definition of a mixed Beauville structure. Suppose first that $h = (x, y, z) \in \Sigma(a, c)$ satisfies $\mathrm{ord}(x) = \mathrm{ord}(y)$: then our Condition 4 implies that $x = y = 1_H$ and (35) shows $h = 1_{H_{[4]}}$.

Let now $g \in H_{[4]}$, $g \notin H_{[2]}$ and $\gamma \in G^0 = H_{[2]}$ be given. Then $g\gamma = (x, y, \pm 1)$ for appropriate $x, y \in H$. We find

$$(g\gamma)^2 = (xy, yx, 2)$$

and the orders of the first two components of $(g\gamma)^2$ are the same. The remark above shows that $(g\gamma)^2 \in \Sigma(a, c)$ implies $(g\gamma)^2 = 1$.

We come now to the fourth condition of our definition of a mixed Beauville quadruple. Let $g \in H_{[4]}$, $g \notin H_{[2]}$ be given, for instance $(1_H, 1_H, 1)$. Conjugation by g interchanges the first two components of an element $h \in H_{[4]}$. Our hypothesis 4 implies the result.

So far we have proved the lemma using Condition 2. Assume that H is a perfect group (this means that H is generated by commutators). Because of Condition 2′ the group H is generated by commutators of words in a_1, c_1. Defining L as before we see again that that $L \cap (H \times H)$ projects surjectively onto the first factor of $H \times H$. The rest of the proof is the same. □

As an application we get

Proposition 4.6. *Let H be one of the following groups:*

1. *the alternating group $\mathcal{A}_n$ for large n,*
2. $\mathbf{SL}(2, \mathbb{F}_p)$ *for $p \neq 2, 3, 5, 17$.*

Then $H_{[4]}$ admits a mixed Beauville structure.

Proof. 1. Fix two triples (n_1, n_2, n_3), $(m_1, m_2, m_3) \in \mathbb{N}^3$ such that neither $T(n_1, n_2, n_3)$ nor $T(m_1, m_2, m_3)$ is one of the nonhyperbolic triangle groups. From [8] we infer that, for large enough $n \in \mathbb{N}$, the group $\mathcal{A}_n$ has systems of generators (a_1, c_1) of type (n_1, n_2, n_3) and (a_2, c_2) of type (m_1, m_2, m_3). Adding the properties that n_1, n_2 are even and $\gcd(n_1 n_2 n_3, m_1 m_2 m_3) = 1$ we find that the $(a_1, c_1; a_2, c_2)$ satisfy the Conditions $1, 2', 3, 4$ of the previous lemma. Since $\mathcal{A}_n$ is, for large n, a simple group the statement follows.

2. The primes $p = 2, 3, 5, 17$ are the only primes with the property that no prime $q \geq 5$ divides $p^2 - 1$. In the other cases we use the system of generators from (40) which is of type $(4, 6, p)$ together with one of the system of generators from (42) or (44) to obtain generators satisfying Conditions $1, 2', 3, 4$ of the previous lemma. Since $\mathbf{SL}(2, \mathbb{F}_p)$ is a perfect group (for $p \neq 2, 3$) the statement follows. □

Questions of reality

Let G be a finite group and $u = (G^0; a, c) \in \mathbb{M}(G) = \mathbb{M}_3(G)$. In analogy with (20) we define

$$\iota(u) := (G^0; a^{-1}, c^{-1}) \tag{36}$$

and infer from Proposition 2.4:

$$S(\iota(u)) = \overline{S(u)}. \tag{37}$$

From Proposition 3.3 we get

Proposition 4.7. *Let G be a finite group and $u \in \mathbb{M}(G)$, then*

1. *$S(u)$ is biholomorphic to $\overline{S(u)}$ if and only if $\iota(u)$ is in the $\mathrm{A}_{\mathbb{M}}(G)$-orbit of u,*
2. *$S(u)$ is real if and only if there exists $\rho \in \mathrm{A}_{\mathbb{M}}(G)$ with $\rho(u) = \iota(u)$ and $\rho(\iota(u)) = u$.*

Observe that if a mixed Beauville surface S is isomorphic to its conjugate, then necessarily the same holds for its natural unmixed double cover S^0.

Now we formulate an algebraic condition on $u \in \mathbb{M}(G)$ which will allow us to show that the associated Beauville surface $S(u)$ is not isomorphic to $\overline{S(u)}$.

Corollary 4.8. *Let G be a finite group, $u = (G^0; a, c) \in \mathbb{M}(G)$ and assume that (a, c) is a strict system of generators for G^0. Then $S(u) \cong \overline{S(u)}$ if and only if there is an automorphism ψ of G such that $\psi(G^0) = G^0$ and $\psi(a) = a^{-1}$, $\psi(c) = c^{-1}$.*

Proof. Follows from Proposition 4.7 in the same way as Corollary 3.13 follows from Proposition 3.11. □

We now give an alternative description of the conclusion of Corollary 4.8.

Remark 4.9. With the assumptions of Corollary 4.8, let $g \in G$ represent the nontrivial coset of G^0 in G. Set $\tau := \tau_g = g^2$ and $\varphi := \varphi_g$. Then $S(u) \cong \overline{S(u)}$ if and only if there is an automorphism β of G^0 such that $\beta(a) = a^{-1}$, $\beta(c) = c^{-1}$, and an element $\gamma \in G^0$ such that $\tau(\beta(\tau^{-1})) = \varphi(\gamma)\gamma$.

Proof. $S \cong \bar{S}$ if and only if $C_1 \times C_2$ admits an antiholomorphism σ which normalizes the action of G. Since there are biholomorphisms of $C_1 \times C_2$ which exchange the factors (and lie in G), we may assume that such an antiholomorphism does not exchange the two factors. Being of product type $\sigma = \sigma_1 \times \sigma_2$, it must normalize the product group $G^0 \times G^0$. We get thus a pair of automorphisms β_1, β_2 of G. Since $\beta_1 \times \beta_2$ leaves the subgroup $\{(\gamma, \varphi(\gamma)) \mid \gamma \in G\}$ invariant , it follows that $\beta_2 = \varphi\beta_1\varphi^{-1}$, and in particular β_2 carries $a' := \varphi(a), c' := \varphi(c)$ to their respective inverses.

Now, $\sigma_1 \times \sigma_2$ normalizes the whole subgroup G if and only if for each $\epsilon \in G^0$ there is $\delta \in G^0$ such that

$$\sigma_1\, \varphi(\epsilon)\sigma_2^{-1} = \varphi(\delta)\sigma_2(\tau\epsilon)\sigma_1^{-1} = \tau\delta.$$

We use now the strictness of the structure: this ensures that both σ_i's are liftings of the standard complex conjugation, whence we easily conclude that there is an element $\gamma \in G^0$ such that $\sigma_2 = \gamma\sigma_1$.

From the second equation we conclude that $\delta = \tau^{-1}\gamma\sigma_1\tau\epsilon\sigma_1^{-1}$, and the first then boils down to $\sigma_1(\varphi(\epsilon))\sigma_1^{-1}\gamma^{-1} = \tau^{-1}(\varphi(\gamma))\gamma\sigma_1\tau(\varphi(\epsilon))\sigma_1^{-1}\gamma^{-1}$.

Since this must hold for all $\epsilon \in G^0$, it is equivalent to require $\sigma_1\tau^{-1}\sigma_1^{-1} = \tau^{-1}(\varphi(\gamma))\gamma$, i.e., $\tau(\beta(\tau^{-1})) = (\varphi(\gamma))\gamma$. □

We now give examples of mixed Beauville structures. In the proofs we use that every automorphism of $\mathbf{SL}(2, \mathbb{F}_p)$ (p a prime) is induced by an inner automorphism of the larger group

$$\mathbf{SL}^{\pm 1}(2, \mathbb{F}_p) := \langle \mathbf{SL}(2, \mathbb{F}_p),\ W := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \rangle. \tag{38}$$

See the appendix of [7] for a proof of this fact. We also use the following lemma which is easy to prove.

Lemma 4.10.

1. *Let H be a perfect group. Every automorphism $\psi : H_{[4]} \to H_{[4]}$ satisfies $\psi(H \times H \times \{0\}) = H \times H \times \{0\}$.*

2. *If H is a nonabelian simple finite group, then every automorphism of $H \times H$ is of product type.*
3. *Let H be $\mathbf{SL}(2, \mathbb{F}_p)$, where p is prime; then every automorphism of $H \times H$ is of product type.*

Proof. 1. $H \times H \times \{0\}$ is the commutator subgroup.

2. the centralizer $C((x, y))$ of an element (x, y) where $x \neq 1, y \neq 1$ does not map surjectively onto H through either of the two product projections, whence every automorphism leaves invariant the unordered pair of subgroups $\{(H \times \{0\}), (\{0\} \times H)\}$.

3. follows by the same argument used for 2. □

We apply the above constructions to obtain some concrete examples.

Proposition 4.11. *Let p be a prime with $p \equiv 3$ mod 4 and $p \equiv 1$ mod 5 and consider the group $H := \mathbf{SL}(2, \mathbb{F}_p)$. Then $H_{[4]}$ admits a mixed Beauville structure u such that $\iota(u)$ does not lie in the $\mathrm{A}_{\mathbb{M}}(H_{[4]})$-orbit of u.*

Proof. Set $a_1 := B$, $c_1 := S$ as defined in (40) and a_2, c_2 one of the systems of generators constructed in Proposition 5.13. That is, the equations

$$\gamma a_2 \gamma^{-1} = a_2^{-1}, \quad \gamma c_2 \gamma^{-1} = c_2^{-1}$$

are solvable with $\gamma \in \mathbf{SL}(2, \mathbb{F}_p)$ but not with $\gamma \in \mathbf{SL}(2, \mathbb{F}_p)W$. Set $a := (a_1, a_2, 2)$, $c := (c_1, c_2, 2)$. By Lemma 4.5 the triple $u := (H_{[2]}, a, c)$ is a mixed Beauville structure on $H_{[4]}$. The type of (a, c) is $(20, 30, 5p)$, hence it is strict.

Suppose that $\iota(u)$ is in the $\mathrm{A}_{\mathbb{M}}(H_{[4]})$-orbit of u. By Corollary 4.8 we have an automorphism $\psi : H_{[4]} \to H_{[4]}$ with $\psi(H_{[2]}) = H_{[2]}$ with $\psi(a) = a^{-1}$ and $\psi(c) = c^{-1}$. From Lemma 4.10 we get two elements $\gamma_1, \gamma_2 \in \mathbf{SL}^{\pm 1}(2, \mathbb{F}_p)$ with

$$\gamma_1 a_1 \gamma_1^{-1} = a_1^{-1},\ \gamma_1 c_1 \gamma_1^{-1} = c_1^{-1},\ \gamma_2 a_2 \gamma_2^{-1} = a_2^{-1},\ \gamma_2 c_2 \gamma_2^{-1} = c_2^{-1}.$$

Since they come from the automorphism $\psi : H_{[4]} \to H_{[4]}$ they have to lie in the same coset of $\mathbf{SL}(2, \mathbb{F}_p)$ in $\mathbf{SL}^{\pm 1}(2, \mathbb{F}_p)$. This is impossible since $\gamma_1 a_1 \gamma_1^{-1} = a_1^{-1}$, $\gamma_1 c_1 \gamma^{-1} = c_1^{-1}$ is only solvable in the coset $\mathbf{SL}(2, \mathbb{F}_p)W$ as a computation shows. □

Proof (of Theorem 1.5). We take the mixed Beauville structure from Proposition 4.11 and let S be the corresponding mixed Beauville surface as constructed in Section 4.1. □

5 Generating groups by two elements

Symmetric groups

In this section we provide a series of intermediate results which lead to the proof of Theorem 1.4. In fact we prove:

Proposition 5.1. *Let $n \in \mathbb{N}$ satisfy $n \geq 8$ and $n \equiv 2$ mod 3, then $\mathcal{S}_n$ has systems of generators (a, c), (a', c') with*

1. *$\Sigma(a, c) \cap \Sigma(a', c') = \{1\}$,*
2. *there is no $\gamma \in \mathcal{S}_n$ with $\gamma a \gamma^{-1} = a^{-1}$ and $\gamma c \gamma^{-1} = c^{-1}$.*

Lemma 5.2. *Let $G = \mathcal{S}_n$ with $n \geq 7$,*

$$a := (5, 4, 1)(2, 6) \quad \text{and} \quad c := (1, 2, 3)(4, 5, 6, \ldots, n).$$

There is no automorphism of G carrying $a \to a^{-1}$, $c \to c^{-1}$.

Proof. Since $n \neq 6$, every automorphism of G is an inner one. If there is a permutation g conjugating a to a^{-1}, c to c^{-1}, g would leave each of the sets $\{1, 2, 3\}$, $\{4, 5, \ldots, n\}$, $\{1, 4, 5\}$, $\{2, 6\}$ invariant. By looking at their intersections we conclude that g leaves the elements $1, 2, 3, 6$ fixed and that the set $\{4, 5\}$ is invariant. But then g conjugates c to $(1, 2, 3)h$, where h is a permutation of $\{4, 5, \ldots, n\}$, which is a different permutation than c^{-1}. □

Lemma 5.3. *The two elements $a := (5, 4, 1)(2, 6)$, $c := (1, 2, 3)(4, 5, 6, \ldots, n)$ generate $\mathcal{S}_n$ if $n \geq 7$ and $n \neq 0$ mod 3.*

Proof. Let G be the subgroup generated by a, c. Then G is generated also by s, α, T, γ, where $s := (2, 6)$, $\alpha := (5, 4, 1)$, $T := (1, 2, 3)$, $\gamma := (4, 5, 6, \ldots, n)$, since these elements are powers of a, c and 3 and $n - 3$ are relatively prime.

Since G contains a transposition, it suffices to show that it is doubly transitive. The transitivity of G being obvious, since the supports of the cyclic permutations s, α, T, γ have the whole set $\{1, 2, \ldots, n\}$ as union. The subgroup $H \subset G$ stabilizing $\{3\}$ contains s, α, γ. Again these are cyclic permutations such that their supports have as union the set $\{1, 2, 4, 5, \ldots, n\}$. Thus G is doubly transitive, whence $G = \mathcal{S}_n$. □

Remark 5.4. • Since $3 \nmid n$, one has $\mathrm{ord}(c) = 3(n - 3)$, while $\mathrm{ord}(a) = 6$.

- We calculate now $\mathrm{ord}(b)$, recalling that $abc = 1$, whence b is the inverse of ca. Since $ca = (1, 6, 3)(4, 2, 7, \ldots n)$ we have $\mathrm{ord}(b) = \mathrm{lcm}(3, n - 4)$.
- Recalling that $a' := \sigma^{-1}$, $c' := \tau\sigma^2$, where $\tau := (1, 2)$ and $\sigma := (1, 2, \ldots, n)$, it follows immediately that a', c' generate the whole symmetric group.
- We have $\mathrm{ord}(b') = \mathrm{ord}(c'a') = \mathrm{ord}(\tau\sigma) = \mathrm{ord}((2, 3, \ldots, n)) = n - 1$, $\mathrm{ord}(a') = \mathrm{ord}(\sigma) = n$.
- If $n = 2m$, then $c' = (1, 2)(1, 3, 5, \ldots, 2m-1)(2, 4, 6, \ldots, 2m)$ is the cyclical permutation $c' = (2, 4, \ldots, 2m, 1, 3, \ldots, 2m - 1)$ and $\mathrm{ord}(c') = n$.
- If $n = 2m + 1$, then $c' = (1, 2)(1, 3, 5, \ldots, 2m + 1, 2, 4, 6, \ldots, 2m) = (1, 3, 5, \ldots, 2m + 1)(2, 4, 6, \ldots, 2m)$ and $\mathrm{ord}(c') = m(m + 1)$.

Proposition 5.5. *Let a, b, c, a', b', c' be as above, then*

$$\Sigma(a, c) \cap \Sigma(a', c') = \{1\}.$$

Proof. We say that a permutation has type $(d_1 \leq \cdots \leq d_k)$, with $d_i \geq 2 \; \forall \, i$, if its cycle decomposition consists of k cycles of respective lengths $d_1, \ldots, d_k$. We say that the type is monochromatic if all the d_i's are equal, and dichromatic if the number of distinct d_i's is exactly two. Two permutations are conjugate to each other iff their types are the same. We say that a type $(p_1 \leq \cdots \leq p_r)$ is derived from $(d_1 \leq \cdots \leq d_k)$ if it is the type of a power of a permutation of type $(d_1 \leq \cdots \leq d_k)$.

Therefore we observe that the types of $\Sigma(a,c)$ are those derived from $(2,3)$, $(3,n-3)$, $(3,n-4)$, while those of $\Sigma(a',c')$ are those derived from $(n), (n-1)$ for n even, and also from $(m,m+1)$ in the case where $n = 2m+1$ is odd.

We use then the following lemma whose proof is straightforward.

Lemma 5.6. *Let g be a permutation of type (d_1, d_2). Then the type of g^h is the reshuffle of h_1-times d_1/h_1 and h_2-times d_2/h_2, where $h_i := \gcd(d_i, h)$. Here, reshuffling means throwing away all the numbers equal to 1 and arranging the others in increasing order.*

In particular, if the type of g^h is dichromatic, (d_1, d_2) are automatically determined. If moreover d_1, d_2 are relatively prime and the type of g^h is monochromatic, then it is derived from type d_1 or from type d_2.

For types in $\Sigma(a',c')$, we get types derived from

$$(n), \; (n-1), \quad \text{or} \quad ((n-1)/2, (n+1)/2).$$

The latter come from relatively prime numbers, whence they can never equal a type in $\Sigma(a,c)$, derived from the pairs $(2,3)$, $(3,n-4)$ and $(3,n-3)$. The monochromatic types in $\Sigma(a,c)$ can only be derived by (3), (2), $(n-4)$, $(n-3)$, since we are assuming that 3 does neither divide n nor $n-1$. □

Alternating groups

In this section we construct certain systems of generators of the alternating groups $\mathcal{A}_n$ $(n \in \mathbb{N})$. Our principal tool is the theorem of Jordan, see [15]. This result says that $\langle a, c \rangle = \mathcal{A}_n$ for any pair $a, c \in \mathcal{A}_n$ which satisfies

- the group $H := \langle a, c \rangle$ acts primitively on $\{1, \ldots, n\}$,
- the group H contains a q-cycle for a prime $q \leq n-3$.

A further result that we shall need is:

Lemma 5.7. *For $n \in \mathbb{N}$ with $n \geq 12$, let $U \leq \mathcal{A}_n$ be a doubly transitive group. If U contains a double-transposition, then $U = \mathcal{A}_n$.*

Proof. The degree $m(\sigma)$ of a permutation $\sigma \in \mathcal{A}_n$ is the number of elements moved by σ. Let $\sigma \in U$ be a double-transposition. We have $m(\sigma) = 4$. Let m now be the minimal degree taken over all nontrivial elements of U. Since U is

also primitive we may apply a result of de Séguier (see [15], page 43) which says that if $m > 3$ (in our situation we would have m=4) then

$$n < \frac{m^2}{4}\log\frac{m}{2} + m\left(\log\frac{m}{2} + \frac{3}{2}\right). \tag{39}$$

For $m = 4$ the right-hand side of (39) is roughly 11.5. Our assumptions imply that $m = 3$. We then apply Jordan's theorem to reach the desired conclusion. □

We now construct the systems of generators required for the constructions of Beauville surfaces. We treat permutations as maps which act from the left and use the notation $g^\gamma := \gamma g \gamma^{-1}$ for the conjugate of an element g.

Proposition 5.8. *Let $n \in \mathbb{N}$ be even with $n \geq 16$ and let $3 \leq p \leq q \leq n-3$ be primes with $n - q \not\equiv 0$ mod p. Then there is a system (a, c) of generators for $\mathcal{A}_n$ of type $(q, p(n-q), n-p+2)$ such that there is no $\gamma \in \mathcal{S}_n$ with $\gamma a \gamma^{-1} = a^{-1}$ and $\gamma c \gamma^{-1} = c^{-1}$.*

Proof. Set $k := n - q$ and define

$$a := (1, 2, \ldots, q),\ c := (q+1, q+2, \ldots, q+k-1, 1)(q+k, p, p-1, \ldots, 2).$$

We compute

$$ca = (1, q+k, p, p+1, \ldots, q+k-1)$$

and the statement about the type is clear. We show that there is no $\gamma \in \mathcal{S}_n$ with the above properties. Otherwise, γ would leave invariant the three sets corresponding to the nontrivial orbits of a, respectively c, and in particular we would have $\gamma(1) = 1$, $\gamma(\{2, \ldots, p\}) = \{2, \ldots, p\}$. But then $\gamma(2) \neq q$, a contradiction. We set $U := \langle a, c\rangle$ and show that $U = \mathcal{A}_n$.

Obviously U is transitive. The stabilizer V of $q + k$ in U contains the elements a, c^p. It is clear that the subgroup generated by these two elements is transitive on $\{1, \ldots, n\} \setminus \{q+k\}$, hence U is doubly transitive. The group U contains the q-cycle a, whence we infer by Jordan's theorem that $U = \mathcal{A}_n$. □

For the applications in the previous sections we need:

Proposition 5.9.

1. *Let $n \in \mathbb{N}$ satisfy $n \geq 16$ with $n \equiv 0 \bmod 4$ and $n \equiv 1 \bmod 3$. There is a pair (a, c) of generators of $\mathcal{A}_n$ of type $(2, 3, 84)$ and an element $\gamma \in \mathcal{S}_n \setminus \mathcal{A}_n$ with $\gamma a \gamma^{-1} = a^{-1}$ and $\gamma c \gamma^{-1} = c^{-1}$.*
2. *Let p be a prime with $p > 5$. Set $n := 3p + 1$. There is a pair (a, c) of generators of $\mathcal{A}_n$ of type $(p, 5p, 2p+3)$ and $\gamma \in \mathcal{S}_n$ with $\gamma a \gamma^{-1} = a^{-1}$ and $\gamma c \gamma^{-1} = c^{-1}$. If $p \equiv 1$ mod 4, then γ can be chosen in $\mathcal{S}_n \setminus \mathcal{A}_n$; if $p \equiv 3$ mod 4, then γ can be chosen in $\mathcal{A}_n$.*

Proof. 1. Take as the reference set $\{0,\dots,n-1\}$ instead of $\{1,\dots,n\}$ and set

$$\gamma := (0)(1)\prod_{i=1}^{\frac{n-4}{6}}(6i-2,6i+1)\cdot(2,3)\cdot\prod_{i=1}^{\frac{n-4}{6}}(6i-1,6i+3)(6i,6i+2),$$

$$a := (0,1)\prod_{i=1}^{\frac{n-4}{6}}(6i-2,6i+1)\cdot\prod_{i=1}^{\frac{n-4}{6}}(6i-4,6i-1)\cdot\prod_{i=1}^{\frac{n-4}{6}}(6i-4,6i-1)^{\gamma}(n-2,n-4),$$

$$c := (0)\prod_{i=1}^{\frac{n-1}{3}}(3i-2,3i-1,3i).$$

Observe that

$$\prod_{i=1}^{\frac{n-4}{6}}(6i-4,6i-1)^{\gamma} = \prod_{i=2}^{\frac{n-4}{6}}(6i-6,6i+3)\cdot(3,9).$$

We have now

$$ca = (0,2,6,13,11,9,1)\cdot(3,7,5)\cdot\prod_{i=1}^{\frac{n-16}{6}}(6i-2,6i+2,6i+6,6i+13,6i+11,6i+9)$$

$$\cdot\,(n-1,n-12,n-8,n-4)\cdot(n-2,n-6).$$

We have $\mathrm{ord}(a)=2$, $\mathrm{ord}(c)=3$, $\mathrm{ord}(ca)=84$, $\gamma a\gamma^{-1}=a^{-1}$ and $\gamma c\gamma^{-1}=c^{-1}$. To apply the theorem of Jordan we note that $(ca)^{12}$ is a 7-cycle. It remains to show that $H:=\langle a,c\rangle$ acts primitively. In Figure 1 we exhibit the orbits of a and c. In the righthand picture we connect two elements of $\{0,\dots,n-1\}$ if they are in the orbit of a, in the lefthand picture similarly for c.

Figure 1 makes it obvious that H acts transitively. Since $(ca)^{12}$ is a 7-cycle, the second condition of Jordan's theorem is fulfilled. Let H_0 be the stabilizer of 0. Then c and $(ca)^7$ are contained in H_0. In Figure 2 we show the orbits of the two elements c and $(ca)^7$. The notation is the same as in Figure 1. A glance at Figure 2 shows that H_0 is transitive on $\{1,\dots,n-1\}$. We infer that H is doubly transitive. Again by Theorem 9.6 of [15] the group H is primitive.

2. Again we take as set of reference $\{0,\dots,n-1\}$ instead of $\{1,\dots,n\}$ and set

$$\gamma := (0)(1)(p+1,2p+1)\cdot\prod_{i=1}^{\frac{p-1}{2}}(1+i,p+1-i)\cdot\prod_{i=1}^{p-1}(p+1+i,3p+1-i),$$

$$a := (0)(1,2,\dots,p)(p+1,p+2,\dots,2p)(2p+1,2p+2,\dots,3p),$$

$$c := (0,p,p-1,p-2,\dots,3,2)(1,p+1,3p,p+2,2p+1).$$

We have

$$ca := (2)(3)\dots(p-1)(0,p,p+1,2p+1,2p+2,\dots,3p-1,p+2,p+3,\dots,2p,3p,1).$$

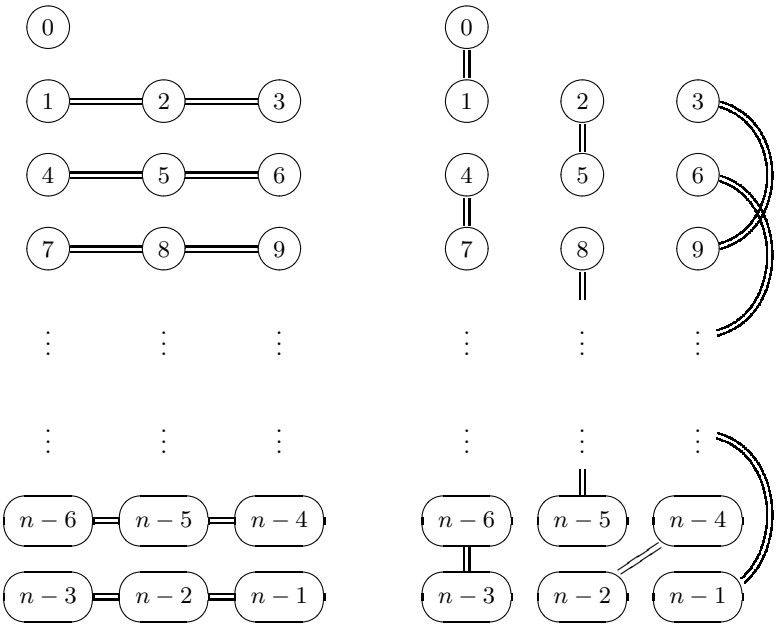


Fig. 1. The orbits of c, a

From this definition we see that $\mathrm{ord}(a) = p$, $\mathrm{ord}(c) = 5p$ and $\mathrm{ord}(ca) = 2p + 3$. The formulae $\gamma a \gamma^{-1} = a^{-1}$ and $\gamma c \gamma^{-1} = c^{-1}$ are also clear.

We verify the conditions of Jordan's theorem. First of all c^5 is a p-cycle. It remains to show that $H := \langle a, c \rangle$ acts primitively. In Figure 3 we exhibit the orbits of a, respectively the orbits of c below. From this it is obvious that H acts transitively.

Let now $\Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_k = \{0, 1, \ldots, 3p\}$ be a block-decomposition for H with $0 \in \Delta_1$. Note that $|\Delta_i | k = 3p + 1$ for all i. Since a is in the stabilizer of 0, we have $|\Delta_1| \in \{1, p + 1, 2p + 1, 3p + 1\}$. Since $|\Delta_1|$ divides $3p + 1$, we infer that H acts primitively. □

Further systems of generators are needed:

Proposition 5.10. *If $n = 2k \geq 16$, then there is a system of generators (a, c) of $\mathcal{A}_n$ of type $(2k - 3, 2k - 2, 2k - 2)$ and $\gamma \in \mathcal{S}_n$ such that*

$$\gamma a \gamma^{-1} = a^{-1}, \quad \gamma c \gamma^{-1} = ac.$$

If k is even, then γ can be chosen in $\mathcal{S}_n \setminus \mathcal{A}_n$; if k is odd, then γ can be chosen in $\mathcal{A}_n$. The system of generators (a, c) has the further property that there is no $\delta \in \mathcal{S}_n$ with

$$\delta a \delta^{-1} = a^{-1} \text{ and } \delta c \delta^{-1} = c^{-1} \quad \text{or} \quad \delta a \delta^{-1} = c^{-1} \text{ and } \delta c \delta^{-1} = a^{-1}.$$

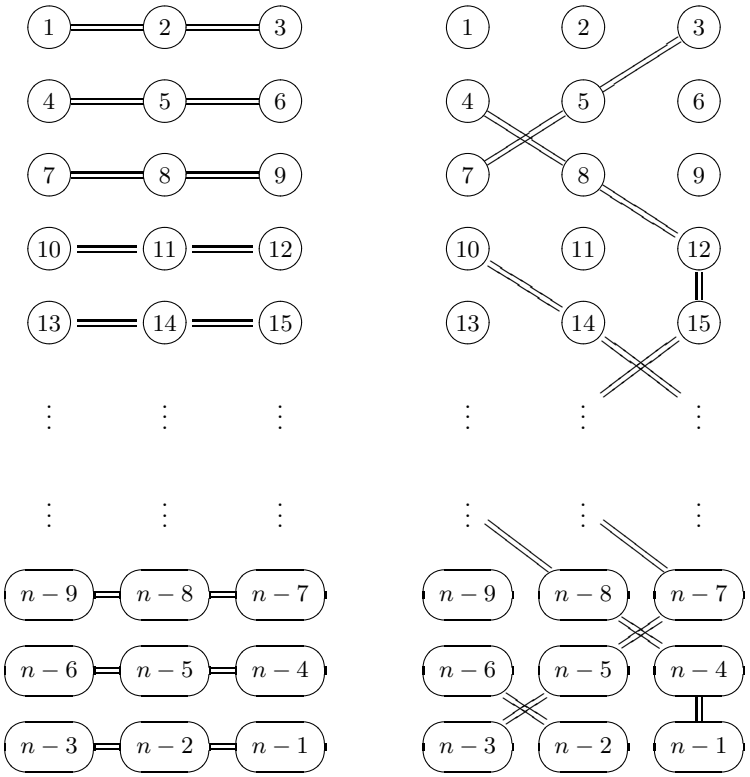


Fig. 2. The orbits of c and $(ca)^7$

Proof. We set

$$a := (1, 2, \ldots, 2k-4, 2k-3),$$

$$d := (1,2,3)(2k-3, 2k-4, 2k-5)(k-1, 2k-1)(2k-2, k-2, 2k, k)$$

and

$$\alpha := (1, 2k-3)(2, 2k-4)(3, 2k-5)\ldots(k-2, k) \cdot (2k-2, 2k).$$

The following are clear:

- d^6 is a double-transposition,
- α is in $\mathcal{A}_n$ if k is odd, and in $\mathcal{S}_n \setminus \mathcal{A}_n$ if k is even,
- $\alpha a \alpha^{-1} = a^{-1}$ and $\alpha d \alpha^{-1} = d$,
- there is no $\delta \in \mathcal{S}_n$ with $\delta a \delta^{-1} = a$ and $\delta d \delta^{-1} = d^{-1}$.

For the last item note that a has $\{2k-2, 2k-1, 2k\}$ as its set of fixed points. A δ with the above property would have to stabilize this set. The element d interchanges $k-1$ and $2k-1$, hence both these elements have to be fixed

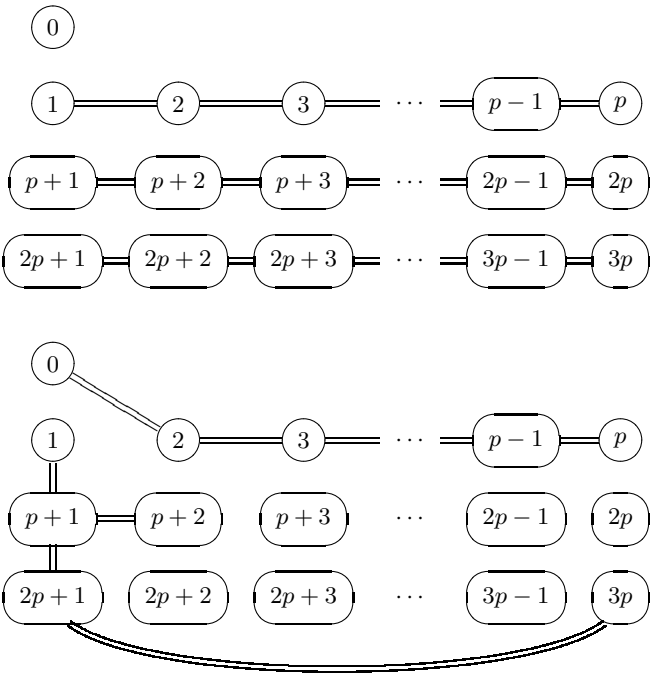


Fig. 3. The orbits of a und c

by δ. The condition $\delta a \delta^{-1} = a$ implies then that δ acts as the identity on $\{1, 2, \ldots, 2k-3\}$, in particular as the identity on the subset $\{1, 2, 3\}$, contradicting $\delta d \delta^{-1} = d^{-1}$. Set

$$c := da^{k-2} = (1, 2k-1, k-1, 2k-4, k-3, 2k-3, 2k, k,$$

$2, 2k-2, k-2, 2k-5, k-4, 2k-6, k-5, 2k-7, \ldots, 5, k+3, 4, k+2)(3, k+1)$.

Note that $\mathrm{ord}(c) = 2k-2$. We find

$$\alpha c \alpha^{-1} = \alpha d a^{k-2} \alpha^{-1} = d a^{-k+2} = c a^{-2k+4} = ca,$$

whence also $\mathrm{ord}(ca) = 2k-2$. Set now $\gamma := a\alpha$, so that γ clearly satisfies the required properties.

Set $U := \langle a, c \rangle$. We shall now show that $U = \mathcal{A}_n$. Clearly, U is a transitive group. Let $V \leq U$ the stabilizer of $2k-1$. The subgroup V contains

$$a,\ d^2,\ da^{k-2} d a^{4-k} d.$$

From the definitions it is clear that these elements generate a group which is transitive on $\{1, \ldots, 2k\} \setminus \{2k-1\}$. Hence U is doubly transitive and contains a double transposition. From Lemma 5.7 we infer that $U = \mathcal{A}_n$.

The last property follows from the above items since in the first case we would have $a \to a^{-1}, d \to (\mathrm{Int}_a)^{-(k-2)}(d^{-1})$, and composing with $\mathrm{Int}_{\alpha \cdot a^{k-2}}$ we contradict the third item.

Whereas, in the second case, just observe that a and c have different order. □

SL(2) and PSL(2) over finite fields

In this section we give systems of generators consisting of two elements of the respective groups $\mathbf{SL}(2, \mathbb{F}_p)$ and $\mathbf{PSL}(2, \mathbb{F}_p)$ which will allow us to construct certain Beauville structures on them.

If p is a prime we denote by $\mathbb{F}_p$, $\mathbb{F}_{p^2}$ the fields with p, respectively p^2 elements and by $\mathbb{F}_p^*$, $\mathbb{F}_{p^2}^*$ the corresponding multiplicative groups. We let

$$N_{\mathbb{F}_{p^2}/\mathbb{F}_p} : \mathbb{F}_{p^2}^* \to \mathbb{F}_p^*$$

be the norm map. We also introduce the matrices

$$B := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \; S := \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}, \; T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \; W := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{40}$$

in $\mathbf{GL}(2, \mathbb{F}_p)$. For $\lambda \in \mathbb{F}_p$ with $\lambda \neq 0$ and $k \in \mathbb{F}_p$ we define

$$D(\lambda) := \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}, \quad M(k) := \begin{pmatrix} 0 & 1 \\ -1 & k \end{pmatrix}.$$

We have

$$B^4 = S^6 = W^2 = 1, \; T = BS, \; WBW^{-1} = B^{-1}, \; WSW^{-1} = S^{-1}.$$

The matrices (B, S) form a system of generators of $\mathbf{SL}(2, \mathbb{F}_p)$ of type $(4, 6, p)$. Their images in $\mathbf{PSL}(2, \mathbb{F}_p)$ form a system of generators of type $(2, 3, p)$.

Proposition 5.11. *Let p be an odd prime and let $q \geq 5$ be a prime with $q|p-1$. Let $\lambda \in \mathbb{F}_p^*$ be of order q: then $\lambda + \lambda^{-1} - 2 \neq 0$, $\lambda^2 + \lambda^{-2} - 2 \neq 0$ and $\lambda + \lambda^{-1} - \lambda^2 - \lambda^{-2} \neq 0$. Set*

$$g := \begin{pmatrix} 1 & b \\ 1 & d \end{pmatrix} \quad \text{with } b := \frac{\lambda + \lambda^{-1} - \lambda^2 - \lambda^{-2}}{\lambda^2 + \lambda^{-2} - 2}, \; d := \frac{\lambda + \lambda^{-1} - 2}{\lambda^2 + \lambda^{-2} - 2}. \tag{41}$$

Then

$$D(\lambda), \quad gD(\lambda)g^{-1} \tag{42}$$

form a system of generators of $\mathbf{SL}(2, \mathbb{F}_p)$ of type (q, q, q). Set further

$$e(\lambda) := \frac{2 - \lambda - \lambda^{-1}}{\lambda + \lambda^{-1} - \lambda^2 - \lambda^{-2}}. \tag{43}$$

There exists $\gamma \in \mathbf{SL}(2, \mathbb{F}_p)$ with $\gamma D(\lambda)\gamma^{-1} = D(\lambda)^{-1}$ and $\gamma g D(\lambda) g^{-1}\gamma^{-1} = gD(\lambda)^{-1}g^{-1}$ if and only if $e(\lambda)$ is a square in $\mathbb{F}_p$. There is $\gamma \in \mathbf{SL}(2, \mathbb{F}_p)W$ satisfying these conditions if and only if $-e(\lambda)$ is a square in $\mathbb{F}_p$.

Proof. Let us prove that $\lambda + \lambda^{-1} - 2 = 0$ leads to a contradiction. In fact, multiplying by λ we find $(\lambda - 1)^2 = 0$ which is impossible since λ has order q. The other two cases are treated similarly.

We see immediately that the determinant of g is equal to 1, furthermore an easy computation shows that the trace of $h := D(\lambda)gD(\lambda)g^{-1}$ is $\lambda + \lambda^{-1}$, hence h has also order q. Notice further that the subgroup H generated by $D(\lambda)$ and $gD(\lambda)g^{-1}$ cannot be solvable because these two elements have no common fixpoint in the action on $\mathbb{P}^1_{\mathbb{F}_p}$. From the list of isomorphism classes of subgroups of $\mathbf{PSL}(2, \mathbb{F}_p)$ given in [10] (Hauptsatz 8.27, page 213) we find that H has to be $\mathbf{SL}(2, \mathbb{F}_p)$ if $q \neq 5$. For $q = 5$ these simple arguments with subgroup orders leave the possibility that H is isomorphic to the binary icosahedral group $2 \cdot \mathcal{A}_5 \leq \mathbf{SL}(2, \mathbb{F}_p)$. But this group does not have a system of generators of type $(5, 5, 5)$, as is easily seen by computer calculation.

The assertion about the simultaneous conjugacy of $D(\lambda)$, $D(\lambda)^{-1}$ and $gD(\lambda)g^{-1}$, $gD(\lambda)^{-1}g^{-1}$ follows by a straightforward computation. In fact we just take a matrix X with indeterminate entries and write down the nine equations resulting from $XD(\lambda) = D(\lambda)^{-1}X$ and $XgD(\lambda)g^{-1} = gD(\lambda)^{-1}g^{-1}X$, $\mathrm{Det}(X) = \pm 1$ and the statement follows by a small manipulation of them. □

Proposition 5.12. *Let p, q be odd primes such that $q \geq 5$ and $q|p + 1$. Let $\lambda \in \mathbb{F}^*_{p^2}$ be of order q (thus $N_{\mathbb{F}_{p^2}/\mathbb{F}_p}(\lambda) = 1$). Consider its trace $k := \lambda + \lambda^{-1} \in \mathbb{F}_p$: then there exists $g \in \mathbf{SL}(2, \mathbb{F}_p)$ such that*

$$M(k), \quad gM(k)g^{-1} \tag{44}$$

form a system of generators of type (q, q, q).

Proof. Choose k as indicated and notice that $k \neq 1$, $k \neq 2$, as we already saw.

Let us now change our perspective and let $R := \mathbb{Z}[r, s, t]$ treating $r = k$ as a variable. Set

$$x := \begin{pmatrix} 0 & 1 \\ -1 & r \end{pmatrix}, \; g := \begin{pmatrix} 1 & s \\ t & 1 + st \end{pmatrix}, \; y := gxg^{-1}, \; z := x \cdot y \in \mathbf{SL}(2, R).$$

The reduction into $\mathbf{SL}(2, \mathbb{F}_p)$ of x and of y have order q for any choice of s, t and $r = k$. We want $z = xy$ also to have order q. This happens if the trace of the reduction into $\mathbf{SL}(2, \mathbb{F}_p)$ of z is equal to k. This follows from $\mathrm{Tr}(z) = r$ which is equivalent to the equation

$$-s^2t^2 + s^2tr - s^2 - st^2r + str^2 - 2st - t^2 + r^2 - r - 2 = 0. \tag{45}$$

We write $C_{p,k}$ for the plane affine curve obtained from (45) by setting $r = k$ and reducing modulo p. Furthermore let $\mathbb{P}C_{p,k}$ be the projective closure of $C_{p,k}$ (with respect to s, t) and $\mathbb{P}C^{\infty}_{p,k}$ its set of points at infinity.

These are immediately seen to be the two points with $s = 0, t = 1$, respectively $t = 0, s = 1$, and in these two points at ∞ the curve has two ordinary double points.

By a computation using a Gröbner-routine over $\mathbb{Z}$ (possible in MAGMA or SINGULAR) one can also verify that the affine curve $C_{p,k}$ is smooth for every $p \geq 11$. To check this fact we projectivize (45), compute derivatives and analyze the ideal in R generated by these homogeneous polynomials. In this step we use $k \neq 1$, $k \neq 2$.

Let now $\tilde{C}_{p,k} \to \mathbb{P}C_{p,k}$ be a nonsingular model of $\mathbb{P}C_{p,k}$. We conclude from the above analysis of the singularities of $\mathbb{P}C_{p,k}$ and from Bézout's theorem that the nonsingular curve $\tilde{C}_{p,k}$ is absolutely irreducible and has genus 1. We may then apply the Hasse–Weil estimate (cf. e.g., the textbook [9], V 1.10, page 368) to obtain

$$||\tilde{C}_{p,k}(\mathbb{F}_p)| - p - 1| \leq 2\sqrt{p},$$

and since there are at most two $\mathbb{F}_p$-points over every singular point of $\mathbb{P}C_{p,k}$ we get at worst

$$||C_{p,k}(\mathbb{F}_p)| - p + 3| \leq 2\sqrt{p}.$$

Hence we have $C_{p,k}(\mathbb{F}_p) \neq \emptyset$ for $p \geq 11$.

Let now (s,t) be in $C_{p,k}(\mathbb{F}_p)$ and x, y, z the corresponding matrices defined above. Notice that all three of them have order q. It can be checked, again by a Gröbner-routine over $\mathbb{Z}$, that we have $y \neq \pm x$ and $y \neq \pm x^{-1}$. Assume that $q > 5$. A glance at the subgroups of $\mathbf{PSL}(2, \mathbb{F}_p)$ ([10]) shows that the subgroup generated by x, y could only be cyclic, which is impossible by the remarks just made. If $q = 5$ we conclude by observing again that $2 \cdot \mathcal{A}_5 \leq \mathbf{SL}(2, \mathbb{F}_p)$ does not have a system of generators of type $(5, 5, 5)$. □

In order to use Proposition 5.11 effectively for our problems we would have to show that the invariant e from (43) takes both square and nonsquare values as λ varies over all elements of order q. This leads to a difficult problem about exponential sums which we could not resolve. In case $q = 5$ we found the following way to treat the problem by a simple trick.

Proposition 5.13. *Let p be a prime with $p \equiv 3$ mod 4 and $p \equiv 1$ mod 5. Then the group $\mathbf{SL}(2, \mathbb{F}_p)$ has a system of generators (a, c) of type $(5, 5, 5)$ such that the equations*

$$\gamma a \gamma^{-1} = a^{-1}, \quad \gamma c \gamma^{-1} = c^{-1} \tag{46}$$

are solvable with $\gamma \in \mathbf{SL}(2, \mathbb{F}_p)$ but not with $\gamma \in \mathbf{SL}(2, \mathbb{F}_p)W$. The same group has another system of generators (a, c) such that (46) is solvable in $\mathbf{SL}(2, \mathbb{F}_p)W$ but not in $\mathbf{SL}(2, \mathbb{F}_p)$.

Proof. Take $\lambda \in \mathbb{F}_p$ with $\lambda^5 = 1$, $\lambda \neq 1$ and consider the system of generators given in (42). Since $p \equiv 3 \bmod 4$ the number -1 is not a square in $\mathbb{F}_p$. Suppose that the invariant $e(\lambda)$ is a square in $\mathbb{F}_p$ whence(46) is solvable in $\mathbf{SL}(2, \mathbb{F}_p)$ and not in $\mathbf{SL}(2, \mathbb{F}_p)W$ (see Proposition 5.11). Then we are done. Suppose instead that the invariant $e(\lambda)$ is not a square in $\mathbb{F}_p$. We replace λ by λ^2 and find by a small computation that $e(\lambda) = -e(\lambda^2)$ up to squares. In this place

we use $\lambda^5 = 1$,whence $(\lambda^2)^2 = \lambda^{-1}$ and the denominator simply changes sign as we replace λ by λ^2. Notice also that

$$2 - \lambda - \lambda^{-1} = -(\mu - \mu^{-1})^2 \qquad (\mu^2 = \lambda)$$

is never a square. We infer that $e(\lambda^2)$ is a square and proceed as before.

The second statement is proved similarly. □

Other groups and more generators

In this subsection we report on computer experiments related to the existence of unmixed or mixed Beauville structures on finite groups. We also try to formulate some conjectures concerning these questions.

We have paid special attention to unmixed Beauville structures on finite nonabelian simple groups.

The smallest of these groups is $\mathcal{A}_5 \cong \mathbf{PSL}(2, \mathbb{F}_5)$. This group cannot have an unmixed Beauville structure. On the one hand it has only elements of orders $1, 2, 3, 5$. It is not solvable hence it cannot be a quotient group of one of the euclidean triangle groups (see Section 6). This implies that any normalized system of generators has type $(n, m, 5)$ with $n, m \in \{2, 3, 5\}$. Finally we note that, by Sylow's theorem, all subgroups of order 5 are conjugate.

There are 47 finite simple nonabelian groups of order ≤ 50000. By computer calculations we have found unmixed Beauville structures on all of them with the exception of $\mathcal{A}_5$. This and the results of Section 3.2 leads us to:

Conjecture 5.14. *All finite simple nonabelian groups except $\mathcal{A}_5$ admit an unmixed Beauville structure.*

We have also checked this conjecture for some bigger simple groups like the Mathieu groups $\mathbf{M}12$, $\mathbf{M}22$ and also matrix groups of size bigger then 2. Furthermore we have proved:

Proposition 5.15. *Let p be an odd prime: then the Suzuki group $\mathbf{Suz}(2^p)$ has an unmixed Beauville structure.*

In the proof, which is not included here, we use in an essential way that the Suzuki groups $\mathbf{Suz}(2^p)$ are minimally simple, that is have only solvable proper subgroups. For the Suzuki groups see [11].

Let us call a type $(r, s, t) \in \mathbb{N}^3$ *hyperbolic* if

$$\frac{1}{r} + \frac{1}{s} + \frac{1}{t} < 1.$$

In this case the triangle group $T(r, s, t)$ is hyperbolic. From our studies also the following looks suggestive:

Conjecture 5.16. *Let (r, s, t), (r', s', t') be two hyperbolic types. Then almost all alternating groups $\mathcal{A}_n$ have an unmixed Beauville structure $v = (a_1, c_1; a_2, c_2)$ where (a_1, c_1) has type (r, s, t) and (a_2, c_2) has type (r', s', t').*

Let us call an unmixed Beauville structure $v = (a_1, c_1; a_2, c_2)$ on the finite group G *strongly real*, if there are $\delta_1, \delta_2 \in G$ and $\psi \in \mathrm{Aut}(G)$ with

$$(\delta_1\psi(a_1)\delta_1^{-1}, \delta_1\psi(c_1)\delta_1^{-1}; \delta_2\psi(a_2)\delta_2^{-1}, \delta_2\psi(c_2)\delta_2^{-1}) = (a_1^{-1}, c_1^{-1}; a_2^{-1}, c_2^{-1}). \tag{47}$$

If the unmixed Beauville structure v is strongly real then the associated surface $S(v)$ is real.

There are 18 finite simple nonabelian groups of order ≤ 15000. By computer calculations we have found strongly real unmixed Beauville structures on all of them with the exceptions of $\mathcal{A}_5$, $\mathbf{PSL}(2, \mathbb{F}_7)$, $\mathcal{A}_6$, $\mathcal{A}_7$, $\mathbf{PSL}(3, \mathbb{F}_3)$, $\mathbf{U}(3, 3)$ and the Mathieu group $\mathbf{M}11$. The alternating group $\mathcal{A}_8$ however has such a structure. This and the results of Section 3 leads us to:

Conjecture 5.17. *All but finitely many finite simple nonabelian groups have a strongly real unmixed Beauville structure.*

Conjectures 5.14, 5.16 and 5.17 are variations of a conjecture of Higman saying that every hyperbolic triangle group surjects onto almost all alternating groups. This conjecture was resolved positively in [8] where a related discussion can be found.

We were unable to find finite 2- or 3-groups having an unmixed Beauville structure. For $p \geq 5$ our construction (26) gives plenty of examples of p-groups having an unmixed Beauville structure.

Finally we report now on two general facts that we have found during our investigations. These are useful in the quest to find Beauville structures on finite groups.

Proposition 5.18. *Let p be an odd prime.*

1. *Let $q_2 > q_1 \geq 5$ be primes with $q_1 q_2 | p - 1$ and let $\lambda_1, \lambda_2 \in \mathbb{F}_p^*$ be of respective orders q_1 and q_2. Then there is an element $g \in \mathbf{SL}(2, \mathbb{F}_p)$ such that*
$$D(\lambda_1), \quad gD(\lambda_2)g^{-1} \tag{48}$$
form a system of generators of type $(q_1, q_2, q_1 q_2)$.
2. *Let $q_2 > q_1 \geq 5$ be primes with $q_1 q_2 | p + 1$ and let $\lambda_{1,2} \in \mathbb{F}_{p^2}^*$ with $N_{\mathbb{F}_{p^2}/\mathbb{F}_p}(\lambda_{1,2}) = 1$ be of respective orders q_1 and q_2. Then their traces $k_{1,2} := \lambda_{1,2} + \lambda_{1,2}^{-1}$ are in $\mathbb{F}_p$ and there is an element $g \in \mathbf{SL}(2, \mathbb{F}_p)$ such that*
$$M(k_1), \quad gM(k_2)g^{-1} \tag{49}$$
form a system of generators of type $(q_1, q_2, q_1 q_2)$.

Surfaces S which are not real but still are biholomorphic to their conjugate $\bar{S}$ are somewhat difficult to find. Our Theorem 1.6 gives examples using the alternating groups. We also have found:

Proposition 5.19. *Let p be an odd prime and assume that there is a prime $q \geq 7$ dividing $p+1$ such that q is not a square modulo p : then there is an unmixed Beauville surface S with group $G = \mathbf{SL}(2, \mathbb{F}_p)$ which is biholomorphic to the complex conjugate surface $\bar{S}$ but is not real.*

For the proof we turn the conditions into polynomial equations and polynomial inequalities (as in Propositions 5.11, 5.12) and then use arithmetic algebraic geometry over finite fields (in a more subtle way) as before. We do not include this here.

Remark 5.20. First examples of primes p satisfying the conditions of Proposition 5.19 are $p = 13$ with $q = 7$, $p = 37$ with $q = 19$ and $p = 41$ with $q = 7$. Let p, q be odd primes. The law of quadratic reciprocity implies that the conditions of Proposition 5.19 are equivalent to $q \equiv 3 \bmod 4$, $p \equiv 1 \bmod 4$ and $p \equiv -1 \bmod q$. Dirichlet's theorem on primes in arithmetic progressions implies that there are infinitely many such pairs (p, q).

6 The wallpaper groups

In this section we analyze finite quotients of the triangular groups

$$T(3,3,3), \qquad T(2,4,4), \qquad T(2,3,6).$$

and we will show that they do not admit any unmixed Beauville structure. We shall give two proofs of this fact, a "geometric" one, and the other in the taste of combinatorial group theory.

These are groups of motions of the euclidean plane, in fact in the classical classification they are the groups **p3**, **p4**, **p6**. Each of them contains a normal subgroup N isomorphic to $\mathbb{Z}^2$ with finite quotient.

In fact, let T be such a triangle group: then T admits a maximal surjective homomorphism onto a cyclic group C_d of order d. Here, d is respectively equal to $3, 4, 6$, and the three generators map to elements of C_d whose order equals their order in T.

It follows that the covering corresponding to T is the universal cover of the compact Riemann surface E corresponding to the surjection onto C_d, and one sees immediately two things:

1. E is an elliptic curve because $\mu(a, c) = 1$.
2. E has multiplication by the group $\mu_d \cong \mathbb{Z}/d\mathbb{Z}$ of d-roots of unity.

Letting $\omega = exp(2/3\pi i)$, we see that

- $T(3,3,3)$ is the group of affine transformations of $\mathbb{C}$ of the form

$$g(z) = \omega^j z + \eta \text{ , for } j \in \mathbb{Z}/3\mathbb{Z}, \eta \in \Lambda_\omega := \mathbb{Z} \oplus \mathbb{Z}\omega.$$

- $T(2,4,4)$ is the group of affine transformations of $\mathbb{C}$ of the form

$$g(z) = i^j z + \eta \text{ , for } j \in \mathbb{Z}/4\mathbb{Z}, \eta \in \Lambda_i := \mathbb{Z} \oplus \mathbb{Z}i.$$

- $T(2,3,6)$ is the group of affine transformations of $\mathbb{C}$ of the form

$$g(z) = (-\omega)^j z + \eta \text{ , for } j \in \mathbb{Z}/6\mathbb{Z}, \eta \in \Lambda_\omega := \mathbb{Z} \oplus \mathbb{Z}\omega.$$

Remark 6.1. Using the above affine representation, we see that N is the normal subgroup of translations, i.e., of the transformations which have no fixed point on $\mathbb{C}$.

Moreover, if an element $g \in T - N$, then the linear part of g is in $\mu_d - \{1\}$, and g has a unique fixed point p_g in $\mathbb{C}$. An immediate calculation shows that indeed this fixed point p_g lies in the lattice Λ, and we obtain in this way that the conjugacy classes of elements $g \in T - N$ are exactly given by their linear parts, so they are in bijection with the elements of $\mu_d - \{1\}$.

Let now $G = T/M$ be a nontrivial finite quotient group of T: then G admits a maximal surjective homomorphism onto a cyclic group C' of order d, where $d \in \{2,3,4,6\}$. Assume that there is an element $g \in T - N$ which lies in the kernel of the composite homomorphism: then the whole conjugacy class of g is in the kernel. Since all transformations in the N - coset of g are in the conjugacy class, it follows that N is in the kernel and G is cyclic, whence isomorphic to C'.

In the case where C' is isomorphic to C, we get that G is a semidirect product $G = K \rtimes C$, where $K = N/N \cap M$, and the action of C on K is induced by the one of C on N. We have thus shown:

Proposition 6.2. *Let G be a nontrivial finite quotient of a triangle group $T = T(3,3,3)$, or $T(2,4,4)$, or $T(2,3,6)$. Then there is a maximal surjective homomorphism of G onto a cyclic group C_d of order $d \leq 6$.*

If moreover G is not isomorphic to C_d, then $d = 3$ for $T(3,3,3)$, for $T(2,4,4)$ $d = 4$, $d = 6$ for $T(2,3,6)$, and G is a semidirect product $G = K \rtimes C_d$, where the action of C is induced by the one of C on N. In particular, let a_1, c_1 and a_2, c_2 by two systems of generators of G: then $|\Sigma(a_1, c_1) \cap \Sigma(a_2, c_2)| \geq 2$.

Proof. Just observe that two elements which have the same image in $C - \{0\}$ belong to the same conjugacy class by our previous remarks. The rest follows right away. □

We give now an alternative proof by purely group theoretical arguments.

In case of $T(3,3,3)$ we have an isomorphism of finitely presented groups

$$\langle a, c \,|\, a^3, c^3, (ac)^3 \rangle \cong \langle x, y, r \,|\, [x,y], r^3, rxr^{-1} = y, ryr^{-1} = x^{-1}y^{-1} \rangle$$

given by $x = ca^{-1}$, $y = cac$, $r = a$. We set $N_3 := \langle x, y \rangle$. The second presentation shows that $\Gamma(3,3,3)$ is isomorphic to the split extension of $N_3 \cong \mathbb{Z}^2$ by the cyclic group (of order 3) generated by r. We have

Proposition 6.3. *Let L be a normal subgroup of finite index in $T(3,3,3)$. If $L \neq T(3,3,3)$, then $L \leq N_3$ and $G := T(3,3,3)/L$ is isomorphic to the split extension of a finite abelian group N by a cyclic group of order 3. The only possible types for a two-generator system of G are (up to permutation) $(3,3,3)$ and $(3,3,l)$ for some divisor l of $|N|$. Let a_1, c_1 and a_2, c_2 be two systems of generators of G, then $|\Sigma(a_1,c_1) \cap \Sigma(a_2,c_2)| \geq 3$.*

Proof. An obvious computation shows that the normal closure of any element $g = ur$ $(u \in N_3)$ contains N_3 and hence is equal to $T(3,3,3)$. This proves the first statement. Let now $L \leq N_3$ and let a_1, a_2 generate $G = T(3,3,3)/L$, then at least one of the cosets a_1, a_2 must contain an element of the form $g = ur^{\pm 1}$ $(u \in N_3)$. By rearrangement both cosets contain an element of this type. A computation shows that every element has order exactly 3 in $T(3,3,3)$. This confirms the statement about the types. Let $g = ur^{\pm 1}$ be as above and let S be the union of the conjugates of the cyclic group generated by g in $\Gamma(3,3,3)$. It is clear that S contains either xr or r or both these elements. □

In case of $T(2,4,4)$ we have an isomorphism of finitely presented groups

$$\langle a, c \,|\, a^2, c^4, (ac)^4 \rangle \cong \langle x, y, r \,|\, [x,y], r^4, rxr^{-1} = y, ryr^{-1} = y^{-1} \rangle$$

given by $x = ac^2$, $y = cac$, $r = c$. We set $N_4 := \langle x, y \rangle$. The second presentation shows that $\Gamma(2,4,4)$ is isomorphic to the split extension of $N_4 \cong \mathbb{Z}^2$ by the cyclic group (of order 4) generated by r. We have

Proposition 6.4. *Let L be a normal subgroup of finite index in $T(2,4,4)$. If the index of L in $T(2,4,4)$ is ≥ 16, then $L \leq N_4$ and $G := T(2,4,4)/L$ is isomorphic to the split extension of a finite abelian group N by a cyclic group of order 4. The only possible types for a two-generator system of G are (up to permutation) $(2,4,4)$ and $(4,4,l)$ for some divisor l of $|N|$. Let a_1, c_1 and a_2, c_2 by two systems of generators of G, then $|\Sigma(a_1,c_1) \cap \Sigma(a_2,c_2)| \geq 2$.*

The proof is analogous to the first proposition of this section.

In case of $T(2,3,6)$ we have an isomorphism of finitely presented groups

$$\langle a, c \,|\, a^2, c^3, (ac)^6 \rangle \cong \langle x, y, r \,|\, [x,y], r^6, rxr^{-1} = y^{-1}x, ryr^{-1} = x \rangle$$

given by $x = cac^{-1}a$, $y = c^{-1}aca$, $r = ac$. We set $N_6 := \langle x, y \rangle$. The second presentation shows that $\Gamma(2,3,6)$ is isomorphic to the split extension of $N_6 \cong \mathbb{Z}^2$ by the cyclic group (of order 6) generated by r. We have

Proposition 6.5. *Let L be a normal subgroup of finite index in $\Gamma(2,3,6)$. If the index of L in $T(2,3,6)$ is ≥ 24, then $L \leq N_6$ and $G := \Gamma(2,3,6)/L$ is isomorphic to the split extension of a finite abelian group N by a cyclic group of order 6. The only possible types for a two-generator system of G are (up to permutation) $(2,3,6)$ and $(6,6,l)$ for some divisor l of $|N|$. Let a_1, c_1 and a_2, c_2 be two systems of generators of G, then $|\Sigma(a_1,c_1) \cap \Sigma(a_2,c_2)| \geq 2$.*

Again the proof is analogous to the first proposition of this section.

References

1. I. Bauer and F. Catanese – "Some new surfaces with $p_g = q = 0$", 2003, `math.AG/0310150`, to appear in the Proceedings of the Fano Conference (Torino 2002), U.M.I. FANO special Volume.
2. A. Beauville – *Surfaces algébriques complexes*, Société Mathématique de France, Paris, 1978, Astérisque, No. 54.
3. B. J. Birch and W. Kuyk (eds.) – *Modular functions of one variable. IV*, Springer-Verlag, Berlin, 1975, Lecture Notes in Mathematics, Vol. 476.
4. F. Catanese – "Fibred surfaces, varieties isogenous to a product and related moduli spaces", *Amer. J. Math.* **122** (2000), no. 1, p. 1–44.
5. — , "Moduli spaces of surfaces and real structures", *Ann. of Math. (2)* **158** (2003), no. 2, p. 577–592.
6. H. S. M. Coxeter and W. O. J. Moser – *Generators and relations for discrete groups*, Second edition. Ergebnisse der Mathematik und ihrer Grenzgebiete, Neue Folge, Band 14, Springer-Verlag, Berlin, 1965.
7. J. Dieudonné – "On the automorphisms of the classical groups", *Mem. Amer. Math. Soc.,* **1951** (1951), no. 2, p. vi+122.
8. B. Everitt – "Alternating quotients of Fuchsian groups", *J. Algebra* **223** (2000), no. 2, p. 457–476.
9. R. Hartshorne – *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52.
10. B. Huppert – *Endliche Gruppen. I*, Die Grundlehren der Mathematischen Wissenschaften, Band 134, Springer-Verlag, Berlin, 1967.
11. B. Huppert and N. Blackburn – *Finite groups. III*, Grundlehren der Mathematischen Wissenschaften, vol. 243, Springer-Verlag, Berlin, 1982.
12. V. S. Kulikov and V. M. Kharlamov – "On real structures on rigid surfaces", *Izv. Ross. Akad. Nauk Ser. Mat.* **66** (2002), no. 1, p. 133–152.
13. R. Miranda – *Algebraic curves and Riemann surfaces*, Graduate Studies in Mathematics, vol. 5, American Mathematical Society, Providence, RI, 1995.
14. M. Suzuki – *Group theory. I*, Grundlehren der Mathematischen Wissenschaften, vol. 247, Springer-Verlag, Berlin, 1982.
15. H. Wielandt – *Finite permutation groups*, Academic Press, New York, 1964.

Couniformization of curves over number fields

Fedor Bogomolov[1] and Yuri Tschinkel[2]

[1] Courant Institute of Mathematical Sciences, N.Y.U., 251 Mercer str., New York, NY 10012, U.S.A.
bogomolo@cims.nyu.edu
[2] Mathematisches Institut, Bunsenstr. 3-5, 37073 Göttingen, Germany
yuri@uni-math.gwdg.de

Summary. We study correspondences between projective curves over $\bar{\mathbb{Q}}$.

1 Introduction

In this note we investigate correspondences between (geometrically irreducible) algebraic curves over number fields. Let C, C' be two such curves. We say that C lies over C' and write

$$C \Rightarrow C'$$

if there exist an étale cover $\tilde{C} \to C$ and a dominant map $\tilde{C} \to C'$. In particular, every curve lies over $\mathbb{P}^1$. Clearly, if $C \Rightarrow C'$ and $C' \Rightarrow C''$, then $C \Rightarrow C''$. We say that a curve C' is *minimal* for some class of curves $\mathcal{C}$ if every $C \in \mathcal{C}$ lies over C'.

Let

$$\mathsf{C}_n \,:\, y^n = x^2 + 1 \tag{1}$$

and $\mathcal{C}$ be the set of such curves. For all $n, m \in \mathbb{N}$ we have the standard, ramified, map

$$\mathsf{C}_{mn} \to \mathsf{C}_n, \;\; y \mapsto y^m.$$

At the same time, $\mathsf{C}_{mn} \Rightarrow \mathsf{C}_n$. Belyi's theorem [1] implies that for every curve C' defined over a number field there exists a curve $\mathsf{C} = \mathsf{C}_n \in \mathcal{C}$ such that $\mathsf{C} \Rightarrow C'$ (see [4] for a simple proof of this corollary). A natural extremal statement is:

Conjecture 1.1. *The curve C_6 lies over every curve C over $\bar{\mathbb{Q}}$.*

Every hyperelliptic curve C of genus $\mathsf{g}(C) \geq 2$ lies over C_6 (see Proposition 2.4 or [4]). The conjecture implies that every hyperbolic hyperelliptic curve lies over any other curve. Our main result towards Conjecture 1.1 is

Theorem 1.2. *For every $m \geq 6$ and $n \in \{2, 3, 5\}$ the curve C_m lies over C_{mn}.*

The relevance of such geometric constructions to number theory comes from a theorem of Chevalley–Weil: if $\pi : \tilde{C} \to C$ is an unramified map of proper algebraic curves over a number field K, then there exists a *finite* extension $\tilde{K}/K$ such that $\pi^{-1}(C(K)) \subset \tilde{C}(\tilde{K})$. Therefore, if $C \Rightarrow C'$, then Mordell's conjecture (Faltings' theorem) for C follows from Mordell's conjecture for C'. Our constructions allow us to control the degree and discriminant of the field $\tilde{K}$ in terms of the coefficients defining the curve. For example, Proposition 2.4 shows that "effective" Mordell for C_6 implies effective Mordell for every hyperelliptic curve (see also [12], [10], [6]).

The proof of this theorem uses certain special properties of modular curves and related elliptic curves. In the construction of unramified covers we need to exhibit maps from various intermediate curves onto $\mathbb{P}^1$ or elliptic curves with simultaneous restrictions on local ramification indices and branching points. This is very close, in spirit, to Belyi's theorem which says that every projective algebraic curve defined over $\bar{\mathbb{Q}}$ has a map onto $\mathbb{P}^1$ ramified in $0, 1, \infty$. In fact, there are many such maps. Our technique involves optimizing the choice of these maps by trading the freedom to impose ramification conditions for the degree of the map. An example of this is given in Section 4 where we prove the first part of Belyi's theorem (reduction to $\mathbb{Q}$-rational branching) under the restriction that the only prime dividing the local ramification indices is 2.

Acknowledgments. We are grateful to E. Bombieri and U. Zannier for their interest and help with references. The first author was partially supported by the NSF Grant DMS-0404715. The second author was partially supported by the NSF Grant DMS-0100277.

2 Minimal curves

Notation 2.1. For a surjective morphism of curves $\pi : C' \to C$ of degree d we denote by $\mathrm{Bran}(\pi) \subset C$ the branching locus of π. For $c \in \mathrm{Bran}(\pi)$ put

$$\mathsf{d}_c := (2^{d_2}, 3^{d_3}, \ldots), \ \sum_i i d_i \leq d,$$

where d_i is the number of points in $\pi^{-1}(c)$ with local ramification index i. Let

$$\mathrm{RD}(\pi) = \{\mathsf{d}_c\}_{c \in \mathrm{Bran}(\pi)}$$

be the *ramification datum*.

Example 2.2. Let $z^n : \mathbb{P}^1 \to \mathbb{P}^1$ be the n-power map $z \mapsto z^n$. Then $\mathrm{Bran}(z^n) = \{0, \infty\}$ and $\mathrm{RD}(z^n) = \{(n)_0, (n)_\infty\}$.

Notation 2.3. Let E be an elliptic curve over $\bar{\mathbb{Q}}$ with a fixed $0 \in E$, $E[n]$ the set of n-torsion points and

$$E[\infty] := \cup_{n=1}^{\infty} E[n] \subset E(\bar{\mathbb{Q}})$$

the set of all torsion points of E. Usually, we write $\sigma : x \to -x$ for the standard involution on E and

$$\pi = \pi_\sigma : E \to E/\sigma = \mathbb{P}^1$$

for the induced map. When we specify the elliptic curve by the branching locus we write $E = E(\mathrm{Bran}(\pi))$.

Proposition 2.4. *The curves* C_6 *and* C_8 *are minimal for the class of hyperbolic hyperelliptic curves.*

Proof. The proof of this proposition and many of the subsequent statements is based on Abhyankar's Lemma (Ramification "cancels" ramification).

Fix a hyperbolic hyperelliptic curve C. Notice that for any such C there exists an étale cover $R_1 \to C$ of degree 2 and a degree 2 surjection $R_1 \to E$ onto an elliptic curve. For example, we can take E to be any elliptic curve ramified in four of the ramification points of the initial hyperelliptic map $C \to \mathbb{P}^1$. Fix such an E.

We use the following simple fact about elliptic curves: Let $\pi : E \to \mathbb{P}^1$ be an elliptic curve. Then $\pi(E[3])$ is (projectively equivalent to) the union of one point from $\mathrm{Bran}(\pi)$ and $\{1, \zeta, \zeta^2, \infty\} \subset \mathbb{P}^1$ (where ζ is a fixed third root of 1). Similarly, $\pi(E[4])$ is (projectively equivalent to)

$$\mathrm{Bran}(\pi) = \{\lambda, \lambda^{-1}, -\lambda, -\lambda^{-1}\} \cup \{1, -1, i, -i, 0, \infty\} \subset \mathbb{P}^1.$$

For $m = 3, 4$ let

$$\varphi_m : E \to E$$

be the (multiplication by m) isogeny, $E_m = \mathsf{C}_m$ and

$$\pi_m : E_m \to \mathbb{P}^1.$$

The map π_m is 2-ramified in $E_m[m]$.

Consider the diagram

$$\begin{array}{ccccccccccccc}
C & \leftarrow & R_1 & \xleftarrow{\tau_2} & R_2 & = & R_2 & \xleftarrow{\tau_3} & R_3 & \xleftarrow{\tau_4} & R_4 & \xleftarrow{\tau_5} & R_5 \\
 & & \downarrow{\scriptstyle \iota_1} & & \downarrow{\scriptstyle \iota_2} & & \downarrow{\scriptstyle \sigma_2} & & \downarrow{\scriptstyle \iota_3} & & \downarrow{\scriptstyle \iota_4} & & \downarrow \\
 & & E & \xleftarrow[\varphi_m]{} & E & \xrightarrow{\pi} & \mathbb{P}^1 & \xleftarrow[\pi_m]{} & E_m & \xleftarrow[\varphi_m]{} & E_m & \xleftarrow[\iota_m]{} & \mathsf{C}_{2m}.
\end{array}$$

Here

- $\mathrm{Bran}(\pi_3) = \{1, \zeta, \zeta^3, \infty\} \subset \mathrm{Bran}(\sigma_2)$;
- $\mathrm{Bran}(\pi_4) = \{1, -1, i, -i\} \subset \mathrm{Bran}(\sigma_2)$;
- $\iota_m : \mathsf{C}_{2m} \to E_m = \mathsf{C}_m$ is the standard map, it is ramified in two points (whose difference is) in $E_m[m]$;
- R_2 is the fiber product $R_1 \times_E E$;
- $\sigma_2 = \pi \circ \iota_2$;
- $R_3 := R_2 \times_{\mathbb{P}^1} E_m$;
- R_4 is an irreducible component of $R_3 \times_{E_m} E_m$;
- $R_5 := R_4 \times_{E_m} \mathsf{C}_{2m}$;

Observe that for $q \in \mathrm{Bran}(\pi_m)$ the local ramification indices in the preimage $\sigma_2^{-1}(q)$ are all even. Therefore, τ_3 is *unramified* and ι_3 has even local ramification indices over (the preimage of) $q \in \{\pi(E[m]) \setminus \mathrm{Bran}(\pi_m)\}$ (such a point exists). Note that $q \in \mathrm{Bran}(\pi)$. The map ι_4 is ramified over the preimages $(\pi_m \circ \varphi_m)^{-1}(q)$, with even local ramification indices, which implies that τ_5 is unramified. Finally, R_5 has a dominant map onto C_{2m} and is unramified over R_4 (and consequently, R_1). This shows that every hyperelliptic curve lies over C_{2m}, for $m = 3, 4$. □

Theorem 2.5. *For all $m \geq 6$ and $\ell \in \{2, 3\}$ one has*

$$\mathsf{C}_m \Rightarrow \mathsf{C}_{m\ell}.$$

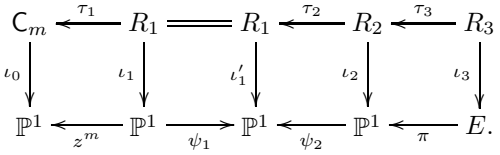
Proof. We first assume that $m = 2n$ is even and ≥ 8, since $\mathsf{C}_6 \Rightarrow \mathsf{C}_8$. First we show that $C := \mathsf{C}_m$ lies over C_{2m}. Consider the diagram:

$$\begin{array}{ccccccccccc}
\mathsf{C}_{2n} & \xleftarrow{\tau_1} & R_1 & \xleftarrow{\tau_2} & R_2 & \xleftarrow{\tau_3} & R_3 & = & R_3 & \xleftarrow{\tau_4} & R_4 \\
\downarrow{\scriptstyle \iota_0} & & \downarrow{\scriptstyle \iota_1} & & \downarrow{\scriptstyle \iota_2} & & \downarrow{\scriptstyle \iota_3} & & \downarrow{\scriptstyle \iota_3'} & & \downarrow \\
\mathbb{P}^1 & \xleftarrow[z^n]{} & \mathbb{P}^1 & \xleftarrow[\pi]{} & E & \xleftarrow[\varphi_2]{} & E & \xrightarrow[\pi']{} & \mathbb{P}^1 & \xleftarrow[\theta]{} & \mathsf{C}_{4n}.
\end{array}$$

Here

- π is a double cover whose branch locus consists of three points in the preimage of 1 under z^n and the preimage of 0;
- R_1 is the fiber product $\mathsf{C}_{2n} \times_{\mathbb{P}^1} \mathbb{P}^1$, note that τ_1 is unramified and that ι_1 is evenly ramified over all points in $\mathrm{Bran}(\pi)$;
- $R_2 = R_1 \times_{\mathbb{P}^1} E$, note that τ_2 is unramified since ι_1 has ramification of order 2 over 0 and even ramification over all $\zeta_n \in \mathbb{P}^1$;
- τ_3 is unramified;
- since $n \geq 4$, the map ι_2 has ramification points of order $2n$ and ι_3 is branched with ramification index $2n$ over all points in $E[2]$;
- π' is the map such that $\mathrm{Bran}(\pi') = \pi'(E[2])$, then $\iota_3' := \pi' \circ \iota_3$ is $4n$-ramified over all points in $\mathrm{Bran}(\pi')$;
- θ is the map branched in three of the above points, in particular, τ_4 is unramified.

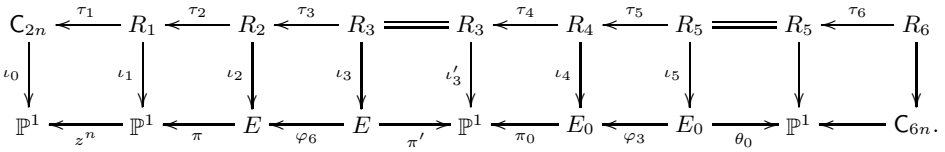
Now we assume that m is odd, $m \geq 5$ and consider the diagram:



Here

- $\psi_1 : z \mapsto (z + z^{-1})/2$, then $\iota_1' = \psi_1 \circ \iota_1 : R_1 \to \mathbb{P}^1$ is 2-ramified over -1, $2m$-ramified over 1 and m-ramified over $\xi_i := (\zeta_m^i + \zeta_m^{-i})/2$;
- $\psi_2 = \sqrt[m]{(z-\xi_1)/(z-\xi_2)}$, it has 2-ramification over all m preimages of -1 and $2m$-ramification over the preimages of 1;
- π is a double cover ramified over (arbitrary) four points in the preimage of -1 under ψ_2, then $\iota_3 : R_3 \to E$ is m-ramified over all other points and we can continue as above.

Now we show that C_m lies over C_{3m} (m even, this suffices for our purposes). Consider:



Here

- π is a double cover whose branch locus consists of three points in the preimage of 1 under z^n and the preimage of 0;
- $R_1 = \mathsf{C}_{2n} \times_{\mathbb{P}^1} \mathbb{P}^1$, note that τ_1 is unramified and that ι_1 is evenly ramified over all points in $\mathrm{Bran}(\pi)$;
- R_2 is the fiber product $R_1 \times_{\mathbb{P}^1} E$, note that τ_2 is unramified since ι_1 has ramification of order two over 0 and even ramification over all $\zeta_n \in \mathbb{P}^1$;
- τ_3 is unramified;
- since $n \geq 4$, the map ι_2 has ramification points of order $2n$ and ι_3 is branched with ramification index $2n$ over all points in $E[6]$;
- $\pi' : E \to \mathbb{P}^1$ is the map such that $\mathrm{Bran}(\pi') = \pi'(E[2])$, then $\iota_3' = \pi' \circ \iota_3$ is $4n$-ramified over all points of $\mathrm{Bran}(\pi')$;
- $\mathrm{Bran}(\pi_3) = \pi'(E[3]) \setminus \pi'(0)$ and the fiber product $R_4 = R_3 \times_{\mathbb{P}^1} \mathsf{C}_3$ is unramified over R_3, since all the preimages of $\mathrm{Bran}(\pi_3)$ in R_3 have even ramifications (for ι_3);
- note that there is a point $q_0 \in E_0$ such that every point in $\iota_4^{-1}(q_0) \in R_4$ has ramification of order $2n$ (for example, take a point q of order exactly 6 in E and take any $q_0 \in \pi_0^{-1}(\pi'(q)) \in E_0$).

- the fiber product $R_5 = R_4 \times_{E_0} E_0$ is unramified over R_4 and the map ι_5 has ramification of order $2n$ over all points in $E_0[3]$;
- now let θ_0 be the triple cover of $\mathbb{P}^1$ ramified in three points of order 3 in E_0, the composition of θ_0 with ι_5 exhibits R_5 as a cover of $\mathbb{P}^1$ so that all local ramification indices over three points in $\mathbb{P}^1$ are multiples of $6n$;
- finally, the fiber product $R_6 = R_5 \times_{\mathbb{P}^1} \mathsf{C}_{6n}$ is unramified over R_5.

□

Proposition 2.6. *We have*

$$\mathsf{C}_6 \Rightarrow \mathsf{C}_5.$$

Proof. Consider the standard action of the alternating group $\mathfrak{A}_5$ on $\mathbb{P}^1$. Choose any $\mathfrak{A}_4 \subset \mathfrak{A}_5$ and let $p_1, \dots, p_{12}$ be the $\mathfrak{A}_4$-orbit of a point fixed by an element of order 5 in $\mathfrak{A}_5$. By Klein (see [7], Ch. 1, 12, p. 58-59), there exists a polynomial identity

$$108t^4 - w^3 + \chi^2 = 0,$$

where

$$\chi \in H^0(\mathbb{P}^1, \mathcal{O}(p_1 + \cdots + p_{12})), t \in H^0(\mathbb{P}^1, \mathcal{O}(6)) \text{ and } w \in H^0(\mathbb{P}^1, \mathcal{O}(8))$$

(the zeroes of t give the vertices of the octahedron, of w the vertices of the cube and of χ the vertices of the icosahedron). An Euler characteristic computation shows that the map $w^3/\chi^2 : \mathbb{P}^1 \to \mathbb{P}^1$ is branched over exactly three points with RD $= \{(3^8), (4^6), (2^{12})\}$.

Consider

$$\begin{array}{ccccccccccccc}
\mathsf{C}_6 & \xleftarrow{\tau_0} & \mathsf{C}_{24} & \xleftarrow{\tau_1} & R_1 & = & R_1 & \xleftarrow{\tau_2} & R_2 & \xleftarrow{\tau_3} & R_3 & \xleftarrow{\tau_4} & R_4 \\
 & & \downarrow{\scriptstyle \iota_0} & & \downarrow{\scriptstyle \iota_1} & & \downarrow{\scriptstyle \iota_1'} & & \downarrow{\scriptstyle \iota_2} & & \downarrow{\scriptstyle \iota_3} & & \downarrow \\
 & & \mathbb{P}^1 & \xleftarrow[w^3/\chi^2]{} & \mathbb{P}^1 & \xrightarrow[\xi_5]{} & \mathbb{P}^1 & \xleftarrow[\pi_2]{} & \mathbb{P}^1 & \xleftarrow[\pi_3]{} & \mathbb{P}^1 & \leftarrow & \mathsf{C}_{30}.
\end{array}$$

Here

- RD$(\iota_0) = \{(24^1), (12^2), (24)^1\}$ and τ_0 is unramified;
- all local ramification indices of ι_1 over all zeroes of χ are divisible by 12.
- $\xi_5 : \mathbb{P}^1 \to \mathbb{P}^1/\mathfrak{A}_5$, the map ι_1 is branched in three points q_0, q_1, q_∞: over q_0 all local ramification indices are even, over q_1 - divisible by 3 and over q_∞ - divisible by 60;
- π_2 is a double cover branched q_0 and q_∞, ι_2 is branched in three points r_0, r_1, r_∞ so that all local ramification indices of ι_2 over r_0, r_1 are divisible by 3 and over r_∞ divisible by 30;
- π_3 is a triple cover, branched in three points so that all local ramification indices of ι_3 are divisible by 30;
- the standard map $\mathsf{C}_{30} \to \mathbb{P}^1$ is ramified over three points with RD $= \{(30^1), (15^2), (30^1)\}$.

Thus $\mathsf{C}_6 \Rightarrow \mathsf{C}_{30} \Rightarrow \mathsf{C}_5$, as claimed. □

Theorem 2.7. *For all $m, p \in \mathbb{N}$ one has*

$$\mathsf{C}_{5m} \Rightarrow \mathsf{C}_{5^p m}.$$

Proof. Let $\pi : E_5 \to \mathbb{P}^1$ be a degree 5 map from an elliptic curve, given by a rational function $f \in \mathbb{C}(E_5)$ with $\mathrm{div}(f) = 5(q_0 - q_\infty)$, and $q_0, q_\infty \in E_5$. Assume that π has cyclic degree 5 ramification over $0 = \pi(q_0)$ and $\infty = \pi(q_\infty)$ and that the (unique) remaining degenerate fiber of π contains two points with local ramification equal to 2 and one point q_1 where π is unramified. (Such a curve can be given as a quotient of the modular curve $X(10)$.)

Note that $5q_0 = 5q_1 = 5q_\infty$ in $\mathrm{Pic}(E_5)$. Since $\mathsf{C}_5 \Rightarrow \mathsf{C}_{20}$ it suffices to consider the diagram

$$\begin{array}{ccccccccc}
\mathsf{C}_{20n} & \xleftarrow{\tau_1} & R_1 & \xleftarrow{\tau_2} & R_2 & = & R_2 & \xleftarrow{\tau_3} & R_3 \\
\downarrow & & \downarrow{\scriptstyle \iota_1} & & \downarrow{\scriptstyle \iota_2} & & \downarrow{\scriptstyle \iota_2'} & & \downarrow{\scriptstyle \iota_3} \\
\mathbb{P}^1 & \xleftarrow[\pi]{} & E_5 & \xleftarrow[\phi_5]{} & E_5 & \xrightarrow{\pi} & \mathbb{P}^1 & \xleftarrow[\theta]{} & \mathsf{C}_{25n}.
\end{array}$$

Here

- $R_1 = \mathsf{C}_{20n} \times_{\mathbb{P}^1} E_5$, and ι_1 has cyclic ramification of order $20n$ over q_1;
- R_2 is (an irreducible component of) the fiber product $R_1 \times_{E_5} E_5$;
- $\iota_2' = \pi \circ \iota_2$ has cyclic $100n$ ramifications over $0, \infty$ and only even local ramification indices over 1;
- θ is the composition of the standard map $\mathsf{C}_{25n} \to \mathbb{P}^1$ with a degree 2 map $\mathbb{P}^1 \to \mathbb{P}^1$ (given by $x \mapsto (x + 1/x) + 1$), so that θ has the following ramification: a unique degree $50n$ cyclic ramification point over 0, two cyclic ramification points of degree $25n$ over ∞ and only degree 2 local ramifications over 1.

Then the (irreducible component of the) fiber product R_3 is unramified over R_2. □

Corollary 2.8. *The subset of minimal curves in the class $\{\mathsf{C}_n\}$ is infinite: if the only prime divisors of n are 2,3 or 5, then C_n is minimal.*

Example 2.9. Let

$$X(7) : x^3y + y^3z + z^3x = 0$$

be Klein's quartic plane curve of genus 3. Using

$$(x, y, z) \mapsto (z^3/x^2y, -z/x)$$

we see that $X(7)$ is isomorphic to the curve $y^7 = x^2(x + 1)$ while C_7 is isomorphic to $y^7 = x(x + 1)$. Thus their fiber product over $\mathbb{P}^1$ is unramified for both projections so that $X(7) \Leftrightarrow \mathsf{C}_7$.

3 A graph on the set of elliptic curves

Axiomatizing the constructions of Section 2, we are lead to consider a certain directed graph structure on the set $\mathcal{E}$ of all elliptic curves defined over $\bar{\mathbb{Q}}$, defined as follows: Write

$$E \rightharpoonup E', \quad \text{resp.} \quad E \rightleftharpoons E',$$

if $\mathrm{Bran}(E', \pi')$ is projectively equivalent to a set of four points in $\pi(E[\infty])$, resp. if E, E' are isogenous. Here π and π' are the standard double covers over $\mathbb{P}^1$. Note that the set

$$\pi(E[\infty]) \subset \mathbb{P}^1(\bar{\mathbb{Q}})$$

depends (up to the action of PGL_2 on $\mathbb{P}^1$) only on E and not on the choice of $0 \in E$.

Definition 3.1. *Let E' be an elliptic curve. A curve C' is called (E', n)-minimal if for every cover $\iota'' : C'' \to E'$ such that all local ramification indices over at least one point in $\mathrm{Bran}(\iota'')$ are divisible by n one has $C'' \Rightarrow C'$.*

Remark 3.2. Note that every curve $\iota' : C' \to E'$ such that

- $\mathrm{Bran}(\iota') \subset E'[\infty]$;
- all local ramification indices of ι' divide n.

is (E', n)-minimal.

Consider the standard action of the icosahedral group $\mathfrak{A}_5$ on $\mathbb{P}^1$. Let

- $\kappa_5 : \mathsf{H}_5 \to \mathbb{P}^1$ be the hyperelliptic curve branched in the 12 five-invariant points;
- $\kappa_3 : \mathsf{H}_3 \to \mathbb{P}^1$ the hyperelliptic curve branched in the 20 three-invariant points;
- $\iota_5 : \mathsf{C}_5 \to \mathbb{P}^1$ the standard curve from (1);
- $\iota : \mathsf{C} \to \mathbb{P}^1$ the degree 4 cover ramified over the primitive 5th roots $\{\zeta^i\}$ of 1, with local ramification indices equal to 2; we have $\mathsf{g}(\mathsf{C}) = 2$.

Proposition 3.3. *We have*

$$\mathsf{H}_5 \Leftrightarrow \mathsf{H}_3 \Leftrightarrow \mathsf{C}_5 \Leftrightarrow \mathsf{C}.$$

Proof. First of all, $\mathsf{H}_5 \Rightarrow \mathsf{C}_5$, since six of the 12 points are projectively equivalent to $\mathrm{Bran}(\iota_5)$ and hence an unramified degree 2 cover of H_5 surjects onto C_5. On the other hand, $\mathsf{C}_{30} \Rightarrow \mathsf{H}_5$, since κ_5 has three ramification points with indices $2, 3, 10$.

Similarly, $\mathsf{C}_{30} \Rightarrow \mathsf{H}_3$, since κ_3 has $2, 6, 5$ as local ramification indices. On the other hand, $\mathsf{H}_3/\mathfrak{C}_5$ is an elliptic curve, and the quotient map is branched at four points with ramification indices equal to 5. Hence $\mathsf{H}_3 \Rightarrow \mathsf{C}_5$. Since κ_5 is 2-ramified over the 5-th roots of unity plus 0, we have $\mathsf{C}_5 \Rightarrow \mathsf{C}$.

Finally, let R be the fiber product of five degree 2-covers $\mathbb{P}^1 \to \mathbb{P}^1$ ramified over, ζ^i, ζ^{i+1}, for $i = 1, \dots, 5$. Then $R \to \mathbb{P}^1$ is a Galois cover, consisting of two components R_1, R_2, each of genus 5, each ramified over $\mathbb{P}^1$ with degree 16 $(32 - 8 \cdot 5 = -8)$. The natural action of the cyclic group $\mathfrak{C}_5$ on R_1 has two invariant points (among the preimages of $0, \infty$), hence $R_1/\mathfrak{C}_5$ is an elliptic curve and, consequently, $R_1 \Rightarrow \mathsf{C}_5$. At the same time, $R_1 \Leftrightarrow \mathsf{C}$. □

Note that C is $(E(\zeta, \zeta^2, \zeta^3, \zeta^4), 2)$-minimal, since its 2-ramifications lie over points of finite order. Similarly, $X(7)$ is 2-minimal with respect to E_7.

Proposition 3.4. *Let C' be an (E', n)-minimal curve and $E \rightharpoonup E'$. Let $\iota : C \to E$ be a cover such that there exists an $e \in E$ with the property that for all $c \in \iota^{-1}(e)$ the local ramification indices are divisible by n. Then $C \Rightarrow C'$.*

Proof. As in Section 2. □

Remark 3.5. Proposition 3.4 explains why we are interested in *minimal* elements of the graph $\mathcal{E}$: curves E' such that for every curve E there is a finite chain

$$E \rightharpoonup E^1 \cdots \rightharpoonup E'$$

ending at E'. We have shown that $\mathcal{E}$ has a minimal element

$$E_0 = \mathsf{C}_3 : \ y^3 = x^2 + 1,$$

(for any E the curve E_0 is ramified over the images of torsion points of order 3 of E in $\mathbb{P}^1$). Thus any curve isogenous to E_0 is also minimal as is any curve E' with $E_0 \rightharpoonup E'$. In particular, every curve $\iota : C \to E_0$ such that $\mathrm{Bran}(\iota) \subset E_0[\infty]$ with local ramification indices equal to products of powers of 2 and 3 is minimal in the sense of Section 2.

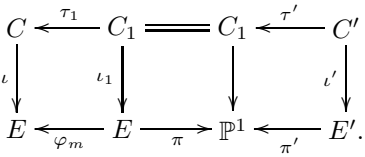
Remark 3.6. Note that $\mathcal{E}$ does not have a *maximal* element, that is, a curve E such that for every elliptic curve E' there is a chain

$$E \rightharpoonup E_1 \rightharpoonup \cdots \rightharpoonup E',$$

(in the class $\mathcal{E}$). This follows from the observation that the Galois groups of fields obtained by adjoining torsion points are contained in iterated extensions of subgroups $\mathrm{GL}_2(\mathbb{Z}/m)$. In particular, fields with simple Galois groups (over the ground field) which have no faithful two-dimensional representations over $\mathbb{F}_p$, for every prime p, cannot be realized.

Lemma 3.7. *Let $E \rightharpoonup E'$ be nonisogenous elliptic curves and let $\iota : C \to E$ be a cover, such that ι has at least one local ramification index divisible by $2n$. Then there is a cover $\iota' : C' \to E'$ from a curve C' such that $C \Rightarrow C'$, and* $\mathrm{Bran}(\iota')$ *includes points in $E'(\bar{\mathbb{Q}}) \setminus E'[\infty]$.*

Proof. Consider the diagram



Here

- m is such that $\mathrm{Bran}(\pi') \subset \pi(E[m])$, it exists since $E \not\to E'$;
- there exists a point $q \in \pi(E[m]) \backslash \mathrm{Bran}(\pi')$ such that the difference between the two preimages of q, under π', in E' is of infinite order in $E'(\bar{\mathbb{Q}})$.

This last claim holds since the set

$$\pi(E[\infty]) \cap \pi'(E'[\infty]) \subset \mathbb{P}^1$$

is finite, provided E is nonisogeneous to E'. Indeed, consider the map

$$\rho : E \times E' \to \mathbb{P}^1 \times \mathbb{P}^1 \supset \Delta(\mathbb{P}^1)$$

of degree 4, induced by π, π'. For nonisogeneous E, E', the genus of the preimage of the diagonal $C := \rho^{-1}(\Delta(\mathbb{P}^1))$ is ≥ 2. By a theorem of Raynaud [11], the set

$$C(\bar{\mathbb{Q}}) \cap (E[\infty] \times E'[\infty])$$

is finite (in fact, one can effectively estimate its cardinality). □

Lemma 3.8. *The set* $\pi(E[\infty]) \cap \mathbb{G}_m[\infty] \subset \mathbb{P}^1(\bar{\mathbb{Q}})$ *is finite.*

Proof. Follows from McQuillan's generalization of a theorem of Raynaud's (see [9], [11], and also [5]). Consider the map

$$(\theta, z^m) : E \times \mathbb{P}^1 \to \mathbb{P}^1 \times \mathbb{P}^1.$$

Then the preimage of the diagonal $(\theta, z^m)^{-1}(\Delta)$ is an affine open curve C of genus > 1. The finiteness of the intersection of C with $(E \times \mathbb{G}_m)_{tors} \subset E \times \mathbb{P}^1$ follows. □

A *cycle* in $\mathcal{E}$ is a finite set of curves $E, E_1, \ldots \in \mathcal{E}$ such that

$$E \not\to E_1 \not\to \cdots \not\to E.$$

Remark 3.9. Lemma 3.7 shows that each nontrivial cycle for E gives new (E, n)-minimal curves, which are n-ramified over points of infinite order in $E(\bar{\mathbb{Q}})$.

We now exhibit several such cycles in $\mathcal{E}$.

Lemma 3.10. *For any $x \in \mathbb{P}^1 \setminus \{0, 1, \infty\}$ one has*

$$E(0, 1, x^2, \infty) \rightharpoonup E(0, 1, x, \infty).$$

Proof. On the curve $E(0, 1, x^2, \infty)$ the preimages of the points $x, -x$ have order 4, since the involution $z \to x^2/z$ maps $0 \to \infty$ and $1 \to x^2$, and has $x, -x$ as invariant points. In particular, by definition,

$$E(0, 1, x^2, \infty) \rightharpoonup E(0, 1, x, \infty) \text{ and } E(0, 1, x^2, \infty) \rightharpoonup E(0, 1, -x, \infty).$$

□

Corollary 3.11. *Let $\zeta = \zeta_{2^n}$ be 2^n-th root of unity. Then there exists a finite chain starting with $E(0, 1, -1, \infty)$ and ending with $E(0, 1, \zeta, \infty)$.*

Corollary 3.12. *Let ℓ be an odd number. Then there exists a finite chain starting with $E(0, 1, \zeta_\ell, \infty)$ and ending with $E(0, 1, \zeta_\ell \cdot \zeta_{2^n}, \infty)$, where ζ_m is an m-th root unity.*

Proof. Some 2^m-th power of $\zeta_\ell \cdot \zeta_{2^n}$ is equal to ζ_ℓ. □

Corollary 3.13. *Let ℓ be an odd number. The set*

$$\{E(0, 1, \zeta_\ell^j, \infty)\}$$

decomposes into $\phi(\ell)/d_\ell$ (nontrivial) cycles of length d_ℓ, where ϕ is the Euler function and d_ℓ is the maximal power of 2 dividing $\phi(\ell)$.

Corollary 3.14. *For any $x \in \mathbb{P}^1 \setminus \{0, 1, \infty\}$ one has*

$$E(0, 1, (x-1)^2, \infty) \rightharpoonup E(0, 1, x, \infty)$$

and similarly,

$$E(0, 1, (2-x)x, \infty) \rightharpoonup E(0, 1, x, \infty).$$

Proof. We use the isomorphism

$$E(0, 1, (1-x), \infty) \sim E(0, 1, x, \infty).$$

□

4 Collecting points

Lemma 4.1. *Let $\mathbb{A}^d$ be the complex affine space of dimension d. For $x \in \mathbb{A}^d(\mathbb{C})$ let S_x be the affine algebraic variety characterized by the property:*

- *$x \in S_x$ and*

- *for every quadratic polynomial* $g \in \mathbb{C}[y], g(y) = g_2y^2 + g_1y + g_0$, *and every* $a = (a_1, \dots, a_d) \in S_x$ *one has* $(g(a_1), \dots, g(a_d)) \in S_x$.

Then S_x *is irreducible and is either equal to* $\mathbb{A}^d$ *or is contained in one of the diagonals* $\Delta_{ij} := \{x_i = x_j,\ i \neq j\}$.

Proof. Note that S_x is built from x as an iteration of vector bundles. At each step we have an irreducible variety. The procedure stabilizes after finitely many steps (by dimension reasons). Thus S_x is irreducible.

We proceed by induction on d. For $d = 1, 2, 3$ the claim is trivial. Assume the claim holds for all $d' < d$. We may also assume that $S_x \subset \mathbb{A}^n$ is a hypersurface not coinciding with a diagonal Δ_{ij}. Otherwise, the projection of S_x onto the first $d-1$ coordinates $\mathbb{A}^{d-1} \subset \mathbb{A}^d$ would not be surjective and hence, by the inductive assumption, contained in one of the diagonals, which would prove our claim.

We see that $\pi_{d-1} : S_x \to \mathbb{A}^{d-1}$ is a generically finite cover. Let

$$T_{d-1} := \{(t_1, \dots, t_{d-1})\} \subset \mathbb{A}^{d-1}$$

be such that all t_j are roots of unity of odd order. The set T_{d-1} is Zariski dense in $\mathbb{A}^{d-1}$. It contains a subset T^0_{d-1} which is Zariski dense in $\mathbb{A}^{d-1}$ and has the property that all fibers of π_{d-1} over T^0_{d-1} are nonempty and finite.

Note that for each $t = (t_j)_{j=1,\dots,d-1} \in T^0_{d-1}$ there exists an $n = n_t \in \mathbb{N}$ such that $t_j^{2^n} = t_j$ for all $j = 1, \dots, d-1$. This implies that the fiber over t is mapped into itself by the map $(a_j)_{j=1,\dots,d-1} \mapsto (a_j^{2^n})_{j=1,\dots,d-1}$. In particular, there is a point $b \in \pi^{-1}_{d-1}t$ and an $n' \geq n$ such that is fixed under the map

$$(b_j)_{j=1,\dots,d-1} \mapsto (b^{2^{n'}})_{j=1,\dots,d-1}.$$

We see that b_j are torsion points in $\mathbb{C}^*$, for all $j = 1, \dots, d-1$.

If $S^0 \subset (\mathbb{C}^*)^d$ is an algebraic subvariety and $T \subset S^0 \cap (\mathbb{C}^*)^d$ the subset of torsion points, then S^0 contains a finite set of translates of subtori by torsion points which contains T (see [8], [5], [13]). If follows that S_x contains a subtorus $(\mathbb{C}^*)^{d-1} \subset (\mathbb{C}^*)^d$ as a Zariski open subvariety.

Thus $S_x \subset \mathbb{A}^d$ is given by an equation

$$\prod_{j \in J} x_j^{n_j} = \prod_{j' \in J'} x_{j'}^{n_{j'}},$$

where $J \cap J' \subset [1, \dots, d]$ and $n_j, n'_j > 0$. The intersection of S_x with every diagonal Δ_{ij} is a proper subset (by assumption) and therefore (by induction) a finite union of subdiagonals (the intersection $S_x \cap \Delta_{ij}$ is stable under quadratic transformations). We may assume that $J \supset \{x_1, x_2\}$ and consider the diagonal $\Delta_{34} := \{x_3 = x_4\}$ (recall that $d \geq 4$). The resulting equation for $S_x \cap \Delta_{34}$ does not define a subset of a union of diagonals. □

Corollary 4.2. *Let $K/\mathbb{Q}$ be a field extension of degree $d = r_1 + 2r_2$, with r_1 real and r_2 (pairs of) complex embeddings, and*

$$K \hookrightarrow \mathbb{R}^{r_1} \oplus \mathbb{C}^{2r_2} \hookrightarrow \mathbb{C}^d = \mathbb{A}^d(\mathbb{C})$$

the corresponding map into the complex affine space. Let $x \in K^*$ be a primitive element (a generator of the field K over $\mathbb{Q}$). For every Zariski closed subset $Z \subset \mathbb{A}^d$ there exists a finite sequence of quadratic polynomials $g_i \in \mathbb{Q}[x]$, $i = 1, \ldots, n$, such that $g_1(g_2(\cdots(g_n(x)))) \notin Z$.

Proof. Since x is primitive, it is not contained in any diagonal in $\mathbb{A}^d$. Therefore, the variety S_x constructed in Lemma 4.1 coincides with $\mathbb{A}^d$. It suffices to observe that the image of x under $\mathbb{Q}$-rational quadratic maps is Zariski dense in $S_x = \mathbb{A}^d$ (at each step of the inductive construction, we get a Zariski dense set of points in the total space of the vector bundle). □

For $q \in \bar{\mathbb{Q}}$ let $\deg(q)$ be the degree of the minimal polynomial $f = f_q(x) \in \mathbb{Q}[x]$ vanishing in q and $K = K_q/\mathbb{Q}$ the field generated by q.

Corollary 4.3. *Let $q \in \bar{\mathbb{Q}}$. Then there exists a sequence of quadratic polynomials $g_i \in \mathbb{Q}[x]$ such that $g := g_1(g_2 \cdots (g_n(x))) \in \mathbb{Q}[x]$ has the property that*

- *$\deg(g(q)) = \deg(q)/2^k$, for some $k \in \mathbb{N}$, and*
- *the derivative of the minimal polynomial $f_{g(q)}(x) \in \mathbb{Q}[x]$ of $g(q) \in \bar{\mathbb{Q}}$ has no multiple roots.*

Proof. The first condition is satisfied, since a $\mathbb{Q}$-rational quadratic map can diminish the degree of the minimal polynomial at most by a factor of 2. The second condition amounts to a Zariski closed condition on the set of points in $K_q \subset \mathbb{A}^{\deg(q)}(\mathbb{C})$. □

Let $f : \mathbb{P}^1 \to \mathbb{P}^1$ be a rational map and $\mathrm{Ram}(f) = \{q \,|\, f'(q) = 0\} \subset \mathbb{P}^1$ the set of ramification points.

Theorem 4.4. *For any finite set $Q \subset \mathbb{P}^1(\bar{\mathbb{Q}})$ there is rational map $f : \mathbb{P}^1 \to \mathbb{P}^1$ such that*

$$\{f(q), q \in Q\} \cup \mathrm{Ram}(f) \subset \mathbb{P}^1(\mathbb{Q}).$$

Moreover, the only prime dividing a local ramification index of f is 2.

Remark 4.5. This is an analog of the first part of Belyi's theorem, with restrictions on the ramification. The proof follows the general line of Belyi's argument.

Proof. We proceed by induction on $m := \max(\deg(q))$, for $q \in Q$. Observe, that for all $f \in \mathbb{Q}[x]$ and all $q \in Q$ we have

$$\deg(f(q)) \leq m.$$

Assume that $m = 2^k$ and let $r \in Q$ be a point with minimal polynomial $f = f_q$ of degree m. If $f' \in \mathbb{Q}[x]$ has no multiple roots, then $f(Q) \bigcap \mathrm{div}_0(f')$ has fewer points of degree m: f maps q to zero and the zeroes of f' have degree $< m$. Moreover, the local ramification indices of f equal 2. If f' has multiple roots, we apply a sequence of $\mathbb{Q}$-rational quadratic maps as in Corollary 4.3, to replace q by $q' := g_1(g_2 \cdots (g_n(q)))$ so that the derivative of the minimal polynomial $f_{q'}(x) \in \mathbb{Q}[x]$ of q' has no multiple roots. The local ramification indices of a sequence of quadratic maps are powers of 2.

Now assume that $2^{k-1} < m < 2^k$, for some $k \in \mathbb{N}$, and put $s = 2^k - m$. Identify the space F_d of monic degree d polynomials with the affine space

$$\mathbb{A}^d = \{f_0 + f_1 x + \cdots + f_{d-1}x^{d-1} + x^d\}$$

and consider the following $\mathbb{Q}$-variety:

$$X \subset F_m \times F_s \times \mathbb{A}^s = \{(a_1, \ldots, a_s)\},$$

given by

$$(f \cdot g)'(a_j) = 0, \quad \text{for all} \quad j = 1, \ldots, s. \tag{2}$$

For fixed $f \in F_m$ and $a \in \mathbb{A}^s$ we get a system of non-homogeneous linear equations, where the variables are the coefficients of g. For generic, in Zariski topology on $F_m \times \mathbb{A}^s$, choices of f and a we get a unique solution, and a $\mathbb{Q}$-birational parametrization of X by $F_m \times \mathbb{A}^s = \mathbb{A}^{m+s}$ (here we use $m > s$). Thus the set of $\mathbb{Q}$-rational triples (f, g, a) subject to the equations (2) is Zariski dense in X.

The natural $\mathbb{Q}$-rational projection

$$X \to F_m \times F_s$$

is surjective (this can be checked over $\mathbb{C}$). In particular, $X(\mathbb{Q})$ is Zariski dense in X. The preimage $Z \subset X$ of the subset of those (f, g) where $(fg)'$ and g have multiple roots is a proper subvariety.

Applying $\mathbb{Q}$-rational quadratic maps as in Lemma 4.1, if necessary, we find a generic $f = f_q \in F_m(\mathbb{Q})$ and, by the argument above, a generic $g \in F_s(\mathbb{Q})$ such that there is a point $(f, g, a) \in (X \setminus Z)(\mathbb{Q})$ over (f, g).

The map $h := fg : \mathbb{P}^1 \to \mathbb{P}^1$ has the following properties:

- $h(q) = 0$ and Q has strictly fewer points of degree m;
- by construction, $(fg)'$ has at least s distinct $\mathbb{Q}$-rational roots so that the degree of points added to Q (the zeroes of $(fg)'$) is strictly less than m;
- all local ramification indices are powers of 2.

This concludes the induction and the proof of the theorem. □

Remark 4.6. A similar statement holds over function fields of any characteristic ($\neq 2$). Using the techniques from [2] one can show the following result: for any affine algebraic variety X over an algebraically closed field there exist a proper finite map $\pi : X \to \mathbb{A}^n$ and a linear projection $\lambda : \mathbb{A}^n \to \mathbb{A}^{n-1}$ such that π is ramified only in the sections of λ and the local ramification indices are powers of 2.

Remark 4.7. The methods of Belyi of collecting $\mathbb{Q}$-points on $\mathbb{P}^1$ produce ramification indices which depend on all pairwise differences between the coordinates of the points (for an exposition, see [3], Chapter 10). They cannot be applied in the construction of maps with restricted ramification.

References

1. G. V. Belyĭ – "Galois extensions of a maximal cyclotomic field", *Izv. Akad. Nauk SSSR Ser. Mat.* **43** (1979), no. 2, p. 267–276, 479.
2. F. A. Bogomolov and T. G. Pantev – "Weak Hironaka theorem", *Math. Res. Lett.* **3** (1996), no. 3, p. 299–307.
3. F. A. Bogomolov and T. Petrov – *Algebraic curves and one-dimensional fields*, Courant Lecture Notes in Mathematics, vol. 8, New York University Courant Institute of Mathematical Sciences, New York, 2002.
4. F. A. Bogomolov and Y. Tschinkel – "Unramified correspondences", Algebraic number theory and algebraic geometry, Contemp. Math., vol. 300, Amer. Math. Soc., Providence, RI, 2002, p. 17–25.
5. E. Bombieri and U. Zannier – "Algebraic points on subvarieties of $\mathbf{G}_m^n$", *Internat. Math. Res. Notices* (1995), no. 7, p. 333–347.
6. N. D. Elkies – "*ABC* implies Mordell", *Internat. Math. Res. Notices* (1991), no. 7, p. 99–109.
7. F. Klein – *Lectures on the icosahedron and the solution of equations of the fifth degree*, revised ed., Dover Publications Inc., New York, N.Y., 1956.
8. M. Laurent – "Équations diophantiennes exponentielles", *Invent. Math.* **78** (1984), no. 2, p. 299–327.
9. M. McQuillan – "Division points on semi-abelian varieties", *Invent. Math.* **120** (1995), no. 1, p. 143–159.
10. L. Moret-Bailly – "Hauteurs et classes de Chern sur les surfaces arithmétiques", *Astérisque* (1990), no. 183, p. 37–58, Séminaire sur les Pinceaux de Courbes Elliptiques (Paris, 1988).
11. M. Raynaud – "Courbes sur une variété abélienne et points de torsion", *Invent. Math.* **71** (1983), no. 1, p. 207–233.
12. L. Szpiro – "Discriminant et conducteur des courbes elliptiques", *Astérisque* (1990), no. 183, p. 7–18, Séminaire sur les Pinceaux de Courbes Elliptiques (Paris, 1988).
13. S. Zhang – "Positive line bundles on arithmetic varieties", *J. Amer. Math. Soc.* **8** (1995), no. 1, p. 187–221.

On the V-filtration of $\mathcal{D}$-modules

Nero Budur

Department of Mathematics, Johns Hopkins University, Baltimore, MD 21218-2686, U.S.A.
nbudur@math.jhu.edu

Summary. In this mostly expository note we give a down-to-earth introduction to the V-filtration of M. Kashiwara and B. Malgrange on $\mathcal{D}$-modules. We survey some applications to generalized Bernstein-Sato polynomials, multiplier ideals, and monodromy of vanishing cycles.

The V-filtration on $\mathcal{D}$-modules was introduced by M. Kashiwara and B. Malgrange to construct vanishing cycles in the category of (regular holonomic) $\mathcal{D}$-modules. Our aim is to give a down-to-earth introduction to this notion and describe some applications. The first application is to the generalized Bernstein-Sato polynomials introduced in [3]. Following G. Lyubeznik, we extend a finiteness result on the set of these polynomials. Then we describe applications to multiplier ideals [4], [3] and to monodromy of vanishing cycles and Hodge spectrum [2], [4].

Acknowledgments. We thank M. Mustaţă who provided us with preliminary notes on the V-filtration. We thank M. Saito for clarifications on many issues. Most of what I learned about the V-filtration is from discussions with them and with L. Ein.

1 Basics

In this section we introduce the filtration V and prove a few consequences assuming its existence. For a complete account of the V-filtration consult [12], [5], [8], [15].

Let X be a smooth complex variety. The sheaf $\mathcal{D}_X$ of algebraic differential operators on X is generated locally by multiplication by functions and by the tangent vector fields. If $X = \mathbb{A}^n$ is the affine n-space, then $\mathcal{D}_X$ is the Weyl algebra

$$A_n(\mathbb{C}) = \mathbb{C}[x_1, \ldots, x_n, \partial_1, \ldots, \partial_n],$$

where $\partial_i = \partial/\partial x_i$ and $\partial_i x_j - x_j \partial_i = \delta_{i,j}$. In these notes we will consider only left $\mathcal{D}_X$-modules, the most important for our applications being $\mathcal{O}_X$. We will frequently work locally without specifically assuming that X is affine.

The V-filtration of M. Kashiwara and B. Malgrange on $\mathcal{D}_X$-modules is defined with respect to some closed subvariety $Z \subset X$.

Case $Z \subset X$ smooth. First we consider smooth closed subvarieties $Z \subset X$. Let $\mathcal{I} \subset \mathcal{O}_X$ denote the ideal of Z. In local coordinates, write

$$X = \{(x,t)\},\ Z = \{x\} = \{t = 0\},$$

with $x = x_1, \ldots, x_n$, and $t = t_1, \ldots, t_r$. Then

$$\mathcal{D}_X = \mathbb{C}[x, t, \partial_x, \partial_t].$$

The V-filtration on $\mathcal{D}_X$ is defined by

$$V^j \mathcal{D}_X = \{\ P \in \mathcal{D}_X \mid P\mathcal{I}^i \subset \mathcal{I}^{i+j} \text{ for all } i \in \mathbb{Z}\ \},$$

with $j \in \mathbb{Z}$ and $\mathcal{I}^i = \mathcal{O}_X$ for $i \leq 0$. Locally,

$$V^j \mathcal{D}_X = \sum_{|\beta| - |\gamma| \geq j} h_{\alpha,\beta,\gamma}(x) \partial_x^\alpha t^\beta \partial_t^\gamma.$$

Here we use vectorial indices for monomials, and $|\beta| = \sum_i \beta_i$. A computation with local coordinates shows:

(i) $V^{j_1}\mathcal{D}_X \cdot V^{j_2}\mathcal{D}_X \subset V^{j_1+j_2}\mathcal{D}_X$, with equality if $j_1, j_2 \geq 0$;
(ii) $V^j \mathcal{D}_X = \mathcal{I}^j \cdot V^0\mathcal{D}_X \cdot \mathcal{D}_{X,-j} = \mathcal{D}_{X,-j} \cdot V^0\mathcal{D}_X \cdot \mathcal{I}^j$, where $\mathcal{D}_{X,j} \subset \mathcal{D}_X$ are the operators of order $\leq j$, and $\mathcal{I}^j = \mathcal{D}_{X,j} = \mathcal{O}_X$ for $j \leq 0$.

Definition 1.1. *The filtration V along Z on a coherent left $\mathcal{D}_X$-module M is an exhaustive decreasing filtration of coherent $V^0\mathcal{D}_X$-submodules $V^\alpha := V^\alpha M$, such that:*

(i) $\{V^\alpha\}_\alpha$ is indexed left-continuously and discretely by rational numbers, i.e., $V^\alpha = \cap_{\beta<\alpha} V^\beta$, every interval contains only finitely many α with $Gr_V^\alpha \neq 0$, and these α must be rational. Here, $Gr_V^\alpha = V^\alpha / V^{>\alpha}$, where $V^{>\alpha} = \cup_{\beta>\alpha} V^\beta$.

(ii) $t_j V^\alpha \subset V^{\alpha+1}$, and $\partial_{t_j} V^\alpha \subset V^{\alpha-1}$ for all $\alpha \in \mathbb{Q}$, i.e., $(V^i\mathcal{D}_X)(V^\alpha M) \subset V^{\alpha+i}M$;

(iii) $\sum_j t_j V^\alpha = V^{\alpha+1}$ for $\alpha \gg 0$;

(iv) the action of $\sum_j \partial_{t_j} t_j - \alpha$ on Gr_V^α is nilpotent on X.

All conditions are independent of the choice of local coordinates.

Theorem 1.2 (M. Kashiwara, B. Malgrange). *The filtration V along Z exists if M is regular holonomic and quasi-unipotent.*

It is beyond our scope to introduce the theory of holonomic systems of differential operators with regular singularities (see [1], [6]). It suffices to say that all the $\mathcal{D}$-modules considered in the applications are regular holonomic and quasi-unipotent.

Proposition 1.3. *The V-filtration along Z is unique.*

Proof. Let $\widetilde{V}$ be another filtration on M satisfying Definition 1.1. By symmetry, it suffices to show that $V^\alpha \subset \widetilde{V}^\alpha$ for every α. Suppose that $\alpha \neq \beta$ and consider

$$V^\alpha \cap \widetilde{V}^\beta / (V^{>\alpha} \cap \widetilde{V}^\beta) + (V^\alpha \cap \widetilde{V}^{>\beta}).$$

Since both filtrations satisfy Definition 1.1-(iv), both $(\sum_j \partial_{t_j} t_j - \alpha)$ and $(\sum_j \partial_{t_j} t_j - \beta)$ are nilpotent on this module. Hence the module is zero.

We show now that for every α we have

$$V^\alpha \subset V^{>\alpha} + \widetilde{V}^\alpha. \tag{1}$$

Fix $w \in V^\alpha$. By exhaustion, there is $\beta \ll 0$ (in particular $\beta < \alpha$) such that $w \in \widetilde{V}^\beta$. By what we have already proved, we may write $w = w_1 + w_2$, with $w_1 \in V^{>\alpha}$ and $w_2 \in V^\alpha \cap \widetilde{V}^{>\beta}$. If we replace w by w_2, then the class in $V^\alpha / V^{>\alpha}$ remains unchanged, but we may choose a larger β. We can repeat the process as long as $\beta < \alpha$. Since the $\widetilde{V}$-filtration is discrete, we can repeat the process until we have $\beta \geq \alpha$. Hence the class of w in $V^\alpha / V^{>\alpha}$ can be represented by an element in $\widetilde{V}^\alpha$, and we get (1).

Since the V-filtration is discrete, a repeated application of (1) shows that for every $\beta \geq \alpha$ we have $V^\alpha \subset V^\beta + \widetilde{V}^\alpha$. We deduce from Definition 1.1-(iii) that if we fix $\beta \gg 0$, then

$$V^\alpha \subset \mathcal{I}^q \cdot V^\beta + \widetilde{V}^\alpha \tag{2}$$

for big enough q, where $\mathcal{I} \subset \mathcal{O}_X$ is the ideal pf Z. By coherence, $V^\beta = \sum V^0 \mathcal{D}_X \cdot u_i$ for finitely many u_i. By exhaustion, there exists some $\gamma \in \mathbb{Z}$ such that $\widetilde{V}^\gamma$ contains the u_i, hence also V^β. By Definition 1.1-(ii), for q with $q + \gamma \geq \alpha$ we have $\mathcal{I}^q \widetilde{V}^\gamma \subset \widetilde{V}^\alpha$. Thus $\mathcal{I}^q V^\beta \subset \widetilde{V}^\alpha$. Hence by (2) we have $V^\alpha \subset \widetilde{V}^\alpha$. □

Case $Z \subset X$ arbitrary. Now let X be a smooth complex variety and $Z \neq X$ a closed subscheme. Suppose $f_1, \ldots, f_r \in \mathcal{O}_X$ generate the ideal $\mathcal{I} \subset \mathcal{O}_X$ of Z. Let $i : X \to X \times \mathbb{A}^r = Y$ be the embedding $x \mapsto (x, f_1(x), \ldots, f_r(x))$. Let $t_j : Y \to \mathbb{A}^1$ be the projection with $t_j \circ i = f_j$.

Let N be a $\mathcal{D}_X$-module and $M = i_* N$, where i_* is the direct image for left D-modules. Working out the definition of the direct image (e.g., [1]), one gets $M = N \otimes \mathbb{C}[\partial_{t_1}, \ldots, \partial_{t_r}]$ with the left $\mathcal{D}_Y$-action given as follows. Let $x_1, \ldots, x_n$ be local coordinates on X. For $g \in \mathcal{O}_X, m \in N$, and $\partial_t^\nu = \partial_{t_1}^{\nu_1} \ldots \partial_{t_r}^{\nu_r}$,

$$g(m\otimes\partial_t^\nu)=gm\otimes\partial_t^\nu,\quad \partial_{x_i}(m\otimes\partial_t^\nu)=\partial_{x_i}m\otimes\partial_t^\nu-\sum_j\frac{\partial f_j}{\partial x_i}m\otimes\partial_{t_j}\partial_t^\nu\ ,$$

$$\partial_{t_j}(m\otimes\partial_t^\nu)=m\otimes\partial_{t_j}\partial_t^\nu,\quad t_j(m\otimes\partial_t^\nu)=f_jm\otimes\partial_t^\nu-\nu_jm\otimes(\partial_t^\nu)_j\ ,$$

where $(\partial_t^\nu)_j$ is obtained from ∂_t^ν by replacing ν_j with ν_j-1. If N satisfies the requirements of Theorem 1.2, then also M does.

Definition 1.4. *The V-filtration along Z on $N(=N\otimes 1)$ is defined by $V^\alpha N=(N\otimes 1)\cap V^\alpha M$, for $\alpha\in\mathbb{Q}$ and V on M taken along $X\times\{0\}$.*

Proposition 1.5. *The definition above depends on the ideal $\mathcal{I}$ of Z in X and not on the particular generators chosen.*

Proof. (cf. [3]-2.7) Suppose $g_1,\dots,g_{r'}\in\mathcal{O}_X$ also generate $\mathcal{I}$, with $g_j=\sum a_{ij}f_i$, $a_{ij}\in\mathcal{O}_X$. Let $i':Y=X\times\mathbb{A}^r\to Y'=X\times\mathbb{A}^r\times\mathbb{A}^{r'}$ be the embedding sending (x,t) to (x,t,t'), where $t'_j=\sum a_{ij}(x)t_i$, $j=1,\dots,r'$. The crucial fact here is that the image of $X\times\{0\}$ is $X\times\{0\}\times\{0\}$. Working locally, we can assume that x,t,u is a local coordinate system on Y' such that $Y=\{u=0\}$, $X=\{t=u=0\}$. Hence $M'=i'_*M$ can be written as $M\otimes\mathbb{C}[\partial_{u_1},\dots,\partial_{u_{r'}}]$ with left $\mathcal{D}_{Y'}$-action as above. Note that some simplifications occur: $\partial_{x_i}(m\otimes\partial_u^\nu)=\partial_{x_i}m\otimes\partial_u^\nu$, $\partial_{t_i}(m\otimes\partial_u^\nu)=\partial_{t_i}m\otimes\partial_u^\nu$, and $u_j(m\otimes\partial_u^\nu)=-\nu_jm\otimes(\partial_u^\nu)_j$, where $m\in M$, $\partial_u^\nu=\partial_{u_1}^{\nu_1}\dots\partial_{u_{r'}}^{\nu_{r'}}$, and $(\partial_u^\nu)_j$ is obtained from ∂_u^ν by replacing ν_j with ν_j-1.

The claim follows if we show that

$$V^\alpha M'=\sum_{\nu\in\mathbb{N}^{r'}}V^{\alpha+|\nu|}M\otimes\mathbb{C}[\partial_u^\nu]$$

is the V-filtration on M' along $X\times\{0\}\times\{0\}$.

Let us check the axioms for the V-filtration. In local coordinates, $V^0\mathcal{D}_{Y'}$ is generated over $\mathcal{O}_{Y'}$ by the ∂_{x_i}, and the $v\partial_w$ with $v,w\in\{t_1,\dots,t_r,u_1,\dots,u_{r'}\}$. From definition, these actions are well-defined on $V^\alpha M'$. To show that $V^\alpha M'$ is coherent over $V^0\mathcal{D}_{Y'}$, it is enough to show that $V^\alpha M'$ is locally finitely generated since $V^0\mathcal{D}_{Y'}$ is coherent. Since $V^\alpha M$ is locally finitely generated over $V^0\mathcal{D}_Y$, we have that for $c\gg 0$, $\sum_{|\nu|\le c}V^{\alpha+|\nu|}M\otimes\mathbb{C}[\partial_u^\nu]$ is locally finitely generated. Also for $c\gg 0$, $V^{\alpha+c+1}M=\sum_i t_iV^\alpha M$ by the axiom (iii) of Definition 1.1. Therefore the rest of $V^\alpha M'$ is recovered from $\sum_{|\nu|\le c}$ through the action of the $t_i\partial_{u_j}$, hence $V^\alpha M'$ is finitely generated.

The axioms (ii) and (iii) of Definition 1.1 follow from the definition of $V^\alpha M'$, the simplifications noted above in the $\mathcal{D}_{Y'}$-action on M', and the same axioms applied to $V^\alpha M$.

The last property to show is the nilpotency of $s-\alpha$ on $\mathrm{Gr}_V^\alpha M'$, where $s=\sum_i\partial_{t_i}t_i+\sum_j\partial_{u_j}u_j$. Let $m\otimes\partial_u^\nu\in V^\alpha M'$ with $m\in V^{\alpha+|\nu|}M$. Then

$$(s-\alpha)(m\otimes\partial_u^\nu)=\left(\sum_i\partial_{t_i}t_i-|\nu|-\alpha\right)m\otimes\partial_u^\nu.$$

Hence $(s-\alpha)^k(m\otimes\partial_u^\nu)\in V^{\alpha+1}M'$ if k is the nilpotency order of $(\sum_i \partial_{t_i}t_i - (\alpha+|\nu|))$ on $\mathrm{Gr}_V^{\alpha+|\nu|}M$. □

Examples will be provided in Section 3.

2 Bernstein-Sato polynomials

The V-filtration can be applied to show the existence of quite general Bernstein-Sato polynomials, [3]. See [6] for an account of the classical version of these polynomials. Following G. Lyubeznik [11], we prove a finiteness result on the set of all polynomials that are Bernstein-Sato polynomials in a sense we make precise later.

We keep the notation from the previous section. Suppose first that $Z\subset X$ is a smooth closed subvariety. To keep this article as concise as possible we take Theorem 1.2 for granted. Then the quickest way to proceed is by means of the following technical tool.

Definition 2.1. *Let M be a coherent left $\mathcal{D}_X$-module. For $u\in M$, the Bernstein-Sato polynomial $b_u(s)$ of u is the monic minimal polynomial of the action of $s=-\sum_j \partial_{t_j}t_j$ on $V^0\mathcal{D}_Xu/V^1\mathcal{D}_Xu$.*

We suppressed from the notation the fact that $b_u(s)$ also depends on Z. Then we can make explicit the V-filtration as follows.

Proposition 2.2 (C. Sabbah [15]). *If the V-filtration along Z exists on M, then $b_u(s)$ exists, it is non-zero for all $u\in M$, and has rational coefficients. Moreover*

$$V^\alpha M=\{\ u\in M\mid \alpha\le c \text{ if } b_u(-c)=0\ \}.$$

Proof. Suppose first that $u\in V^\alpha M$. Recall that $\sum_j\partial_{t_j}t_j-\beta$ is nilpotent on $V^\beta/V^{>\beta}$ and V is indexed discretely. Then, for a given β there is a polynomial $b(s)$ depending on β, having all roots $\le-\alpha$ (and rational), and such that $b(-\sum_j\partial_{t_j}t_j)\cdot u\in V^\beta$. Hence it is enough to show that there is β such that $V^\beta\cap V^0\mathcal{D}_Xu\subset V^1\mathcal{D}_Xu$.

Let $A=\bigoplus_{i\ge0}V^i\mathcal{D}_X\tau^{-i}$ and define $F_k(A)=\bigoplus_{i\ge0}(V^i\mathcal{D}_X\cap\mathcal{D}_{X,k})\tau^{-i}$. Then by Lemma 2.3, A is a noetherian ring. Now $\bigoplus_{i\ge0}V^iM$ is coherent over A because by axiom (iii) of Definition 1.1, there exists i_0 such that V^iM is recovered from $V^{i_0}M$ if $i\ge i_0$. Denote by N the $V^0\mathcal{D}_X$-submodule $V^0\mathcal{D}_Xu$, and let $U^i=V^i\cap N$ for $i\ge0$. Then $\bigoplus_{i\ge0}U^iN$ is also coherent over A since A is noetherian. It follows that $\bigoplus_{i\ge0}\mathrm{Gr}_U^iN$ is coherent over $\bigoplus_{i\ge0}\mathrm{Gr}_V^i\mathcal{D}_X$, in particular locally finitely generated. If i is big compared with the degrees of local generators, we see that $U^iN\subset V^1\mathcal{D}_Xu$.

Conversely, fix an element $u\in M$ and suppose that $\alpha\le c$ whenever $b_u(-c)=0$. Let $\alpha_u=\max\{\beta\mid u\in V^\beta\}$. We need to show that $\alpha\le\alpha_u$. It is

enough to show that $b_u(-\alpha_u) = 0$. For $\beta \neq \alpha_u$, $(\sum_j \partial_{t_j} t_j - \beta)$ is invertible on $V^{\alpha_u}/V^{>\alpha_u}$. But $b_u(-\sum_j \partial_{t_j} t_j)u \in V^{>\alpha_u}$. Hence we must have $b_u(-\alpha_u) = 0$. □

Lemma 2.3 ([6]-A.29). *Let A be a be a filtered ring (sheaf on X). Assume that $F_0(A)$ and $Gr^F(A)$ are noetherian rings, and that $Gr^F_k(A)$ are (locally) finitely generated $F_0(A)$-modules for all k. Then A is noetherian.*

Now let $Z \subset X$ be an arbitrary closed subset. Let $f_1, \ldots, f_r$ be generators of the ideal of Z, where $f_j \neq 0$ for any j. Then $\mathcal{D}_X$ acts naturally on $\mathcal{O}_X[\prod_i f_i^{-1}, s_1, \ldots, s_r] \prod_i f_i^{s_i}$, where the s_i are independent variables. Define a $\mathcal{D}_X$-linear action of t_i by $t_i(s_j) = s_j + 1$ if $i = j$, and $t_i(s_j) = s_j$ otherwise. Let $s_{ij} = s_i t_i^{-1} t_j$, and $s = \sum_i s_i$. We will see in Lemma 2.6 that under a well-defined isomorphism the t_i's here correspond to the t_i's introduced in the second part of Section 1.

Definition 2.4 ([3]). *The Bernstein-Sato polynomial $b_f(s)$ of $f := (f_1, \ldots, f_r)$ is defined to be the monic polynomial of the lowest degree in s satisfying the relation*

$$b_f(s) \prod_i f_i^{s_i} = \sum_j (P_j f_j \prod_i f_i^{s_i}), \tag{3}$$

where the P_j belong to the ring generated by $\mathcal{D}_X$ and the s_{ij}. For $h \in \mathcal{O}_X$, define similarly $b_{f,h}(s)$ with $\prod_i f_i^{s_i}$ replaced by $\prod_i f_i^{s_i} h$.

Example 2.5.

(i) $f = x^2 + y^3$. Then $b_f(s) = (s+1)(s+5/6)(s+7/6)$ and

$$P = (\partial_y^3/27 + y\partial_x^2\partial_y/6 + x\partial_x^3/8).$$

(ii) $f = (x_2x_3, x_1x_3, x_1x_2)$. Then $b_f(s) = (s+3/2)(s+2)^2$ and the s_{ij} cannot be avoided by the operators P_j in the above definition (see [3]).

The polynomial $b_Z(s) := b_f(s-r)$ with $r = \operatorname{codim}_X Z$ is shown in [3] to depend only on Z and not on f. The existence of non-zero $b_{f,h}(s)$ follows from the following.

Lemma 2.6. *With the notation as in Definition 1.4, if $M = i_*\mathcal{O}_X$, $u = h \otimes 1$ with $h \in \mathcal{O}_X$, and the V-filtration is taken along $X \times \{0\}$, then $b_u(s) = b_{f,h}(s)$.*

Proof. It suffices to show that $b_u(s)$ is the minimal polynomial of the action of $s = \sum_j s_j$ on

$$\mathcal{D}_X[s_{ij}] \prod_j f_j^{s_j} h / \sum_k \mathcal{D}_X[s_{ij}] f_k \prod_j f_j^{s_j} h,$$

a quotient of submodules of $\mathcal{O}_X[\prod_i f_i^{-1}, s_1, \ldots, s_r] \prod_i f_i^{s_i} h$. We can check that $\mathcal{D}_X[s_{ij}] \prod_j f_j^{s_j} h$ and $\sum_k \mathcal{D}_X[s_{ij}] f_k \prod_j f_j^{s_j} h$ are isomorphic to $V^0\mathcal{D}_Y u$ and $V^1\mathcal{D}_Y u$. The action of t_j is defined by $s_j \mapsto s_j + 1$, $\prod_j f_j^{s_j} h$ corresponds to u, s_j corresponds to $-\partial_{t_j} t_j$, and $s_{ij} = s_i t_i^{-1} t_j$. □

Hence it follows from Proposition2.2 that $b_{f,h}(s)$ are polynomials with rational coefficients. One is allowed to change the field of definition in the coefficients of the f_j's.

Proposition 2.7. *There exist non-zero Bernstein-Sato polynomials $b_{f,h}(s)$ even if in Definition 2.4 one replaces X, $\mathcal{O}_X$, and $\mathcal{D}_X$ with $\mathbb{A}^n_k$, $k[x_1,\dots,x_n]$, and $A_n(k)$ (the Weyl algebra) respectively, for k a field of characteristic zero.*

Proof. First, suppose that the coefficients of the f_j's lie in a subfield K of $\mathbb{C}$. Then also the scalar coefficients of the P_j's can be assumed to lie in K. Indeed, (3) implies, after equating coefficients of monomials in s_i's and x_i's, that certain K-linear relations (*) hold among the scalar coefficients of the P_j's. Let L be the field generated by the coefficients of the P_j's. Fix a basis S of L/K containing 1 and such that every scalar coefficient c which appears in a P_j can be written as a unique K-linear combination of a finite number of elements of S. Let $c_1 \in K$ be the coefficient of 1 in c under this basis, and let $P_{j,1}$ be the induced operator. Then the K-linear relations (*) hold with c_1 replacing c, and so (3) holds with $P_{j,1}$ replacing P_j.

Now, going back to our proposition, the conclusion follows from the Lefschetz principle. Indeed, let K be a subfield of k generated over $\mathbb{Q}$ by the coefficients of the f_j's. Since $\mathbb{C}$ has infinite transcendental dimension over $\mathbb{Q}$, K can be embedded into $\mathbb{C}$. Then the coefficients of the P_j are in $K \subset k$. □

We extend a result of G. Lyubeznik [11] to the case of these more general Bernstein-Sato polynomials. The proof follows closely his proof.

Proposition 2.8. *Fix n and d positive integers. The set of all polynomials which are of the form $b_f(s)$ for some $f = f_1,\dots,f_r \in k[x_1,\dots,x_n]$ with $\deg f_i \le d$ is finite even if k is varying over all the fields of characteristic zero.*

Proof. Let N be the number of monomials in $x_1,\dots,x_n$ of degree $\le d$. Then the f's are the closed k-rational points of $[\mathbb{A}^N_k]^{\times r}$. Let $P = \sum_{|\alpha|\le d} c_\alpha x^\alpha$ be the polynomial of n variables of degree d with undetermined coefficients. Then $B_k = k[\text{ the } c_\alpha\text{'s}]^{\otimes r}$ is the coordinate ring of $[\mathbb{A}^N_k]^{\times r}$. Define

$$F_i = 1 \otimes \cdots \otimes P \otimes \cdots \otimes 1 \in B_{\mathbb{Q}}[x_1,\dots,x_n]$$

by placing P in the i-th position. Here $\times$ and $\otimes$ mean over $\mathbb{Q}$.

Let Y be a reduced and irreducible closed subset of $[\mathbb{A}^N_{\mathbb{Q}}]^{\times r}$. Denote by $G = (G_i)_i$ the image of $F = (F_i)_i$ under the natural $\mathbb{Q}$-algebra homomorphism $B_{\mathbb{Q}}[x_1,\dots,x_n] \to \mathbb{Q}(Y)[x_1,\dots,x_n]$, where $\mathbb{Q}(Y)$ is the function field of Y. Let $\mathbb{Q}[Y]$ be the coordinate ring of Y. Then, we have a functional equation

$$b_G(s)\prod_i G_i^{s_i} = \sum_j P_j G_j \prod_i G_i^{s_i} \tag{4}$$

with $s = \sum s_i$ and $P_j \in A_n(\mathbb{Q}(Y))[$ the s_{ij}'s$]$. Denote by $c \in \mathbb{Q}[Y]$ the common denominator. Denote by $U(c) \subset Y$ the subscheme whose coordinate ring is $\mathbb{Q}[Y]_c$. By specializing (4), the f's given by the closed k-rational points of $U(c) \times k$ have b-functions $b_f(s)$ dividing $b_G(s)$. Hence they are only finitely many such $b_f(s)$ even if k varies.

We proceed now by induction on the dimension of Y proving that the k-rational points of $Y \times k$ give only finitely many b-functions even if k varies. For dimension zero, $c = 1$ and so $U(c) = Y$. In higher dimensions, $Y \setminus U(c)$ is the union of reduced and irreducible closed subsets of smaller dimension. □

3 Multiplier ideals

The multiplier ideals introduced by A. Nadel [14] encode the complexity of singularities via their resolutions. It turns out that they are essentially the same as the V filtration on $\mathcal{O}_X$.

Let X be a smooth complex variety and $Z \neq X$ a closed subscheme. Let $\mu : X' \to X$ be a log resolution of (X, Z). That is μ is proper birational, X' is smooth, and $\mathrm{Ex}(\mu) \cup \mu^{-1}Z$ is a divisor with simple normal crossings. Here $\mathrm{Ex}(\mu)$ denotes the exceptional locus of μ. Let $\mathcal{I} \subset \mathcal{O}_X$ denote the ideal of Z. Let H be the effective divisor on X' such that $\mu^{-1}(\mathcal{I}) \cdot \mathcal{O}_{X'} = \mathcal{O}_{X'}(-H)$.

Definition 3.1. *For $\alpha > 0$, the multiplier ideal of $(X, \alpha \cdot Z)$ is defined as*

$$\mathcal{J}(\alpha \cdot Z) = \mu_*(\omega_{X'/X} \otimes \mathcal{O}_{X'}(-\llcorner \alpha \cdot H \lrcorner)).$$

Here $\omega_{X'/X} = \det \Omega^1_{X'} \otimes \mu^*(\det \Omega^1_X)^\vee$ is the sheaf of relative top-dimensional forms, and $\llcorner . \lrcorner$ rounds down the coefficients of the irreducible divisors. One can extend this definition to a formal combination of closed subschemes $\sum_i \alpha_i \cdot Z_i$ by replacing $\alpha \cdot H$ with $\sum_i \alpha_i \cdot H_i$.

The original analytic definition of multiplier ideals is, locally,

$$\mathcal{J}(\alpha \cdot Z) = \{\ h \in \mathcal{O}_X \mid |h|^2/(\sum_{1 \leq i \leq r} |f_i|^2)^\alpha \in L^1_{loc}\ \},$$

where $f_1, \ldots, f_r$ generate $\mathcal{I}$. The first definition shows the second is independent of the choice of generators, and the second definition shows the first is independent of the choice of resolution. See [9] for more on multiplier ideals.

The multiplier ideals measure how singular Z is. The intuition here is that smaller multiplier ideals means worse singularities. For example, varying the coefficient in front of Z, one obtains a decreasing family $\{\mathcal{J}(\alpha \cdot Z)\}_{\alpha \in \mathbb{Q}}$. Because of the rounding-down of coefficients in the construction of $\mathcal{J}(\alpha \cdot Z)$ there exist positive rational numbers $0 < \alpha_1 < \alpha_2 < \cdots$ such that

$$\mathcal{J}(\alpha_j \cdot Z) = \mathcal{J}(\alpha \cdot Z) \neq \mathcal{J}(\alpha_{j+1} \cdot Z)$$

for $\alpha_j \leq \alpha < \alpha_{j+1}$ where $\alpha_0 = 0$. These numbers $\alpha_j\,(j > 0)$ are called the jumping numbers of the multiplier ideals associated to (X, Z). The log-canonical threshold of (X, Z) is the smallest non-zero jumping number, $\mathrm{lc}(X, Z) = \alpha_1$. Equivalently, $\mathrm{lc}(X, Z)$ is the number α such that $\mathcal{J}(\alpha \cdot Z) \neq \mathcal{O}_X$, but $\mathcal{J}((\alpha - \epsilon)Z) = \mathcal{O}_X$ for $0 < \epsilon \ll 1$.

Example 3.2.

(i) $Z = \{\, x^2 + y^3 = 0 \,\} \subset \mathbb{A}^2$. Then $\mathcal{J}(\alpha \cdot Z)$ is equal to $\mathcal{O}_X$ if $0 < \alpha < 5/6$, and it is the maximal ideal at $(0, 0)$ if $5/6 \leq \alpha < 1$.

(ii) $Z = \{\, x_1x_2 = x_2x_3 = x_1x_3 = 0 \,\} \subset \mathbb{A}^3$. Then $\mathcal{J}(\alpha \cdot Z)$ is equal to $\mathcal{O}_X$ if $0 < \alpha < 3/2$, and it is the ideal (x_1, x_2, x_3) if $3/2 \leq \alpha < 2$. This follows from [9]-III.9.3.4 which gives the formula for multiplier ideals of monomial ideals.

(iii) The multiplier ideals of hyperplane arrangements, and more generally, stratified locally conical divisors are determined in [13], and respectively [16].

Theorem 3.3 ([3], [4]). *For $\alpha > 0$, $V^\alpha\mathcal{O}_X = \mathcal{J}((\alpha - \epsilon) \cdot Z)$, where $0 < \epsilon \ll 1$ and the filtration V of $\mathcal{O}_X$ is taken along Z as in Definition 1.4.*

The relation with Bernstein-Sato polynomials is then given by Proposition 2.2 and Lemma 2.6:

Corollary 3.4. *For $\alpha > 0$,*

$$\mathcal{J}(\alpha \cdot Z) = \{h \in \mathcal{O}_X \mid \alpha < c \text{ if } b_{f,h}(-c) = 0 \},$$

where $f = f_1, \ldots, f_r$ is any set of generators of the ideal $\mathcal{I} \subset \mathcal{O}_X$ of Z.

In particular,

$$\mathrm{lc}(X, Z) = -(\text{biggest root of } b_f(s)),$$

since $b_f(s) = b_{f,1}(s)$ by definition [7], [10].

4 Monodromy of vanishing cycles

The initial scope of the V-filtration of M. Kashiwara and B. Malgrange was to construct vanishing cycles in the category of (regular holonomic) $\mathcal{D}$-modules.

Let X be a smooth complex variety. Denote by $M_{rh}(\mathcal{D}_X)$ the abelian category of regular holonomic $\mathcal{D}_X$-modules, and by $D^b_{rh}(\mathcal{D}_X)$ the derived category of bounded complexes of $\mathcal{D}_X$-modules with regular holonomic cohomology. By A. Beilinson, this is equivalent with the bounded derived category of $M_{rh}(\mathcal{D}_X)$. Let $D^b_c(X)$ be the derived category of bounded complexes of sheaves (in the analytic topology of X) of $\mathbb{C}$-vector spaces with constructible cohomology. The Riemann-Hilbert correspondence generalizing the analogy between the $\mathcal{D}_X$-module $\mathcal{O}_X$ and the constant sheaf $\mathbb{C}_X$ states (see [1]):

Theorem 4.1 (M. Kashiwara, Z. Mebkhout). *Let X be a smooth complex variety. There is a well-defined functor*

$$DR : D^b_{rh}(\mathcal{D}_X) \longrightarrow D^b_c(X)$$

which is an equivalence of categories commuting with the usual six functors. DR also defines an equivalence $M_{rh}(\mathcal{D}_X) \to Perv(X)$, where $Perv(X) \subset D^b_c(X)$ is the subcategory of perverse sheaves.

Let $f \in \mathcal{O}_X$ be a regular function. The vanishing cycles functor ϕ_f on $D^b_c(X)$ and the monodromy action T on it should then have a meaning only in terms of $\mathcal{D}$-modules since the shift $\phi_f[-1]$ restricts as a functor to Perv (X). Let M be a regular holonomic $\mathcal{D}_X$-module. Let $\tilde{M}$ be its direct image under the graph of f, as in section 2. If M is also quasi-unipotent, then there exists a V-filtration indexed by $\mathbb{Q}$ on $\tilde{M}$ along $X \times \{0\}$. If M is not quasi-unipotent, a close version of the following still holds:

Theorem 4.2 (M. Kashiwara, B. Malgrange). *Let $\alpha \in [0,1)$ be a rational number. $Gr^\alpha_V \tilde{M}$ corresponds to the* $\exp(-2\pi i\alpha)$*-eigenspace of $\phi_f[-1](DR(M))$ with respect to the action of the semisimple part T_s of the monodromy.*

Combining this result with Theorem 3.3 and with additional structures such as mixed Hodge modules, one obtains a relation between multiplier ideals and the Hodge spectrum of hypersurface singularities. Let $f : X \to \mathbb{A}^1$ be a regular function. Recall that if $i_x : x \hookrightarrow f^{-1}(0)$ is a point, the Milnor fiber of f at x is is the Milnor fiber of the corresponding holomorphic germ $f : (\mathbb{C}^m, 0) \to (\mathbb{C}, 0)$,

$$M_{f,x} = \{z \in \mathbb{C}^m \mid |z| < \epsilon \text{ and } f(z) = t\}$$

for a fixed t with $0 < |t| < \epsilon \ll 1$. Then

$$H^i(i_x^* \phi_f \mathbb{C}_X) = \tilde{H}^i(M_{f,x}, \mathbb{C}), \tag{5}$$

where $\tilde{H}$ stands for reduced cohomology. These vector spaces are endowed with the monodromy action T and with mixed Hodge structures on which T_s acts as automorphism. Indeed, the mixed Hodge module theory of M. Saito on the left-hand side of (5) recovers the mixed Hodge structure of V. Navarro-Aznar from the right-hand side. As numerical invariants encoding the behaviour of the Hodge filtration F under T_s one has the generalized equivariant Euler characteristics

$$n(i,\alpha) = \sum_j (-1)^j \dim \mathrm{Gr}^i_F \tilde{H}^j(M_{f,x}, \mathbb{C})_\alpha,$$

where $\alpha \in \mathbb{Q} \cap [0,1)$, $i \in \{0, \ldots, m-1\}$, and the subscript α stands for the eigenspace of T_s with eigenvalue $exp(2\pi i\alpha)$. These invariants form the Hodge spectrum of f introduced by J. Steenbrink [17]. For $\alpha \in (0,1]$, let

$$n_{\alpha,x}(f) = (-1)^{n-1} n(m-1, 1-\alpha),$$

so that $n_{\alpha,x}(f)$ describe the spectrum for the smallest piece of the Hodge filtration.

Example 4.3. If $f = x^2 + y^3$ and $x = (0,0) \in \mathbb{A}^2$, then $n_{\alpha,x}(f)$ is zero for $\alpha \notin \{5/6, 7/6\}$, and is 1 otherwise.

On the other hand, for every jumping number $\alpha \in (0,1]$ of (X, Z) where Z is the zero set of a regular function f, define the inner jumping multiplicity at x

$$n_{\alpha,x}(Z) = \dim \mathcal{J}((1-\epsilon)\alpha \cdot Z)/\mathcal{J}((1-\epsilon)\alpha \cdot Z) + \delta \cdot x),$$

where $0 < \epsilon \ll \delta \ll 1$. It is proved in [2] that $n_{\alpha,x}(Z)$ is finite and does not depend on ϵ and δ. Let $\widetilde{\mathcal{O}_X}$ be the $\mathcal{D}$-module direct image of $\mathcal{O}_X$ under the graph of f. In connection with Theorem 3.3 it is crucial to observe that the smallest piece of the Hodge filtration on $V^\alpha \widetilde{\mathcal{O}_X}$ is exactly $V^\alpha \mathcal{O}_X$. Then the above arguments lead to:

Theorem 4.4 ([2], [4]). *For $\alpha \in (0,1]$,*

$$n_{\alpha,x}(f) = n_{\alpha,x}(Z).$$

References

1. A. Borel, P.-P. Grivel, B. Kaup, A. Haefliger, B. Malgrange and F. Ehlers – *Algebraic D-modules*, Perspectives in Mathematics, vol. 2, Academic Press Inc., Boston, MA, 1987.
2. N. Budur – "On Hodge spectrum and multiplier ideals", *Math. Ann.* **327** (2003), no. 2, p. 257–270.
3. N. Budur, M. Mustaţă and M. Saito – "Bernstein-Sato polynomials of arbitrary varieties", 2004, preprint.
4. N. Budur and M. Saito – "Multiplier ideals, V-filtration, and spectrum", to appear in J. Algebraic Geom.
5. M. Kashiwara – "Vanishing cycle sheaves and holonomic systems of differential equations", Algebraic geometry (Tokyo/Kyoto, 1982), Lecture Notes in Math., vol. 1016, Springer, Berlin, 1983, p. 134–142.
6. M. Kashiwara – *D-modules and microlocal calculus*, Translations of Mathematical Monographs, vol. 217, American Mathematical Society, Providence, RI, 2003.
7. J. Kollár – "Singularities of pairs", Algebraic geometry—Santa Cruz 1995, Proc. Sympos. Pure Math., vol. 62, Amer. Math. Soc., Providence, RI, 1997, p. 221–287.
8. G. Laumon – "Transformations canoniques et spécialisation pour les D-modules filtrés", *Astérisque* (1985), no. 130, p. 56–129, Differential systems and singularities (Luminy, 1983).
9. R. Lazarsfeld – "Positivity in algebraic geometry", book to appear in 2004.

10. B. Lichtin – "Poles of $|f(z,w)|^{2s}$ and roots of the b-function", *Ark. Mat.* **27** (1989), no. 2, p. 283–304.
11. G. Lyubeznik – "On Bernstein-Sato polynomials", *Proc. Amer. Math. Soc.* **125** (1997), no. 7, p. 1941–1944.
12. B. Malgrange – "Polynômes de Bernstein-Sato et cohomologie évanescente", Analysis and topology on singular spaces, II, III (Luminy, 1981), Astérisque, vol. 101, Soc. Math. France, Paris, 1983, p. 243–267.
13. M. Mustaţă – "Multiplier ideals of hyperplane arrangements", 2004, `math.AG/0402232`.
14. A. M. Nadel – "Multiplier ideal sheaves and Kähler-Einstein metrics of positive scalar curvature", *Ann. of Math. (2)* **132** (1990), no. 3, p. 549–596.
15. C. Sabbah – "D-modules et cycles évanescents (d'après B. Malgrange et M. Kashiwara)", Géométrie algébrique et applications, III (La Rábida, 1984), Travaux en Cours, vol. 24, Hermann, Paris, 1987, p. 53–98.
16. M. Saito – "Multiplier ideals, b-function, and spectrum", 2004.
17. J. H. M. Steenbrink – "The spectrum of hypersurface singularities", *Astérisque* (1989), no. 179-180, p. 11, 163–184, Actes du Colloque de Théorie de Hodge (Luminy, 1987).

Hecke orbits on Siegel modular varieties

Ching-Li Chai

Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19003, U.S.A.
chai@math.upenn.edu

Summary. We sketch a proof of the Hecke orbit conjecture for the Siegel modular variety $\mathcal{A}_{g,n}$ over $\overline{\mathbb{F}_p}$, where p is a prime number, fixed throughout this article. We also explain several techniques developed for the Hecke orbit conjecture, including a generalization of the Serre–Tate coordinates.

1 Introduction

In this article we give an overview of the proof of a conjecture of F. Oort that every prime-to-p Hecke orbit in the moduli space $\mathcal{A}_g$ of principally polarized abelian varieties over $\overline{\mathbb{F}_p}$ is dense in the *leaf* containing it. See Conjecture 4.1 for a precise statement, Definition 2.1 for the definition of Hecke orbits, and Definition 3.1 for the definition of a leaf. Roughly speaking, a *leaf* is the locus in $\mathcal{A}_g$ consisting of all points s such that the principally quasi-polarized Barsotti–Tate group attached to s belongs to a fixed isomorphism class, while the prime-to-p Hecke orbit of a closed point x consists of all closed points y such that there exists a prime-to-p quasi-isogeny from A_x to A_y which preserves the polarizations. Here $(A_x, \lambda_x), (A_y, \lambda_y)$ denote the principally polarized abelian varieties attached to x, y respectively; a prime-to-p quasi-isogeny is the composition of a prime-to-p isogeny with the inverse of a prime-to-p isogeny.

For clarity in logic, it is convenient to separate the *prime-to-p Hecke orbit conjecture*, or the *Hecke orbit conjecture* for short, into two parts (see Conjecture4.1):

(i) the continuous part, which asserts that the Zariski closure of a prime-to-p Hecke orbit has the same dimension as the dimension of the leaf containing it, and

(ii) the discrete part, which asserts that the prime-to-p Hecke correspondences operate transitively on the set of irreducible components of every leaf; see Conjecture 4.1.

The prime-to-p Hecke correspondences on $\mathcal{A}_g$ form a large family of symmetries on $\mathcal{A}_g$. In characteristic 0, each prime-to-p Hecke orbit is dense in the metric topology of $\mathcal{A}_g(\mathbb{C})$, because of complex uniformization. In characteristic p, it is reasonable to expect that every prime-to-p Hecke orbit is "as large as possible". The decomposition of $\mathcal{A}_g$ into the disjoint union of leaves constitutes a "fine" geometric structure of $\mathcal{A}_g$, existing only in characteristic p and called *foliation* in [26]. The prime-to-p Hecke orbit conjecture says, in particular, that the foliation structure on $\mathcal{A}_g$ over $\overline{\mathbb{F}_p}$ is determined by the Hecke symmetries.

The prime-to-p Hecke orbit $\mathcal{H}^{(p)}(x)$ of a point x is a countable subset of $\mathcal{A}_g$. Experience indicates that determining the Zariski closure of a countable subset of an algebraic variety in positive characteristic is often difficult. We developed a number of techniques to deal with the Hecke orbit conjecture. They include

(M) the ℓ-adic monodromy of leaves,

(C) the theory of canonical coordinates on leaves, generalizing Serre–Tate parameters on the local moduli spaces of ordinary abelian varieties,

(R) a rigidity result for p-divisible formal groups,

(S) a trick "splitting at supersingular point",

(H) hypersymmetric points,

and will be described in §5, §7, §8, §11, and §10 respectively. We hope that the above techniques will also be useful in other situations. Among them, the most significant is perhaps the theory of canonical coordinates on leaves, which generalizes the Serre–Tate coordinates for the local moduli space of ordinary abelian varieties. At a non-ordinary closed point $x \in \mathcal{A}_g(\overline{\mathbb{F}_p})$, there is no description of the formal completion $\mathcal{A}_g^{/x}$ of $\mathcal{A}_g$ at x comparable to what the Serre–Tate theory provides. But if we restrict to the leaf $\mathcal{C}(x)$ passing through x, then there is a "good" structure theory for the formal completion $\mathcal{C}(x)^{/x}$. To get an idea, the simplest situation is when the Barsotti–Tate group $A_x[p^\infty]$ is isomorphic to a direct product $X \times Y$, where X, Y are isoclinic Barsotti–Tate groups over $\overline{\mathbb{F}_p}$ of Frobenius slopes $\mu_{_X}$, $\mu_{_Y}$ respectively, and $\mu_{_X} < \mu_{_Y} = 1-\mu_{_X}$. In this case, $\mathcal{C}(x)^{/x}$ has a natural structure as an isoclinic p-divisible formal group of height $\frac{g(g+1)}{2}$, Frobenius slope $\mu_{_Y}-\mu_{_X}$, and dimension $\dim(\mathcal{C}(x)^{/x}) = (\mu_{_Y}-\mu_{_X}) \cdot \frac{g(g+1)}{2}$. Moreover, there is a natural isomorphism of V-isocrystals

$$\mathrm{M}(\mathcal{C}(x)^{/x}) \otimes_{\mathbb{Z}} \mathbb{Q} \xrightarrow{\sim} \mathrm{Hom}^{\mathrm{sym}}_{W(\overline{\mathbb{F}_p})}(\mathrm{M}(X), \mathrm{M}(Y)) \otimes_{\mathbb{Z}} \mathbb{Q}\,,$$

where $\mathrm{M}(\mathcal{C}(x)^{/x}), \mathrm{M}(X), \mathrm{M}(Y)$ denote the Cartier–Dieudonné modules of $\mathcal{C}(x)^{/x}$, X, Y respectively, $W(\overline{\mathbb{F}_p})$ is the ring of p-adic Witt vectors, and the right-hand side of the formula denotes the symmetric part of the internal Hom,

with respect to the involution induced by the principal polarization on A_x. In the general case, $\mathcal{C}(x)^{/x}$ is built up from a successive system of fibrations, and each fibration has a natural structure of a torsor for a suitable p-divisible formal group.

The fundamental idea underlying our method is to exploit the action of the local stabilizer subgroups. Recall that the prime-to-p Hecke correspondences come from the action of the group $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ on the prime-to-p tower of the moduli space $\mathcal{A}_g$. Here the symplectic group Sp_{2g} in $2g$ variables is viewed as a split group scheme over $\mathbb{Z}$, and $\mathbb{A}_f^{(p)}$ denotes the restricted product of $\mathbb{Q}_\ell$'s, where ℓ runs through all primes not equal to p. Suppose that $Z \subset \mathcal{A}_g$ is a closed subscheme of $\mathcal{A}_g$ which is stable under all prime-to-p Hecke correspondences. It is clear that for any closed point $x \in Z(k)$, the subscheme Z is stable under the set $\mathrm{Stab}(x)$ consisting of all prime-to-p Hecke correspondences having x as a fixed point. This is an elementary fact, referred to as the *local stabilizer principle*, and will be rephrased in a more usable form below.

The stabilizer $\mathrm{Stab}(x)$ comes from the unitary group G_x over $\mathbb{Q}$ attached to the pair $(\mathrm{End}_k(A_x)\otimes_{\mathbb{Z}}\mathbb{Q}, *_x)$, where $*_x$ denotes the Rosati involution on the semisimple algebra $\mathrm{End}_k(A_x)\otimes_{\mathbb{Z}}\mathbb{Q}$. Notice that $G_x = \mathrm{U}(\mathrm{End}_k(A_x)\otimes_{\mathbb{Z}}\mathbb{Q}, *_x)$ has a natural $\mathbb{Z}$-model attached to the $\mathbb{Z}$-lattice $\mathrm{End}_k(A_x) \subset \mathrm{End}_k(A_x)\otimes_{\mathbb{Z}}\mathbb{Q}$, and we denote by $G_x(\mathbb{Z}_p)$ the group of $\mathbb{Z}_p$-valued points for that $\mathbb{Z}$-model. The group $G_x(\mathbb{Z}_p)$ is a subgroup of the p-adic group $\mathrm{U}(\mathrm{End}_k(A_x[p^\infty]), *_x)$; the latter operates naturally on the formal completion $\mathcal{A}_g^{/x}$ by deformation theory. With the help of the weak approximation theorem, applied to G_x, the local stabilizer principle then says that the formal completion $Z^{/x}$ of Z at x, as a closed formal subscheme of $\mathcal{A}_g^{/x}$, is stable under the action of $G_x(\mathbb{Z}_p)$. See §6 for details.

The tools (C), (R), (H) mentioned above allows us to use the local stabilizer principle effectively. A useful consequence is that, if Z is a closed subscheme of $\mathcal{A}_g$ stable under all prime-to-p Hecke correspondences, and x is a *split hypersymmetric* point of Z, then Z contains an irreducible component of the leaf passing through x; see Theorem 10.6. Here a *split* point of $\mathcal{A}_g$ is a point y of $\mathcal{A}_g$ such that A_y is isogenous to a product of abelian varieties where each factor has at most two slopes, while a *hypersymmetric* point of $\mathcal{A}_g$ is a point y of $\mathcal{A}_g$ such that $\mathrm{End}_k(A_y)\otimes_{\mathbb{Z}}\mathbb{Z}_p \xrightarrow{\sim} \mathrm{End}_k(A_y[p^\infty])$. It should not come as a surprise that the local stabilizer principle gives us a lot of information at a hypersymmetric point, where the local stabilizer subgroup is quite large.

Let $x \in \mathcal{A}_g(\overline{\mathbb{F}_p})$ be a closed point of $\mathcal{A}_g$. Let $\overline{\mathcal{H}^{(p)}(x)}$ be the Zariski closure of the prime-to-p Hecke orbit $\mathcal{H}^{(p)}(x)$ of x, and let $\overline{\mathcal{H}^{(p)}(x)}^0 := \overline{\mathcal{H}^{(p)}(x)}\cap\mathcal{C}(x)$.[1] The conclusion of the last paragraph tells us that, to show that $\overline{\mathcal{H}^{(p)}(x)}$ is

[1] In fact $\overline{\mathcal{H}^{(p)}(x)}^0$ is the open subscheme of $\overline{\mathcal{H}^{(p)}(x)}$ consisting of all points y of $\overline{\mathcal{H}^{(p)}(x)}$ such that the Newton polygon of A_y is equal to the Newton polygon of A_x.

irreducible, it suffices to show that $\overline{\mathcal{H}^{(p)}(x)}^0$ contains a split hypersymmetric point. The result that $\overline{\mathcal{H}^{(p)}(x)}^0$ contains a split hypersymmetric point is accomplished through what we call the *Hilbert trick* and the *splitting at supersingular points*.

The Hilbert trick refers to a special property of $\mathcal{A}_g$: Up to an isogeny correspondence, there exists a Hilbert modular subvariety of maximal dimension passing through any given $\overline{\mathbb{F}_p}$-valued point of $\mathcal{A}_g$; see §9. To elaborate a bit, let x be a given point of $\mathcal{A}_g(\overline{\mathbb{F}_p})$. The Hilbert trick tells us that there exists an isogeny correspondence f, from a g-dimensional Hilbert modular subvariety $\mathcal{M}_E \subset \mathcal{A}_g$ to $\mathcal{A}_g$, whose image contains x. The Hilbert modular variety above is attached to a commutative semisimple subalgebra E of $\mathrm{End}_{\overline{\mathbb{F}_p}}(A_x) \otimes_{\mathbb{Z}} \mathbb{Q}$, such that $[E : \mathbb{Q}] = g$ and E is fixed by the Rosati involution. There are Hecke correspondences on $\mathcal{M}_E$ coming from the semisimple algebraic group $\mathrm{SL}(2, E)$ over $\mathbb{Q}$, and $\mathrm{SL}(2, E)$ can be regarded as a subgroup of the symplectic group Sp_{2g}. The isogeny correspondence f above respects the prime-to-p Hecke correspondences. So, among other things, the Hilbert trick tells us that, for an $\overline{\mathbb{F}_p}$-point x of $\mathcal{A}_g$ as above, the Hecke orbit $\mathcal{H}^{(p)}(x)$ contains the f-image of a prime-to-p Hecke orbit $\mathcal{H}_E^{(p)}(\tilde{x})$ on the Hilbert modular variety $\mathcal{M}_E$, where $\tilde{x}$ is a pre-image of x under the isogeny correspondence f.

A consequence of the Hilbert trick and the local stabilizer principle, is the following trick of "splitting at supersingular points"; see Theorem 11.3. This "splitting trick" says that, in the interior of the Zariski closure of a given Hecke orbit, there exists a point y such that A_y is a *split* abelian variety. The last clause means that A_y is isogenous to a product of abelian varieties, where each factor abelian variety has at most two slopes.

One can formulate the notion of leaves and the Hecke orbit conjecture for Hilbert modular varieties. It turns out that the prime-to-p Hecke orbit conjecture for Hilbert modular varieties is easier to solve than Siegel modular varieties, reflecting the fact that a Hilbert modular variety comes from a reductive group G over $\mathbb{Q}$ such that every $\mathbb{Q}$-simple factor of the adjoint group G^{ad} has $\mathbb{Q}$-rank one. The trick "splitting at supersingular points" and a standard technique in algebraic geometry implies that, when one tries to prove the prime-to-p Hecke orbit conjecture, one may assume that the point x of $\mathcal{A}_g$ is defined over $\overline{\mathbb{F}_p}$ and the abelian variety A_x is split. Now we apply the Hilbert trick to x. To simplify the exposition, we will assume, for simplicity, that we have a Hilbert modular variety $\mathcal{M}_E$ in $\mathcal{A}_g$ passing through the point x, suppressing the isogeny correspondence f. We will also assume (or "pretend") that the leaf $\mathcal{C}_E(x)$ on $\mathcal{M}_E$ passing through x is the intersection of $\mathcal{C}(x)$ with $\mathcal{M}_E$. (The last assumption is not far from the truth, if we interpret "intersection" as a suitable fiber product.) Notice that the commutative semisimple algebra E is a product of totally real number fields F_i, $i = 1, \ldots, m$, and $F_i \otimes \mathbb{Q}_p$ is a field for each i, because the abelian variety A_x is split.

It is easy to see that every leaf in $\mathcal{M}_E$ contains a hypersymmetric point y of $\mathcal{A}_g$. Moreover A_y is split because A_x is split. So if we can prove the Hecke orbit conjecture for $\mathcal{M}_E$, then we will know that the Zariski closure of the Hecke orbit $\mathcal{H}^{(p)}(x)$ in $\mathcal{C}(x)$ contains a split hypersymmetric point y. Therefore the prime-to-p Hecke orbit conjecture for Hilbert modular varieties implies the continuous part of the prime-to-p Hecke orbit conjecture for $\mathcal{A}_g$.

The general methods we developed, when applied to a Hilbert modular variety $\mathcal{M}_E$, produce a proof of the continuous part of the prime-to-p Hecke orbit conjecture for $\mathcal{M}_E$. So the prime-to-p Hecke orbit conjecture for $\mathcal{A}_g$ is reduced to the discrete part of the prime-to-p Hecke orbit conjecture for both $\mathcal{A}_g$ and the Hilbert modular varieties.

The discrete part of the Hecke orbit conjecture is equivalent to the statement that every non-supersingular leaf is irreducible, see Theorem 5.1; the same holds for Hilbert modular varieties. Generally such irreducibility statements do not come by easily; so far there is no unified approach which works for all modular varieties of PEL-type. Using the techniques (H) and (M), one can reduce the discrete part of the Hecke orbit conjecture for $\mathcal{A}_g$ to the statement that the prime-to-p Hecke correspondences operate transitively on the set of irreducible components of every non-supersingular Newton polygon stratum in $\mathcal{A}_g$. Happily the results of Oort in [24], [25] can be applied to settle the latter irreducibility statement; see Theorem 13.1, [21], and references cited in 13.1.

The discrete part of the Hecke orbit conjecture for the Hilbert modular varieties, however, requires a different approach, based on the Lie-alpha stratification of Hilbert modular varieties, and the following property of Hilbert modular varieties: For each slope datum ξ for $\mathcal{M}_E$, there exists a Lie-alpha stratum $\mathcal{N}_{\underline{e},\underline{a}} \subset \mathcal{M}_E$, contained in the Newton polygon stratum in $\mathcal{M}_E$ attached to the given slope datum ξ, and a dense open subset $\mathcal{U}_{\underline{e},\underline{a}}$ of $\mathcal{N}_{\underline{e},\underline{a}}$ such that $\mathcal{U}_{\underline{e},\underline{a}}$ is a leaf in $\mathcal{M}_E$. Here a slope datum for $\mathcal{M}_E$ is a function which to each prime ideal $\wp$ of $\mathcal{O}_E/p\mathcal{O}_E$ attaches a set of the form $\{\mu_\wp, 1-\mu_\wp\}$, where $0 \le \mu_\wp \le \frac{1}{2}$, and the denominator of $\mu_\wp$ divides $2[E_\wp : \mathbb{Q}_p]$. There is a natural *slope stratification* of $\mathcal{M}_E$, indexed by the set of slope data for $\mathcal{M}_E$. The *Lie-alpha stratification* of $\mathcal{M}_E$ is defined in terms of the Lie type and alpha type of the $\mathcal{O}_E$-abelian varieties attached to points of $\mathcal{M}_E$; the Lie type (resp. alpha type) of an $\mathcal{O}_E$ abelian variety A over $\overline{\mathbb{F}_p}$ refers to the (semi-simplification of) the linear representation of the algebra $\mathcal{O}_E \otimes_{\mathbb{F}_p} \overline{\mathbb{F}_p}$ on the vector space $\mathrm{Lie}(A)$ (resp. $\mathrm{Hom}(\alpha_p, A)$) over $\mathbb{F}_p$. A critical step in the proof of the discrete part of the Hecke orbit conjecture for Hilbert modular varieties, due to C.-F. Yu, is the construction of "enough" deformations for understanding the incidence relation of the Lie-alpha stratification.

Details of the proof of the Hecke orbit conjecture will appear in a manuscript with F. Oort. All unattributed results are due to suitable subsets of {Oort, Yu, Chai}. The author is responsible for all errors and imprecisions.

Acknowledgments. It is a pleasure to thank F. Oort for many stimulating discussions on the Hecke orbit conjecture over the last ten years, and for generously sharing his insights on the foliation structure. The author would like to thank C.-F. Yu for the enjoyable collaboration on the Hecke orbit conjecture for Hilbert modular varieties; a conversation with him in the spring of 2002 led to the discovery of the canonical coordinates on leaves. The author thanks the referee for a very careful reading and many suggestions. This article was completed when the author visited the National Center for Theoretical Sciences in Taipei, from January to August of 2004. The author thanks both NCTS/TPE-Math and the Department of Mathematics of the National Taiwan University for hospitality.

This work was partially supported by a grant from the National Science Council of Taiwan and by grant DMS01-00441 from the National Science Foundation.

2 Hecke orbits

Let p be a prime number, fixed throughout this article. Let $\mathbb{Z}_f^{(p)} = \prod_{\ell \neq p} \mathbb{Z}_\ell$, where ℓ runs through all prime numbers different from p. Let $\mathbb{A}_f^{(p)}$ be the restricted product $\prod'_{\ell \neq p} \mathbb{Q}_\ell$ of $\mathbb{Q}_\ell$'s for $\ell \neq p$, naturally isomorphic to $\mathbb{Z}_f^{(p)} \otimes_{\mathbb{Z}} \mathbb{Q}$ and known as the ring of prime-to-p finite adèles attached to $\mathbb{Q}$.

Let k be an algebraically closed field of characteristic p. Choose and fix an isomorphism $\zeta : \mathbb{Z}_f^{(p)} \xrightarrow{\sim} \mathbb{Z}_f^{(p)}(1)$ over k, i.e., a compatible system of isomorphisms $\zeta_m : \mathbb{Z}/m\mathbb{Z} \simeq \mu_m(k)$, where m runs through all positive integers which are not divisible by p. For any natural number g and any integer $n \geq 3$ with $(n, p) = 1$, denote by $\mathcal{A}_{g,n}$ the moduli space over k classifying g-dimensional principally polarized abelian varieties with a symplectic level-n structure with respect to ζ.

For any two integers $n_1, n_2 \geq 3$, such that $(p, n_1 n_2) = 1$ and $n_1 \mid n_2$, there is a canonical map $\mathcal{A}_{g,n_2} \to \mathcal{A}_{g,n_1}$. Denote by $\mathcal{A}_{g,(p)}$ the resulting projective system of the moduli spaces $\mathcal{A}_{g,n}$, where n runs through all integers $n \geq 3$ with $(p, n) = 1$. By definition, a geometric point of $\mathcal{A}_{g,(p)}(k)$ corresponds to a triple (A, λ, η), where A is a g-dimensional principally polarized abelian variety over k, λ is a principal polarization on A, and η is a level-$\mathbb{Z}_f^{(p)}$ structure on A, i.e., η is a symplectic isomorphism from $\prod_{\ell \neq p} A[\ell^\infty]$ to $(\mathbb{Z}_f^{(p)})^{2g}$, where the free $\mathbb{Z}_f^{(p)}$-module $(\mathbb{Z}_f^{(p)})^{2g}$ is endowed with the standard symplectic pairing.

From the definition of $\mathcal{A}_{g,(p)}$ we see that there is a natural action of $\mathrm{Sp}_{2g}(\mathbb{Z}_f^{(p)})$ on $\mathcal{A}_{g,(p)}$, operating as covering transformations over the moduli stack $\mathcal{A}_g$. Moreover there is a natural action of the group $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ on $\mathcal{A}_{g,(p)}$, extending the action of $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ and gives a much larger collection of symmetries on the tower $\mathcal{A}_{g,(p)}$. The automorphism h_γ of $\mathcal{A}_{g,(p)}$ attached to an element $\gamma \in \mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ is characterized by the following property. There is a prime-to-p isogeny α_γ from the universal abelian scheme A to $h_\gamma^* A$ such that

$$\eta \circ \alpha_\gamma[(p)] = \gamma \circ \eta\,,$$

where $\alpha_\gamma[(p)]$ denotes the prime-to-p quasi-isogeny induced by α_γ, between the prime-to-p-divisible groups attached to A and $h_\gamma^* A$ respectively. On each individual moduli space $\mathcal{A}_{g,n}$, the action of $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ induces algebraic correspondences to itself; they are the classical Hecke correspondences on the Siegel moduli spaces.

Definition 2.1. *Let $n \geq 3$ be an integer, $(n,p) = 1$. Let $x \in \mathcal{A}_{g,n}(k)$ be a geometric point of $\mathcal{A}_{g,n}$, and let $\tilde{x} \in \mathcal{A}_{g,(p)}(k)$ be a geometric point of the tower $\mathcal{A}_{g,(p)}$ above x.*

(i) *The* prime-to-p Hecke orbit *of x in $\mathcal{A}_{g,n}$, denoted by $\mathcal{H}^{(p)}(x)$, or $\mathcal{H}(x)$ for short, is the image of the subset $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})\cdot\tilde{x}$ of $\mathcal{A}_{g,(p)}$ under the projection map $\pi_n : \mathcal{A}_{g,(p)} \to \mathcal{A}_{g,n}$.*

(ii) *Let ℓ be a prime number, $\ell \neq p$. The* ℓ-adic Hecke orbit *of x in $\mathcal{A}_{g,n}$, denoted by $\mathcal{H}_\ell(x)$, is the image of $\mathrm{Sp}_{2g}(\mathbb{Q}_\ell)\cdot\tilde{x}$ under $\pi : \mathcal{A}_{g,(p)} \to \mathcal{A}_{g,n}$.*

Remark 2.2.

(i) It is easy to see that the definition of $\mathcal{H}_\ell(x)$ does not depend on the choice of $\tilde{x}$. One can also use the ℓ-adic tower above $\mathcal{A}_{g,n}$ to define the ℓ-adic Hecke orbits.

(ii) Explicitly, the countable set $\mathcal{H}^{(p)}(x)$ (resp. $\mathcal{H}_\ell(x)$) consists of all points $y \in \mathcal{A}_{g,n}(k)$ such that there exists an abelian variety B over k and two prime-to-p isogenies (resp. ℓ-power isogenies) $\alpha : B \to A_x$, $\beta : B \to A_y$ such that $\alpha^*(\lambda_x) = \beta^*(\lambda_y)$.

(iii) The moduli stack $\mathcal{A}_g$ over k has a natural pro-étale $\mathrm{GSp}_{2g}(\mathbb{Z}_f^{(p)})$ cover; and the group $\mathrm{GSp}_{2g}(\mathbb{A}_f^{(p)})$ operate on the projective limit. Then for any geometric point $x \in \mathcal{A}_{g,n}(k)$, we can define the $\mathrm{GSp}_{2g}(\mathbb{A}_f^{(p)})$-orbit of x and the $\mathrm{GSp}_{2g}(\mathbb{Q}_\ell)$-orbit of x as in Definition 2.1 using the pro-étale $\mathrm{GSp}_{2g}(\mathbb{Z}_f^{(p)})$-tower. Explicitly, the $\mathrm{GSp}_{2g}(\mathbb{A}_f^{(p)})$-orbit of x (resp. the $\mathrm{GSp}_{2g}(\mathbb{Q}_\ell)$-orbit of x) on $\mathcal{A}_{g,n}$ for a geometric point $x \in \mathcal{A}_{g,n}(k)$ can be explicitly described as follows. It consists of all points $y \in \mathcal{A}_{g,n}(k)$ such that there

exists a prime-to-p isogeny (resp. an ℓ-power isogeny) $\beta : A_x \to A_y$ such that $\beta^*(\lambda_y) = m(\lambda_x)$, where m is a prime-to-p positive integer (resp. a non-negative integer power of ℓ.)

Remark 2.3. In Definition 2.1 we used the group $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ to define the prime-to-p Hecke orbits of a closed point x in $\mathcal{A}_{g,n} \to \mathrm{Spec}(k)$. Geometrically that means to consider the orbit of x under all prime-to-p symplectic quasi-isogenies. One can also consider the orbit of x under all symplectic quasi-isogenies, or, as a slight variation, the orbit of x under all quasi-isogenies which preserve the polarization up to a multiple. The latter was used in [22, 15.A]. We considered only the prime-to-p Hecke correspondences in this article, since they are finite étale correspondences on $\mathcal{A}_{g,n}$, and reflect well the underlying group-theoretic properties.

For any totally real number field F and any integer $n \geq 3$, $(n, p) = 1$, denote by $\mathcal{M}_{F,n}$ the Hilbert modular variety over k attached to F as defined in [8]. Just as in the case of Siegel modular varieties, the varieties $\mathcal{M}_{F,n}$ over k form a projective system, with a natural action by the group $\mathrm{SL}_2(F \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)})$. The prime-to-$p$ Hecke orbit $\mathcal{H}_F^{(p)}(x)$ and the ℓ-adic Hecke orbit $\mathcal{H}_{F,\ell}(x)$ of a geometric point $x \in \mathcal{M}_{F,n}(k)$ are, by definition, the image in $\mathcal{M}_{F,n}(k)$ of $\mathrm{SL}_2(F \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)}) \cdot \tilde{x}$ and $\mathrm{SL}_2(F \otimes_{\mathbb{Q}} \mathbb{Q}_\ell) \cdot \tilde{x}$ respectively, where $\tilde{x}$ is a k-valued point, lying above x, of the projective system $\mathcal{M}_{F,(p)} := \{\mathcal{M}_{F,m} \,:\, (m, p) = 1\}$.

More generally, if $E = F_1 \times \cdots \times F_r$ is a product of totally real number fields, and $n \geq 3$ is a positive integer not divisible by p, we can define the Hilbert modular variety $\mathcal{M}_E$ over k attached to E, in the same fashion as in [8], with $\mathcal{O}_E := \mathcal{O}_{F_1} \times \cdots \times \mathcal{O}_{F_r}$, as follows. For any k-scheme S, $\mathcal{M}_E(S)$ is the set of isomorphism classes of triples of the form

$$(A \to S,\ \alpha : \mathcal{O}_E \to \mathrm{End}_S(A),\ \phi : A \otimes_{\mathcal{O}_E} L \xrightarrow{\sim} A^t)\,,$$

where α is a ring homomorphism, L is an invertible $\mathcal{O}_E$ module with a notion of positivity $L^+ \subset L \otimes_{\mathbb{Q}} \mathbb{R}$, and ϕ is an isomorphism of abelian varieties such that for each element $\lambda \in L$, the homomorphism $\phi_\lambda : A \to A^t$ attached to λ is symmetric, and ϕ_λ is a polarization of A if λ is positive. Then we have a canonical isomorphism $\mathcal{M}_E = \mathcal{M}_{F_1} \times \cdots \times \mathcal{M}_{F_r}$. The notion of Hecke orbits generalizes in the obvious way to the present situation.

Remark 2.4. The notion of prime-to-p Hecke orbits can be generalized to other modular varieties over k of PEL-type in a natural way. Furthermore, one expects that the notion of prime-to-p Hecke orbits can be generalized to the reduction over k of a Shimura variety X, with satisfactory properties.

3 Leaves

In this section we work over an algebraically closed field k of characteristic $p > 0$. The modular varieties $\mathcal{A}_{g,n}$ and $\mathcal{M}_{E,n}$ are considered over the fixed based field k.

Theorem 3.1 (Oort). *Let $n \geq 3$ be an integer, $(n, p) = 1$. Let $x \in \mathcal{A}_{g,n}(k)$ be a geometric point of $\mathcal{A}_{g,n}$.*

(i) *There exists a unique reduced constructible subscheme $\mathcal{C}(x)$ of $\mathcal{A}_{g,n}$, called the* leaf *passing through x, characterized by the following property. For every algebraically closed field $K \supseteq k$, $\mathcal{C}(x)(K)$ consists of all elements $y \in \mathcal{A}_{g,n}(K)$ such that*

$$(A_x[p^\infty], \lambda_x[p^\infty]) \times_{\mathrm{Spec}(k)} \mathrm{Spec}(K) \simeq (A_y[p^\infty], \lambda_y[p^\infty]),$$

where $\lambda_x[p^\infty], \lambda_y[p^\infty]$ are the principal quasi-polarizations induced by the principal polarizations λ_x, λ_y on the Barsotti–Tate groups $A_x[p^\infty], A_y[p^\infty]$ respectively.

(ii) *The leaf $\mathcal{C}(x)$ is a locally closed subscheme of $\mathcal{A}_{g,n}$. Moreover it is smooth over k.*

Remark 3.2.

(i) Theorem 3.1 is proved in [26, 3.3, 3.14]. The claim that the subset of $\mathcal{A}_{g,n}(k)$ consisting of all geometric points y such that $(A_y[p^\infty], \lambda_y[p^\infty])$ is isomorphic to $(A_x[p^\infty], \lambda_x[p^\infty])$ is the set of geometric points of a constructible subset of $\mathcal{A}_{g,n}$, follows from the following fact, proved in Manin's thesis [16]: A Barsotti–Tate group over k of a given height h is determined, up to non-unique isomorphism, by its truncation modulo a sufficiently high level $N \geq N(h)$.

(ii) T. Zink showed, in a letter to C.-L. Chai dated May 1, 1999, the following generalization of Manin's result: A *crystal* M over k is determined, up to non-unique isomorphisms, by its quotient modulo p^N, for some suitable $N > 0$ depending only on the height of M and the maximum among the slopes of M.

(iii) In [26], $\mathcal{C}(x)$ is called the *central leaf* passing through x.

(iv) It is clear from the definition that each leaf in $\mathcal{A}_{g,n}$ is stable under all prime-to-p Hecke correspondences. In particular, the Hecke orbit $\mathcal{H}^{(p)}(x)$ is contained in the leaf $\mathcal{C}(x)$ passing through x.

(v) Every leaf is contained in a *Newton polygon stratum* of $\mathcal{A}_{g,n}$, and every Newton polygon stratum is a disjoint union of leaves. Recall that a *Newton polygon stratum* $W_\xi(\mathcal{A}_{g,n})$ in $\mathcal{A}_{g,n}$ over k is, by definition, the subset of $\mathcal{A}_{g,n}$ such that $W_\xi(\mathcal{A}_{g,n})(K)$ consists of all K-valued points y of $\mathcal{A}_{g,n}$ such

that the Newton polygon of $A_y[p^\infty]$ is equal to ξ, for all fields $K \supset k$.[2] By Grothendieck–Katz, $W_\xi(\mathcal{A}_{g,n})$ is a locally closed subset of $\mathcal{A}_{g,n}$; see [14] for a proof. There are infinitely many leaves in $\mathcal{A}_{g,n}$ if $g \geq 2$. In particular the decomposition of $\mathcal{A}_{g,n}$ into a disjoint union of leaves is not a stratification in the usual sense: There are infinitely many leaves, and the closure of some leaves contain infinitely many leaves.

Examples 3.3.

(i) The ordinary locus of $\mathcal{A}_{g,n}$, that is the largest open subscheme of $\mathcal{A}_{g,n}$ over which each geometric fiber of the universal abelian scheme is an ordinary abelian variety, is a leaf.

(ii) The "almost ordinary" locus of $\mathcal{A}_{g,n}$, or, the locus consisting of all geometric points x such that the maximal étale quotient of the attached Barsotti–Tate group $A_x[p^\infty]$ has height $g-1$, is a leaf.

(iii) Every supersingular leaf in $\mathcal{A}_{g,n}$ is finite over k. Hence there are infinitely many supersingular leaves in $\mathcal{A}_{g,n}$ if $g \geq 2$.

(iv) Consider the Newton polygon stratum $\mathcal{W}_\xi(\mathcal{A}_{3,n})$ in $\mathcal{A}_{3,n}$, where the Newton polygon ξ has slopes $(\frac{1}{3}, \frac{2}{3})$. Every leaf $\mathcal{C}$ contained in $\mathcal{W}_\xi(\mathcal{A}_{3,n})$ is two-dimensional, while $\dim(\mathcal{W}_\xi(\mathcal{A}_{3,n})) = 3$.

Proposition 3.4. *Let $\mathcal{C}$ be a leaf in $\mathcal{A}_{g,n}$. For each integer $N \geq 1$, denote by $A[p^N] \to \mathcal{C}$ the p^N-torsion subgroup scheme of the restriction to $\mathcal{C}$ of the universal abelian scheme. Then there exists a finite surjective morphism $f : S \to \mathcal{C}$ such that $(A[p^N], \lambda[p^N]) \times_{\mathcal{C}} S$ is a constant principally polarized truncated Barsotti–Tate group over S.*

Proof. See [26, 1.3]. □

Remark 3.5. Using Proposition 3.4, one can show that there exist finite surjective isogeny correspondences between any two leaves lying in the same Newton polygon stratum; see [26, Lemma 3.14]. In particular, any two leaves in the same Newton polygon stratum have the same dimension.

Remark 3.6. In this article we have focused our attention on leaves in $\mathcal{A}_{g,n}$ over k. The notion of leaves can be extended to other modular varieties of PEL-type in a similar way, and the basic properties of leaves, including Theorem 3.1 and Propositions 3.4, 3.7, can all be generalized; some of the generalized statements become a little stronger. It is expected that the notion of leaves can be defined on reduction over k of a Shimura variety X, with nice properties.

[2] Some author use the notation $W^0_\xi(\mathcal{A}_{g,n})$ instead of $W^0_\xi(\mathcal{A}_{g,n})$, and call it an "open Newton polygon stratum"; then they denote by $W_\xi(\mathcal{A}_{g,n})$ the closure of $W^0_\xi(\mathcal{A}_{g,n})$ in $\mathcal{A}_{g,n}$ and call it a Newton polygon stratum.

Proposition 3.7. *Let $\mathcal{C}$ be a leaf in $\mathcal{A}_{g,n}$. Denote by $A[p^\infty] \to \mathcal{C}$ the Barsotti–Tate group attached to the restriction to $\mathcal{C}$ of the universal abelian scheme. Then there exists a slope filtration on $A[p^\infty] \to \mathcal{C}$. More precisely, there exist Barsotti–Tate subgroups*

$$0 = G_0 \subset G_1 \subset G_2 \subset \cdots \subset G_m = A[p^\infty]$$

of $A[p^\infty] \to \mathcal{C}$ over the leaf $\mathcal{C}$ such that G_i/G_{i-1} is a Barsotti–Tate group over $\mathcal{C}$ with a single Frobenius slope μ_i, $i = 1, \ldots, m$, and $\mu_1 > \mu_2 > \cdots > \mu_m$. Moreover each Barsotti–Tate group $G_i/G_{i-1} \to \mathcal{C}$ is geometrically fiberwise constant, for $i = 1, \ldots, m$. In other words, any two geometric fibers of $G_i/G_{i-1} \to \mathcal{C}$ are isomorphic after base extension to a common algebraically closed overfield.

Remark 3.8.

(i) The statement that $H_i := G_i/G_{i-1}$ has Frobenius slope μ_i means that there exist constants $c, d > 0$ such that

$$\mathrm{Ker}([p^{\lfloor N\mu_i - c \rfloor}]_{H_i}) \subseteq \mathrm{Ker}(\mathrm{Fr}_{H_i}^{(p^N)}) \subseteq \mathrm{Ker}([p^{\lfloor N\mu_i + d \rfloor}]_{H_i})$$

for all $N \gg 0$. Here $\mathrm{Fr}_{H_i}^{(p^N)} : H_i \to H_i^{(p^N)}$ denotes the relative p^N-Frobenius for $H_i \to \mathcal{C}$, also called the N-th iterate of the relative Frobenius by some authors, while $\mathrm{Ker}([p^{\lfloor N\mu_i - c \rfloor}]_{H_i})$ (resp. $\mathrm{Ker}([p^{\lfloor N\mu_i + d \rfloor}]_{H_i})$) is the kernel of multiplication by $p^{\lfloor N\mu_i - c \rfloor}$ (resp. by $p^{\lfloor N\mu_i + d \rfloor}$) on H_i.

(ii) The Frobenius slopes of a Barsotti–Tate group X measures divisibility property of iterates of the Frobenius map on X. A Barsotti–Tate group X is isoclinic with Frobenius slope μ if $(\mathrm{Fr}_X)^N/p^{\mu N}$ and $p^{\mu N}/(\mathrm{Fr}_X)^N$ are both bounded as $N \to \infty$. In the literature the terminology "slope" is sometimes also used to measure the divisibility of the Verschiebung, hence we use "Frobenius slope" to avoid possible confusion.

(iii) When all fibers of $\mathcal{A}[p^\infty]$ at points of $\mathcal{C}$ are *completely slope divisible*, the existence of the slope filtration was proved by in [32, Proposition 14]; see also [27, Proposition 2.3]. The statement of Proposition 3.7 has not appeared in the literature, but the following stronger statement can be deduced from [32, Theorem 7] and [27, Theorem 2.1]: If $S \to \mathrm{Spec}(\mathbb{F}_p)$ is an integral Noetherian normal scheme of characteristic p, and G is a Barsotti–Tate group over S which is geometrically fiber-wise constant, then $G \to S$ admits a slope filtration.

(iv) The slope filtration on a leaf holds the key to the theory of canonical coordinates on a leaf; see §7.

(v) It is clear that on a Barsotti–Tate group over a reduced base scheme S over k, there exists at most one slope filtration.

(vi) One can construct a Barsotti–Tate group G over a smooth base scheme S over k, for instance $\mathbb{P}^1$, such that G does not have a slope filtration.

Denote by $\Pi_0(\mathcal{C}(x))$ the scheme of geometrically *irreducible* components of $\mathcal{C}(x)$, or equivalently, the set of geometrically *connected* components of $\mathcal{C}(x)$, since $\mathcal{C}(x)$ is smooth over k. The scheme $\Pi_0(\mathcal{C}(x))$ is finite and étale over k; this assertion holds even if the base field k is not assumed to be algebraically closed.

Let $E = F_1 \times \cdots \times F_r$ be the product of totally real number fields $F_1, \ldots, F_r$, and let $n \geq 3$ be an integer with $(n, p) = 1$. The notion of *leaves* can be extended to the Hilbert modular variety $\mathcal{M}_{E,n}$ over k, as follows. Let $x \in \mathcal{M}_{E,n}(k)$ be a geometric point of the Hilbert modular variety $\mathcal{M}_{E,n}(k)$. The *leaf* in $\mathcal{M}_{E,n}$ passing through x is the smooth locally closed subscheme $\mathcal{C}_E(x)$, characterized by the property that $\mathcal{C}_E(x)(K)$ consists of all geometric points $y \in \mathcal{M}_{E,n}(K)$ such that there exists an $\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p$-linear isomorphism from $A_y[p^\infty]$ to $A_x[p^\infty]$ compatible with the $\mathcal{O}_E$-polarizations, for every algebraically closed field $K \supset k$.

Just as in the case of Siegel modular varieties, each leaf in $\mathcal{M}_{E,n}$ is stable under all prime-to-p Hecke correspondences on $\mathcal{M}_{E,n}$.

The slope filtration on the Barsotti–Tate group over a leaf in $\mathcal{M}_{E,n}$ takes the following form. Let $\mathcal{C}_E$ be a leaf in $\mathcal{M}_{E,n}$, and denote by G the Barsotti–Tate group attached to the restriction to $\mathcal{C}_E$ of the universal abelian scheme over $\mathcal{C}_E$. Write $\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p = \prod_{j=1}^{s} \mathcal{O}_{E_{\wp_j}}$, where each $\mathcal{O}_{E_{\wp_j}}$ is a complete discrete valuation ring. The natural action of $\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p$ on G gives a decomposition

$$G = G_1 \times \cdots \times G_s \,,$$

where each G_j is a Barsotti–Tate group over $\mathcal{C}_E$, with action by $\mathcal{O}_{E_{\wp_j}}$, and the height of G_j is equal to $2\,[\mathcal{O}_{E_{\wp_j}} : \mathbb{Z}_p]$. Moreover, for $j \in \{1, \ldots, s\}$ and G_j *not* isoclinic of slope $\frac{1}{2}$, there exists a Barsotti–Tate subgroup $H_j \subset G_j$ over $\mathcal{C}_E$, stable under the action of $\mathcal{O}_{E_{\wp_j}}$, such that

- the height of H_j is equal to $[\mathcal{O}_{E_{\wp_j}} : \mathbb{Z}_p]$,
- both H_j and G_j/H_j are isoclinic, of Frobenius slopes μ_j, μ'_j respectively, and
- $\mu_j > \mu'_j$ and $\mu_j + \mu'_j = 1$.

4 The Hecke orbit conjecture

Let k be an algebraically closed field of characteristic p, and let $n \geq 3$ be an integer, $(n, p) = 1$.

Conjecture 4.1. *Denote by $\mathcal{A}_{g,n}$ the moduli space of g-dimensional principally polarized abelian varieties over k with symplectic level-n structures as before.*

(**HO**) *For any geometric point x of $\mathcal{A}_{g,n}$, the Hecke orbit $\mathcal{H}^{(p)}(x)$ is dense in $\mathcal{C}(x)$.*

$(\mathbf{HO})_{\mathrm{ct}}$ *For any geometric point x of $\mathcal{A}_{g,n}$, we have $\dim(\overline{\mathcal{H}^{(p)}(x)}) = \dim(\mathcal{C}(x))$, where $\overline{\mathcal{H}^{(p)}(x)}$ denotes the Zariski closure of the countable subset $\mathcal{H}^{(p)}(x)$ in $\mathcal{A}_{g,n}$. Equivalently, $\overline{\mathcal{H}^{(p)}(x)}$ contains the irreducible component of $\mathcal{C}(x)$ passing through x.*

$(\mathbf{HO})_{\mathrm{dc}}$ *For any geometric point x of $\mathcal{A}_{g,n}$, the canonical map*

$$\Pi_0(\overline{\mathcal{H}^{(p)}(x)}^{\circ}) \to \Pi_0(\mathcal{C}(x))$$

is surjective, where $\overline{\mathcal{H}^{(p)}(x)}^{\circ} := \overline{\mathcal{H}^{(p)}(x)} \cap \mathcal{C}(x)$ denotes the Zariski closure of the Hecke orbit $\mathcal{H}^{(p)}(x)$ in the leaf $\mathcal{C}(x)$. In other words, the prime-to-p Hecke correspondences operate transitively on the set $\Pi_0(\mathcal{C}(x))$ of geometrically irreducible components of $\mathcal{C}(x)$.

Remark 4.2.

(i) Conjecture (HO) is due to Oort, see [26, 6.2]. It implies Conjecture 15.A in [22], which asserts that the orbit of a point x of $\mathcal{A}_{g,n}(k)$ under all Hecke correspondences, including all purely inseparable ones, is Zariski dense in the Newton polygon stratum containing x.

(ii) It is clear that conjecture (HO) is equivalent to the conjunction of $(\mathrm{HO})_{\mathrm{ct}}$ and $(\mathrm{HO})_{\mathrm{dc}}$. We call $(\mathrm{HO})_{\mathrm{ct}}$ (resp. $(\mathrm{HO})_{\mathrm{dc}}$) the continuous (resp. discrete) part of the Hecke orbit conjecture (HO).

(iii) Conjecture $(\mathrm{HO})_{\mathrm{dc}}$ is essentially an irreducibility statement; see Theorem 5.1.

(iv) We can also formulate an ℓ-adic version of the Hecke orbit conjecture, $(\mathrm{HO})_{\ell}$, for any prime number $\ell \neq p$. It asserts that $\mathcal{H}_{\ell}(x)$ is dense in $\mathcal{C}(x)$. One can define the continuous part $(\mathrm{HO})_{\ell,\mathrm{ct}}$, and the discrete part $(\mathrm{HO})_{\ell,\mathrm{dc}}$ of $(\mathrm{HO})_{\ell}$ as in 4.1. Clearly, $(\mathrm{HO})_{\ell} \iff (\mathrm{HO})_{\ell,\mathrm{ct}} + (\mathrm{HO})_{\ell,\mathrm{dc}}$.

(v) Theorem 5.1 tells us that $(\mathrm{HO})_{\ell,\mathrm{dc}} \iff (\mathrm{HO})_{\mathrm{dc}}$, and $(\mathrm{HO})_{\ell} \iff (\mathrm{HO})$. So, although $(\mathrm{HO})_{\ell}$ appears to be a stronger statement than (HO), it is essentially equivalent to it. Strictly speaking, Theorem 5.1 gives the implications when the Hecke orbit in question is not supersingular, however the supersingular case can be dealt with directly, using the weak approximation theorem.

Let E be a finite product of totally real number fields, and let $\mathcal{M}_E$ be the Hilbert modular variety over k attached to E. Then we can formulate the Hecke orbit conjectures for $\mathcal{M}_n$ as in Conjecture 4.1, and will use $(\mathrm{HO})_E$,

$(\mathrm{HO})_{E,\mathrm{ct}}$, and $(\mathrm{HO})_{E,\mathrm{dc}}$ to denote the Hecke orbit conjecture for $\mathcal{M}_n$ and its two parts. Remark 4.2 (ii), (iii), (iv) hold in the present context.

Remark 4.3. The Hecke orbit conjecture(s) can be formulated for other modular varieties of PEL-type, and the reduction over k of any Shimura variety X if one is optimistic. It should be noted, however, that the statement in Remark 4.2 (iii) needs to be modified, because the last sentence of Theorem 5.1 depends on the fact that Sp_{2g} is simply connected. The remedy is to use the $G^{\mathrm{der}}(\mathbb{A}_f^{(p)})$-orbit instead of the $G(\mathbb{A}_f^{(p)})$-orbit, where G is the connected reductive group over $\mathbb{Q}$ in the input data of the Shimura variety X.

Theorem 4.4. *The Hecke orbit conjectures* (HO),$(\mathrm{HO})_\ell$ *hold for the Siegel modular varieties. In other words, every prime-to-p Hecke orbit is Zariski dense in the leaf containing it; the same is true for every ℓ-adic Hecke orbit, for every prime number ℓ with $(\ell, p) = 1$.*

In the rest of this article we present an outline of the proof of Theorem 4.4. We have already seen that Theorem 5.1 on ℓ-adic monodromy groups is helpful in clarifying the discrete Hecke orbit conjecture, and for the equivalence between $(\mathrm{HO})_\ell$ and (HO). The foundation underlying our approach is the *local stabilizer principle*, to be explained in §6; this principle is quite general and can be applied to all PEL-type modular varieties. We will also use a special property of the Siegel modular varieties, called the *Hilbert trick*, to be explained in §9. That property holds for modular varieties of PEL-type C, but not for PEL-type A or D. Both the local stabilizer principle and the Hilbert trick were used in [6]; the former was used not only for points of the ordinary locus, but also the zero-dimensional cusps and supersingular points.

There are several techniques, listed as items (C), (R), (S), (H) in the fourth paragraph of §1, which make the local stabilizer principle more potent. Among them, the methods (C), (R), (H) can be generalized to all modular varieties of PEL-type, while (S) depends on the Hilbert trick, therefore applies only to modular varieties of PEL-type C.

The Hecke orbit conjecture for the Hilbert modular varieties enters the proof of $(\mathrm{HO})_{\mathrm{ct}}$ for $\mathcal{A}_{g,n}$ at a critical point, through the Hilbert trick.

Theorem 4.5. *The Hecke orbit conjecture holds for Hilbert modular varieties. In other words, every prime-to-p Hecke orbit in a Hilbert modular variety is Zariski dense in the leaf containing it.*

5 ℓ-adic monodromy of leaves

Theorem 5.1 below explores the relation between the Hecke symmetries and the ℓ-adic monodromy. It asserts that the ℓ-adic monodromy of any non-supersingular leaf on $\mathcal{A}_g$ is maximal. A byproduct of Theorem 5.1, from a

group theoretic consideration, is an irreducibility statement. The irreducibility statement implies that for a non-supersingular leaf $\mathcal{C}$ in $\mathcal{A}_g$, the discrete part $(\mathrm{HO})_{\mathrm{dc}}$ of the Hecke orbit conjecture holds for $\mathcal{C}$ if and only if $\mathcal{C}$ is irreducible.

Theorem 5.1. *Let k be an algebraically closed field of characteristic p. Let $n \geq 3$ be a natural number which is prime to p. Let ℓ be a prime number $\ell \nmid pn$. Let Z be a smooth locally closed subvariety of $\mathcal{A}_{g,n}$ over k. Assume that Z is stable under all ℓ-adic Hecke correspondences coming from $\mathrm{Sp}_{2g}(\mathbb{Q}_\ell)$, and that the ℓ-adic Hecke correspondences operate transitively on the set of irreducible components of Z. Let $A \to Z$ be the restriction to Z of the universal abelian scheme. Let Z_0 be an irreducible component of Z, and let $\bar{\eta}$ be a geometric generic point of Z_0. Assume that $A_{\bar{\eta}}$ is not supersingular. Then the image $\rho_{A,\ell}(\pi_1(Z_0, \overline{\eta}))$ of the ℓ-adic monodromy representation of $A \to Z_0$ is equal to $\mathrm{Sp}(\mathrm{T}_\ell, \langle\ ,\ \rangle_\ell) \cong \mathrm{Sp}_{2g}(\mathbb{Z}_\ell)$, where $\mathrm{T}_\ell = \mathrm{T}_\ell(A_{\bar{\eta}}) = \varprojlim_n A[\ell^n](\bar{\eta})$ denotes the ℓ-adic Tate module of $A_{\bar{\eta}}$. Moreover $Z = Z_0$, i.e., Z is irreducible, and Z is stable under all prime-to-p Hecke correspondences on $\mathcal{A}_{g,n}$.*

Remark 5.2.

(i) Theorem 5.1 is handy when one tries to prove the irreducibility of certain subvarieties of $\mathcal{A}_g$. For instance, if one wants to show that a leaf or a Newton polygon stratum in $\mathcal{A}_g$ is irreducible, Theorem 5.1 tells us that it suffices to show that the the prime-to-p Hecke correspondences operate transitively on the set of irreducible components of the given leaf or Newton polygon stratum. The latter statement be approached by the standard degeneration argument in algebraic geometry.

(ii) Theorem 5.1 is the main result of [4]. The proof of Theorem 5.1 can be generalized to other modular varieties of PEL-type, but one has to make suitable modification of the statement if the derived group of G is not simply connected.

(iii) The assumption that Z is stable under all ℓ-adic Hecke correspondences coming from $\mathrm{Sp}_{2g}(\mathbb{Q}_\ell)$ means that the closed points of Z is a union of ℓ-adic Hecke correspondences. See Section 2 for the action of $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ on the tower $\mathcal{A}_{g,(p)}$ of modular varieties. The action of the subgroup $\mathrm{Sp}_{2g}(\mathbb{Q}_\ell)$ of $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ induces the ℓ-adic Hecke correspondences on $\mathcal{A}_{g,n}$.

(iv) The proof of Theorem 5.1 is mostly group-theoretic; the algebro-geometric input is the semisimplicity of the ℓ-adic monodromy group.

6 The action of the local stabilizer subgroup

Let k be an algebraically closed field of characteristic p. Let $n \geq 3$ be an integer, $(n, p) = 1$. Let ℓ be a prime number, $\ell \neq p$. Let $Z \subset \mathcal{A}_{g,n}$ be a reduced closed subscheme stable under all ℓ-adic Hecke correspondences. In

other words, Z is a union of ℓ-adic Hecke orbits. Let $x = ([A_x, \lambda_x]) \in Z(k)$ be a closed point of Z. Let $E = \mathrm{End}_k(A_x) \otimes_{\mathbb{Z}} \mathbb{Q}_p$, and let $*$ be the Rosati involution of E induced by the principal polarization λ_x. Let

$$H = \{u \in E^\times \mid u \cdot u^* = u^* \cdot u = 1\}$$

be the unitary group attached to the pair $(E \otimes_{\mathbb{Q}} \mathbb{Q}_p, *)$. Define the *local stabilizer subgroup* U_x at $x \in \mathcal{A}_g(k)$ by $U_x := H \cap \mathrm{End}_k(A_x[p^\infty])^\times$.

Similarly, let $\tilde{E} := \mathrm{End}_k(A_x[p^\infty]) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$, and let $\tilde{*}$ be the involution on $\tilde{E}$ induced by λ_x. Denote by $\tilde{H}$ the unitary group attached to the pair $(\tilde{E}, \tilde{x})$, and let $\tilde{U}_x = \tilde{H} \cap \mathrm{End}_k(A_x[p^\infty])^\times$. The group $\tilde{U}_x$ operates naturally on $\mathcal{A}_{g,n}^{/x}$ by deformation theory. Since there is a natural inclusion $U_x \hookrightarrow \tilde{U}_x$, the subgroup U_x inherits an action on $\mathcal{A}_{g,n}^{/x}$.

Proposition 6.1 (local stabilizer principle). *Notation as above. Then the closed formal subscheme $Z^{/x}$ of $\mathcal{A}_{g,n}^{/x}$ is stable under the action of the local stabilizer subgroup U_x on $\mathcal{A}_{g,n}^{/x}$*

Proof (Sketch). Let $\underline{U}$ be the unitary group attached to the pair $(E, *)$; it is a reductive linear algebraic group over $\mathbb{Q}$. In particular the weak approximation theorem holds for $\underline{U}$. Choose and fix a "standard embedding"

$$\underline{U}(\mathbb{A}_f^{(p)}) \hookrightarrow \mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$$

coming from a choice of a symplectic level-$\mathbb{Z}_f^{(p)}$ structure of A_x. Then every element of the subgroup $\underline{U}(\mathbb{A}_f^{(p)})$ of $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ gives rise to a prime-to-p Hecke correspondence having x as a fixed point. For any given element $\gamma_p \in U_x$, choose an element $\gamma \in \underline{U}(\mathbb{Q})$ close to γ_p in $\underline{U}(\mathbb{Q}_p)$. Note that the image of γ in $\underline{U}(\mathbb{A}_f^{(p)})$ gives rise to a prime-to-p Hecke correspondence, which has x as a fixed point and sends the formal subscheme $Z^{/x}$ of $\mathcal{A}_{g,n}^{/x}$ into $Z^{/x}$ itself. Interpreted in terms of deformation theory, the last assertion implies that a formal neighborhood $\mathrm{Spec}\left(\mathcal{O}_{Z^{/x}}/\mathfrak{m}_x^N\right)$ of x in $Z^{/x}$, as a formal subscheme of $\mathcal{A}_{g,n}^{/x}$, is stable under the natural action of γ_p, where $\mathfrak{m}_x$ is the maximal ideal of $\mathcal{O}_{Z^{/x}}$, and $N = N(\gamma_p, \gamma)$ depends on how close γ is to γ_p, $N(\gamma_p, \gamma) \to \infty$ as $\gamma \to \gamma_p$. Taking the limit as γ goes to γ_p, we see that $Z^{/x}$ is stable under the action of γ_p. □

Remark 6.2.

(i) The action of the local stabilizer subgroup on the deformation space goes back to Lubin and Tate in [15].

(ii) In [6], the local stabilizer principle was applied to the zero-dimensional cusps of $\mathcal{A}_{g,n}$, and also to points of $\mathcal{A}_{g,n}$ defined over finite fields. The calculation of [6, Proposition 2, p. 454] at the zero-dimensional cusps is a bit complicated, and can be avoided, using "Larsen's example" on page 443 of [6] instead.

(iii) The bigger the local stabilizer subgroup U_x, the more information the action of U_x on $\mathcal{A}_{g,n}^{/x}$ contains. The size of $\underline{U}$, the linear algebraic group over $\mathbb{Q}_p$ such that U_x is open in $\underline{U}(\mathbb{Q}_p)$, is maximal when the abelian variety A_x is supersingular. If x is a supersingular point, then $\underline{U}$ is an inner twist of Sp_{2g}, so in some sense almost all information about the prime-to-p Hecke correspondences on $\mathcal{A}_{g,n}$ are encoded in the action of U_x on $\mathcal{A}_{g,n}^{/x}$. The challenge, however, is to dig the buried information out of this action; the success stories include Theorem 11.3, and [6, §5, Proposition 7].

7 Canonical coordinates for leaves

Let k be an algebraically closed field of characteristic p. Let $\mathcal{C}$ be a leaf on $\mathcal{A}_{g,n}$, where $n \geq 3$ is a natural number relatively prime to p. Let $x \in \mathcal{C}(k)$ be a closed point of $\mathcal{C}$. Recall that the leaf $\mathcal{C}$ is defined by a point-wise property, namely, a point $y \in \mathcal{C}(k)$ is in $\mathcal{C} = \mathcal{C}(x)$ if and only if the principally quasi-polarized Barsotti–Tate groups $(A_y[p^\infty], \lambda_y[p^\infty])$ and $(A_x[p^\infty], \lambda_x[p^\infty])$ are isomorphic. One can also use the same point-wise property to define leaves (on the base scheme) for a (principally quasi-polarized) Barsotti–Tate group over a Noetherian integral base scheme over k; see [26].

From the definition it is not immediately clear how to "compute" the formal completion $\mathcal{C}^{/x}$ of the leaf $\mathcal{C}$ at x. However this turns out to be possible, and the resulting theory is a generalization of the classical Serre–Tate theory for the local moduli of ordinary abelian varieties. Some highlights of the description of $\mathcal{C}^{/x}$ will be explained in this section. More details can be found in [7], [2].

Recall that the deformation theory of (A_x, λ_x) is the same as that of the associated principally quasi-polarized Barsotti–Tate group $(A_x[p^\infty], \lambda_x[p^\infty])$. Let

$$0 = G_0 \subset G_1 \subset G_2 \subset \cdots \subset G_m = A_{\mathcal{C}}[p^\infty]$$

be the slope filtration of the restriction to $\mathcal{C}$ of the Barsotti–Tate group attached to the universal abelian scheme, so that each G_i/G_{i-1} is a Barsotti–Tate group over $\mathcal{C}$ with slope μ_i, $i = 1, 2, \ldots, m$, and $\mu_1 > \mu_2 > \cdots > \mu_m$. Moreover, each subquotient G_i/G_{i-1} is constant over the formal completion $\mathcal{C}^{/x}$ of $\mathcal{C}$ at x, because it is geometrically fiberwise constant over the complete strictly henselian base formal scheme $\mathcal{C}^{/x}$.

Let $\mathfrak{Def}(A_x) = \mathfrak{Def}(A_x[p^\infty])$ be the local deformation space of A_x over k, or equivalently the local deformation space of $A_x[p^\infty]$ over k; it is a g^2-dimensional smooth formal scheme over k. A basic phenomenon here is that $\mathcal{C}^{/x}$ is determined by the slope filtration on $A[p^\infty] \to \mathcal{C}^{/x}$. More precisely,

the formal subscheme $\mathcal{C}^{/x} \subset \mathcal{A}_{g,n}^{/x} \subset \mathfrak{Def}(A_x)$ is contained in the "extension part" $\mathfrak{MDE}(A_x[p^\infty])$ of $\mathfrak{Def}(A_x)$, where $\mathfrak{MDE}(A_x[p^\infty])$ is the maximal closed formal subscheme of the local deformation space $\mathfrak{Def}(A_x) = \mathfrak{Def}(A_x[p^\infty])$ such that the restriction to $\mathfrak{MDE}(A_x[p^\infty])$ of the universal Barsotti–Tate group is a successive extension of constant Barsotti–Tate groups

$$(G_i/G_{i-1})_x \times_{\mathrm{Spec}(k)} \mathfrak{MDE}(A_x[p^\infty])\,,$$

extending the slope filtration of $A_x[p^\infty]$. For each Artinian local k-algebra R, $\mathfrak{MDE}(R)$ is the set of isomorphism classes of tuples

$$\Big(0 = \tilde{G}_0 \subset \tilde{G}_1 \subset \cdots \subset \tilde{G}_m; \alpha_1, \ldots, \alpha_m; \beta_1, \ldots, \beta_m\Big)\,,$$

such that

- $\tilde{G}_i$ is a Barsotti–Tate group over R for each i,
- each quotient $\tilde{G}_i/\tilde{G}_{i-1}$ is a Barsotti–Tate group over R, $i = 1, \ldots, m$,
- α_i is an isomorphism from $\tilde{G}_i \times_{\mathrm{Spec}(R)} \mathrm{Spec}(k)$ to $(G_i)_x$, for $i = 1, \ldots, m$,
- β_i is an isomorphism from $\tilde{G}_i/\tilde{G}_{i-1}$ to $(G_i/G_{i-1})_x \times_{\mathrm{Spec}(k)} \mathrm{Spec}(R)$, for $i = 1, \ldots, m$,
- the inclusion maps $G_i \hookrightarrow G_{i+1}$ and $\tilde{G}_i \hookrightarrow \tilde{G}_{i+1}$, for $i = 1, \ldots, m-1$, are compatible with the isomorphisms $\alpha_1, \ldots, \alpha_m$
- the isomorphisms $\beta_1, \ldots, \beta_m$ are compatible with $\alpha_1, \ldots, \alpha_m$.

Our theory of canonical coordinates provides a description of the closed formal subscheme $\mathcal{C}^{/x}$ of $\mathfrak{MDE}(A_x[p^\infty])$ in terms of the structure of $\mathfrak{MDE}(A_x[p^\infty])$, independent of the notion of leaves. If the abelian variety A_x is ordinary, then $m = 2$, G_1 is toric, G_2/G_1 is étale, and the theory reduces to the classical Serre–Tate coordinates.

The computation of $\mathcal{C}^{/x}$ can be reduced to the following two "essential cases". In both cases we have two p-Barsotti–Tate groups X and Y over k; X has slope $\mu_{_X}$, while Y has slope $\mu_{_Y}$. We assume that $\mu_{_X} < \mu_{_Y}$. Let $\mathrm{Spf}(R)$ be the equi-characteristic deformation space of $X \times Y$. Let $G \to \mathrm{Spf}(R)$ be the universal deformation of $X \times Y$. For each $s \geq 1$, since $G[p^s]$ is a finite locally free group scheme over $\mathrm{Spf}(R)$, it is the formal completion of a unique finite locally free group scheme over $\mathrm{Spec}(R)$, denoted by $G[p^s]' \to \mathrm{Spec}(R)$. The inductive system of finite locally free group schemes $G[p^s]' \to \mathrm{Spec}(R)$ form a Barsotti–Tate group over $\mathrm{Spec}(R)$, denoted by $G \to \mathrm{Spec}(R)$, abusing the notation.

- (unpolarized case) In this case, our goal is to compute the leaf in $\mathrm{Spec}(R)$, passing through the closed point of $\mathrm{Spec}(R)$, for the Barsotti–Tate group $G \to \mathrm{Spec}(R)$. This leaf will be denoted by $\mathcal{C}^{\wedge}_{\mathrm{up}}$.

- (polarized case) Suppose that λ is a principal quasi-polarization on the product $X \times Y$. This assumption implies that $\mu_X + \mu_Y = 1$. The equi-characteristic deformation space of $(X \times Y, \lambda)$ is a closed formal subscheme $\mathrm{Spf}(R/I)$ of $\mathrm{Spf}(R)$. We would like to compute the leaf in $\mathrm{Spec}(R/I)$, passing through the closed point of $\mathrm{Spec}(R/I)$, for the principally polarized Barsotti–Tate group $G \to \mathrm{Spec}(R/I)$; denote this leaf by $\mathcal{C}^\wedge$.

Our starting point in the computation of $\mathcal{C}^{/x}_{\mathrm{up}}$ and $\mathcal{C}^{/x}$ is the following observation. There is a closed formal subscheme $\mathfrak{DE}(X,Y)$ of the deformation space $\mathrm{Spf}(R)$, maximal with respect to the property that the restriction to $\mathfrak{DE}(X,Y)$ of the universal deformation of $X \times Y$ is an extension of the constant group $X \times_{\mathrm{Spec}(k)} \mathfrak{DE}(X,Y)$ by the constant group $Y \times_{\mathrm{Spec}(k)} \mathfrak{DE}(X,Y)$. It is not difficult to see that $\mathfrak{DE}(X,Y)$ is formally smooth over k. The existence of the canonical filtration of the restriction of G to the leaves implies that both $\mathcal{C}^\wedge_{\mathrm{up}}$ and $\mathcal{C}^\wedge$ are closed formal subschemes of $\mathfrak{DE}(X,Y)$. On the other hand, the Baer sum for extensions produces a group law on $\mathfrak{DE}(X,Y)$, so that $\mathfrak{DE}(X,Y)$ has a natural structure as a smooth formal group over k.

Theorem 7.1.

(i) *In the unpolarized case, the leaf $\mathcal{C}^\wedge_{\mathrm{up}}$ is naturally isomorphic to the maximal p-divisible formal subgroup $\mathfrak{DE}(X,Y)_{p\text{-div}}$ of $\mathfrak{DE}(X,Y)$. The p-divisible group $\mathfrak{DE}(X,Y)_{p\text{-div}}$ has slope $\mu_Y - \mu_X$.*

(ii) *In the polarized case, the principal quasi-polarization λ on $X \times Y$ induces an involution on $\mathfrak{DE}(X,Y)_{p\text{-div}}$, and $\mathcal{C}^\wedge$ is equal to the maximal subgroup $\mathfrak{DE}(X,Y)^{\mathrm{sym}}_{p\text{-div}}$ of $\mathfrak{DE}(X,Y)_{p\text{-div}}$ which is fixed under the involution. Again, $\mathfrak{DE}(X,Y)^{\mathrm{sym}}_{p\text{-div}}$ is a p-divisible formal group with slope $\mu_Y - \mu_X$.*

Remark 7.2.

(i) Theorem 7.1 gives a structural characterization of the leaves $\mathcal{C}^\wedge_{\mathrm{up}}$ and $\mathcal{C}^\wedge$ in the formal subscheme $\mathfrak{DE}(X,Y)$ of the deformation space $\mathrm{Spf}(R)$ of $X \times Y$. In Theorem 7.7 and Proposition 7.8, we will see a structural characterization of a leaf $\mathcal{C}(\mathfrak{Def}(G))$ in the equi-characteristic deformation space $\mathfrak{Def}(G)$ of a general Barsotti–Tate group G over k, in a similar spirit. The above characterization deals with the differential property of leaves, and complements the global point-wise definition of leaves.

(ii) The statement in Theorem 7.1 (ii) follows quickly from 7.1 (i). The last sentence of 7.1 (i) can be proved by comparing the effect of iterates of the relative Frobenius on $\mathfrak{DE}(X,Y)_{p\text{-div}}$ with suitable powers p, assuming without loss of generality that X and Y are both minimal.

(iii) We have a natural inclusion $\mathcal{C}^\wedge_{\mathrm{up}} \subset \mathfrak{DE}(X,Y)$. To prove that $\mathcal{C}^\wedge_{\mathrm{up}}$ contains $\mathfrak{DE}(X,Y)_{p\text{-div}}$, one shows that the pull-back of the universal extension of X by Y over $\mathfrak{DE}(X,Y)$ to the perfection of $\mathfrak{DE}(X,Y)_{p\text{-div}}$ splits. To prove that $\mathcal{C}^\wedge_{\mathrm{up}} \subseteq \mathfrak{DE}(X,Y)_{p\text{-div}}$, one shows that for every complete Noetherian

local domain S over k and every S-valued point $f : S \to \mathfrak{DE}(X,Y)_{\rm unip}$ of the maximal unipotent part of $\mathfrak{DE}(X,Y)$, if the extension of $X \times_{{\rm Spec}(k)} S$ by $Y \times_{{\rm Spec}(k)} S$ attached to f becomes trivial over the perfection $S^{\rm perf}$ of S, then f corresponds to the trivial extension over S.

Theorem 7.3. *Let* $\mathrm{M}(X), \mathrm{M}(Y)$ *be the covariant Dieudonné module of* X, Y *respectively. Let* $B(k)$ *be the fraction field of* $W(k)$*. The* $B(k)$*-vector space*

$$\mathrm{Hom}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y)) \otimes_{W(k)} B(k)$$

has a natural structure as a V*-isocrystal.*

(i) *Let* $\mathrm{M}(\mathfrak{DE}(X,Y)_{p\text{-div}})$ *be the covariant Diedonné module of*

$$\mathcal{C}^{\wedge}_{\rm up} = \mathfrak{DE}(X,Y)_{p\text{-div}}.$$

Then there exists a natural isomorphism of V*-isocrystals*

$$\mathrm{M}(\mathfrak{DE}(X,Y)_{p\text{-div}}) \otimes_{W(k)} B(k) \xrightarrow{\sim} \mathrm{Hom}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y)) \otimes_{W(k)} B(k)\,.$$

(ii) *Suppose that* λ *is a principal quasi-polarization* λ *on* $X \times Y$*. Let* ι *be the involution on* $\mathrm{Hom}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y)) \otimes_{W(k)} B(k)$ *induced by* λ *and* $\mathrm{M}(\mathfrak{DE}(X,Y)^{\rm sym}_{p\text{-div}})$ *the covariant Diedonné module of* $\mathcal{C}^{\wedge} = \mathfrak{DE}(X,Y)^{\rm sym}_{p\text{-div}}$*. Then there exists a natural isomorphism of* V*-isocrystals*

$$\mathrm{M}(\mathfrak{DE}(X,Y)^{\rm sym}_{p\text{-div}}) \otimes_{W(k)} B(k) \xrightarrow{\sim} \mathrm{Hom}^{\rm sym}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y)) \otimes_{W(k)} B(k)\,,$$

where the right-hand side is the subspace of $\mathrm{Hom}_{W(k)}(\mathrm{M}(X), \mathrm{M}(Y)) \otimes_{W(k)} B(k)$ *fixed under the involution* ι*.*

Remark 7.4.

(i) See [2] for a proof of Theorem 7.3. The set $\mathrm{Cart}_p(k[[t]])$ of all formal curves in the functor of reduced Cartier ring for algebras over $\mathbb{Z}_{(p)}$ plays a crucial role in the proof of Theorem 7.3; it is denoted by $\mathrm{BC}_p(k)$ in [2]. The set $\mathrm{BC}_p(k)$ has a natural $(\mathrm{Cart}_p(k), \mathrm{Cart}_p(k))$-bimodule structure, because $\mathrm{Cart}_p(k)$ is a subring of $\mathrm{Cart}_p(k[[t]])$. Moreover $\mathrm{Cart}_p(k[[t]])$ has an "extra" $\mathrm{Cart}_p(k)$-module structure, compatible with the above bimodule structure; it comes from the Cartier theory, because the functor Cart_p is a commutative smooth formal group. The Cartier module of $\mathfrak{MDE}(X,Y)$ is canonically isomorphic to

$$\mathrm{Ext}^1_{\mathrm{Cart}_p(k)}\left(\mathrm{M}(X), \mathrm{BC}_p(k) \otimes_{\mathrm{Cart}_p(k)} \mathrm{M}(Y)\right)$$

where the extension functor is computed using the left $\mathrm{Cart}_p(k)$-module structure in the bimodule structure, and the action of $\mathrm{Cart}_p(k)$ on $\mathfrak{MDE}(X,Y)$ comes from the "extra" $\mathrm{Cart}_p(k)$-module structure of $\mathrm{BC}_p(k)$ mentioned above. It follows that the covariant V-isocrystal attached to $\mathfrak{MDE}(X,Y)_{p\text{-div}}$ is canonically isomorphic to

$$\mathrm{Ext}^1_{\mathrm{Cart}_p(k)}\left(\mathrm{M}(X), \mathrm{BC}_p(k) \otimes_{\mathrm{Cart}_p(k)} \mathrm{M}(Y)\right) \otimes_{W(k)} B(k)\,.$$

(ii) Theorem 7.3 is a generalization of the appendix of [18]. In [18] the authors dealt with the case when Y is the formal completion of $\mathbb{G}_m$. In that case $\mathfrak{MDE}(X,Y)$ is already a p-divisible formal group, and the natural map in the last displayed formula in Theorem 7.3 (i) preserves the natural integral structures, giving a formula for the Cartier module of $\mathfrak{MDE}(X,Y)$. The proof of Theorem 7.3 (i) begins by choosing a finite free resolution of $\mathrm{M}(X)$ of length 1, then using the resolution to write down the canonical map in Theorem 7.3 (i). The main technical ingredient is an approximation of $\mathrm{BC}_p(k)\otimes_{\mathbb{Z}}\mathbb{Q}$ by $\mathrm{Cart}_p(k)\hat{\otimes}_{W(k)}\mathrm{Cart}_p(k)\otimes_{\mathbb{Z}}\mathbb{Q}$, where $\mathrm{Cart}_p(k)\hat{\otimes}_{W(k)}\mathrm{Cart}_p(k)$ denotes a completed tensor product. The statement 7.3 (ii) follows easily from the proof of 7.3 (i).

(iii) The method of the proof of Theorem 7.3 can be regarded as a generalization of §4 and §5 of Mumford's seminal paper [17]. It may be interesting to note that the set denoted by $\tilde{A}_R$ on pages 316–317 of [17], together with its (A_R, A_R)-bimodule structure is essentially the same as the set $\mathrm{BC}_p(k)\otimes_{\mathrm{Cart}_p(k)} M(\widehat{\mathbb{G}_m})$ in our notation, with two structures of left $(\mathrm{Cart}_p(k)$-modules that commute with each other. The first action of $\mathrm{Cart}_p(k))$ comes from the left action of $\mathrm{Cart}_p(k))$ on $\mathrm{BC}_p(k)$, while the second left action of $\mathrm{Cart}_p(k))$ comes from the "extra" $\mathrm{Cart}_p(k)$-module structure of $\mathrm{BC}_p(k)$.

(iv) We do not know a convenient characterization of the the p-divisible formal group $\mathfrak{DE}(X,Y)_{p\text{-div}}$ inside its isogeny class, in terms of the Dieudonné modules $\mathrm{M}(X), \mathrm{M}(Y)$. When both X and Y are *minimal* in the sense of [20], i.e., the endomorphism $\mathbb{Z}_p$-algebra of X, Y are maximal orders, we expect that $\mathfrak{DE}(X,Y)_{p\text{-div}}$ is also minimal. It is easy to check that this conjectural statement holds when the denominators of the Brauer invariant of X and Y are relatively prime.

Corollary 7.5. *Let $h(X), h(Y)$ be the height of X, Y respectively.*

(i) *In the unpolarized case, the height of $\mathfrak{DE}(X,Y)_{p\text{-div}}$ is equal to $h(X)\cdot h(Y)$, and*

$$\dim(\mathfrak{DE}(X,Y)_{p\text{-div}}) = (\mu_{_Y} - \mu_{_X})\cdot h(X)\cdot h(Y).$$

(ii) *In the polarized case, we have $h(X) = h(Y)$, the height of $\mathfrak{DE}(X,Y)^{\mathrm{sym}}_{p\text{-div}}$ is equal to $\frac{h(X)\cdot(h(X)+1)}{2}$, and*

$$\dim(\mathfrak{DE}(X,Y)^{\mathrm{sym}}_{p\text{-div}}) = \frac{1}{2}(\mu_{_Y} - \mu_{_X})\cdot h(X)\cdot(h(X)+1).$$

Remark 7.6. The formulae (i), (ii) in Corollary 7.5 are quite similar to the formulae for the dimension of the deformation space of an h-dimensional abelian variety and the dimension of $\mathcal{A}_h$ respectively, except that there is an "extra factor" $\mu_{_Y} - \mu_{_X}$.

We go back to the general case. Just as in Theorem 7.1, it is convenient to consider the leaves in the local deformation space for the (unpolarized) Barsotti–Tate group $A_x[p^\infty]$. Denote by $\mathcal{C}(\mathfrak{Def}(A_x[p^\infty]))$ the leaf in the deformation space $\mathfrak{Def}(A_x[p^\infty])$ of the Barsotti–Tate group $A_x[p^\infty]$. Just as in Proposition 3.7, there exists a slope filtration

$$0 = G_0 \subset G_1 \subset \cdots \subset G_m = A_{\mathcal{C}(\mathfrak{Def}(A_x[p^\infty]))}[p^\infty]$$

on the universal Barsotti–Tate group over $\mathcal{C}(\mathfrak{Def}(A_x[p^\infty]))$, where each graded piece G_i/G_{i-1} is an isoclinic Barsotti–Tate group over $\mathcal{C}(\mathfrak{Def}(A_x[p^\infty]))$ with slope μ_i, $\mu_1 > \cdots > \mu_m$. Therefore the leaf $\mathcal{C}(\mathfrak{Def}(A_x[p^\infty]))$ is contained in $\mathfrak{MDE}(A_x[p^\infty])$, the maximal closed formal subscheme of $\mathfrak{Def}(A_x[p^\infty])$ such that the restriction to $\mathfrak{MDE}(A_x[p^\infty])$ of the universal Barsotti–Tate group has a slope filtration extending the slope filtration of $A_x[p^\infty]$. We would like to have a structural description of the leaf $\mathcal{C}(\mathfrak{Def}(A_x[p^\infty]))$ as a closed formal subscheme of $\mathfrak{MDE}(A_x[p^\infty])$, independent of the "point-wise" definition of the leaf. This will be achieved inductively, allowing us to understand how $\mathcal{C}(\mathfrak{Def}(A_x[p^\infty]))$ is "built up" from the p-divisible formal groups $\mathfrak{DE}(G_i/G_{i-1}, G_j/G_{j-1})_{p\text{-div}}$, for $1 \le j < i \le m$.

For each Barsotti–Tate group G over k, we can consider the leaf $\mathcal{C}(\mathfrak{Def}(G))$ in the deformation space $\mathfrak{Def}(G)$ over k, and we know that $\mathcal{C}(\mathfrak{Def}(G))$ is contained in $\mathfrak{MDE}(G)$, the maximal closed formal subscheme of $\mathfrak{Def}(G)$ such that the restriction to $\mathfrak{MDE}(G)$ of the universal Barsotti–Tate group has a slope filtration extending the slope filtration of G.

Let $0 = G_0 \subset G_1 \subset \cdots \subset G_m$ be the slope filtration of a Barsotti–Tate group G over k. Suppose that $0 \le j_1 \le j_2 < i_2 \le i_1 \le m$. Then there exists a natural formally smooth morphism

$$\pi_{[j_2,i_2],[j_1,i_1]} : \mathfrak{MDE}(G_{i_1}/G_{j_1}) \to \mathfrak{MDE}(G_{i_2}/G_{j_2})\,.$$

These morphisms form a finite projective system, that is

$$\pi_{[j_3,i_3],[j_2,i_2]} \circ \pi_{[j_2,i_2],[j_1,i_1]} = \pi_{[j_3,i_3],[j_1,i_1]}$$

if $0 \le j_1 \le j_2 \le j_3 < i_3 \le i_2 \le i_1 \le m$. Moreover, using the theory of biextensions of Mumford and Grothendieck in [17] and [11], one can show that the morphism

$$\mathfrak{MDE}(G_i/G_j) \longrightarrow \mathfrak{MDE}(G_{i-1}/G_j) \times_{\mathfrak{MDE}(G_{i-1}/G_{j+1})} \mathfrak{MDE}(G_i/G_{j+1})$$

attached to the pair of morphisms $(\pi_{[j,i-1],[j,i]}, \pi_{[j+1,i],[j,i]})$ has a natural structure as a torsor for the formal group $\mathfrak{DE}(G_i/G_{i-1}, G_j/G_{j-1})$.

Theorem 7.7.

(i) *If* $1 \leq i \leq m-1$, *then* $\mathcal{C}(\mathfrak{Def}(G_{i+1}/G_{i-1}))$ *is a torsor for the p-divisible formal group* $\mathfrak{DE}(G_{i+1}/G_i, G_i/G_{i-1})_{p\text{-div}}$.

(ii) *If* $0 \leq j_1 \leq j_2 < i_2 \leq i_1 \leq m$, *then the restriction of* $\pi_{[j_2,i_2],[j_1,i_1]}$ *to the closed formal subscheme* $\mathcal{C}(\mathfrak{Def}(G_{i_1}/G_{j_1}))$ *of* $\mathfrak{MDE}(G_{i_1}/G_{j_1})$ *factors through* $\mathcal{C}(\mathfrak{Def}(G_{i_2}/G_{j_2})) \hookrightarrow \mathfrak{MDE}(G_{i_2}/G_{j_2})$, *and induces a formally smooth morphism*

$$\pi_{[j_2,i_2],[j_1,i_1]} : \mathcal{C}(\mathfrak{Def}(G_{i_1}/G_{j_1})) \to \mathcal{C}(\mathfrak{Def}(G_{i_2}/G_{j_2}))\,.$$

(iii) *If* $1 \leq i, j \leq m$, $i \geq j+2$, *then the morphism*

$$\mathcal{C}(\mathfrak{Def}(G_i/G_j)) \longrightarrow \mathcal{C}(\mathfrak{Def}(G_{i-1}/G_j)) \times_{\mathcal{C}(\mathfrak{Def}(G_{i-1}/G_{j+1}))} \mathcal{C}(\mathfrak{Def}(G_i/G_{j+1}))$$

attached to the pair of morphisms $(\pi_{[j,i-1],[j,i]}, \pi_{[j+1,i],[j,i]})$ *is a torsor for the p-divisible formal group* $\mathfrak{DE}(G_i/G_{i-1}, G_j/G_{j-1})_{p\text{-div}}$, *respecting the* $\mathfrak{DE}(G_i/G_{i-1}, G_j/G_{j-1})$*-torsor structure of*

$$\mathfrak{MDE}(G_i/G_j) \longrightarrow \mathfrak{MDE}(G_{i-1}/G_j) \times_{\mathfrak{MDE}(G_{i-1}/G_{j+1})} \mathfrak{MDE}(G_i/G_{j+1}).$$

Proposition 7.8. *The properties* (i), (ii), (iii) *in Theorem 7.7 determine uniquely the family of formal schemes* $\{\mathcal{C}(\mathfrak{Def}(G_i/G_j)) : 0 \leq j < i \leq m\}$, *where each member* $\mathcal{C}(\mathfrak{Def}(G_i/G_j))$ *of the family is considered as a closed formal subscheme of* $\mathfrak{Def}(G_i/G_j)$.

Remark 7.9. It is possible to do better than what was stated in Prop. 7.8. Namely, one can actually *construct* closed subschemes $\mathfrak{MDE}(G_i/G_j)_{p\text{-div}}$ of $\mathfrak{MDE}(G_i/G_j)$, satisfying the properties (i), (ii), (iii) in Theorem 7.7, using structural properties of the formal schemes $\mathfrak{MDE}(G_i/G_j)$, without the concept of leaves, in an inductive way. An important ingredient of the construction uses the theory of biextensions due to Mumford [17] and Grothendieck [11]. Of course, $\mathfrak{MDE}(G_i/G_j)_{p\text{-div}}$ is canonically isomorphic to $\mathcal{C}(\mathfrak{Def}(G_i/G_j))$ by Proposition 7.8. However that construction is a bit complicated, so we do not give further indication here.

Corollary 7.10. *Notation as in Thm. 7.7. Then*

$$\dim(\mathcal{C}(\mathfrak{Def}(G))) = \sum_{1 \leq j < i \leq m} (\mu_i - \mu_j) \cdot h_i \cdot h_j\,,$$

where μ_i *is the slope of* G_i/G_{i-1} *and* h_i *is the height of* G_i/G_{i-1}, *for* $i = 1, \ldots, m$.

Proposition 7.11. *Let* G *be a Barsotti–Tate group over* k, *with a principal quasi-polarization* λ. *Then* λ *induces an involution on* $\mathfrak{MDE}(G)_{p\text{-div}}$. *Denote by* $\mathfrak{MDE}(G)^{\text{sym}}_{p\text{-div}}$ *the maximal closed subscheme of* $\mathfrak{MDE}(G)_{p\text{-div}}$ *which is fixed by the involution. Then* $\mathfrak{MDE}(G)^{\text{sym}}_{p\text{-div}}$ *is the largest closed*

formal subscheme of $\mathfrak{MDE}(G)_{p\text{-div}}$ *such that* λ *extends to a principal quasi-polarization on the restriction to* $\mathfrak{MDE}(G)^{\text{sym}}_{p\text{-div}}$ *of the universal Barsotti–Tate group over* $\mathfrak{MDE}(G)_{p\text{-div}} \subset \mathfrak{Def}(G)$. *If* $(G,\lambda) = (A_x[p^\infty], \lambda_x[p^\infty])$ *for some point* $x \in \mathcal{A}_{g,n}(k)$, *then there is a natural isomorphism of formal schemes from* $\mathfrak{MDE}(G)^{\text{sym}}_{p\text{-div}}$ *to* $\mathcal{C}^{/x}$, *where* $\mathcal{C}$ *is the leaf in* $\mathcal{A}_{g,n}$ *passing through* x.

Proposition 7.12. *Let* A_x *be a* g*-dimensional principally polarized abelian variety over* k. *Suppose that* $A_x[p^\infty]$ *has Frobenius slopes* $\mu_1 < \mu_2 < \cdots < \mu_m$, *so that* $\mu_i + \mu_{m-i+1} = 1$ *for* $i = 1, \ldots, m$. *Let* h_i *be the multiplicity of* μ_i, *so that* $h_i = h_{m-i+1}$ *for all* i, $\sum_{i=1}^m h_i = 2g$, $\sum_{i=1}^m h_i\mu_i = g$. *Then*

$$\dim(\mathcal{C}(x)) = \frac{1}{2} \sum_{i<j,\, i+j\neq 1} (\mu_j - \mu_i)\cdot h_i \cdot h_j + \frac{1}{2} \sum_{2i\leq m} (1-2\mu_i)\cdot h_i(h_i+1)\,.$$

Remark 7.13. Proposition 7.12 follows from Proposition 7.11 and Corollary 7.5; see [2].

Remark 7.14. Historically, the formula for the dimension of a leaf $\mathcal{C}(x)$ in $\mathcal{A}_{g,n}$ (resp. the dimension of the leaf $\mathcal{C}(\mathfrak{Def}(G))$ in the deformation space of a Barsotti–Tate group G) were first conjectured by Oort, in terms of the number of lattice points inside suitable regions under the Newton polygon of A_x (resp. G), after suggestions by B. Poonen. See [7] for the original proofs of Propositions 7.10 and 7.12, which depend on the following fact, proved in [20]: If G_1, G_2 are Barsotti–Tate groups over k, G_1 is minimal, and $G_1[p]$ is isomorphic to $G_2[p]$, then G_2 is isomorphic to G_1.

Remark 7.15. The theory of canonical coordinates inspires a conjectural group-theoretic formula for the dimension of leaves in the reduction over k of a Shimura variety. That formula will be explained in a future article with C.-F. Yu, and verified for modular varieties of PEL-type.

8 A rigidity result for p-divisible formal groups

Let k be an algebraically closed field of characteristic p. Let X be a p-divisible formal group over k. Then $\text{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ is a semisimple algebra of finite dimension over $\mathbb{Q}_p$, and $\text{End}_k(X)$ is an order in $\text{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$. Let H be a connected reductive linear algebraic group over $\mathbb{Q}_p$. Let ρ be a $\mathbb{Q}_p$ rational homomorphism $H(\mathbb{Q}_p) \to (\text{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p)^\times$, i.e., ρ comes from a $\mathbb{Q}_p$-homomorphism from H to the linear algebraic group over $\mathbb{Q}_p$ whose R-valued points is $(\text{End}_k(X) \otimes_{\mathbb{Z}_p} R)^\times$ for every commutative $\mathbb{Q}_p$-algebra R. Let $U \subset H(\mathbb{Q}_p)$ be an open subgroup of $H(\mathbb{Q}_p)$ such that $\rho(U) \subseteq \text{End}_k(X)^\times$, so that U operates on X via ρ.

Theorem 8.1. *Notation as above. Let* Z *be an irreducible closed formal subscheme of* X *which is stable under the action of* U. *Let* $\mathbf{r}_x$ *be the left regular*

representation of $(\mathrm{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p)^\times$ *on* $\mathrm{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$*, viewed as a* $\mathbb{Q}_p$*-linear representation of the group* $(\mathrm{End}_k(X) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p)^\times$ *on a finite-dimensional* $\mathbb{Q}_p$*-vector space. We assume that the composition* $\mathbf{r}_X \circ \rho$ *of* ρ *with* $\mathbf{r}_X$ *does not contain the trivial representation of* H *as a subquotient. Then* Z *is a* p*-divisible formal subgroup of* X*.*

Remark 8.2.

(i) Theorem 8.1 is a considerable strengthening of [6, §4, Proposition 4], in several aspects. There, the p-divisible formal group is a formal torus, and the formal subvariety is assumed to be formally smooth. The most significant part is that, in [6, §4, Proposition 4], the symmetry group $\mathcal{O}_{\wp_1}^\times \times \cdots \times \mathcal{O}_{\wp_r}^\times$ has about the same size as the formal torus

$$\left(Y_1 \otimes_{\mathbb{Z}_p} \widehat{\mathbb{G}}_m\right) \times \cdots \times \left(Y_r \otimes_{\mathbb{Z}_p} \widehat{\mathbb{G}}_m\right)$$

in some sense, while the symmetry group H in Theorem 8.1 can be quite small compared with the p-divisible formal group X.

(ii) A "typical" special case of Theorem 8.1 is to take $H = \mathbb{G}_m$, $U = \mathbb{Z}_p^\times$, with each $u \in \mathbb{Z}_p^\times$ operating as $[u]_X$ on X, the map "multiplication by u" on X. Any argument which proves this special case is likely to be strong enough to prove Theorem 8.1 as well.

(iii) The proof of Theorem 8.1 in [5] is elementary, in the sense that it is mostly manipulation of power series.

9 The Hilbert trick

Notation 9.1. Let $n \geq 3$ be an integer prime to p. Let $x \in \mathcal{A}_{g,n}(\overline{\mathbb{F}_p})$ be an $\overline{\mathbb{F}_p}$-valued point of $\mathcal{A}_{g,n}$. Let $B = \mathrm{End}_{\overline{\mathbb{F}_p}}(A_x) \otimes_{\mathbb{Z}} \mathbb{Q}$, and let $*$ be the involution of B induced by λ_x. Let $E = F_1 \times \cdots \times F_m$ be a product of totally real number fields contained in B, fixed under the involution $*$, such that $\dim_{\mathbb{Q}}(E) = g$. Let $\mathcal{O}_E = \mathcal{O}_{F_1} \times \cdots \times \mathcal{O}_{F_m}$. Denote by $\mathrm{SL}(2, E)$ the linear algebraic group over $\mathbb{Q}$ whose set of R-valued points is $\mathrm{SL}_2(E \otimes_{\mathbb{Q}} R)$ for every $\mathbb{Q}$-algebra R. There exists a "standard embedding" $h : \mathrm{SL}(2, E) \hookrightarrow \mathrm{Sp}_{2g}$, well-defined up to conjugation.

We will use the following variant of the definition of Hilbert modular varieties in [31], slightly different from the definition in [8]. Denote by $\mathcal{M}_{E,n}$ the Hilbert modular scheme attached to $\mathcal{O}_E$, such that for every $\overline{\mathbb{F}_p}$-scheme S, $\mathcal{M}_{E,n}(S)$ is the set of isomorphism classes of $(A \to S, \lambda, \iota, \eta)$, where $A \to S$ is an abelian scheme, $\iota : \mathcal{O}_E \to \mathrm{End}_S(A)$ is an injective ring homomorphism,

λ is an $\mathcal{O}_E$-linear principal polarization of $A \to S$ of degree prime to p, and η is a level-n structure on $A \to S$. See [31, §5]. The modular scheme $\mathcal{M}_{E,n}$ is locally of finite type over k, and every irreducible component of $\mathcal{M}_{E,n}$ is of finite type over $\overline{\mathbb{F}_p}$. There is a set of algebraic correspondences on $\mathcal{M}_{E,n}$, coming from the adelic group $\mathrm{SL}_2(E \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)})$, called the prime-to-$p$ Hecke correspondences on the Hilbert modular scheme $\mathcal{M}_{E,n}$.

Proposition 9.2 (Hilbert trick). *Notation as above. Then there exists*

- *a non-empty open-and-closed subscheme $\mathcal{M}_0$ of $\mathcal{M}_{E,n_1}$ for some natural number $n_1 \geq 3$ not divisible by p,*
- *a finite morphism $\mathcal{M}_0 \to \mathcal{M}_{E,n}$,*
- *a point $y \in \mathcal{M}_0(\overline{\mathbb{F}_p})$, and*
- *a finite morphism $f : \mathcal{M}_0 \to \mathcal{A}_{g,n}$*

such that

(i) $f(y) = x$,

(ii) *f is compatible with the prime-to-p Hecke correspondences on $\mathcal{M}_0$ and $\mathcal{A}_{g,n}$, coming from the embedding $h : \mathrm{SL}(2, E) \hookrightarrow \mathrm{Sp}_{2g}$, and*

(iii) *the pull-back by f of the universal abelian scheme over $\mathcal{A}_{g,n}$ is isogenous to the universal abelian scheme over $\mathcal{M}_0$.*

The idea of the proof of Proposition 9.2 is as follows. It is well-known that every abelian variety defined over a finite field has "sufficiently many complex multiplications". Hence every maximal commutative semisimple subalgebra L of B stable under the Rosati involution $*$ is a product of CM-fields, and the subalgebra of L fixed under $*$ is a product of totally real number fields. In particular this shows the existence of subalgebras E with the required properties in Notation 9.1. If $\mathrm{End}_{\overline{\mathbb{F}_p}}(A_x)$ contains $\mathcal{O}_E$, then we obtain a natural morphism $\mathcal{M}_{E,n} \to \mathcal{A}_{g,n}$ passing through $x = [(A_x, \lambda_x, \eta_x)] \in \mathcal{A}_{g,n}(\overline{\mathbb{F}_p})$. In general $E \cap \mathrm{End}_{\overline{\mathbb{F}_p}}(A_x)$ is an order of $\mathcal{O}_E$, and we have to use an isogeny correspondence to conclude the proof of Proposition 9.2.

Remark 9.3. The local stabilizer principle and Theorem 8.1, applied to a point y of a Hilbert modular variety $\mathcal{M}_{E,n}$ over $\overline{\mathbb{F}_p}$, implies that there are only a finite number of possibilities of $\overline{\mathcal{H}_{E,n}(y)}^{/y}$, as a closed formal subscheme of $\mathcal{M}_{E,n}$ over k, where $\mathcal{H}_{E,n}(y)$ denotes the prime-to-p Hecke orbit of y in $\mathcal{M}_{E,n}$. The possibilities are parametrized by non-empty subsets of the finite set $\mathrm{Spec}(\mathcal{O}_E/p\mathcal{O}_E)$ of maximal ideals of $\mathcal{O}_E$ containing p. Then one can verify that the subset of $\mathrm{Spec}(\mathcal{O}_E/p\mathcal{O}_E)$ attached to $\overline{\mathcal{H}_{E,n}(y)}^{/y}$ must be equal to $\mathrm{Spec}(\mathcal{O}_E/p\mathcal{O}_E)$ itself. That proves the continuous part of the Hecke orbit conjecture for Hilbert modular varieties. The above line of argument is possible because $\mathrm{SL}_2(E)$ is "small" in some sense, for instance each semisimple factor of $\mathrm{SL}_2(E)$n has $\mathbb{Q}$-rank one. So Hilbert modular varieties are "not too big"

either, and it turns out that one can understand the incidence relation of the Lie-alpha strata to prove the discrete part of the Hecke orbit conjecture for Hilbert modular varieties. With the Hecke orbit conjecture for Hilbert modular varieties known, the Hilbert trick becomes an effective tool for the Hecke orbit conjecture for the Siegel modular varieties $\mathcal{A}_{g,n}$, as well as other modular varieties of PEL-type.

Remark 9.4. In this section the base field is $\overline{\mathbb{F}_p}$, because every abelian variety over $\overline{\mathbb{F}_p}$ has sufficiently many complex multiplications. So it seems that if we use the Hilbert trick, we would be able to deal with the Hecke orbit conjecture (HO) "only" in the case when the algebraically closed base field k is equal to $\overline{\mathbb{F}_p}$. However every closed subvariety of $\mathcal{A}_{g,n}$ over k is finitely presented over k, and a standard argument in algebraic geometry shows that the validity of (HO) over $\overline{\mathbb{F}_p}$ implies the validity of (HO) over every algebraically closed field k. See the beginning of §3 of [6] for details.

10 Hypersymmetric points

Let k be an algebraically closed field of characteristic p as before.

Definition 10.1. *An abelian variety A over k is* hypersymmetric *if the natural map*

$$\mathrm{End}_k(A) \otimes_{\mathbb{Z}} \mathbb{Z}_p \to \mathrm{End}_k(A[p^\infty])$$

is an isomorphism. An equivalent condition is that the canonical map

$$\mathrm{End}_k(A) \otimes_{\mathbb{Z}} \mathbb{Q}_p \to \mathrm{End}_k(A[p^\infty]) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$$

is an isomorphism.

Remark 10.2. It is clear from the definition that the abelian variety A_x has sufficiently many complex multiplications for any hypersymmetric point x. Therefore a theorem of Grothendieck tells us that A_x is isogenous to an abelian variety defined over $\overline{\mathbb{F}_p}$; see [23] for a proof of Grothendieck's theorem.

Examples 10.3.

(i) A g-dimensional ordinary abelian variety over k is hypersymmetric if and only if it is isogenous to a g-fold self-product $E \times \cdots \times E$, where E is an ordinary elliptic curve defined over $\overline{\mathbb{F}_p}$.

(ii) Let A be an abelian variety over k such that $A[p^\infty]$ has exactly two slopes, $g = \dim(A)$. It is hypersymmetric if and only if $\mathrm{End}_k(A) \otimes_{\mathbb{Z}} \mathbb{Q}$ is a central simple algebra over an imaginary quadratic field, and

$$\dim_{\mathbb{Q}}(\mathrm{End}_k(A) \otimes_{\mathbb{Z}} \mathbb{Q}) = 2g^2.$$

The assertions in the two examples can be verified using Honda–Tate theory for abelian varieties over finite fields. See [30] and [29] for the Honda–Tate theory.

Remark 10.4. In every given Newton polygon stratum W_ξ in $\mathcal{A}_{g,n}$ over k, there exists a hypersymmetric point $x \in W_\xi(k)$. This statement follows easily from the Honda–Tate theory; see [19] for a proof.

Remark 10.5. Let $E = F_1 \times \cdots \times F_r$ be a totally real number field such that there is only one place of F_i above p for $i = 1, \ldots, r$. Let $\mathcal{M}_E$ be the Hilbert modular variety over k attached to $\mathcal{M}_E$. Then there exists a hypersymmetric point in every given Newton polygon stratum of $\mathcal{M}_E$. Similarly, there exists a hypersymmetric point in every given leaf of $\mathcal{M}_E$. This statement can be derived from the Honda–Tate theory and the "foliation structure" on $\mathcal{M}_E$.

Theorem 10.6. *Let $[(A_x, \lambda_x)]$ be a point of $\mathcal{A}_g(k)$ such that*

- *A_x is hypersymmetric, and*
- *A_x is split, i.e., A_x is isomorphic to a product $B_1 \times \cdots \times B_m$, where each B_i is an abelian variety over k, and each B_i has at most two slopes.*

Then Zariski closure in $\mathcal{A}_g$ of the the prime-to-p Hecke orbit $\mathcal{H}^{(p)}(x)$ contains the irreducible component of the leaf $\mathcal{C}(x)$ passing through x.

Remark 10.7. A special case of Theorem 10.6 is an example of M. Larsen; see [6, §1].

The proof of Theorem 10.6 uses Proposition 6.1, Theorem 8.1 and the theory of canonical coordinates. Here we sketch a proof of the special case when $A_x[p^\infty]$ is isomorphic to a product $X \times Y$, where X, Y are isoclinic Barsotti–Tate groups of height g, with slopes $\mu_X < \mu_Y$, $\mu_X + \mu_Y = 1$. The principal polarization λ_x induces an isomorphism between X and the Serre dual of Y. The theory of canonical coordinates tells us that $\mathcal{C}(x)^{/x}$ is isomorphic to the maximal subgroup $\mathfrak{DE}(X,Y)^{\mathrm{sym}}_{p\text{-div}}$ of the Barsotti–Tate group $\mathfrak{DE}(X,Y)_{p\text{-div}}$ fixed under the involution induced by the principal polarization λ_x. Let $Z(x)$ be the Zariski closure of the Hecke orbit $\mathcal{H}(x)$ in $\mathcal{C}(x)$. Notice that $Z(x)$ is smooth over k: it contains a dense open subset U smooth over k by generic smoothness, and $Z(x)$ is equal to the the union of Hecke-translates of U.

Let $Z(x)^{/x}$ be the formal completion of $Z(x)$ at x. Clearly $Z(x)^{/x}$ is irreducible, because it is formally smooth over k. The local stabilizer principle says that the closed formal subscheme $Z(x)^{/x}$ of $\mathcal{C}(x)^{/x}$ is stable under the natural action of the local stabilizer U_x. By Theorem 8.1, $Z(x)^{/x}$ is a Barsotti–Tate subgroup of the Barsotti–Tate group $\mathfrak{DE}(X,Y)^{\mathrm{sym}}_{p\text{-div}}$.

Now we are ready to use Dieudonné theory and translate the last assertion into a statement in linear algebra. Let $V_X = \mathrm{M}(X) \otimes_{W(k)} B(k)$, $V_Y = \mathrm{M}(Y) \otimes_{W(k)} B(k)$. The principal polarization λ_x induces a duality pairing between V_X and V_Y. Theorem 7.3 tells us that $\mathrm{M}(\mathfrak{DE}(X,Y)_{p\text{-div}}) \otimes_{W(k)} B(k)$

is naturally isomorphic to $\mathrm{Hom}^{\mathrm{sym}}_{B(k)}(V_X, V_Y)$, the symmetric part of the internal Hom. The group U_x operates naturally on $\mathrm{M}(X) \otimes_{W(k)} B(k)$ and $\mathrm{M}(Y) \otimes_{W(k)} B(k)$. One checks that, after passing to the algebraic closure $\overline{B(k)}$ of $B(k)$, the Zariski closure of U_x operating on $V_Y \otimes_{B(k)} \overline{B(k)}$ is isomorphic to the standard representation of GL_g, and the Zariski closure of U_x operating on $V_X \otimes_{B(k)} \overline{B(k)}$ is isomorphic to the dual of the standard representation of GL_g. So the action of the Zariski closure of U_x on $\mathrm{Hom}^{\mathrm{sym}}_{B(k)}(V_X, V_Y) \otimes_{B(k)} \overline{B(k)}$ is isomorphic to the second symmetric product of the standard representation of GL_g. The last representation is absolutely irreducible; in fact it is one of the fundamental representations. Since $\mathrm{M}(Z(x)^{/x}) \otimes_{W(k)} B(k)$ is a non-trivial subrepresentation of the absolutely irreducible representation $\mathrm{Hom}^{\mathrm{sym}}_{B(k)}(V_X, V_Y)$ of U_x, we conclude that $\mathrm{M}(Z(x)^{/x}) \otimes_{W(k)} B(k)$ is equal to $\mathrm{Hom}^{\mathrm{sym}}_{B(k)}(V_X, V_Y)$, therefore $Z(x)^{/x} = \mathcal{C}(x)^{/x}$. □

Remark 10.8. A weaker form of Theorem 8.1, in which the closed formal subscheme is assumed to be formally smooth instead of being irreducible, would be sufficient for the proof of Theorem 10.6.

Proposition 10.9. *Let $\mathcal{C}^+$ be an irreducible component of a leaf $\mathcal{C}$ in $\mathcal{A}_{g,n}$, and let W_ξ be the Newton polygon stratum in $\mathcal{A}_{g,n}$ containing $\mathcal{C}^+$. Assume that W_ξ is irreducible. Then for every point $y \in W_\xi(k)$, there exists a point $x \in \mathcal{C}^+(k)$ such that there exists an isogeny from A_x to A_y, which respects the polarizations up to a multiple.*

Idea of Proof. Proposition 10.9 is an immediate consequence of the "almost product structure" on each irreducible component of a Newton polygon stratum W_ξ; see [26, Theorem 5.3]. We sketch the proof below.

Using Proposition 3.4, one constructs a finite surjective morphism $f : S \to \mathcal{C}^+$, a scheme T over k, and a morphism $g : S \times_{\mathrm{Spec}(k)} T \to W_\xi$ such that

(i) For any $s_1, s_2 \in S(k)$, $t_1, t_2 \in T(k)$, if $f(s_1) = f(s_2)$, then there exists an isogeny from $A_{g(s_1,t_1)}$ to $A_{g(s_2,t_2)}$, which respects the polarizations up to a multiple.

(ii) The image of g, in the naive sense, is a union of irreducible components of W_ξ.

So far we have not used the assumption that W_ξ is irreducible. The irreducibility of W_ξ implies that f is surjective. Proposition 10.9 follows.

Proposition 10.10. *Let $\mathcal{C}$ be a leaf in $\mathcal{A}_{g,n}$, and let W_ξ be the Newton polygon stratum in $\mathcal{A}_{g,n}$ containing $\mathcal{C}$. Assume that W_ξ is irreducible. Then the prime-to-p Hecke correspondences operate transitively on $\pi_0(\mathcal{C})$. Consequently $\mathcal{C}$ is irreducible if W_ξ is not the supersingular locus of $\mathcal{A}_{g,n}$.*

Idea of Proof. Let y be a hypersymmetric point of W_ξ; such a point exists by Remark 10.4. By Proposition 10.9, for each irreducible component $\mathcal{C}_j^+$ of $\mathcal{C}$, there exists a hypersymmetric point x_j in $\mathcal{C}_j^+$, related to y by a (possibly inseparable) isogeny which preserves the polarizations up to a multiple. Using the weak approximation theorem, one sees that the x_j's are related by suitable prime-to-p Hecke correspondences. This shows that the prime-to-p Hecke correspondences operate transitively on the irreducible components of the leaf $\mathcal{C}$. The last statement follows from Theorem 5.1.

We would like to discuss an emerging picture about the leaves and the hypersymmetric points. In many ways each non-supersingular leaf in $\mathcal{A}_{g,n}$ has properties similar to those for the Siegel modular variety in characteristic 0, of genus g and with symplectic level-n structures. The Hecke orbit conjecture (HO) is an example of this phenomenon, so is Theorem 5.1. Borrowing an idea from Hindu mythology, one might want to think of the decomposition of $\mathcal{A}_{g,n}$ into leaves as Indra-inspired.

For a leaf $\mathcal{C}$ in $\mathcal{A}_{g,n}$, the hypersymmetric points of $\mathcal{C}$ serve as an analogue of the notion of *special points* (or CM points) on a Shimura variety in characteristic 0. The following is an analogue of the André–Oort conjecture in characteristic p. Let $\mathcal{C}$ be a leaf of $\mathcal{A}_{g,n}$ over k, and let Z be a closed irreducible subvariety in $\mathcal{C}$. Assume that there is a subset $S \subset Z(k)$ such that S is dense in Z, and every point of S is hypersymmetric. Then there is a closed subvariety $X \subset \mathcal{A}_{g,n}$ which is the reduction over k of a Shimura subvariety such that Z is an irreducible component of $\mathcal{C} \cap X$. This conjecture seems to be more difficult than the André–Oort conjecture.

In another direction, one expects that the p-adic monodromy of a subvariety Z in a leaf $\mathcal{C} \subset \mathcal{A}_{g,n}$ can be described in terms of the canonical coordinates and the *naive* p-adic monodromy of Z; see the first paragraph of §14 for the definition of naive p-adic monodromy. The case when $\mathcal{C}$ is the ordinary locus of $\mathcal{A}_{g,n}$ has been considered in [3], and one expects that the general phenomenon is similar. In particular, there should be a more global theory of canonical coordinates on a leaf, and we hope to carry out such a project in the near future.

11 Splitting at supersingular points

Proposition 11.1. *Let k be an algebraically closed field of characteristic p. Let x be a point of $\mathcal{A}_{g,n}$ over k, and let $\mathcal{H}^{(p)}(x)$ be the prime-to-p Hecke orbit of x. Then there exists a point z_0 in the Zariski closure of $\mathcal{H}^{(p)}(x)$ such that A_{z_0} is a supersingular abelian variety over k.*

Remark 11.2.

(i) Similarly, every prime-to-p Hecke orbit in a Hilbert modular variety has a supersingular point in its closure.

(ii) One can replace "prime-to-p" by "ℓ-adic" in Proposition 11.1, and also in (i) above.

(iii) See [6, Proposition 6] for a proof of Proposition 11.1 and (i), (ii) above. A key ingredient is the fact that every Ekedahl–Oort stratum in $\mathcal{A}_{g,n}$ is quasi-affine; see [25].

Theorem 11.3. *Let $x \in \mathcal{A}_{g,n}(\overline{\mathbb{F}_p})$ be an $\overline{\mathbb{F}_p}$-valued point of $\mathcal{A}_{g,n}$. Let Z be the Zariski closure in $\mathcal{A}_{g,n}$ of the prime-to-p Hecke orbit $\mathcal{H}^{(p)}(x)$ of x, and let Z^0 be the intersection of Z with the leaf $\mathcal{C}(x)$ passing through x. Then there exists*

- *a point $y \in Z^0(\overline{\mathbb{F}_p})$,*
- *totally real number fields $L_1, \ldots, L_s$, and*
- *an injective ring homomorphism $\beta : L_1 \times \cdots \times L_s \longrightarrow \mathrm{End}_k(A_y) \otimes_{\mathbb{Z}} \mathbb{Q}$*

such that

(i) $[L_1 : \mathbb{Q}] + \cdots + [L_s : \mathbb{Q}] = g$,

(ii) *$\beta(L_1 \times \cdots \times L_s)$ is fixed by the Rosati involution on $\mathrm{End}_{\overline{\mathbb{F}_p}}(A_y) \otimes_{\mathbb{Z}} \mathbb{Q}$ induced by λ_y, and*

(iii) *there is only one maximal ideal in $\mathcal{O}_{L_j}$ which contains p, for $j = 1, \ldots, s$.*

In particular, there exists a point $y \in Z^0(\overline{\mathbb{F}_p})$ and abelian varieties $B_1, \ldots, B_s$ over $\overline{\mathbb{F}_p}$ such that A_y is isogenous to $B_1 \times \cdots \times B_s$, and each B_j has at most two slopes, $j = 1, \ldots, s$.

Remark 11.4. Theorem 11.3 depends crucially on the fact that x is an $\overline{\mathbb{F}_p}$-valued point. However we have seen in Remark 9.4 that we may assume that the base field k is $\overline{\mathbb{F}_p}$ when considering the Hecke orbit conjecture (HO).

We sketch a proof of Proposition 11.3, which uses the action of the local stabilizer subgroup at a supersingular point in the closure of $\mathcal{C}$ and the Hilbert trick.

We may and do assume that there exists a product $E = F_1 \times \cdots \times F_r$ of totally real number fields, $[E : \mathbb{Q}] = g$, such that there exists an embedding $\iota : \mathcal{O}_E \hookrightarrow \mathrm{End}_{\overline{\mathbb{F}_p}}(A_x)$ of rings, and $\iota(\mathcal{O}_E)$ is fixed under the Rosati involution. This means that we have a natural morphism $f : \mathcal{M}_{E,m} \longrightarrow \mathcal{A}_{g,n}$ passing through x, compatible with the Hecke correspondences, for some m prime to p, such that for every geometric point $u \in \mathcal{M}_{E,m}(\overline{\mathbb{F}_p})$, the map induced by f on the strict henselizations

$$f^{(u)} : \mathcal{M}_{E,m}^{(u)} \to \mathcal{A}_{g,n}^{(f(u))}$$

is a closed embedding. Here $\mathcal{M}_{E,m}^{(u)}$ denotes the henselization of $\mathcal{M}_{E,m}$ at u, and $\mathcal{A}_{g,n}^{(f(u))}$ denotes the henselization of $\mathcal{A}_{g,n}$ at $f(u)$. Let W be the Zariski closure of the prime-to-p Hecke orbit $\mathcal{H}_E^{(p)}(x)$ in $\mathcal{M}_{E,n}$.

By Remark 11.2 (i), there exists a supersingular point $z \in W(k)$. The local stabilizer principle tells us that the formal subscheme $Z^{/z} \subset \mathcal{A}_{g,n}$ is stable under the natural action of the local stabilizer subgroup U_z attached to z. Recall that U_z is a subgroup of $\mathrm{End}_{\overline{\mathbb{F}_p}}(A_z[p^\infty])^\times$ by definition.

One checks that there exists an element $\gamma \in U_z$ such that the subring

$$\mathrm{Ad}(\gamma)(E \otimes_{\mathbb{Q}} \mathbb{Q}_p) = \gamma \cdot (E \otimes_{\mathbb{Q}} \mathbb{Q}_p) \cdot \gamma^{-1}$$

of $\mathrm{End}_{\overline{\mathbb{F}_p}}(A_z) \otimes_{\mathbb{Z}} \mathbb{Q}_p$ is equal to the $\mathbb{Q}_p$-linear span of

$$E' := (\mathrm{Ad}(\gamma)(E) \otimes_{\mathbb{Q}} \mathbb{Q}_p) \cap \left(\mathrm{End}_{\overline{\mathbb{F}_p}}(A_z) \otimes_{\mathbb{Z}} \mathbb{Q} \right),$$

and E' is a product of totally real number fields $L_1 \times \cdots \times L_s$, such that there is only one maximal ideal in $\mathcal{O}_{L_j}$ above p for $j = 1, \ldots, s$.

Denote by $\gamma^{/z}$ the automorphism of $\mathcal{A}_{g,n}^{/z}$ attached to γ. The fact that $\gamma^{/z}(W^{/z}) \subset Z^{/z}$ tells us, in the case when

$$\mathrm{Ad}(\gamma)(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p) \subset \mathrm{End}_{\overline{\mathbb{F}_p}}(A_z[p^\infty]) = \mathrm{End}_{\overline{\mathbb{F}_p}}(A_z) \otimes_{\mathbb{Z}} \mathbb{Z}_p,$$

that there is a natural finite morphism

$$f_1 : \mathcal{M}_{E',m} \longrightarrow \mathcal{A}_{g,n}$$

with the following properties:

(1) There exists a point $z_1 \in \mathcal{M}_{E',m}(\overline{\mathbb{F}_p})$ such that $f_1(z_1) = z$.

(2) For every point $u \in \mathcal{M}_{E',m}(\overline{\mathbb{F}_p})$, the morphism f_1 induces a closed embedding, from the henselization $\mathcal{M}_{E',m}^{(u)}$ of $\mathcal{M}_{E',m}$ at u, to the henselization $\mathcal{A}_{g,n}^{(f_1(u))}$ of $\mathcal{A}_{g,n}$ at $f_1(u)$.

(3) $\gamma^{/z}(W^{/z}) \subset f_1^{/z_1} \left(\mathcal{M}_{E',n}^{/z_1} \right) \cap Z^{/z}$.

Hence the fiber product $\mathcal{M}_{E',m} \times_{\mathcal{A}_{g,n}} Z^0$ is not empty. Pick a point $\tilde{y} \in (\mathcal{M}_{E',m} \times_{\mathcal{A}_{g,n}} Z^0)(\overline{\mathbb{F}_p})$, and let y be the image of $\tilde{y}$ in $Z^0(\overline{\mathbb{F}_p})$. It is easy to see that y has the property stated in Theorem 11.3, and we are done. In general $(\mathrm{Ad}(\gamma)(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p)) \cap \left(\mathrm{End}_{\overline{\mathbb{F}_p}}(A_z) \otimes_{\mathbb{Z}} \mathbb{Z}_p \right)$ is of finite index in $\mathrm{Ad}(\gamma)(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Q}_p)$ and may not be equal to $\mathrm{Ad}(\gamma)(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p)$, and we have to use an isogeny correspondence to conclude the proof.

Remark 11.5. The last sentence in the statement of Theorem 11.3 follows from the properties (i), (ii), (iii) of A_y stated in 11.3.

12 Logical interdependencies

Let k be an algebraically closed field of characteristic p as before. We summarize the logical interdependencies of various statements.

12.1. We have seen that

$$(\mathrm{HO}) \iff (\mathrm{HO})_{\mathrm{ct}} + (\mathrm{HO})_{\mathrm{dc}}.$$

12.2. If $x \in \mathcal{A}_g(k)$ is not supersingular. then Thm. 5.1 shows that

$$(\mathrm{HO})_{\mathrm{dc}} \text{ for x} \iff \mathcal{C}(\mathrm{x}) \text{ is irreducible.}$$

12.3. Suppose that $x, y \in \mathcal{A}_g(k)$, and there is an isogeny from A_x to A_y which preserves the polarizations up to multiples. Then

$$(\mathrm{HO})_{\mathrm{ct}} \text{ for x} \iff (\mathrm{HO})_{\mathrm{ct}} \text{ for y.}$$

This is a consequence of Remark 3.5, which depends on Proposition 3.4.

12.4. Suppose that $x, y \in \mathcal{A}_g(k)$, and there is an isogeny from A_x to A_y which preserves the polarizations up to multiples. Then

$$(\mathrm{HO})_{\mathrm{dc}} \text{ for x} \iff (\mathrm{HO})_{\mathrm{dc}} \text{ for y.}$$

The proof of the above statement is similar to the argument of Proposition 10.10, using hypersymmetric points.

12.5. Let W_ξ be a non-supersingular Newton polygon stratum on $\mathcal{A}_g$, and let $\mathcal{C}$ be a leaf in W_ξ. Then

$$W_\xi \text{ is irreducible} \implies \mathcal{C} \text{ is irreducible}.$$

See Proposition 10.10.

12.6. The implication

$$(\mathrm{HO}) \text{ for Hilbert modular varieties} \implies (\mathrm{HO})_{\mathrm{ct}}$$

holds.

Here is a sketch of the proof of 12.6. Assume the Hecke orbit conjecture for Hilbert modular varieties. As remarked in Remark 9.4, we may and do assume that the base field is $\overline{\mathbb{F}_p}$. Apply the trick "splitting at supersingular points" to get a point $y \in \mathcal{A}_{g,n}(\overline{\mathbb{F}_p})$ contained in $\overline{\mathcal{H}^{(p)}(x)} \cap \mathcal{C}(x)$ as in Theorem 11.3. The Hilbert trick and the Hecke orbit conjecture for Hilbert modular varieties show that there exists a point $y_2 \in \left(\overline{\mathcal{H}^{(p)}(x)} \cap \mathcal{C}(x)\right)(\overline{\mathbb{F}_p})$ such that A_{y_2} is hypersymmetric and split. Here we used Remark 10.5 on the existence of hypersymmetric points on every leaf of the Hilbert modular subvariety in $\mathcal{A}_{g,n}$ passing through the point y. Apply Theorem 10.6; the continuous part of the Hecke orbit conjecture for a Siegel modular variety $\mathcal{A}_{g,n}$ follows.

13 Outline of the proof of the Hecke orbit conjecture

Theorem 13.1. *Every non-supersingular Newton polygon stratum in $\mathcal{A}_{g,n}$ is irreducible.*

Proof. See [21]. The proof uses Theorem 5.1 and the results in [24], [13], [25]. □

We have seen in Proposition 10.10 and 12.5 that $(\mathrm{HO})_{\mathrm{dc}}$ follows from Theorem 13.1. We are left with the continuous part $(\mathrm{HO})_{\mathrm{ct}}$ of the Hecke orbit conjecture.

The continuous part $(\mathrm{HO})_{\mathrm{ct}}$ of the Hecke orbit conjecture for Hilbert modular varieties uses Theorem 8.1 and the argument in [3, §8]; the latter depends on the main result of [12] by de Jong. It is also possible to avoid de Jong's theorem in [12], using instead the local stabilizer principle at a supersingular point, similar to the argument of [6, §5, Proposition 7]. But the argument will not be as clean.

By 12.6, to complete the proof of the Hecke orbit conjecture for the Siegel modular varieties $\mathcal{A}_{g,n}$, it suffices to prove the discrete part of the Hecke orbit conjecture for Hilbert modular varieties.

The proof of the discrete part of the Hecke orbit conjecture for Hilbert modular varieties uses the Lie-alpha stratification on Hilbert modular varieties. See [31] for some properties of the Lie-alpha stratification; see also [10] for the case when p is unramified in the totally real number field, and [1] for the case when p is totally ramified in the totally real number field. The starting point is the fact that for each given Newton polygon stratum W_ξ on a given Hilbert modular variety $\mathcal{M}_F$, there exists a leaf $\mathcal{C}$ contained in W_ξ which is an open subset of some Lie-alpha stratum of $\mathcal{M}_F$. A standard degeneration argument shows that it suffices to prove that the closure of every Lie-alpha stratum contains a superspecial point of a specific type. This observation allows us to bring in deformation theory. The last and the most crucial step was done by C.-F. Yu, who constructed enough deformations to facilitate an induction on the partial ordering on the family of irreducible components of Lie-alpha strata induced by the incidence relation.

14 *p*-adic monodromy of leaves

In this last section we mention a maximality property of the naive p-adic monodromy group. By definition, the naive p-adic monodromy representation of a leaf $\mathcal{C}(x)$ passing through a point $x \in \mathcal{A}_{g,n}(\overline{\mathbb{F}_p})$ is the natural action of the Galois group of the function field of $\mathcal{C}(x)$ on the product

$$\prod_{i=1}^{m} \mathrm{Hom}\left((G_i/G_{i-1})_x, (G_i/G_{i-1})_{\bar{\eta}}\right) ,$$

where $0 = G_0 \subset G_1 \subset G_2 \subset \cdots \subset G_m = A[p^\infty]$ is the slope filtration of $A[p^\infty] \to \mathcal{C}(x)$ denotes the slope filtration as in Proposition 3.7, and $\bar{\eta}$ is a geometric generic point of $\mathcal{C}(x)$. The naive p-adic monodromy group is the image of the naive p-adic monodromy representation. The notion of hypersymmetric points plays an important role in the proof of Theorem 14.1.

Theorem 14.1. *Let x be a hypersymmetric point such that $A_x[p^\infty]$ is minimal, i.e., the ring $\mathrm{End}_k(A_x[p^\infty])$ of endomorphisms is a maximal order of $\mathrm{End}_k(A_x[p^\infty])\otimes_{\mathbb{Z}_p}\mathbb{Q}_p$. Then the naive p-adic monodromy group of the leaf $\mathcal{C}(x)$ is maximal. In other words, if we use x as the base point, then the image of the naive p-adic monodromy group is equal to the intersection of $\mathrm{Aut}(A_x[p^\infty])$ with the unitary group attached to the pair $(\mathrm{End}_k(A_x[p^\infty])\otimes_{\mathbb{Z}_p}\mathbb{Q}_p, *)$, where $*$ denotes the involution on the semisimple algebra $\mathrm{End}_k(A_x[p^\infty])\otimes_{\mathbb{Z}_p}\mathbb{Q}_p$ over $\mathbb{Q}_p$ induced by the principal polarization λ_x on A_x.*

Corollary 14.2. *Let $x \in \mathcal{A}_{g,n}(k)$ be a closed point of $\mathcal{A}_{g,n}$ such that the ring $\mathrm{End}_k(A_x[p^\infty])$ is a maximal order of $\mathrm{End}_k(A_x[p^\infty])\otimes_{\mathbb{Z}_p}\mathbb{Q}_p$. Then the naive p-adic monodromy group of the leaf $\mathcal{C}(x)$ is maximal.*

The idea of the proof of Theorem 14.1 is the following. First we prove an analogous statement for the naive p-adic monodromy group using Ribet's method in [28], [9]. Use a hypersymmetric point x with the properties in the statement of Theorem 14.1 as the base point for computing the p-adic monodromy group. This allows us to overcome the usual sticky issues related to different choices of base points, and reduce Theorem 14.1 to showing that the conjugates of the p-adic monodromy group of a leaf in a Hilbert modular subvariety already generates the target group of the naive p-adic monodromy representation. The last group-theoretic statement is elementary and can be verified directly.

References

1. F. Andreatta and E. Z. Goren – "Hilbert modular varieties of low dimension", To appear in *Geometric Aspects of Dwork's Theory, A Volume in memory of Bernard Dwork*, A. Adolphson, F. Baldassarri, P. Berthelot, N. Katz, and F. Loeser (Eds.), 62 pages.
2. C.-L. Chai – "Canonical coordinates on leaves of p-divisible groups", Part II. Cartier theory, preprint, 28 pp., September 2003.
3. — , "Families of ordinary abelian varieties: canonical coordinates, p-adic monodromy, Tate-linear subvarieties and hecke orbits", Preprint, 55 pp., July 2003, available from `http://www.math.upenn.edu/~chai/`.

4. — , "Monodromy of Hecke-invariant subvarieties", Preprint, 10 pp., April, 2003, available from `http://www.math.upenn.edu/∼chai/`.
5. — , "A rigidity result of p-divisible formal groups", Preprint, 11 pp., July, 2003, available from `http://www.math.upenn.edu/∼chai/`.
6. — , "Every ordinary symplectic isogeny class in positive characteristic is dense in the moduli", *Invent. Math.* **121** (1995), no. 3, p. 439–479.
7. C.-L. Chai and F. Oort – "Canonical coordinates on leaves of p-divisible groups", Part I. General properties, preprint, 24 pp., August 2003.
8. P. Deligne and G. Pappas – "Singularités des espaces de modules de Hilbert, en les caractéristiques divisant le discriminant", *Compositio Math.* **90** (1994), no. 1, p. 59–79.
9. P. Deligne and K. A. Ribet – "Values of abelian L-functions at negative integers over totally real fields", *Invent. Math.* **59** (1980), no. 3, p. 227–286.
10. E. Z. Goren and F. Oort – "Stratifications of Hilbert modular varieties", *J. Algebraic Geom.* **9** (2000), no. 1, p. 111–154.
11. A. Grothendieck – *Groupes de monodromie en géométrie algébrique. I*, Séminaire de Géométrie Algébrique du Bois-Marie 1967–1969 (SGA 7 I), Lecture Notes in Math., vol. 288, Springer-Verlag, Berlin, 1972.
12. A. J. de Jong – "Homomorphisms of Barsotti-Tate groups and crystals in positive characteristic", *Invent. Math.* **134** (1998), no. 2, p. 301–333.
13. A. J. de Jong and F. Oort – "Purity of the stratification by Newton polygons", *J. Amer. Math. Soc.* **13** (2000), no. 1, p. 209–241.
14. N. M. Katz – "Slope filtration of F-crystals", Journées de Géométrie Algébrique de Rennes (Rennes, 1978), Vol. I, Astérisque, vol. 63, Soc. Math. France, Paris, 1979, p. 113–163.
15. J. Lubin and J. Tate – "Formal complex multiplication in local fields", *Ann. of Math. (2)* **81** (1965), p. 380–387.
16. Y. I. Manin – "Theory of commutative formal groups over fields of finite characteristic", *Uspehi Mat. Nauk* **18** (1963), no. 6 (114), p. 3–90.
17. D. Mumford – "Bi-extensions of formal groups", Algebraic Geometry (Internat. Colloq., Tata Inst. Fund. Res., Bombay, 1968), Oxford Univ. Press, London, 1969, p. 307–322.
18. T. Oda and F. Oort – "Supersingular abelian varieties", *Proceedings of the International Symposium on Algebraic Geometry (Kyoto Univ., Kyoto, 1977)* (Tokyo), Kinokuniya Book Store, 1978, p. 595–621.
19. F. Oort – "Hypersymmetric abelian varieties", preliminary version, Feb. 12, 2004, available from `http://www.math.uu.nl/people/oort/`.
20. — , "Minimal p-divisible groups", to appear in *Ann. of Math.*, available from `http://www.math.uu.nl/people/oort/`.
21. — , "Monodromy, Hecke orbits and Newton polygon strata", Seminar at MPI, Bonn, Feb. 14, 2004, available from `http://www.math.uu.nl/people/oort/`.
22. — , "Some questions in algebraic geometry", Preliminary version, June 1995, available from `http://www.math.uu.nl/people/oort/`.
23. — , "The isogeny class of a CM-type abelian variety is defined over a finite extension of the prime field", *J. Pure Appl. Algebra* **3** (1973), p. 399–408.
24. — , "Newton polygons and formal groups: conjectures by Manin and Grothendieck", *Ann. of Math. (2)* **152** (2000), no. 1, p. 183–206.
25. — , "A stratification of a moduli space of abelian varieties", Moduli of abelian varieties (Texel Island, 1999), Progr. Math., vol. 195, Birkhäuser, Basel, 2001, p. 345–416.

26. — , "Foliations in moduli spaces of abelian varieties", *J. Amer. Math. Soc.* **17** (2004), no. 2, p. 267–296 (electronic).
27. F. Oort and T. Zink – "Families of p-divisible groups with constant Newton polygon", appear in *Documenta Mathematica*.
28. K. A. Ribet – "p-adic interpolation via Hilbert modular forms", Algebraic geometry (Proc. Sympos. Pure Math., vol. 29, Humboldt State Univ., Arcata, Calif., 1974), Amer. Math. Soc., Providence, R. I., 1975, p. 581–592.
29. J. Tate – "Classes d'isogeny de variétés abéliennes sur un corps fini (d'apès T. Honda", *Séminaire Bourbaki 1968/69*, Exposé 352, Lecture Notes in Math., vol. 179, Springer, Berlin, 1971, p.. 95–110.
30. J. Tate – "Endomorphisms of abelian varieties over finite fields", *Invent. Math.* **2** (1966), p. 134–144.
31. C.-F. Yu – "On reduction of Hilbert-Blumenthal varieties", *Ann. Inst. Fourier (Grenoble)* **53** (2003), no. 7, p. 2105–2154.
32. T. Zink – "On the slope filtration", *Duke Math. J.* **109** (2001), no. 1, p. 79–95.

Ax–Kochen–Eršov Theorems for p-adic integrals and motivic integration

Raf Cluckers[1] and François Loeser[2]

[1] Katholieke Universiteit Leuven, Departement wiskunde, Celestijnenlaan 200B, B-3001 Leuven, Belgium. Current address: École Normale Supérieure, Département de mathématiques et applications, 45 rue d'Ulm, 75230 Paris Cedex 05, France
raf.cluckers@wis.kuleuven.ac.be

[2] École Normale Supérieure, Département de mathématiques et applications, 45 rue d'Ulm, 75230 Paris Cedex 05, France (UMR 8553 du CNRS)
Francois.Loeser@ens.fr

Summary. We express the Lefschetz number of iterates of the monodromy of a function on a smooth complex algebraic variety in terms of the Euler characteristic of a space of truncated arcs.

1 Introduction

This paper is concerned with extending classical results à la Ax–Kochen–Eršov to p-adic integrals in a motivic framework. The first section is expository, starting from Artin's conjecture and the classical work of Ax, Kochen, and Eršov and ending with recent work of Denef and Loeser giving a motivic version of the results of Ax, Kochen, and Eršov. In that section we have chosen to adopt a quite informal style, since the reader will find precise technical statements of more general results in later sections. We also explain the cell decomposition Theorem of Denef–Pas and how it leads to a quick proof of the results of Ax, Kochen, and Eršov. In sections 3, 4 and 5, we present our new, general construction of motivic integration, in the framework of constructible motivic functions. This has been announced in [2] and [3] and is developed in the paper [1]. In the last two sections we explain the relation to p-adic integration and we announce general Ax–Kochen–Eršov Theorems for integrals depending on parameters. We conclude the paper by discussing briefly the relevance of our results to the study of orbital integrals and the Fundamental Lemma.

Acknowledgments. The present text is an expanded and updated version of a talk given by the senior author at the Miami Winter School "Geometric

Methods in Algebra and Number Theory". We would like to thank the organizers for providing such a nice and congenial opportunity for presenting our work.

During the realization of this project, the first author was a postdoctoral fellow of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.).

2 From Ax–Kochen–Eršov to motives

2.1 Artin's conjecture

Let i and d be integers. A field K is said to be $C_i(d)$ if every homogeneous polynomial of degree d with coefficients in K in $d^i + 1$ (effectively appearing) variables has a non-trivial zero in K. Note we could replace "in d^i+1 variables" by "in at least $d^i + 1$ variables" in that definition. When the field K is $C_i(d)$ for every d we say it is C_i. For instance for a field K to be C_0 means to be algebraically closed, and all finite fields are C_1, thanks to the Chevalley–Warning Theorem. Also, one can prove without much trouble that if the field K is C_i, then the fields $K(X)$ and $K((X))$ are C_{i+1}. It follows in particular that the fields $\mathbf{F}_q((X))$ are C_2.

Conjecture 2.2 (Artin). *The p-adic fields $\mathbf{Q}_p$ are C_2.*

In 1965 Terjanian [24] gave an example of a homogeneous form of degree 4 in $\mathbf{Q}_2$ in $18 > 4^2$ variables having only trivial zeroes in $\mathbf{Q}_2$, thus giving a counterexample to Artin's Conjecture. Let us briefly recall Terjanian's construction, referring to [24] and [7] for more details. The basic idea is the following: if f is a homogeneous polynomial of degree 4 in 9 variables with coefficients in $\mathbf{Z}$, such that, for every x in $\mathbf{Z}^9$, if $f(x) \equiv 0 \bmod 4$, then 2 divides x, then the polynomial in 18 variables $h(x, y) = f(x) + 4f(y)$ will have no non-trivial zero in $\mathbf{Q}_2$. An example of such a polynomial f is given by

$$f = n(x_1, x_2, x_3) + n(x_4, x_5, x_6) + n(x_7, x_8, x_9) \tag{1}$$

with

$$n(X, Y, Z) = X^2YZ + XY^2Z + XYZ^2 + X^2 + Y^2 + Z^2 - X^4 - Y^4 - Z^4. \tag{2}$$

At about the same time, Ax and Kochen proved that, if not true, Artin's conjecture is asymptotically true in the following sense:

Theorem 2.3 (Ax–Kochen). *An integer d being fixed, all but finitely many fields $\mathbf{Q}_p$ are $C_2(d)$.*

2.4 Some Model Theory

In fact, Theorem 2.3 is a special instance of the following, much more general, statement:

Theorem 2.5 (Ax–Kochen–Eršov). *Let φ be a sentence in the language of rings. For all but finitely many prime numbers p, φ is true in $\mathbf{F}_p((X))$ if and only if it is true in $\mathbf{Q}_p$. Moreover, there exists an integer N such that, for any two local fields K, K' with isomorphic residue fields of characteristic $> N$, one has that φ is true in K if and only if it is true in K'.*

By a sentence in the language of rings, we mean a formula, without free variables, built from symbols 0, $+$, $-$, 1, $\times$, symbols for variables, logical connectives $\wedge$, $\vee$, $\neg$, quantifiers $\exists$, $\forall$ and the equality symbol $=$. It is very important that in this language, any given natural number can be expressed – for instance 3 as $1+1+1$ – but that quantifiers running for instance over natural numbers are not allowed. Given a field k, we may interpret any such formula φ in k by letting the quantifiers run over k, and, when φ is a sentence, we may say whether φ is true in k or not. Since for a field to be $C_2(d)$ for a fixed d may be expressed by a sentence in the language of rings, we see that Theorem 2.3 is a special case of Theorem 2.5. On the other hand, it is for instance impossible to express by a single sentence in the language of rings that a field is algebraically closed.

In fact, it is natural to introduce here the language of valued fields. It is a language with two sorts of variables. The first sort of variables will run over the valued field and the second sort of variables will run over the value group. We shall use the language of rings over the valued field variables and the language of ordered abelian groups $0, +, -, \geq$ over the value group variables. Furthermore, there will be an additional functional symbol ord , going from the valued field sort to the value group sort, which will be interpreted as assigning to a non-zero element in the valued field its valuation.

Theorem 2.6 (Ax–Kochen–Eršov). *Let K and K' be two henselian valued fields of residual characteristic zero. Assume their residue fields k and k' and their value groups Γ and Γ' are elementary equivalent, that is, they have the same set of true sentences in the rings, resp. ordered abelian groups, language. Then K and K' are elementary equivalent, that is, they satisfy the same set of formulas in the valued fields language.*

We shall explain a proof of Theorem 2.5 after Theorem 2.10. Let us sketch how Theorem 2.5 also follows from Theorem 2.6. Indeed this follows directly from the classical ultraproduct construction. Let φ be a given sentence in the language of valued fields. Suppose by contradiction that for each r in $\mathbf{N}$ there exist two local fields K_r, K_r' with isomorphic residue field of characteristic $> r$ and such that φ is true in K_r and false in K_r'. Let U be a non-principal ultrafilter on $\mathbf{N}$. Denote by $\mathbf{F}_U$ the corresponding ultraproduct of the residue

fields of K_r, r in $\mathbf{N}$. It is a field of characteristic zero. Now let K_U and K'_U be respectively the ultraproduct relative to U of the fields K_r and K'_r. They are both henselian with residue field $\mathbf{F}_U$ and value group $\mathbf{Z}_U$, the ultraproduct over U of the ordered group $\mathbf{Z}$. Hence certainly Theorem 2.6 applies to K_U and K'_U. By the very ultraproduct construction, φ is true in K_U and false in K'_U, which is a contradiction.

2.7 Cell decomposition

In this paper, we shall in fact consider, instead of the language of valued fields, what we call a language of Denef–Pas, $\mathcal{L}_{\mathrm{DP}}$. It it is a language with three sorts, running respectively over valued field, residue field, and value group variables. For the first two sorts, the language is the ring language and for the last sort, we take any extension of the language of ordered abelian groups. For instance, one may choose for the last sort the Presburger language $\{+, 0, 1, \leq\} \cup \{\equiv_n \mid n \in \mathbf{N},\ n > 1\}$, where $\equiv_n$ denote equivalence modulo n. We denote the corresponding Denef–Pas language by $\mathcal{L}_{\mathrm{DP,P}}$. We also have two additional symbols, ord as before, and a functional symbol ac, going from the valued field sort to the residue field sort.

A typical example of a structure for that language is the field of Laurent series $k((t))$ with the standard valuation $\mathrm{ord} : k((t))^\times \to \mathbf{Z}$ and ac defined by $\mathrm{ac}(x) = xt^{-\mathrm{ord}\,(x)} \mathrm{mod} t$ if $x \neq 0$ in $k((t))$ and by $\mathrm{ac}(0) = 0$.[3] Also, we shall usually add to the language constant symbols in the first, resp. second, sort for every element of $k((t))$ resp. k, thus considering formulas with coefficients in $k((t))$, resp. k, in the valued field, resp. residue field, sort. Similarly, any finite extension of $\mathbf{Q}_p$ is naturally a structure for that language, once a uniformizing parameter ϖ has been chosen; one just sets $\mathrm{ac}(x) = x\varpi^{-\mathrm{ord}\,(x)} \mathrm{mod} \varpi$ and $\mathrm{ac}(0) = 0$. In the rest of the paper, for $\mathbf{Q}_p$ itself, we shall always take $\varpi = p$.

We now consider a valued field K with residue field k and value group $\mathbf{Z}$. We assume k is of characteristic zero, K is henselian and admits an angular component map, that is, a map $\mathrm{ac} : K \to k$ such that $\mathrm{ac}(0) = 0$, ac restricts to a multiplicative morphism $K^\times \to k^\times$, and on the set $\{x \in K, \mathrm{ord}\,(x) = 0\}$, ac restricts to the canonical projection to k. We also assume that $(K, k, \Gamma, \mathrm{ord}, \mathrm{ac})$ is a structure for the language $\mathcal{L}_{\mathrm{DP}}$.

We call a subset C of $K^m \times k^n \times \mathbf{Z}^r$ definable if it may be defined by an $\mathcal{L}_{\mathrm{DP}}$-formula. We call a function $h : C \to K$ definable if its graph is definable.

Definition 2.8. *Let $D \subset K^m \times k^{n+1} \times \mathbf{Z}$ and $c : K^m \times k^n \to K$ be definable. For ξ in k^n, we set*

$$A(\xi) = \Big\{(x,t) \in K^m \times K \;\Big|\; (x, \xi, \mathrm{ac}(t - c(x,\xi)), \mathrm{ord}_0(t - c(x,\xi))) \in D\Big\}, \quad (3)$$

[3] Technically speaking, any function symbol of a first order language must have as domain a product of sorts; a concerned reader may choose an arbitrary extension of ord to the whole field K; sometimes we will use $\mathrm{ord}_0 : K \to \mathbf{Z}$ which sends 0 to 0 and non-zero x to $\mathrm{ord}\,(x)$.

where $\mathrm{ord}_0(x) = \mathrm{ord}\,(x)$ *for* $x \neq 0$ *and* $\mathrm{ord}_0(0) = 0$. *If for every* ξ *and* ξ' *in* k^n *with* $\xi \neq \xi'$, *we have* $A(\xi) \cap A(\xi') = \emptyset$, *then we call*

$$A = \bigcup_{\xi \in k^n} A(\xi) \tag{4}$$

a cell in $K^m \times K$ *with parameters* ξ *and center* $c(x, \xi)$.

Now we can state the following version of the cell decomposition theorem of Denef and Pas:

Theorem 2.9 (Denef–Pas [21]). *Consider functions* $f_1(x,t), \ldots, f_r(x,t)$ *on* $K^m \times K$ *which are polynomials in* t *with coefficients definable functions from* K^m *to* K. *Then,* $K^m \times K$ *admits a finite partition into cells* A *with parameters* ξ *and center* $c(x,\xi)$, *such that, for every* ξ *in* k^n, (x,t) *in* $A(\xi)$, *and* $1 \leq i \leq r$, *we have,*

$$\mathrm{ord}_0 f_i(x,t) = \mathrm{ord}_0 h_i(x,\xi)(t - c(x,\xi))^{\nu_i} \tag{5}$$

and

$$\mathrm{ac} f_i(x,t) = \xi_i, \tag{6}$$

where the functions $h_i(x,\xi)$ *are definable and* ν_i, n *are in* $\mathbf{N}$ *and where* $\mathrm{ord}_0(x) = \mathrm{ord}\,(x)$ *for* $x \neq 0$ *and* $\mathrm{ord}_0(0) = 0$.

Using Theorem 2.9 it is not difficult to prove by induction on the number of valued field variables the following quantifier elimination result (in fact, Theorems 2.9 and 2.10 have a joint proof in [21]):

Theorem 2.10 (Denef–Pas [21]). *Let* K *be a valued field satisfying the above conditions. Then, every formula in* $\mathcal{L}_{\mathrm{DP}}$ *is equivalent to a formula without quantifiers running over the valued field variables.*

Let us now explain why Theorem 2.5 follows easily from Theorem 2.10. Let U be a non-principal ultrafilter on $\mathbf{N}$. Let K_r and K'_r be local fields for every r in $\mathbf{N}$, such that the residue field of K_r is isomorphic to the residue field of K'_r and has characteristic $> r$. We consider again the fields K_U and K'_U that are respectively the ultraproduct relative to U of the fields K_r and K'_r. By the argument we already explained it is enough to prove that these two fields are elementary equivalent. Clearly they have isomorphic residue fields and isomorphic value groups (isomorphic as ordered groups). Furthermore they both satisfy the hypotheses of Theorem 2.10. Consider a sentence true for K_U. Since it is equivalent to a sentence with quantifiers running only over the residue field variables and the value group variables, it will also be true for K'_U, and vice versa.

Note that the use of cell decomposition to prove Ax–Kochen–Eršov type results goes back to P.J. Cohen [5].

2.11 From sentences to formulas

Let φ be a formula in the language of valued fields or, more generally, in the language $\mathcal{L}_{\mathrm{DP,P}}$ of Denef–Pas. We assume that φ has m free valued field variables and no free residue field nor value group variables. For every valued field K which is a structure for the language $\mathcal{L}_{\mathrm{DP}}$, we denote by $h_\varphi(K)$ the set of points $(x_1, \ldots, x_m)$ in K^m such that $\varphi(x_1, \ldots, x_m)$ is true.

When $m = 0$, φ is a sentence and $h_\varphi(K)$ is either the one point set or the empy set, depending on whether φ is true in K or not. Having Theorem 2.5 in mind, a natural question is to compare $h_\varphi(\mathbf{Q}_p)$ with $h_\varphi(\mathbf{F}_p((t)))$.

An answer is provided by the following statement:

Theorem 2.12 (Denef–Loeser [13]). *Let φ be a formula in the language $\mathcal{L}_{\mathrm{DP,P}}$ with m free valued field variables and no free residue field nor value group variables. There exists a virtual motive M_φ, canonically attached to φ, such that, for almost all prime numbers p, the volume of $h_\varphi(\mathbf{Q}_p)$ is finite if and only if the volume of $h_\varphi(\mathbf{F}_p((t)))$ is finite, and in this case they are both equal to the number of points of M_φ in $\mathbf{F}_p$.*

Here we have chosen to state Theorem 2.12 in an informal, non-technical way. A detailed presentation of more general results we recently obtained is given in § 7. A few remarks are necessary in order to explain the statement of Theorem 2.12. Firstly, what is meant by volume? Let d be an integer such that for almost all p, $h_\varphi(\mathbf{Q}_p)$ is contained in $X(\mathbf{Q}_p)$, for some subvariety of dimension d of $\mathbf{A}^m_{\mathbf{Q}}$. Then the volume is taken with respect to the canonical d-dimensional measure (cf. § 6 and 7). Implicit in the statement of the theorem is the fact that $h_\varphi(\mathbf{Q}_p)$ and $h_\varphi(\mathbf{F}_p((t)))$ are measurable (at least for almost all p for the later one). Originally, cf. [13] [14] [11], the virtual motive M_φ lies in a certain completion of the ring $K_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q}$ explained in Section 5.7 (in particular, $K_0^{\mathrm{mot}}(\mathrm{Var}_k)$ is a subring of the Grothendieck ring of Chow motives with rational coefficients), but it now follows from the new construction of motivic integration developed in [1] that we can take M_φ in the ring obtained from $K_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q}$ by inverting the Lefschetz motive $\mathbf{L}$ and $1 - \mathbf{L}^{-n}$ for $n > 0$.

One should note that even for $m = 0$, Theorem 2.12 gives more information than Theorem 2.5, since it says that for almost all p the validity of φ in $\mathbf{Q}_p$ and $\mathbf{F}_p((t))$ is governed by the virtual motive M_φ. Finally, let us note that Theorem 2.5 naturally extends to integrals of definable functions as will be explained in § 7.

The proof of Theorem 2.12 is based on motivic integration. In the next sections we shall give a quick overview of the new general construction of motivic integration given in [1], that allows one to integrate a very general class of functions, constructible motivic functions. These results have already been announced in a condensed way in the notes [2] and [3]; here, we are given the opportunity to present them more leisurely and with some more details.

3 Constructible motivic functions

3.1 Definable subassignments

Let φ be a formula in the language $\mathcal{L}_{\mathrm{DP,P}}$ with coefficients in $k((t))$, resp. k, in the valued field, resp. residue field, sort, having say respectively m, n, and r free variables in the various sorts. To such a formula φ we assign, for every field K containing k, the subset $h_\varphi(K)$ of $K((t))^m \times K^n \times \mathbf{Z}^r$ consisting of all points satisfying φ. We shall call the datum of such subsets for all K definable (sub)assignments. In analogy with algebraic geometry, where the emphasis is not put anymore on equations but on the functors they define, we consider instead of formulas the corresponding subassignments (note $K \mapsto h_\varphi(K)$ is in general not a functor). Let us make these definitions more precise.

First, we recall the definition of subassignments, introduced in [13]. Let $F : \mathcal{C} \to \mathrm{Ens}$ be a functor from a category $\mathcal{C}$ to the category of sets. By a subassignment h of F we mean the datum, for every object C of $\mathcal{C}$, of a subset $h(C)$ of $F(C)$. Most of the standard operations of elementary set theory extend trivially to subassignments. For instance, given subassignments h and h' of the same functor, one defines subassignments $h \cup h'$, $h \cap h'$ and the relation $h \subset h'$, etc. When $h \subset h'$ we say h is a subassignment of h'. A morphism $f : h \to h'$ between subsassignments of functors F_1 and F_2 consists of the datum for every object C of a map $f(C) : h(C) \to h'(C)$. The graph of f is the subassignment $C \mapsto \mathrm{graph}(f(C))$ of $F_1 \times F_2$.

Next, we explain the notion of definable subassignments. Let k be a field and consider the category F_k of fields containing k. We denote by $h[m,n,r]$ the functor $F_k \to \mathrm{Ens}$ given by $h[m,n,r](K) = K((t))^m \times K^n \times \mathbf{Z}^r$. In particular, $h[0,0,0]$ assigns the one point set to every K. To any formula φ in $\mathcal{L}_{\mathrm{DP,P}}$ with coefficients in $k((t))$, resp. k, in the valued field, resp. residue field, sort, having respectively m, n, and r free variables in the various sorts, we assign a subsassignment h_φ of $h[m,n,r]$, which associates to K in F_k the subset $h_\varphi(K)$ of $h[m,n,r](K)$ consisting of all points satisfying φ. We call such subassignments definable subassignements. We denote by Def_k the category whose objects are definable subassignments of some $h[m,n,r]$, morphisms in Def_k being morphisms of subassignments $f : h \to h'$ with h and h' definable subassignments of $h[m,n,r]$ and $h[m',n',r']$ respectively such that the graph of f is a definable subassignment. Note that $h[0,0,0]$ is the final object in this category.

If S is an object of Def_k, we denote by Def_S the category of morphisms $X \to S$ in Def_k. If $f : X \to S$ and $g : Y \to S$ are in Def_S, we write $X \times_S Y$ for the product in Def_S defined as $K \mapsto \{(x,y) \in X(K) \times Y(K) | f(x) = g(y)\}$, with the natural morphism to S. When $S = h[0,0,0)]$, we write $X \times Y$ for $X \times_S Y$. We write $S[m,n,r]$ for $S \times h[m,n,r]$, hence, $S[m,n,r](K) = S(K) \times K((t))^m \times K^n \times \mathbf{Z}^r$. By a point x of S we mean a pair (x_0, K) with K in F_k and x_0 a point of $S(K)$. We denote by $|S|$ the set of points of S. For such x we then set $k(x) = K$. Consider a morphism $f : X \to S$, with X and S

respectively definable subassignments of $h[m,n,r]$ and $h[m',n',r']$. Let $\varphi(x,s)$ be a formula defining the graph of f in $h[m+m',n+n',r+r']$. Fix a point (s_0,K) of S. The formula $\varphi(x,s_0)$ defines a subassignment in Def_K. In this way we get for s a point of S a functor "fiber at s" $i_s^* : \mathrm{Def}_S \to \mathrm{Def}_{k(s)}$.

3.2 Constructible motivic functions

In this subsection we define, for S in Def_k, the ring $\mathcal{C}(S)$ of constructible motivic functions on S. The main goal of this construction is that, as we will see in section 4, motivic integrals with parameters in S are constructible motivic functions on S. In fact, in the construction of a measure, as we all know since studying Lebesgue integration, positive functions often play a basic fundamental role. This the reason why we also introduce the semiring $\mathcal{C}_+(S)$ of positive[4] constructible motivic functions. A technical novelty occurs here: $\mathcal{C}(S)$ is the ring associated to the semiring $\mathcal{C}_+(S)$, but the canonical morphism $\mathcal{C}_+(S) \to \mathcal{C}(S)$ has in general no reason to be injective.

Basically, $\mathcal{C}_+(S)$ and $\mathcal{C}(S)$ are built up from two kinds of functions. The first type consists of elements of a certain Grothendieck (semi)ring. Recall that in "classical" motivic integration as developed in [12], the Grothendieck ring $K_0(\mathrm{Var}_k)$ of algebraic varieties over k plays a key role. In the present setting the analogue of the category of algebraic varieties over k is the category of definable subassignments of $h[0,n,0]$, for some n, when $S = h[0,0,0]$. Hence, for a general S in Def_k, it is natural to consider the subcategory RDef_S of Def_S whose objects are definable subassignments Z of $S \times h[0,n,0]$, for variable n, the morphism $Z \to S$ being induced by the projection on S. The Grothendieck semigroup $SK_0(\mathrm{RDef}_S)$ is the quotient of the free semigroup on isomorphism classes of objects $[Z \to S]$ in RDef_S by relations $[\emptyset \to S] = 0$ and $[(Y \cup Y') \to S] + [(Y \cap Y') \to S] = [Y \to S] + [Y' \to S]$. We also denote by $K_0(\mathrm{RDef}_S)$ the corresponding abelian group. Cartesian product induces a unique semiring structure on $SK_0(\mathrm{RDef}_S)$, resp. ring structure on $K_0(\mathrm{RDef}_S)$.

There are some easy functorialities. For every morphism $f : S \to S'$, there is a natural pullback by $f^* : SK_0(\mathrm{RDef}_{S'}) \to SK_0(\mathrm{RDef}_S)$ induced by the fiber product. If $f : S \to S'$ is a morphism in $\mathrm{RDef}_{S'}$, composition with f induces a morphism $f_! : SK_0(\mathrm{RDef}_S) \to SK_0(\mathrm{RDef}_{S'})$. Similar constructions apply to K_0. That one can view elements of $SK_0(\mathrm{RDef}_S)$ as functions on S (which we even would like to integrate), is illustrated in section 6 on p-adic integration and in the introduction of [1], in the part on integration against Euler characteristic over the reals.

The second type of functions are certain functions with values in the ring

$$A = \mathbf{Z}\Big[\mathbf{L}, \mathbf{L}^{-1}, \Big(\frac{1}{1-\mathbf{L}^{-i}}\Big)_{i>0}\Big], \tag{7}$$

[4] Or maybe better, non-negative.

where, for the moment, $\mathbf{L}$ is just considered as a symbol. Note that a definable morphism $\alpha : S \to h[0,0,1]$ determines a function $|S| \to \mathbf{Z}$, also written α, and a function $|S| \to A$ sending x to $\mathbf{L}^{\alpha(x)}$, written $\mathbf{L}^{\alpha}$. We consider the subring $\mathcal{P}(S)$ of the ring of functions $|S| \to A$ generated by constants in A and by all functions α and $\mathbf{L}^{\alpha}$ with $\alpha : S \to \mathbf{Z}$ definable morphisms. Now we should define positive functions with values in A. For every real number $q > 1$, let us denote by $\vartheta_q : A \to \mathbf{R}$ the morphism sending $\mathbf{L}$ to q. We consider the subsemigroup A_+ of A consisting of elements a such that $\vartheta_q(a) \geq 0$ for all $q > 1$ and we define $\mathcal{P}_+(S)$ as the semiring of functions in $\mathcal{P}(S)$ taking their values in A_+.

Now we explain how to put together these two types of functions. For Y a definable subassignment of S, we denote by $\mathbf{1}_Y$ the function in $\mathcal{P}(S)$ taking the value 1 on Y and 0 outside Y. We consider the subring $\mathcal{P}^0(S)$ of $\mathcal{P}(S)$, resp. the subsemiring $\mathcal{P}^0_+(S)$ of $\mathcal{P}_+(S)$, generated by functions of the form $\mathbf{1}_Y$ with Y a definable subassignment of S, and by the constant function $\mathbf{L}-1$. We have canonical morphisms $\mathcal{P}^0(S) \to K_0(\mathrm{RDef}_S)$ and $\mathcal{P}^0_+(S) \to SK_0(\mathrm{RDef}_S)$ sending $\mathbf{1}_Y$ to $[Y \to S]$ and $\mathbf{L}-1$ to the class of $S \times (h[0,1,0] \setminus \{0\})$ in $K_0(\mathrm{RDef}_S)$ and in $SK_0(\mathrm{RDef}_S)$, respectively. To simplify notation we shall denote by $\mathbf{L}$ and $\mathbf{L}-1$ the class of $S[0,1,0]$ and $S \times (h[0,1,0] \setminus \{0\})$ in $K_0(\mathrm{RDef}_S)$ and in $SK_0(\mathrm{RDef}_S)$.

We may now define the semiring of positive constructible functions as

$$\mathcal{C}_+(S) = SK_0(\mathrm{RDef}_S) \otimes_{\mathcal{P}^0_+(S)} \mathcal{P}_+(S) \tag{8}$$

and the ring of constructible functions as

$$\mathcal{C}(S) = K_0(\mathrm{RDef}_S) \otimes_{\mathcal{P}^0(S)} \mathcal{P}(S). \tag{9}$$

If $f : S \to S'$ is a morphism in Def_k, one shows in [1] that the morphism f^* may naturally be extended to a morphism

$$f^* : \mathcal{C}_+(S') \longrightarrow \mathcal{C}_+(S). \tag{10}$$

If, furthermore, f is a morphism in $\mathrm{RDef}_{S'}$, one shows that the morphism $f_!$ may naturally be extended to

$$f_! : \mathcal{C}_+(S) \longrightarrow \mathcal{C}_+(S'). \tag{11}$$

Similar functorialities exist for $\mathcal{C}$.

3.3 Constructible motivic "Functions"

In fact, we shall need to consider not only functions as we just defined, but functions defined almost everywhere in a given dimension, that we call Functions. (Note the capital in Functions.)

We start by defining a good notion of dimension for objects of Def_k. Heuristically, that dimension corresponds to counting the dimension only in the valued field variables, without taking in account the remaining variables. More

precisely, to any algebraic subvariety Z of $\mathbf{A}^m_{k((t))}$ we assign the definable subassignment h_Z of $h[m,0,0]$ given by $h_Z(K) = Z(K((t)))$. The Zariski closure of a subassignment S of $h[m,0,0]$ is the intersection W of all algebraic subvarieties Z of $\mathbf{A}^m_{k((t))}$ such that $S \subset h_Z$. We define the dimension of S as $\dim S := \dim W$. In the general case, when S is a subassignment of $h[m,n,r]$, we define $\dim S$ as the dimension of the image of S under the projection $h[m,n,r] \to h[m,0,0]$.

One can prove, using Theorem 2.9 and results of van den Dries [15], the following result, which is by no means obvious:

Proposition 3.4. *Two isomorphic objects of* Def_k *have the same dimension.*

For every non-negative integer d, we denote by $\mathcal{C}_+^{\leq d}(S)$ the ideal of $\mathcal{C}_+(S)$ generated by functions $\mathbf{1}_Z$ with Z definable subassignments of S with $\dim Z \leq d$. We set $C_+(S) = \oplus_d C_+^d(S)$ with $C_+^d(S) := \mathcal{C}_+^{\leq d}(S)/\mathcal{C}_+^{\leq d-1}(S)$. It is a graded abelian semigroup, and also a $\mathcal{C}_+(S)$-semimodule. Elements of $C_+(S)$ are called positive constructible Functions on S. If φ is a function lying in $\mathcal{C}_+^{\leq d}(S)$ but not in $\mathcal{C}_+^{\leq d-1}(S)$, we denote by $[\varphi]$ its image in $C_+^d(S)$. One defines similarly $C(S)$ from $\mathcal{C}(S)$.

One of the reasons why we consider functions which are defined almost everywhere originates in the differentiation of functions with respect to the valued field variables: one may show that a definable function $c : S \subset h[m,n,r] \to h[1,0,0]$ is differentiable (in fact even analytic) outside a definable subassignment of S of dimension $< \dim S$. In particular, if $f : S \to S'$ is an isomorphism in Def_k, one may define a function $\mathrm{ordjac}f$, the order of the jacobian of f, which is defined almost everywhere and is equal almost everywhere to a definable function, so we may define $\mathbf{L}^{-\mathrm{ordjac}f}$ in $C_+^d(S)$ when S is of dimension d. In Section 5.2, we shall define $\mathbf{L}^{-\mathrm{ordjac}f}$ using differential forms.

4 Construction of the general motivic measure

Let k be a field of characteristic zero. Given S in Def_k, we define S-integrable Functions and construct pushforward morphisms for these:

Theorem 4.1. *Let k be a field of characteristic zero and let S be in* Def_k. *There exists a unique functor $Z \mapsto \mathrm{I}_S C_+(Z)$ from* Def_S *to the category of abelian semigroups, the functor of S-integrable Functions, assigning to every morphism $f : Z \to Y$ in* Def_S *a morphism $f_! : \mathrm{I}_S C_+(Z) \to \mathrm{I}_S C_+(Y)$ such that for every Z in* Def_S, $\mathrm{I}_S C_+(Z)$ *is a graded subsemigroup of $C_+(Z)$ and $\mathrm{I}_S C_+(S) = C_+(S)$, satisfying the following list of axioms (A1)-(A8).*

(A1a) **(Naturality)**
If $S \to S'$ is a morphism in Def_k and Z is an object in Def_S, then any S'-integrable Function φ in $C_+(Z)$ is S-integrable and $f_!(\varphi)$ is the same, considered in $\mathrm{I}_{S'}$ or in I_S.

(A1b) **(Fubini)**
A positive Function φ on Z is S-integrable if and only if it is Y-integrable and $f_!(\varphi)$ is S-integrable.

(A2) **(Disjoint union)**
If Z is the disjoint union of two definable subassignments Z_1 and Z_2, then the isomorphism $C_+(Z) \simeq C_+(Z_1) \oplus C_+(Z_2)$ induces an isomorphism $\mathrm{I}_S C_+(Z) \simeq \mathrm{I}_S C_+(Z_1) \oplus \mathrm{I}_S C_+(Z_2)$, under which $f_! = f_{|Z_1!} \oplus f_{|Z_2!}$.

(A3) **(Projection formula)**
For every α in $\mathcal{C}_+(Y)$ and every β in $\mathrm{I}_S C_+(Z)$, $\alpha f_!(\beta)$ is S-integrable if and only if $f^*(\alpha)\beta$ is, and then $f_!(f^*(\alpha)\beta) = \alpha f_!(\beta)$.

(A4) **(Inclusions)**
If $i : Z \hookrightarrow Z'$ is the inclusion of definable subassignments of the same object of Def_S, $i_!$ is induced by extension by zero outside Z and sends injectively $\mathrm{I}_S C_+(Z)$ to $\mathrm{I}_S C_+(Z')$.

(A5) **(Integration along residue field variables)**
Let Y be an object of Def_S and denote by π the projection $Y[0,n,0] \to Y$. A Function $[\varphi]$ in $C_+(Y[0,n,0])$ is S-integrable if and only if, with notation of Equation 11, $[\pi_!(\varphi)]$ is S-integrable and then $\pi_!([\varphi]) = [\pi_!(\varphi)]$.

Basically this axiom means that integrating with respect to variables in the residue field just amounts to taking the pushforward induced by composition at the level of Grothendieck semirings.

(A6) **(Integration along Z-variables)** Basically, integration along the **Z**-variables corresponds to summing over the integers, but to state precisely (A6), we need to perform some preliminary constructions.

Consider a function φ in $\mathcal{P}(S[0,0,r])$, hence φ is a function $|S| \times \mathbf{Z}^r \to A$. We shall say φ is S-integrable if for every $q > 1$ and every x in $|S|$, the series $\sum_{i \in \mathbf{Z}^r} \vartheta_q(\varphi(x,i))$ is summable. One proves that if φ is S-integrable, there exists a unique function $\mu_S(\varphi)$ in $\mathcal{P}(S)$ such that $\vartheta_q(\mu_S(\varphi)(x))$ is equal to the sum of the previous series for all $q > 1$ and all x in $|S|$. We denote by $\mathrm{I}_S \mathcal{P}_+(S[0,0,r])$ the set of S-integrable functions in $\mathcal{P}_+(S[0,0,r])$ and we set

$$\mathrm{I}_S \mathcal{C}_+(S[0,0,r]) = \mathcal{C}_+(S) \otimes_{\mathcal{P}_+(S)} \mathrm{I}_S \mathcal{P}_+(S[0,0,r]). \tag{12}$$

Hence $\mathrm{I}_S \mathcal{P}_+(S[0,0,r])$ is a sub-$\mathcal{C}_+(S)$-semimodule of $\mathcal{C}_+(S[0,0,r])$ and μ_S may be extended by tensoring to

$$\mu_S : \mathrm{I}_S \mathcal{C}_+(S[0,0,r]) \to \mathcal{C}_+(S). \tag{13}$$

Now we can state (A6):
Let Y be an object of Def_S and denote by π the projection $Y[0,0,r] \to Y$. A Function $[\varphi]$ in $C_+(Y[0,0,r])$ is S-integrable if and only if there exists φ' in

$\mathcal{C}_+(Y[0,0,r])$ with $[\varphi'] = [\varphi]$ which is Y-integrable in the previous sense and such that $[\mu_Y(\varphi')]$ is S-integrable. We then have $\pi_!([\varphi]) = [\mu_Y(\varphi')]$.

(A7) **(Volume of balls)** It is natural to require (by analogy with the p-adic case) that the volume of a ball $\{z \in h[1,0,0] | \mathrm{ac}(z-c) = \alpha, \mathrm{ac}(z-c) = \xi\}$, with α in $\mathbf{Z}$, c in $k((t))$ and ξ non-zero in k, should be $\mathbf{L}^{-\alpha-1}$. (A7) is a relative version of that statement:
Let Y be an object in Def_S and let Z be the definable subassignment of $Y[1,0,0]$ defined by $\mathrm{ord}\,(z - c(y)) = \alpha(y)$ and $\mathrm{ac}(z - c(y)) = \xi(y)$, with z the coordinate on the $\mathbf{A}^1_{k((t))}$-factor and α, ξ, c definable functions on Y with values respectively in $\mathbf{Z}$, $h[0,1,0] \setminus \{0\}$, and $h[1,0,0]$. We denote by $f : Z \to Y$ the morphism induced by projection. Then $[\mathbf{1}_Z]$ is S-integrable if and only if $\mathbf{L}^{-\alpha-1}[\mathbf{1}_Y]$ is, and then $f_!([\mathbf{1}_Z]) = \mathbf{L}^{-\alpha-1}[\mathbf{1}_Y]$.

(A8) **(Graphs)** This last axiom expresses the pushforward for graph projections. It relates volume and differentials and is a special case of the change of variables Theorem 4.2.
Let Y be in Def_S and let Z be the definable subassignment of $Y[1,0,0]$ defined by $z - c(y) = 0$ with z the coordinate on the $\mathbf{A}^1_{k((t))}$-factor and c a morphism $Y \to h[1,0,0]$. We denote by $f : Z \to Y$ the morphism induced by projection. Then $[\mathbf{1}_Z]$ is S-integrable if and only if $\mathbf{L}^{(\mathrm{ordjac}f)\circ f^{-1}}$ is, and then $f_!([\mathbf{1}_Z]) = \mathbf{L}^{(\mathrm{ordjac}f)\circ f^{-1}}$.

Once Theorem 4.1 is proved, one may proceed as follows to extend the constructions from C_+ to C . One defines $\mathrm{I}_S C(Z)$ as the subgroup of $C(Z)$ generated by the image of $\mathrm{I}_S C_+(Z)$. One shows that if $f : Z \to Y$ is a morphism in Def_S, the morphism $f_! : \mathrm{I}_S C_+(Z) \to \mathrm{I}_S C_+(Y)$ has a natural extension $f_! : \mathrm{I}_S C(Z) \to \mathrm{I}_S C(Y)$.

The relation of Theorem 4.1 with motivic integration is the following. When S is equal to $h[0,0,0]$, the final object of Def_k, one writes $\mathrm{I}C_+(Z)$ for $\mathrm{I}_S C_+(Z)$ and we shall say integrable for S-integrable, and similarly for C. Note that $\mathrm{I}C_+(h[0,0,0]) = C_+(h[0,0,0]) = SK_0(\mathrm{RDef}_k) \otimes_{\mathbf{N}[\mathbf{L}-1]} A_+$ and that $\mathrm{I}C(h[0,0,0]) = K_0(\mathrm{RDef}_k) \otimes_{\mathbf{Z}[\mathbf{L}]} A$. For φ in $\mathrm{I}C_+(Z)$, or in $\mathrm{I}C(Z)$, one defines the motivic integral $\mu(\varphi)$ by $\mu(\varphi) = f_!(\varphi)$ with f the morphism $Z \to h[0,0,0]$. Working in the more general framework of Theorem 4.1 to construct μ appears to be very convenient for inductions occuring in the proofs. Also, it is not clear how to characterize μ alone by existence and unicity properties. Note also that one reason for the statement of Theorem 4.1 to look somewhat cumbersome, is that we have to define at once the notion of integrability and the value of the integral.

The proof of Theorem 4.1 is quite long and involved. In a nutshell, the basic idea is the following. Integration along residue field variables is controlled by (A5) and integration along $\mathbf{Z}$-variables by (A6). Integration along valued field variables is constructed one variable after the other. To integrate with

respect to one valued field variable, one may, using (a variant of) the cell decomposition Theorem 2.9 (at the cost of introducing additional new residue field and **Z**-variables), reduce to the case of cells which is covered by (A7) and (A8). An important step is to show that this is independent of the choice of a cell decomposition. When one integrates with respect to more than one valued field variable (one after the other) it is crucial to show that it is independent of the order of the variables, for which we use a notion of bicells.

In this new framework, we have the following general form of the change of variables theorem, generalizing the corresponding statements in [12] and [13].

Theorem 4.2. *Let $f : X \to Y$ be an isomorphism between definable subassignments of dimension d. For every function φ in $\mathcal{C}_+^{\leq d}(Y)$ having a non-zero class in $C_+^d(Y)$, $[f^*(\varphi)]$ is Y-integrable and $f_![f^*(\varphi)] = \mathbf{L}^{(\mathrm{ordjac}f)\circ f^{-1}}[\varphi]$. A similar statement holds in C.*

4.3 Integrals depending on parameters

One pleasant feature of Theorem 4.1 is that it generalizes readily to the relative setting of integrals depending on parameters.

Indeed, let us fix Λ in Def_k playing the role of a parameter space. For S in Def_Λ, we consider the ideal $\mathcal{C}^{\leq d}(S \to \Lambda)$ of $\mathcal{C}_+(S)$ generated by functions $\mathbf{1}_Z$ with Z definable subassignment of S such that all fibers of $Z \to \Lambda$ are of dimension $\leq d$. We set

$$C_+(S \to \Lambda) = \bigoplus_d C_+^d(S \to \Lambda) \tag{14}$$

with

$$C_+^d(S \to \Lambda) := \mathcal{C}_+^{\leq d}(S \to \Lambda)/\mathcal{C}_+^{\leq d-1}(S \to \Lambda). \tag{15}$$

It is a graded abelian semigroup (and also a $\mathcal{C}_+(S)$-semimodule). If φ belongs to $\mathcal{C}_+^{\leq d}(S \to \Lambda)$ but not to $\mathcal{C}_+^{\leq d-1}(S \to \Lambda)$, we write $[\varphi]$ for its image in $C_+^d(S \to \Lambda)$. The following relative analogue of Theorem 4.1 holds.

Theorem 4.4. *Let k be a field of characteristic zero, let Λ be in* Def_k*, and let S be in* Def_Λ*. There exists a unique functor $Z \mapsto \mathrm{I}_S C_+(Z \to \Lambda)$ from* Def_S *to the category of abelian semigroups, assigning to every morphism $f : Z \to Y$ in* Def_S *a morphism $f_{!\Lambda} : \mathrm{I}_S C_+(Z \to \Lambda)) \to \mathrm{I}_S C_+(Y \to \Lambda))$ satisfying properties analogous to (A0)-(A8) obtained by replacing $C_+(_)$ by $C_+(_ \to \Lambda)$ and* ordjac *by its relative analogue* ordjac_Λ[5]*.*

Note that $C_+(\Lambda \to \Lambda) = \mathcal{C}_+(\Lambda)$ (and also $\mathrm{I}_\Lambda C_+(\Lambda \to \Lambda) = C_+(\Lambda \to \Lambda)$. Hence, given $f : Z \to \Lambda$ in Def_Λ, we may define the relative motivic measure with respect to Λ as the morphism

[5]Defined similarly as ordjac, but using relative differential forms.

$$\mu_\Lambda := f_{!\Lambda} : \mathrm{I}_\Lambda C_+(Z \to \Lambda) \longrightarrow \mathcal{C}_+(\Lambda). \tag{16}$$

By the following statement, μ_Λ indeed corresponds to integration along the fibers over Λ:

Proposition 4.5. *Let φ be a Function in $C_+(Z \to \Lambda)$. It belongs to $\mathrm{I}_\Lambda C_+(Z \to \Lambda)$ if and only if for every point λ in Λ, the restriction φ_λ of φ to the fiber of Z at λ is integrable. The motivic integral of φ_λ is then equal to $i_\lambda^*(\mu_\Lambda(\varphi))$, for every λ in Λ.*

Similarly as in the absolute case, one can also define the relative analogue $C(S \to \Lambda)$ of $C(S)$, and extend the notion of integrability and the construction of $f_{!\Lambda}$ to this setting.

5 Motivic integration in a global setting and comparison with previous constructions

5.1 Definable subassignments on varieties

Objects of Def_k are by construction affine, being subassignments of functors $h[m,n,r] : F_k \to \mathrm{Ens}$ given by $K \mapsto K((t))^m \times K^n \times \mathbf{Z}^r$. We shall now consider their global analogues and extend the previous constructions to the global setting.

Let $\mathcal{X}$ be a variety over $k((t))$, that is, a reduced and separated scheme of finite type over $k((t))$, and let X be a variety over k. For r an integer ≥ 0, we denote by $h[\mathcal{X}, X, r]$ the functor $F_k \to \mathrm{Ens}$ given by $K \mapsto \mathcal{X}(K((t))) \times X(K) \times \mathbf{Z}^r$. When $X = \operatorname{Spec} k$ and $r = 0$, we write $h[\mathcal{X}]$ for $h[\mathcal{X}, X, r]$. If $\mathcal{X}$ and X are affine and if $i : \mathcal{X} \hookrightarrow \mathbf{A}^m_{k((t))}$ and $j : X \hookrightarrow \mathbf{A}^n_k$ are closed immersions, we say a subassignment h of $h[\mathcal{X}, X, r]$ is definable if its image by the morphism $h[\mathcal{X}, X, r] \to h[m,n,r]$ induced by i and j is a definable subassignment of $h[m,n,r]$. This definition does not depend on i and j. More generally, we shall say a subassignment h of $h[\mathcal{X}, X, r]$ is definable if there exist coverings $(\mathcal{U}_i)$ and (U_j) of $\mathcal{X}$ and X by affine open subsets such that $h \cap h[\mathcal{U}_i, U_j, r]$ is a definable subassignment of $h[\mathcal{U}_i, U_j, r]$ for every i and j. We get in this way a category GDef_k whose objects are definable subassignments of some $h[\mathcal{X}, X, r]$, morphisms being definable morphisms, that is, morphisms whose graphs are definable subassignments.

The category Def_k is a full subcategory of GDef_k. Dimension as defined in Section 3.3 may be directly generalized to objects of GDef_k and Proposition 3.4 still holds in GDef_k. Also, if S is an object in GDef_k, our definitions of RDef_S, $\mathcal{C}_+(S)$, $\mathcal{C}(S)$, $C_+(S)$ and $C(S)$ extend.

5.2 Definable differential forms and volume forms

In the global setting, one does not integrate functions anymore, but volume forms. Let us start by introducing differential forms in the definable framework. Let h be a definable subassignment of some $h[\mathcal{X}, X, r]$. We denote by $\mathcal{A}(h)$ the ring of definable morphisms $h \to h[\mathbf{A}^1_{k((t))}]$. Let us define, for i in $\mathbf{N}$, the $\mathcal{A}(h)$-module $\Omega^i(h)$ of definable i-forms on h. Let $\mathcal{Y}$ be the closed subset of $\mathcal{X}$, which is the Zariski closure of the image of h under the projection $\pi : h[\mathcal{X}, X, r] \to h[\mathcal{X}]$. We denote by $\Omega^i_{\mathcal{Y}}$ the sheaf of algebraic i-forms on $\mathcal{Y}$, by $\mathcal{A}_{\mathcal{Y}}$ the Zariski sheaf associated to the presheaf $U \mapsto \mathcal{A}(h[U])$ on $\mathcal{Y}$, and by $\Omega^i_{h[\mathcal{Y}]}$ the sheaf $\mathcal{A}_{\mathcal{Y}} \otimes_{\mathcal{O}_{\mathcal{Y}}} \Omega^i_{\mathcal{Y}}$. We set

$$\Omega^i(h) := \mathcal{A}(h) \otimes_{\mathcal{A}(h[\mathcal{Y}])} \Omega^i_{h[\mathcal{Y}]}(\mathcal{Y}), \tag{17}$$

the $\mathcal{A}(h[\mathcal{Y}])$-algebra structure on $\mathcal{A}(h)$ given by composition with π.

We now assume h is of dimension d. We denote by $\mathcal{A}^<(h)$ the ideal of functions in $\mathcal{A}(h)$ that are zero outside a definable subassignment of dimension $< d$. There is a canonical morphism of abelian semi-groups $\lambda : \mathcal{A}(h)/\mathcal{A}^<(h) \to C^d_+(h)$ sending the class of a function f to the class of $\mathbf{L}^{-\operatorname{ord} f}$, with the convention $\mathbf{L}^{-\operatorname{ord} 0} = 0$. We set $\tilde{\Omega}^d(h) = \mathcal{A}(h)/\mathcal{A}^<(h) \otimes_{\mathcal{A}(h)} \Omega^d(h)$, and we define the set $|\tilde{\Omega}|_+(h)$ of definable positive volume forms as the quotient of the free abelian semigroup on symbols (ω, g) with ω in $\tilde{\Omega}^d(h)$ and g in $C^d_+(h)$ by relations $(f\omega, g) = (\omega, \lambda(f)g)$, $(\omega, g + g') = (\omega, g) + (\omega, g')$ and $(\omega, 0) = 0$, for f in $\mathcal{A}(h)/\mathcal{A}^<(h)$. We write $g|\omega|$ for the class (ω, g), in order to have $g|f\omega| = g\mathbf{L}^{-\operatorname{ord} f}|\omega|$. The $\mathcal{C}_+(h)$-semimodule structure on $C^d_+(h)$ induces after passing to the quotient a structure of semiring on $C^d_+(h)$ and $|\tilde{\Omega}|_+(h)$ is naturally endowed with a structure of $C^d_+(h)$-semimodule. We shall call an element $|\omega|$ in $|\tilde{\Omega}|_+(h)$ a gauge form if it is a generator of that semimodule. One should note that in the present setting gauge forms always exist, which is certainly not the case in the usual framework of algebraic geometry. Indeed, gauge forms always exist locally (that is, in suitable affine charts), and in our definable world there is no difficulty in gluing local gauge forms to global ones. One may define similarly $|\tilde{\Omega}|(h)$, replacing C^d_+ by C^d, but we shall only consider $|\tilde{\Omega}|_+(h)$ here.

If h is a definable subassignment of dimension d of $h[m, n, r]$, one may construct, similarly as Serre [23] in the p-adic case, a canonical gauge form $|\omega_0|_h$ on h. Let us denote by $x_1, \ldots, x_m$ the coordinates on $\mathbf{A}^m_{k((t))}$ and consider the d-forms $\omega_I := dx_{i_1} \wedge \cdots \wedge dx_{i_d}$ for $I = \{i_1, \ldots, i_d\} \subset \{1, \ldots, m\}$, $i_1 < \cdots < i_d$, and their image $|\omega_I|_h$ in $|\tilde{\Omega}|_+(h)$. One may check there exists a unique element $|\omega_0|_h$ of $|\tilde{\Omega}|_+(h)$, such that, for every I, there exists definable functions with integral values α_I, β_I on h, with β_I only taking as values 1 and 0, such that $\alpha_I + \beta_I > 0$ on h, $|\omega_I|_h = \beta_I \mathbf{L}^{-\alpha_I}|\omega_0|_h$ in $|\tilde{\Omega}|_+(h)$, and such that $\inf_I \alpha_I = 0$.

If $f : h \to h'$ is a morphism in GDef_k with h and h' of dimension d and all fibers of dimension 0, there is a mapping $f^* : |\tilde{\Omega}|_+(h') \to |\tilde{\Omega}|_+(h)$ induced by

pullback of differential forms. This follows from the fact that f is "analytic" outside a definable subassignment of dimension $d-1$ of h. If, furthermore, h and h' are objects in Def_k, one defines $\mathbf{L}^{-\mathrm{ordjac}f}$ by

$$f^*|\omega_0|_{h'} = \mathbf{L}^{-\mathrm{ordjac}f}|\omega_0|_h. \tag{18}$$

If $\mathcal{X}$ is a $k((t))$-variety of dimension d, and $\mathcal{X}^0$ is a $k[[t]]$-model of $\mathcal{X}$, it is possible to define an element $|\omega_0|$ in $|\tilde{\Omega}|_+(h[\mathcal{X}])$, which depends only on $\mathcal{X}^0$, and which is characterized by the following property: for every open U^0 of $\mathcal{X}^0$ on which the $k[[t]]$-module $\Omega^d_{U^0|k[[t]]}(U^0)$ is generated by a non-zero form ω, $|\omega_0|_{|h[U^0\otimes \mathrm{Spec}\, k((t))]} = |\omega|$ in $|\tilde{\Omega}|_+(h[U^0 \otimes \mathrm{Spec}\, k((t))])$.

5.3 Integration of volume forms and Fubini's Theorem

Now we are ready to construct motivic integration for volume forms. In the affine case, using canonical gauge forms, one may pass from volume forms to Functions in top dimension, and vice versa. More precisely, let $f : S \to S'$ be a morphism in Def_k, with S of dimension s and S' of dimension s'. Every positive form α in $|\tilde{\Omega}|_+(S)$ may be written $\alpha = \psi_\alpha|\omega_0|_S$ with ψ_α in $C^s_+(S)$. We shall say α is f-integrable if ψ_α is f-integrable and we then set

$$f^{\mathrm{top}}_!(\alpha) := \{f_!(\psi_\alpha)\}_{s'}|\omega_0|_{S'}, \tag{19}$$

$\{f_!(\psi_\alpha)\}_{s'}$ denoting the component of $f_!(\psi_\alpha)$ lying in $C^{s'}_+(S')$.

Consider now a morphism $f : S \to S'$ in GDef_k. The previous construction may be globalized as follows. Assume there exist isomorphisms $\varphi : T \to S$ and $\varphi' : T' \to S'$ with T and T' in Def_k. We denote by $\tilde{f}$ the morphism $T \to T'$ such that $\varphi' \circ \tilde{f} = f \circ \varphi$. We shall say α in $|\tilde{\Omega}|_+(S)$ is f-integrable if $\varphi^*(\alpha)$ is $\tilde{f}$-integrable and we define then $f^{\mathrm{top}}_!(\alpha)$ by the relation

$$\tilde{f}^{\mathrm{top}}_!(\varphi^*(\alpha)) = \varphi'^*(f^{\mathrm{top}}_!(\alpha)). \tag{20}$$

It follows from Theorem 4.2 that this definition is independent of the choice of the isomorphisms φ and φ'. By additivity, using affine charts, the previous construction may be extended to any morphism $f : S \to S'$ in GDef_k, in order to define the notion of f-integrability for a volume form α in $|\tilde{\Omega}|_+(S)$, and also, when α is f-integrable, the fiber integral $f^{\mathrm{top}}_!(\alpha)$, which belongs to $|\tilde{\Omega}|_+(S')$. When $S = h[0,0,0]$, we shall say integrable instead of f-integrable, and we shall write $\int_S \alpha$ for $f^{\mathrm{top}}_!(\alpha)$.

In this framework, one may deduce from (A1b) in Theorem 4.1 the following general form of Fubini's Theorem for motivic integration:

Theorem 5.4 (Fubini's Theorem). *Let $f : S \to S'$ be a morphism in GDef_k. Assume S is of dimension s, S' is of dimension s', and that the fibers of f are all of dimension $s - s'$. A positive volume form α in $|\tilde{\Omega}|_+(S)$ is integrable if and only if it is f-integrable and $f^{\mathrm{top}}_!(\alpha)$ is integrable. When this holds, then*

$$\int_S \alpha = \int_{S'} f^{\mathrm{top}}_!(\alpha). \tag{21}$$

5.5 Comparison with classical motivic integration

In the definition of Def_k, RDef_k and GDef_k, instead of considering the category F_k of all fields containing k, one could as well restrict to the subcategory ACF_k of algebraically closed fields containing k and define categories $\mathrm{Def}_{k,\mathrm{ACF}_k}$, etc. In fact, it is a direct consequence of Chevalley's constructibility theorem that $K_0(\mathrm{RDef}_{k,\mathrm{ACF}_k})$ is nothing else than the Grothendieck ring $K_0(\mathrm{Var}_k)$ considered in [12]. It follows that there is a canonical morphism $SK_0(\mathrm{RDef}_k) \to K_0(\mathrm{Var}_k)$ sending $\mathbf{L}$ to the class of $\mathbf{A}^1_k$, which we shall still denote by $\mathbf{L}$. One can extend this morphism to a morphism $\gamma : SK_0(\mathrm{RDef}_k) \otimes_{\mathbf{N}[\mathbf{L}-1]} A_+ \to K_0(\mathrm{Var}_k) \otimes_{\mathbf{Z}[\mathbf{L}]} A$. By considering the series expansion of $(1-\mathbf{L}^{-i})^{-1}$, one defines a canonical morphism $\delta : K_0(\mathrm{Var}_k) \otimes_{\mathbf{Z}[\mathbf{L}]} A \to \widehat{\mathcal{M}}$, with $\widehat{\mathcal{M}}$ the completion of $K_0(\mathrm{Var}_k)[\mathbf{L}^{-1}]$ considered in [12].

Let X be an algebraic variety over k of dimension d. Set $\mathcal{X}^0 := X \otimes_{\mathrm{Spec}\, k} \mathrm{Spec}\, k[[t]]$ and $\mathcal{X} := \mathcal{X}^0 \otimes_{\mathrm{Spec}\, k[[t]]} \mathrm{Spec}\, k((t))$. Consider a definable subassignment W of $h[\mathcal{X}]$ in the language $\mathcal{L}_{\mathrm{DP,P}}$, with the restriction that constants in the valued field sort that appear in formulas defining W in affine charts defined over k belong to k (and not to $k((t))$). We assume $W(K) \subset \mathcal{X}(K[[t]])$ for every K in F_k. With the notation of [12], formulas defining W in affine charts define a semialgebraic subset of the arc space $\mathcal{L}(X)$ in the corresponding chart, by Theorem 2.10 and Chevalley's constructibility theorem. In this way we assign canonically to W a semialgebraic subset $\tilde{W}$ of $\mathcal{L}(X)$. Similarly, let α be a definable function on W taking integral values and satisfying the additional condition that constants in the valued field sort, appearing in formulas defining α can only belong to k. To any such function α we may assign a semialgebraic function $\tilde{\alpha}$ on $\tilde{W}$.

Theorem 5.6. *Under the former hypotheses, $|\omega_0|$ denoting the canonical volume form on $h[\mathcal{X}]$, for every definable function α on W with integral values satisfying the previous conditions and bounded below, $\mathbf{1}_W \mathbf{L}^{-\alpha}|\omega_0|$ is integrable on $h[\mathcal{X}]$ and*

$$(\delta \circ \gamma)\Big(\int_{h[\mathcal{X}]} \mathbf{1}_W \mathbf{L}^{-\alpha}|\omega_0|\Big) = \int_{\tilde{W}} \mathbf{L}^{-\tilde{\alpha}} d\mu', \tag{22}$$

μ' denoting the motivic measure considered in [12].

It follows from Theorem 5.6 that, for semialgebraic sets and functions, the motivic integral constructed in [12] in fact already exists in $K_0(\mathrm{Var}_k) \otimes_{\mathbf{Z}[\mathbf{L}]} A$, or even in $SK_0(\mathrm{Var}_k) \otimes_{\mathbf{N}[\mathbf{L}-1]} A_+$, with $SK_0(\mathrm{Var}_k) = SK_0(\mathrm{RDef}_{k,\mathrm{ACF}_k})$, the Grothendieck semiring of varieties over k.

5.7 Comparison with arithmetic motivic integration

Similarly, instead of ACF_k, we may also consider the category PFF_k of pseudo-finite fields containing k. Let us recall that a pseudo-finite field is a perfect field

F having a unique extension of degree n for every n in a given algebraic closure and such that every geometrically irreducible variety over F has an F-rational point. By restriction from F_k to PFF_k we can define categories $\mathrm{Def}_{k,\mathrm{PFF}_k}$, etc. In particular, the Grothendieck ring $K_0(\mathrm{RDef}_{k,\mathrm{PFF}_k})$ is nothing else but what is denoted by $K_0(\mathrm{PFF}_k)$ in [14] and [11].

In the paper [13], arithmetic motivic integration was taking its values in a certain completion $\hat{K}_0^v(\mathrm{Mot}_{k,\bar{\mathbf{Q}}})_{\mathbf{Q}}$ of a ring $K_0^v(\mathrm{Mot}_{k,\bar{\mathbf{Q}}})_{\mathbf{Q}}$. Somewhat later it was remarked in [14] and [11] that one can restrict to the smaller ring $K_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q}$, the definition of which we shall now recall.

The field k being of characteristic 0, there exists, by [16] and [17], a unique morphism of rings $K_0(\mathrm{Var}_k) \to K_0(\mathrm{CHMot}_k)$ sending the class of a smooth projective variety X over k to the class of its Chow motive. Here $K_0(\mathrm{CHMot}_k)$ denotes the Grothendieck ring of the category of Chow motives over k with rational coefficients. By definition, $K_0^{\mathrm{mot}}(\mathrm{Var}_k)$ is the image of $K_0(\mathrm{Var}_k)$ in $K_0(\mathrm{CHMot}_k)$ under this morphism. [Note that the definition of $K_0^{\mathrm{mot}}(\mathrm{Var}_k)$ given in [14] is not clearly equivalent and should be replaced by the one given above.] In [14] and [11], the authors have constructed, using results from [13], a canonical morphism $\chi_c : K_0(\mathrm{PFF}_k) \to K_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q}$ as follows:

Theorem 5.8 (Denef–Loeser [14] [11]). *Let k be a field of characteristic zero. There exists a unique ring morphism*

$$\chi_c : \mathrm{K}_0(\mathrm{PFF}_k) \longrightarrow \mathrm{K}_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q} \tag{23}$$

satisfying the following two properties:

(i) For any formula φ which is a conjunction of polynomial equations over k, the element $\chi_c([\varphi])$ equals the class in $\mathrm{K}_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q}$ of the variety defined by φ.

(ii) Let X be a normal affine irreducible variety over k, Y an unramified Galois cover of X, that is, Y is an integral étale scheme over X with $Y/G \cong X$, where G is the group of all endomorphisms of Y over X, and C a cyclic subgroup of the Galois group G of Y over X. For such data we denote by $\varphi_{Y,X,C}$ a ring formula whose interpretation, in any field K containing k, is the set of K-rational points on X that lift to a geometric point on Y with decomposition group C (i.e., the set of points on X that lift to a K-rational point of Y/C, but not to any K-rational point of Y/C' with C' a proper subgroup of C). Then

$$\chi_c([\varphi_{Y,X,C}]) = \frac{|C|}{|\mathrm{N}_G(C)|}\chi_c([\varphi_{Y,Y/C,C}]),$$

where $\mathrm{N}_G(C)$ is the normalizer of C in G.

Moreover, when k is a number field, for almost all finite places $\mathfrak{P}$, the number of rational points of $(\chi_c([\varphi]))$ in the residue field $k(\mathfrak{P})$ of k at $\mathfrak{P}$ is equal to the cardinality of $h_\varphi(k(\mathfrak{P}))$.

The construction of χ_c has been recently extended to the relative setting by J. Nicaise [19].

The arithmetical measure takes values in some completion $\hat{K}_0^{\mathrm{mot}}(\mathrm{Var}_k)\otimes\mathbf{Q}$ of the localisation of $K_0^{\mathrm{mot}}(\mathrm{Var}_k)\otimes\mathbf{Q}$ with respect to the class of the affine line. There is a canonical morphism

$$\hat{\gamma} : SK_0(\mathrm{RDef}_k)\otimes_{\mathbf{N}[\mathbf{L}-1]} A_+ \to K_0(\mathrm{PFF}_k)\otimes_{\mathbf{Z}[\mathbf{L}]} A.$$

Considering the series expansion of $(1-\mathbf{L}^{-i})^{-1}$, the map χ_c induces a canonical morphism $\tilde{\delta} : K_0(\mathrm{PFF}_k)\otimes_{\mathbf{Z}[\mathbf{L}]} A \to \hat{K}_0^{\mathrm{mot}}(\mathrm{Var}_k)\otimes\mathbf{Q}$.

Let X be an algebraic variety over k of dimension d. Set

$$\mathcal{X}^0 := X\otimes_{\mathrm{Spec}\,k}\mathrm{Spec}\,k[[t]],\quad \mathcal{X} := \mathcal{X}^0\otimes_{\mathrm{Spec}\,k[[t]]}\mathrm{Spec}\,k((t)),$$

and consider a definable subassignment W of $h[\mathcal{X}]$ satisfying the conditions in Section 5.5. Formulas defining W in affine charts allow us to define, in the terminology and with the notation in [13], a definable subassignment of $h_{\mathcal{L}(X)}$ in the corresponding chart, and we may assign canonically to W a definable subassignment $\tilde{W}$ of $h_{\mathcal{L}(X)}$ in the sense of [13].

Theorem 5.9. *Under the previous hypotheses and with the previous notation, $\mathbf{1}_W|\omega_0|$ is integrable on $h[\mathcal{X}]$ and*

$$(\tilde{\delta}\circ\hat{\gamma})\Big(\int_{h[\mathcal{X}]}\mathbf{1}_W|\omega_0|\Big) = \nu(\tilde{W}), \tag{24}$$

ν denoting the arithmetical motivic measure as defined in [13].

In particular, Theorem 5.9 implies that in the present setting the arithmetical motivic integral constructed in [13] already exists in $K_0(\mathrm{PFF}_k)\otimes_{\mathbf{Z}[\mathbf{L}]}A$ (or even in $SK_0(\mathrm{PFF}_k)\otimes_{\mathbf{N}[\mathbf{L}-1]}A_+$), without completing further the Grothendieck ring and without considering Chow motives (and even without inverting additively all elements of the Grothendieck semiring).

6 Comparison with *p*-adic integration

In the next two sections we present new results on specialization to p-adic integration and Ax–Kochen–Eršov Theorems for integrals with parameters. We plan to give complete details in a future paper.

6.1 *P*-adic definable sets

We fix a finite extension K of $\mathbf{Q}_p$ together with an uniformizing parameter ϖ_K. We denote by R_K the valuation ring and by k_K the residue field, $k_K \simeq \mathbf{F}_{q(K)}$ for some power $q(K)$ of p. Let φ be a formula in the language $\mathcal{L}_{\mathrm{DP,P}}$ with

coefficients in K in the valued field sort and coefficients in k_K in the residue field sort, with m free variables in the valued field sort, n free variables in the residue field sort and r free variables in the value group sort. The formula φ defines a subset Z_φ of $K^m \times k_K^n \times \mathbf{Z}^r$ (recall that since we have chosen ϖ_K, K is endowed with an angular component mapping). We call such a subset a p-adic definable subset of $K^m \times k_K^n \times \mathbf{Z}^r$. We define morphisms between p-adic definable subsets similarly as before: if S and S' are p-adic definable subsets of $K^m \times k_K^n \times \mathbf{Z}^r$ and $K^{m'} \times k_K^{n'} \times \mathbf{Z}^{r'}$ respectively, a morphism $f : S \to S'$ will be a function $f : S \to S'$ whose graph is p-adic definable.

6.2 *P*-adic dimension

By the work of Scowcroft and van den Dries [22], there is a good dimension theory for p-adic definable subsets of K^m. By Theorem 3.4 of [22], a p-adic definable subset A of K^m has dimension d if and only its Zariski closure has dimension d in the sense of algebraic geometry. For S a p-adic definable subset of $K^m \times k_K^n \times \mathbf{Z}^r$, we define the dimension of S as the dimension of its image S' under the projection $\pi : S \to K^m$.

More generally if $f : S \to S'$ is a morphism of p-adic definable subsets, one defines the relative dimension of f to be the maximum of the dimensions of the fibers of f.

6.3 Functions

Let S be a p-adic definable subset of $K^m \times k_K^n \times \mathbf{Z}^r$. We shall consider the $\mathbf{Q}$-algebra $\mathcal{C}_K(S)$ generated by functions of the form α and q^α with α a $\mathbf{Z}$-valued p-adic definable function on S. For $S' \subset S$ a p-adic definable subset, we write $\mathbf{1}_{S'}$ for the characteristic function of S' in $\mathcal{C}_K(S)$.

For $d \geq 0$ an integer, we denote by $\mathcal{C}_K^{\leq d}(S)$ the ideal of $\mathcal{C}(S)$ generated by all functions $\mathbf{1}_{S'}$ with S' a p-adic definable subset of S of dimension $\leq d$. Similarly to what we did before, we set

$$C_K^d(S) := \mathcal{C}_K^{\leq d}(S)/\mathcal{C}_K^{\leq d-1}(S) \quad \text{and} \quad C_K(S) := \bigoplus_d C_K^d(S). \tag{25}$$

Also, similarly as before, we have relative variants of the above definitions. If $f : Z \to S$ is a morphism between p-adic definable subsets, we define $\mathcal{C}_K^{\leq d}(Z \to S)$, $C_K^d(Z \to S)$ and $C_K(Z \to S)$ by replacing dimension by relative dimension.

6.4 *P*-adic measure

Let S be a p-adic definable subset of K^m of dimension d. By the construction of [25] based on [23], bounded p-adic definable subsets A of S have a canonical d-dimensional volume $\mu_K^d(A)$ in $\mathbf{R}$.

Now let S be a p-adic definable subset of $K^m \times k_K^n \times \mathbf{Z}^r$ of dimension d and S' its image under the projection $\pi : S \to K^m$. We define the measure μ_d on S as the measure induced by the product measure on $S' \times k_K^n \times \mathbf{Z}^r$ of the d-dimensional volume μ_K^d on the factor S' and the counting measure on the factor $k_K^n \times \mathbf{Z}^r$. When S is of dimension $< d$ we declare μ_K^d to be identically zero.

We call φ in $\mathcal{C}_K(S)$ integrable on S if φ is integrable against μ_d and we denote the integral by $\mu_K^d(\varphi)$.

One defines $\mathrm{IC}_K^d(S)$ as the abelian subgroup of $C_K^d(S)$ consisting of the classes of integrable functions in $\mathcal{C}_K(S)$. The measure μ_K^d induces a morphism of abelian groups $\mu_K^d : \mathrm{IC}_K^d(S) \to \mathbf{R}$.

More generally if $\varphi = \varphi \mathbf{1}_{S'}$, where S' has dimension $i \leq d$, we say φ is i-integrable if its restriction φ' to S' is integrable and we set $\mu_K^i(\varphi) := \mu_K^i(\varphi')$. One defines $\mathrm{IC}_K^i(S)$ as the abelian subgroup of $C_K^i(S)$ of the classes of i-integrable functions in $\mathcal{C}_K(S)$. The measure μ_K^i induces a morphism of abelian groups $\mu_K^i : \mathrm{IC}_K^i(S) \to \mathbf{R}$. Finally we set $\mathrm{IC}_K(S) := \bigoplus_i \mathrm{IC}_K^i(S)$ and we define $\mu_K : \mathrm{IC}_K(S) \to \mathbf{R}$ to be the sum of the morphisms μ_K^i. We call elements of $C_K(S)$, resp. $\mathrm{IC}_K(S)$, constructible Functions, resp. integrable constructible Functions on S.

Also, if $f : S \to \Lambda$ is a morphism of p-adic definable subsets, we shall say an element φ in $C_K(S \to \Lambda)$ is integrable if the restriction of φ to every fiber of f is an integrable constructible Function and we denote by $\mathrm{IC}_K(S \to \Lambda)$ the set of such Functions.

We may now reformulate Denef's basic theorem on p-adic integration (Theorem 1.5 in [10], see also [8]):

Theorem 6.5 (Denef). *Let $f : S \to \Lambda$ be a morphism of p-adic definable subsets. For every integrable constructible Function φ in $C_K(S \to \Lambda)$, there exists a unique function $\mu_{K,\Lambda}(\varphi)$ in $\mathcal{C}(\Lambda)$ such that, for every point λ in Λ,*

$$\mu_{K,\Lambda}(\varphi)(\lambda) = \mu_K(\varphi_{|f^{-1}(\lambda)}). \tag{26}$$

Strictly speaking, this is not the statement that one finds in [10], but the proof sketched there extends to our setting.

6.6 Pushforward

It is possible to define, for every morphism $f : S \to S'$ of p-adic definable subsets, a natural pushforward morphism

$$f_! : \mathrm{IC}_K(S) \longrightarrow \mathrm{IC}_K(S') \tag{27}$$

satisfying similar properties as in Theorem 4.1. This may be done along similar lines as what we did in the motivic case using Denef's p-adic cell decomposition [9] instead of Denef–Pas cell decomposition. Note however that much less work is required in this case, since one already knows what the p-adic measure is!

In particular, when f is the projection on the one point definable subset, one recovers the p-adic measure μ_K. Also in the relative setting we have natural pushforward morphisms

$$f_{!\Lambda} : \mathrm{IC}_K(S \to \Lambda) \longrightarrow \mathrm{IC}_K(S' \to \Lambda), \tag{28}$$

for $f : S \to S'$ over Λ, and one recovers the relative p-adic measure $\mu_{K,\Lambda}$ when f is the projection to Λ.

6.7 Comparison with p-adic integration

Let k be a number field with ring of integers $\mathcal{O}$. Let $\mathcal{A}_{\mathcal{O}}$ be the collection of all the p-adic completions of k and of all finite field extensions of k. In this section and in section 7.2 we let $\mathcal{L}_{\mathcal{O}}$ be the language $\mathcal{L}_{\mathrm{DP,P}}(\mathcal{O}[[t]])$, that is, the language $\mathcal{L}_{\mathrm{DP,P}}$ with coefficients in k for the residue field sort and coefficients in $\mathcal{O}[[t]]$ for the valued field sort, and, all definable subassignments, definable morphisms, and motivic constructible functions will be with respect to this language. To stress the fact that our language is $\mathcal{L}_{\mathcal{O}}$ we use the notation $\mathrm{Def}(\mathcal{L}_{\mathcal{O}})$ for Def, and similarly for $\mathcal{C}(S, \mathcal{L}_{\mathcal{O}})$, $\mathrm{Def}_S(\mathcal{L}_{\mathcal{O}})$ and so on.

For K in $\mathcal{A}_{\mathcal{O}}$ we write k_K for its residue field with $q(K)$ elements, R_K for its valuation ring and ϖ_K for a uniformizer of R_K.

Let us choose for a while, for every definable subassignment S in $\mathrm{Def}(\mathcal{L}_{\mathcal{O}})$, an $\mathcal{L}_{\mathcal{O}}$-formula ψ_S defining S. We shall write $\tau(S)$ to denote the datum (S, ψ_S). Similarly, for any element φ of $\mathcal{C}(S)$, $C(s)$, $IC(S)$, and so on, we choose a finite set $\psi_{\varphi,i}$ of formulas needed to determine φ and we write $\tau(\varphi)$ for $(\varphi, \{\psi_{\varphi,i}\}_i)$.

Let S be a definable subassignment of $h[m, n, r]$ in $\mathrm{Def}(\mathcal{L}_{\mathcal{O}})$ with $\tau(S) = (S, \psi_S)$. Let K be in $\mathcal{A}_{\mathcal{O}}$. One may consider K as an $\mathcal{O}[[t]]$-algebra via the morphism

$$\lambda_{\mathcal{O},K} : \mathcal{O}[[t]] \to K : \sum_{i \in \mathbf{N}} a_i t^i \mapsto \sum_{i \in \mathbf{N}} a_i \varpi_K^i, \tag{29}$$

hence, if one interprets elements a of $\mathcal{O}[[t]]$ as $\lambda_{\mathcal{O},K}(a)$, the formula ψ_S defines a p-adic definable subset $S_{K,\tau}$ of $K^m \times k_K^n \times \mathbf{Z}^r$.

If now $\tau(S) = (S, \psi_S)$ is replaced by $\tau'(S) = (S, \psi'_S)$ with ψ'_S another $\mathcal{L}_{\mathcal{O}}$-formula defining S, it follows, from a small variant of Proposition 5.2.1 of [13] (a result of Ax–Kochen–Eršov type that uses ultraproducts and follows from the theorem of Denef–Pas), that there exists an integer N such that $S_{K,\tau} = S_{K,\tau'}$ for every K in $\mathcal{A}_{\mathcal{O}}$ with residue field characteristic $\mathrm{char} k_K \geq N$. (Note however that this number N can be arbitrarily large for different τ'.)

Let us consider the quotient

$$\prod_{K \in \mathcal{A}_{\mathcal{O}}} \mathcal{C}_K(S_{K,\tau}) / \sum_N \prod_{\substack{K \in \mathcal{A}_{\mathcal{O}} \\ \mathrm{char} k_K < N}} \mathcal{C}_K(S_{K,\tau}), \tag{30}$$

consisting of families indexed by K of elements of $\mathcal{C}_K(S_{K,\tau})$, two such families being identified if for some $N > 0$ they coincide for $\mathrm{char} k_K \geq N$. It follows

from the above remark that it is independent of τ (more precisely all these quotients are canonically isomorphic), so we may denote it by

$$\prod{}' \mathcal{C}_K(S_K). \tag{31}$$

One defines similarly $\prod{}' C_K(S_K)$, $\prod{}' \mathrm{IC}_K(S_K)$, etc.

Now take W in $\mathrm{RDef}_S(\mathcal{L}_\mathcal{O})$. It defines a p-adic definable subset $W_{K,\tau}$ of $S_{K,\tau} \times (k_K)^\ell$, for some ℓ, for every K in $\mathcal{A}_\mathcal{O}$. We may now consider the function $\psi_{W,K,\tau}$ on $S_{K,\tau}$ assigning to a point x the number of points mapping to it in $W_{K,\tau}$, that is, $\psi_{W,K,\tau}(x) = \mathrm{card}(W_{K,\tau} \cap (\{x\} \times k_K^\ell))$. Similarly as before, if we take another function τ', we have $\psi_{W,K,\tau} = \psi_{W,K,\tau'}$ for every K in $\mathcal{A}_\mathcal{O}$ with residue field characteristic $\mathrm{char} k_K \geq N$, hence we get in this way an arrow $\mathrm{RDef}_S(\mathcal{L}_\mathcal{O}) \to \prod{}' \mathcal{C}_K(S_K)$ which factorizes through a ring morphism $K_0(\mathrm{RDef}_S(\mathcal{L}_\mathcal{O})) \to \prod{}' \mathcal{C}_K(S_K)$. If we send $\mathbf{L}$ to $q(K)$, one can extend uniquely this morphism to a ring morphism

$$\mathit{\Gamma} : \mathcal{C}(S, \mathcal{L}_\mathcal{O}) \longrightarrow \prod{}' \mathcal{C}_K(S_K). \tag{32}$$

Since $\mathit{\Gamma}$ preserves the (relative) dimension of support on those factors K with $\mathrm{char} k_K$ big enough, $\mathit{\Gamma}$ induces the morphisms

$$\mathit{\Gamma} : C(S, \mathcal{L}_\mathcal{O}) \longrightarrow \prod{}' C_K(S_K) \tag{33}$$

and

$$\mathit{\Gamma} : C(S \to \mathit{\Lambda}, \mathcal{L}_\mathcal{O}) \longrightarrow \prod{}' C_K(S_K \to \mathit{\Lambda}_K), \tag{34}$$

for $S \to \mathit{\Lambda}$ a morphism in $\mathrm{Def}_K(\mathcal{L}_\mathcal{O})$.

The following comparison theorem says that the morphism $\mathit{\Gamma}$ commutes with pushforward. In more concrete terms, given an integrable function φ in $C(S \to \mathit{\Lambda}, \mathcal{L}_\mathcal{O})$, for almost all p, its specialization φ_K to any finite extension K of $\mathbf{Q}_p$ in $\mathcal{A}_\mathcal{O}$ is integrable, and the specialization of the pushforward of φ is equal to the pushforward of φ_K.

Theorem 6.8. *Let $\mathit{\Lambda}$ be in $\mathrm{Def}_K(\mathcal{L}_\mathcal{O})$ and let $f : S \to S'$ be a morphism in $\mathrm{Def}_\mathit{\Lambda}(\mathcal{L}_\mathcal{O})$. The morphism*

$$\mathit{\Gamma} : C(S \to \mathit{\Lambda}, \mathcal{L}_\mathcal{O}) \to \prod{}' C_K(S_K \to \mathit{\Lambda}_K) \tag{35}$$

induces a morphism

$$\mathit{\Gamma} : \mathrm{IC}(S \to \mathit{\Lambda}, \mathcal{L}_\mathcal{O}) \to \prod{}' \mathrm{IC}_K(S_K \to \mathit{\Lambda}_K) \tag{36}$$

(and similarly for S'), and the following diagram is commutative:

$$\begin{array}{ccc} \mathrm{IC}(S \to \mathit{\Lambda}, \mathcal{L}_\mathcal{O}) & \xrightarrow{\mathit{\Gamma}} & \prod{}' \mathrm{IC}_K(S_K \to \mathit{\Lambda}_K) \\ \Big\downarrow{\scriptstyle f_{!\mathit{\Lambda}}} & & \Big\downarrow{\scriptstyle \prod{}' f_{K,\mathit{\Lambda}_K !}} \\ \mathrm{IC}(S' \to \mathit{\Lambda}, \mathcal{L}_\mathcal{O}) & \xrightarrow{\mathit{\Gamma}} & \prod{}' \mathrm{IC}_K(S'_K \to \mathit{\Lambda}_K), \end{array} \tag{37}$$

with $f_K : S_K \to S'_K$ the morphism induced by f and where the map $\prod' f_{K,\Lambda_K!}$ is induced by the maps $f_{K,\Lambda_K!} : \mathrm{IC}_K(S_K \to \Lambda_K) \to \mathrm{IC}_K(S'_K \to \Lambda_K)$.

Proof (Sketch of proof). The image of φ in $\mathrm{IC}(S \to \Lambda, \mathcal{L}_\mathcal{O})$ under $f_{!\Lambda}$ can be calculated by taking an appropriate cell decomposition of the occurring sets, adapted to the occurring functions (as in [1] and inductively applied to all valued field variables). Such calculation is independent of the choice of cell decomposition by the unicity statement of Theorem 4.1. By the Ax–Kochen–Eršov principle for the language $\mathcal{L}_\mathcal{O}$ implied by Theorem 2.10, this cell decomposition determines, for K in $\mathcal{A}_\mathcal{O}$ with $\mathrm{char} k_K$ sufficiently large, a cell decomposition à la Denef (in the formulation of Lemma 4 of [4]) of the K-component of these sets, adapted to the K-component of the functions occuring here, where thus the same calculation can be pursued. That this calculation is actually the same follows from the fact that p-adic integration satisfies properties analogous to the axioms of Theorem 4.1. □

In particular, we have the following statement, which says that, given an integrable function φ in $C(S \to \Lambda, \mathcal{L}_\mathcal{O})$, for almost all p, its specialization φ_F to any finite extension F of $\mathbf{Q}_p$ in $\mathcal{A}_\mathcal{O}$ is integrable, and the specialization of the motivic integral $\mu(\varphi)$ is equal to the p-adic integral of φ_F:

Theorem 6.9. *Let $f : S \to \Lambda$ be a morphism in $\mathrm{Def}_K(\mathcal{L}_\mathcal{O})$. The following diagram is commutative:*

$$\begin{array}{ccc} \mathrm{IC}(S \to \Lambda, \mathcal{L}_\mathcal{O}) & \xrightarrow{\Gamma} & \prod{}' \mathrm{IC}_K(S_K \to \Lambda_K) \\ \downarrow{\scriptstyle \mu_\Lambda} & & \downarrow{\scriptstyle \prod' \mu_{K,\Lambda_K}} \\ \mathcal{C}(\Lambda, \mathcal{L}_\mathcal{O}) & \xrightarrow{\Gamma} & \prod{}' \mathcal{C}_K(\Lambda_K). \end{array}$$

7 Reduction mod p and a motivic Ax–Kochen–Eršov Theorem for integrals with parameters

7.1 Integration over $\mathbf{F}_q((t))$

Consider now the field $K = \mathbf{F}_q((t))$ with valuation ring R_K and residue field $k_K = \mathbf{F}_q$ with $q = q(K)$ a prime power. One may define $\mathbf{F}_q((t))$-definable sets similarly as in Section 6.1. Little is known about the structure of these $\mathbf{F}_q((t))$-definable sets, but, for any subset A of K^m, not necessarily definable, we may still define the dimension of A as the dimension of its Zariski closure. Similarly as in Section 6.2, one extends that definition to any subset A of $K^m \times k_K^n \times \mathbf{Z}^r$ and define the relative dimension of a mapping $f : A \to \Lambda$, with Λ any subset of $K^{m'} \times k_K^{n'} \times \mathbf{Z}^{r'}$. When A is $\mathbf{F}_q((t))$-definable, one can define a $\mathbf{Q}$-algebra $\mathcal{C}_K(A)$ as in Section 6.3, but since no analogue of Theorem 6.5 is

known in this setting, we shall consider, for A any subset of $K^m \times k_K^n \times \mathbf{Z}^r$, the $\mathbf{Q}$-algebra $\mathcal{F}_K(A)$ of all functions $A \to \mathbf{Q}$. For $d \geq 0$ an integer, we denote by $\mathcal{F}_K^{\leq d}(A)$ the ideal of functions with support of dimension $\leq d$. We set $F_K^d(A) := \mathcal{F}_K^{\leq d}(A)/\mathcal{F}_K^{\leq d-1}(A)$ and $F_K(A) := \oplus_d F_K^d(A)$. One defines similarly relative variants $\mathcal{F}_K^{\leq d}(A \to \Lambda)$, $F_K^d(A \to \Lambda)$ and $F_K(A \to \Lambda)$, for $f : A \to A'$ as above.

Let A be a subset of K^m with Zariski closure $\bar{A}$ of dimension d. We consider the canonical d-dimensional measure μ_K^d on $\bar{A}(K)$ as in [20]. We say a function φ in $\mathcal{F}_K(A)$ is integrable if it is measurable and integrable with respect to the measure μ_K^d. Now we may proceed as in Section 6.4 to define, for A a subset of $K^m \times k_K^n \times \mathbf{Z}^r$, $\mathrm{I}F_K(A)$ and $\mu_K : \mathrm{I}F_K(A) \to \mathbf{R}$. Also, if $f : A \to \Lambda$ is a mapping as before, one defines $\mathrm{I}F_K(A \to \Lambda)$ as Functions whose restrictions to all fibers lie in $\mathrm{I}F_K$. Let $\mu_{K,\Lambda}$ be the unique mapping $\mathrm{I}F_K(A \to \Lambda) \to \mathcal{F}(\Lambda)$ such that, for every φ in $\mathrm{I}F_K(A \to \Lambda)$ and every point λ in Λ, $\mu_{K,\Lambda}(\varphi)(\lambda) = \mu_K(\varphi_{|f^{-1}(\lambda)})$.

7.2 Reduction mod p

We go back to the notation of Section 6.7. In particular, k denotes a number field with ring of integers $\mathcal{O}$, $\mathcal{A}_\mathcal{O}$ denotes the set of all p-adic completions of k and of all the finite field extensions of k, and $\mathcal{L}_\mathcal{O}$ stands for the language $\mathcal{L}_{\mathrm{DP,P}}(\mathcal{O}[[t]])$. We also use the map τ as defined in section 6.7.

Let $\mathcal{B}_\mathcal{O}$ be the set of all local fields over $\mathcal{O}$ of positive characteristic. As for $\mathcal{A}_\mathcal{O}$, we use for every K in $\mathcal{B}_\mathcal{O}$ the notation k_K for its residue field with $q(K)$ elements, R_K for its valuation ring and ϖ_K for a uniformizer of R_K.

Let S be a definable subassignment of $h[m,n,r]$ in $\mathrm{Def}(\mathcal{L}_\mathcal{O})$ and let $\tau(S)$ be (S, ψ_S) with ψ_S a $\mathcal{L}_\mathcal{O}$-formula. Similarly as for $\mathcal{A}_\mathcal{O}$, since every K in $\mathcal{B}_\mathcal{O}$ is an $\mathcal{O}[[t]]$-algebra under the morphism

$$\lambda_{\mathcal{O},K} : \mathcal{O}[[t]] \to K : \sum_{i \in \mathbf{N}} a_i t^i \mapsto \sum_{i \in \mathbf{N}} a_i \varpi_K^i, \tag{38}$$

interpreting any element a of $\mathcal{O}[[t]]$ as $\lambda_{\mathcal{O},K}(a)$, ψ_S defines a K-definable subset $S_{K,\tau}$ of $K^m \times k_K^n \times \mathbf{Z}^r$. Again by a small variant of Proposition 5.2.1 of [13], for any other τ' we have for every K in $\mathcal{B}_\mathcal{O}$ with $\mathrm{char} k_K$ big enough that $S_{K,\tau} = S_{K,\tau'}$, hence, may define, similarly as in Section 6.7,

$$\prod{}' \mathcal{F}_K(S_K) \tag{39}$$

to be the quotient

$$\prod_{K \in \mathcal{B}_\mathcal{O}} \mathcal{F}_K(S_{K,\tau}) / \sum_N \prod_{\substack{K \in \mathcal{B}_\mathcal{O} \\ \mathrm{char} k_K < N}} \mathcal{F}_K(S_{K,\tau}), \tag{40}$$

and similarly for $\prod' F_K(S_K)$, $\prod' \mathrm{I}F_K(S_K)$, etc.

Similarly as in Section 6.7, one may define ring morphisms

$$\hat{\Gamma} : \mathcal{C}(S, \mathcal{L}_{\mathcal{O}}) \longrightarrow \prod{}' \mathcal{F}_K(S_K), \tag{41}$$

$$\hat{\Gamma} : C(S, \mathcal{L}_{\mathcal{O}}) \longrightarrow \prod{}' F_K(S_K) \tag{42}$$

and

$$\hat{\Gamma} : C(S \to \Lambda, \mathcal{L}_{\mathcal{O}}) \longrightarrow \prod{}' F_K(S_K \to \Lambda_K), \tag{43}$$

for $S \to \Lambda$ a morphism in $\mathrm{Def}_K(\mathcal{L}_{\mathcal{O}})$.

The following statement is a companion to Theorem 6.9 and has an essentially similar proof.

Theorem 7.3. *Let $f : S \to \Lambda$ be a morphism in $\mathrm{Def}_K(\mathcal{L}_{\mathcal{O}})$. The morphism*

$$\hat{\Gamma} : C(S \to \Lambda, \mathcal{L}_{\mathcal{O}}) \to \prod{}' F_K(S_K \to \Lambda_K) \tag{44}$$

induces a morphism

$$\hat{\Gamma} : \mathrm{I}C(S \to \Lambda, \mathcal{L}_{\mathcal{O}}) \to \prod{}' \mathrm{I}F_K(S_K \to \Lambda_K) \tag{45}$$

and the following diagram is commutative:

$$\begin{array}{ccc} \mathrm{I}C(S \to \Lambda, \mathcal{L}_{\mathcal{O}}) & \xrightarrow{\hat{\Gamma}} & \prod{}' \mathrm{I}F_K(S_K \to \Lambda_K) \\ \downarrow{\scriptstyle \mu_\Lambda} & & \downarrow{\scriptstyle \prod{}' \mu_{K,\Lambda_K}} \\ \mathcal{C}(\Lambda, \mathcal{L}_{\mathcal{O}}) & \xrightarrow{\hat{\Gamma}} & \prod{}' \mathcal{F}_K(\Lambda_K). \end{array}$$

7.4 Ax–Kochen–Eršov Theorems for motivic integrals

We keep the notation of Sections 6.7 and 7.2. Let S, resp. Λ, be definable subassignments of $h[m, n, r]$, resp. $h[m', n', r']$, in the language $\mathcal{L}_{\mathcal{O}}$ and consider a definable (in the language $\mathcal{L}_{\mathcal{O}}$) morphism $f : S \to \Lambda$. Since we are interested in integrals along the fibers of f, there is no restriction in assuming, and we shall do so, that $\Lambda = h[m', n', r']$. We set $\Lambda(\mathcal{O}) := \mathcal{O}[[t]]^{m'} \times k^{n'} \times \mathbf{Z}^{r'}$.

A first attempt to get Ax–Kochen–Eršov Theorems for motivic integrals is by comparing values. This is achieved as follows. To every point λ in $\Lambda(\mathcal{O})$ we may assign, for all K in $\mathcal{A}_{\mathcal{O}} \cup \mathcal{B}_{\mathcal{O}}$, a point λ_K in $(R_K)^{m'} \times k_K^{n'} \times \mathbf{Z}^{r'}$, by using the maps $\lambda_{\mathcal{O},K}$ on the $\mathcal{O}[[t]]^{m'}$-factor and reduction modulo $\mathrm{char}k_K$ for the $k^{n'}$-factor.

Let φ be in $C(S \to \Lambda, \mathcal{L}_{\mathcal{O}})$. With a slight abuse of notation, we shall write φ_K for the component at K in $C_K(S_K \to \Lambda_K)$ of $\Gamma(\varphi)$, resp. of $\hat{\Gamma}(\varphi)$, for K with $\mathrm{char}k_K$ big, in $\mathcal{A}_{\mathcal{O}}$, resp. in $\mathcal{B}_{\mathcal{O}}$. We shall use similar notation for φ in $\mathcal{C}(S, \mathcal{L}_{\mathcal{O}})$.

Over the final subassignment $h[0,0,0]$ the morphisms Γ and $\hat{\Gamma}$ have quite simple descriptions. Indeed, the morphism $\hat{\gamma}$ of Section 5.7 induces a ring morphism

$$\gamma' : \mathcal{C}(h_{\operatorname{Spec} k}, \mathcal{L}_{\mathcal{O}}) \longrightarrow K_0(\mathrm{PFF}_K) \otimes_{\mathbf{Z}[\mathbf{L}]} A. \tag{46}$$

One the other hand, note that $\mathcal{C}_K(\text{point}) \simeq \mathcal{F}_{K'}(\text{point}) \simeq \mathbf{Q}$ for K in $\mathcal{A}_{\mathcal{O}}$ and K' in $\mathcal{B}_{\mathcal{O}}$. The morphism $\chi_c : K_0(\mathrm{PFF}_k) \to K_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q}$ from Section 5.7 induces a ring morphism

$$\delta' : K_0(\mathrm{PFF}_K) \otimes_{\mathbf{Z}[\mathbf{L}]} A \to K_0^{\mathrm{mot}}(\mathrm{Var}_k) \otimes \mathbf{Q} \otimes_{\mathbf{Z}[\mathbf{L}]} A. \tag{47}$$

Note that, for K in $\mathcal{A}_{\mathcal{O}}$, resp. in $\mathcal{B}_{\mathcal{O}}$, with $\mathrm{char}K$ big enough, the K-component of $\Gamma(\alpha)$, resp. of $\hat{\Gamma}(\alpha)$, is equal to the trace of the Frobenius at k_K acting on an étale realisation of $(\delta' \circ \gamma')(\alpha)$, for α in $\mathcal{C}(h_{\operatorname{Spec} k}, \mathcal{L}_{\mathcal{O}})$. In particular, one deduces the following statement:

Lemma 7.5. *Let ψ be a function in $\mathcal{C}(\Lambda, \mathcal{L}_{\mathcal{O}})$. Then, for every λ in $\Lambda(\mathcal{O})$, there exists an integer N such that, for every K_1 in $\mathcal{A}_{\mathcal{O}}$, K_2 in $\mathcal{B}_{\mathcal{O}}$ with $k_{K_1} \simeq k_{K_2}$ and $\mathrm{char}k_{K_1} > N$,*

$$\psi_{K_1}(\lambda_{K_1}) = \psi_{K_2}(\lambda_{K_2}), \tag{48}$$

which also is equal to $(i_\lambda^*(\psi))_{K_1}$ and to $(i_\lambda^*(\psi))_{K_2}$.

From Lemma 7.5, Theorem 6.9 and Theorem 7.3 one deduces immediately:

Theorem 7.6. *Let $f : S \to \Lambda$ be as above and let φ be a Function in $\mathrm{IC}(S \to \Lambda, \mathcal{L}_{\mathcal{O}})$. Then, for every λ in $\Lambda(\mathcal{O})$, there exists an integer N such that for all K_1 in $\mathcal{A}_{\mathcal{O}}$, K_2 in $\mathcal{B}_{\mathcal{O}}$ with $k_{K_1} \simeq k_{K_2}$ and $\mathrm{char}k_{K_1} > N$,*

$$\mu_{K_1}(\varphi_{K_1|f_{K_1}^{-1}(\lambda_{K_1})}) = \mu_{K_2}(\varphi_{K_2|f_{K_2}^{-1}(\lambda_{K_2})}), \tag{49}$$

which also equals $(\mu_\Lambda(\varphi))_{K_1}(\lambda_{K_1})$ and $(\mu_\Lambda(\varphi))_{K_2}(\lambda_{K_2})$.

Note that Theorem 2.12 is a corollary of Theorem 7.6 when $(m', n', r') = (0,0,0)$. In fact, Theorem 7.6 is not really satisfactory when $(m', n', r') \neq (0,0,0)$, since it is not uniform with respect to λ. The following example shows that this is unavoidable: take $k = \mathbf{Q}$, $S = \Lambda = h[1,0,0]$, f the identity and $\varphi = \mathbf{1}_{S \smallsetminus \{0\}}$ in $\mathrm{IC}(S \to \Lambda) = \mathcal{C}(S)$. Take K_1 in $\mathcal{A}_{\mathcal{O}}$ and K_2 in $\mathcal{B}_{\mathcal{O}}$. We have $\varphi_{K_1}(\lambda_{K_1}) = \varphi_{K_2}(\lambda_{K_2})$ for $\lambda \neq 0$ in $\mathbf{Z}$ only if the characteristic of K_2 does not divide λ.

Hence, instead of comparing values of integrals depending on parameters, we better compare the integrals as functions, which is done as follows:

Theorem 7.7. *Let $f : S \to \Lambda$ be as above and let φ be a Function in $\mathrm{IC}(S \to \Lambda, \mathcal{L}_{\mathcal{O}})$. Then, there exists an integer N such that for all K_1 in $\mathcal{A}_{\mathcal{O}}$, K_2 in $\mathcal{B}_{\mathcal{O}}$ with $k_{K_1} \simeq k_{K_2}$ and $\mathrm{char}k_{K_1} > N$,*

$$\mu_{K_1,\Lambda_{K_1}}(\varphi_{K_1}) = 0 \quad \textit{if and only if} \quad \mu_{K_2,\Lambda_{K_2}}(\varphi_{K_2}) = 0. \tag{50}$$

Proof. Follows directly from Theorem 6.9, Theorem 7.3, and Theorem 7.8. □

Theorem 7.8. *Let ψ be in $\mathcal{C}(\Lambda, \mathcal{L}_{\mathcal{O}})$. Then, there exists an integer N such that for all K_1 in $\mathcal{A}_{\mathcal{O}}$, K_2 in $\mathcal{B}_{\mathcal{O}}$ with $k_{K_1} \simeq k_{K_2}$ and* $\mathrm{char} k_{K_1} > N$

$$\psi_{K_1} = 0 \quad \textit{if and only if} \quad \psi_{K_2} = 0. \tag{51}$$

Remark 7.9. Thanks to results of Cunningham and Hales [6], Theorem 7.7 applies to the orbital integrals occuring in the Fundamental Lemma. Hence, it follows from Theorem 7.7 that the Fundamental Lemma holds over function fields of large characteristic if and only if it holds for p-adic fields of large characteristic. (Note that the Fundamental Lemma is about the equality of two integrals, or, which amounts to the same, their difference to be zero.) In the special situation of the Fundamental Lemma, a more precise comparison result has been proved by Waldspurger [26] by representation theoretic techniques. Let us recall that the Fundamental Lemma for unitary groups has been proved recently by Laumon and Ngô [18] for function fields.

References

1. R. Cluckers and F. Loeser – "Constructible motivic functions and motivic integration", in preparation.
2. — , "Fonctions constructibles et intégration motivique I", `math.AG/0403349`, to appear in C. R. Acad. Sci. Paris Sér. I Math.
3. — , "Fonctions constructibles et intégration motivique II", `math.AG/0403350`, to appear in C. R. Acad. Sci. Paris Sér. I Math.
4. R. Cluckers – "Classification of semi-algebraic p-adic sets up to semi-algebraic bijection", *J. Reine Angew. Math.* **540** (2001), p. 105–114.
5. P. J. Cohen – "Decision procedures for real and p-adic fields", *Comm. Pure Appl. Math.* **22** (1969), p. 131–151.
6. C. Cunningham and T. Hales – "Good orbital integrals", `math.RT/0311353`.
7. F. Delon – "Some p-adic model theory", European women in mathematics (Trieste, 1997), Hindawi Publ. Corp., Stony Brook, NY, 1999, p. 63–76.
8. J. Denef – "On the evaluation of certain p-adic integrals", Séminaire de théorie des nombres, Paris 1983–84, Progr. Math., vol. 59, Birkhäuser Boston, Boston, MA, 1985, p. 25–47.
9. — , "p-adic semi-algebraic sets and cell decomposition", *J. Reine Angew. Math.* **369** (1986), p. 154–166.
10. — , "Arithmetic and geometric applications of quantifier elimination for valued fields", Model theory, algebra, and geometry, Math. Sci. Res. Inst. Publ., vol. 39, Cambridge Univ. Press, Cambridge, 2000, p. 173–198.
11. J. Denef and F. Loeser – "On some rational generating series occuring in arithmetic geometry", `math.NT/0212202`.
12. — , "Germs of arcs on singular algebraic varieties and motivic integration", *Invent. Math.* **135** (1999), no. 1, p. 201–232.

13. — , "Definable sets, motives and p-adic integrals", *J. Amer. Math. Soc.* **14** (2001), no. 2, p. 429–469 (electronic).
14. — , "Motivic integration and the Grothendieck group of pseudo-finite fields", *Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002)* (Beijing), Higher Ed. Press, 2002, p. 13–23.
15. L. van den Dries – "Dimension of definable sets, algebraic boundedness and Henselian fields", *Ann. Pure Appl. Logic* **45** (1989), no. 2, p. 189–209, Stability in model theory, II (Trento, 1987).
16. H. Gillet and C. Soulé – "Descent, motives and K-theory", *J. Reine Angew. Math.* **478** (1996), p. 127–176.
17. F. Guillén and V. Navarro Aznar – "Un critère d'extension des foncteurs définis sur les schémas lisses", *Publ. Math. Inst. Hautes Études Sci.* (2002), no. 95, p. 1–91.
18. G. Laumon and B. C. Ngô – "Le lemme fondamental pour les groupes unitaires", `math.AG/0404454`.
19. J. Nicaise – "Relative motives and the theory of pseudo-finite fields", `math.AG/0403160`.
20. J. Oesterlé – "Réduction modulo p^n des sous-ensembles analytiques fermés de $\mathbf{Z}_p^N$", *Invent. Math.* **66** (1982), no. 2, p. 325–341.
21. J. Pas – "Uniform p-adic cell decomposition and local zeta functions", *J. Reine Angew. Math.* **399** (1989), p. 137–172.
22. P. Scowcroft and L. van den Dries – "On the structure of semialgebraic sets over p-adic fields", *J. Symbolic Logic* **53** (1988), no. 4, p. 1138–1164.
23. J.-P. Serre – "Quelques applications du théorème de densité de Chebotarev", *Inst. Hautes Études Sci. Publ. Math.* (1981), no. 54, p. 323–401.
24. G. Terjanian – "Un contre-exemple à une conjecture d'Artin", *C. R. Acad. Sci. Paris Sér. A-B* **262** (1966), p. A612.
25. W. Veys – "Reduction modulo p^n of p-adic subanalytic sets", *Math. Proc. Cambridge Philos. Soc.* **112** (1992), no. 3, p. 483–486.
26. J.-L. Waldspurger – "Endoscopie et changement de caractéristiques", 2004, preprint.

Nested sets and Jeffrey–Kirwan residues

Corrado De Concini and Claudio Procesi

Dip. Mat. Castelnuovo, Univ. di Roma La Sapienza, Rome, Italy
deconcin@mat.uniroma1.it
procesi@mat.uniroma1.it

Summary. For the complement of a hyperplane arrangement we construct a dual homology basis to the no-broken-circuit basis of cohomology. This is based on the theory of wonderful embeddings and nested sets developed in [4]. Our result allows us to express the so-called Jeffrey–Kirwan residues in terms of integration on some explicit geometric cycles.

1 Introduction

In this paper we discuss some new notions in the theory of hyperplane arrangements. The paper grew out of our plan to give an improved and simplified version of some of the results of Szenes–Vergne [6].

We start from a complex vector space U of finite dimension r and a finite central hyperplane arrangement in U^*, given by a finite set $\Delta \subset U$ of linear equations. From these data one constructs the partially ordered set of subspaces obtained by intersection of the given hyperplanes and the open set $\mathcal{A}_\Delta$ complement of the union of the hyperplanes of the arrangement.

This paper consists of three parts. Part 1 is a recollection of the results in [4]. In Part 2 we present three new results. The first, of combinatorial nature, establishes a canonical bijective correspondence between the set of no-broken-circuit bases and maximal nested sets which satisfy a condition called properness.

Next we associate to each proper maximal nested set $\mathbb{M}$ a geometric cycle $c_{\mathbb{M}}$ of dimension r in $\mathcal{A}_\Delta$. We show that integration of a top degree differential form over this cycle is done, by a simple algorithm, taking a multiple residue with respect to a system of local coordinates. The last result is the proof that, under the duality given by integration, the basis of cohomology given by the forms associated to the no-broken-circuit bases is dual to the basis of homology determined by the cycles $c_{\mathbb{M}}$. Section 3 is dedicated to the application relevant for the computations of [6], that is to say the Jeffrey–Kirwan residues.

Acknowledgments. We wish to thank M. Vergne for explaining to us some of the theory and for various discussions and suggestions. The authors are partially supported by the Cofin 40 %, MIUR.

1.1 Notation

With the notation of the introduction, let U be a complex vector space of dimension r, $\Delta \subset U$ a totally ordered finite set of vectors $\Delta = \{\alpha_1, \dots, \alpha_m\}$. These vectors are the linear equations of a hyperplane arrangement in U^*. For simplicity we also assume that Δ spans U and any two distinct elements in Δ are linearly independent.

An example is a (complete) set of positive roots in a root system ordered by any total order which refines the reverse dominance order.

In the A_{n-1} case we could say that $x_i - x_j \geq x_h - x_k$ if $k - h \geq j - i$ and, if they are equal, if $i \leq h$.

We want to recall briefly the main points of the theory (cf. [5]). Let $\Omega_i(\mathcal{A}_\Delta)$ denote the space of rational differential forms of degree i on $\mathcal{A}_\Delta$.

We shall use implicitly the *formality*, that is the fact that the $\mathbb{Z}$-subalgebra of differential forms on $\mathcal{A}_\Delta$ generated by the linear forms $\frac{1}{2\pi i} d\log\alpha, \alpha \in \Delta$ is isomorphic (via De Rham theory) to the integral cohomology of $\mathcal{A}_\Delta$.

Formality implies in particular that $\Omega_r(\mathcal{A}_\Delta) = H^r \oplus d\Omega_{r-1}(\mathcal{A}_\Delta)$, for top degree forms. Here $H^r \equiv H^r(\mathcal{A}_\Delta, \mathbb{C})$ is the $\mathbb{C}$-span of the top degree forms $\omega_\sigma := d\log\gamma_1 \wedge \cdots \wedge d\log\gamma_r$ for all bases $\sigma := \{\gamma_1, \dots, \gamma_r\}$ extracted from Δ. The forms ω_σ satisfy a set of linear relations generated by the following ones. Given $r+1$ elements $\gamma_i \in \Delta$ spanning U, we have:

$$\sum_{i=1}^{r+1} (-1)^i d\log\gamma_1 \wedge \cdots d\check{\log}\gamma_i \cdots \wedge d\log\gamma_r = 0.$$

Recall that a *no-broken-circuit* in Δ (with respect to the given total ordering) is an ordered linearly independent subsequence $\{\alpha_{i_1}, \dots, \alpha_{i_t}\}$ such that, for each $1 \leq \ell \leq t$, there is no $j < i_\ell$ such that the vectors $\alpha_j, \alpha_{i_\ell}, \dots, \alpha_{i_t}$ are linearly dependent. In other words α_{i_ℓ} is the minimum element of $\Delta \cap \langle \alpha_{i_\ell}, \dots, \alpha_{i_t} \rangle$. In [5] it is proved that the elements

$$(\frac{1}{2\pi i})^r \omega_\sigma := (\frac{1}{2\pi i})^r d\log\gamma_1 \wedge \cdots \wedge d\log\gamma_r,$$

where $\sigma = \{\gamma_1, \dots, \gamma_r\}$ runs over all ordered bases of V which are no-broken-circuits, give a linear $\mathbb{Z}$-basis of the integral cohomology of $\mathcal{A}_\Delta$.

1.2 Irreducibles

Let us now recall some notions from [4]. Given a subset $S \subset \Delta$ we shall denote by U_S the space spanned by S.

Definition 1. *Given a subset $S \subset \Delta$, the completion $\overline{S}$ of S equals $U_S \cap \Delta$. S is called complete if $S = \overline{S}$.*

A complete subset $S \subset \Delta$ is called reducible if we can find a partition $S = S_1 \dot\cup S_2$, called a **decomposition** *such that $U_S = U_{S_1} \oplus U_{S_2}$, irreducible otherwise.*

Equivalently we say that the space U_S is reducible. Notice that, in the reducible case, $S = S_1 \dot\cup S_2$, also S_1 and S_2 are complete.

From this definition it is easy to see [4]:

Lemma 1.1. *Given complete sets $A \subset S$ and a decomposition $S = S_1 \dot\cup S_2$ of S we have that $A = (A \cap S_1) \dot\cup (A \cap S_2)$ is a decomposition of A. Let $S \subset \Delta$ be complete. Then there is a sequence (unique up to reordering) $S_1, \ldots, S_m$ of irreducible subsets in S such that*

- *$S = S_1 \cup \cdots \cup S_m$ as disjoint union.*
- *$U_S = U_{S_1} \oplus \cdots \oplus U_{S_m}$.*

The S_i's are called the irreducible components of S and the decomposition $S = S_1 \cup \cdots \cup S_m$, the irreducible decomposition of S.

In the example of root systems, a complete set S is irreducible if and only if $S \cup -S$ is an irreducible root system.

We shall denote by $\mathcal{I}$ the family of all irreducible subsets in Δ.

1.3 A minimal model

In [4] we have constructed a minimal smooth variety X_Δ containing $\mathcal{A}_\Delta$ as an open set with complement a normal crossings divisor, plus a proper map $\pi : X_\Delta \to U^*$ extending the identity of $\mathcal{A}_\Delta$. The smooth irreducible components of the boundary are indexed by the *irreducible subsets*. To describe the intersection pattern between these divisors, in [4] we developed the general theory of nested sets. Maximal nested sets correspond to special points at infinity, intersections of these boundary divisors. In the papers [7] and [6], implicitly the authors use the points at infinity coming from complete flags which correspond, in the philosophy of [4], to a *maximal* model with normal crossings. It is thus not a surprise that by passing from a maximal to a minimal model the combinatorics gets simplified and the constructions become more canonical.

Let us recall the main construction of [4]. For each $S \in \mathcal{I}$ we have a subspace $S^\perp \subset U^*$ where $S^\perp = \{a \in U^* \mid s(a) = 0,\ \forall s \in S\}$. We have the projective space $\mathbb{P}(U^*/S^\perp)$ of lines in $U^*/S^\perp$ a map $i : \mathcal{A}_\Delta \to U^* \times_{S \in \mathcal{I}} \mathbb{P}(U^*/S^\perp)$. Set X_Δ equal to the closure of the image $i(\mathcal{A}_\Delta)$ in this product. In [4] we have seen that X_Δ is a smooth variety containing a copy of $\mathcal{A}_\Delta$ and the complement of $\mathcal{A}_\Delta$ in X_Δ is a union of smooth irreducible divisors D_S, having transversal intersection, indexed by the elements $S \in \mathcal{I}$.

1.4 Nested sets

Still in [4] we showed that a family D_{S_i} of divisors indexed by irreducibles S_i has non-empty intersection (which is then smooth irreducible) if and only if the family is *nested* according to:

Definition 2. *A subfamily $\mathbb{M} \subset \mathcal{I}$ is called nested if, given any subfamily $\{S_1, \dots, S_m\} \subset \mathbb{M}$ with the property that for no $i \neq j$, $S_i \subset S_j$, then $S := S_1 \cup \cdots \cup S_m$ is complete and the S_i's are the irreducible components of S.*

Lemma 1.2.

1) Let $\mathbb{M} = \{S_1, \dots, S_m\}$ be a nested set. Then $S := \cup_{i=1}^m S_i$ is complete. The irreducible components of S are the maximal elements of $\mathbb{M}$.
2) Any nested set is the set of irreducible components of the elements of a flag $A_1 \supset A_2 \supset \cdots \supset A_k$, where each A_i is complete.

Proof. 1) By definition of nested set, the maximal elements of $\mathbb{M}$ decompose their union which is complete.

2) It is clear that, if $A \subset B$, the irreducible components of A are contained each in an irreducible component of B. From this follows that the irreducible components of the sets of a flag form a nested set. Conversely let $\mathbb{M} = \{S_1, \dots, S_m\}$ be a nested set. Set $A_1 = \cup_{i=1}^m S_i$. Next remove from $\mathbb{M}$ the irreducible components of A_1 (in $\mathbb{M}$ by part 1)). We have a new nested set to which we can apply the same procedure. Working inductively we construct a flag of which $\mathbb{M}$ is the decomposition. □

One way of using the previous result is the following. Given a basis $\sigma := \{\gamma_1, \dots, \gamma_r\} \subset \Delta$, one can associate to σ a maximal flag $F(\sigma)$ by setting $A_i(\sigma) := \Delta \cap \langle \gamma_i, \dots, \gamma_r \rangle$. Clearly the maps from bases to flags and from flags to maximal nested sets are both surjective. We thus obtain a surjective map from bases to maximal nested sets. We will see that this map induces a bijection between the set of no-broken-circuit bases and that of proper maximal nested sets (see below for their definition).

Proposition 1.3.

1) Let $A_1 \supsetneq A_2 \cdots \supsetneq A_k$, be a maximal flag of complete non-empty sets. Then $k = r$ and for each i, A_i spans a subspace of codimension $i - 1$.
2) Let $\Delta = S_1 \cup \cdots \cup S_t$ be the irreducible decomposition of Δ.
i) Then the S_i's are the maximal elements in $\mathcal{I}$.
ii) Every maximal nested set contains each of the elements S_i, $i = 1, \dots, t$ and is a union of maximal nested sets in the sets S_i.
3) Let $\mathbb{M}$ be a maximal nested set, $A \in \mathbb{M}$ and $B_1, \dots, B_r \in \mathbb{M}$ maximal among the elements in $\mathbb{M}$ properly contained in A.
Then the subspaces U_{B_i} form a direct sum and

$$\dim(\oplus_{i=1}^k U_{B_i}) + 1 = \dim U_A.$$

4) A maximal nested set always has r elements.

Proof. 1) By definition $A_1 = \Delta$ spans U. If $\alpha \in A_i - A_{i+1}$ the completion of $A_{i+1} \cup \{\alpha\}$ must be A_i by the maximality of the flag. On the other hand by definition α is not in the subspace spanned by A_{i+1}, hence we have that $\dim U_{A_i} = \dim U_{A_{i+1}} + 1$ which implies 1). Claim 2) is immediate from the definitions. As for 3), by definition the subspaces U_{B_i} form a direct sum and since A is irreducible, $\oplus_{i=1}^k U_{B_i} \subsetneq U_A$. Let $\alpha \in A - \cup_{i=1}^k B_i$ and B be the completion of $\{\alpha\} \cup \cup_{i=1}^k B_i$. We must have $B = A$, otherwise we can add the irreducible components of B to $\mathbb{M}$ which remains nested, contradicting the maximality. Thus $\dim(\oplus_{i=1}^k U_{B_i}) + 1 = \dim U_A$. Statement 4) follows from 3) and an easy induction. □

A maximal nested set $\mathbb{M}$ corresponds thus to a set of r divisors in X_Δ which, by [4], intersect transversally in a single point $P_{\mathbb{M}}$. Let us explicit the example of the positive roots of type A_{n-1}. We think of such a root as a pair (i, j) with $1 \leq i < j \leq n$. The irreducible subsets are indexed by subsets (which we display as sequences) $(i_1, \ldots, i_s)$, $s > 1$, with $1 \leq i_1 < \cdots < i_s \leq n$. To such a sequence corresponds the set S of pairs supported in the sequence. A family of subsets is nested if any two of them are either disjoint or one is contained in the other.

In this case a maximal nested set $\mathbb{M}$ has the following property. If $A \in \mathbb{M}$ has k elements and $k > 2$, we have two possibilities; either the maximal elements of $\mathbb{M}$ reduce to one subset with $k - 1$ elements or to two disjoint subsets A_1, A_2 with $A = A_1 \cup A_2$.

We define a map $\phi : \mathcal{I} \to \Delta$ by associating to each $S \in \mathcal{I}$ its minimum $\phi(S) := <(a \in S)$ with respect to the given ordering.

For example, in the root system case, with the ordering given before, we have that $\phi(S)$ is the highest root in S.

We come to the main new definition:

Definition 3. *A maximal nested set $\mathbb{M}$ is called proper if the set $\phi(\mathbb{M}) \subset \Delta$ is a basis of V.*

Example 1.4. In the A_n case with the previous ordering $\phi(i_1, \ldots, i_s) = x_{i_1} - x_{i_s} = (i_1, i_s)$.

A proper maximal nested set $\mathbb{M}$ is thus encoded by a sequence of $n - 1$ subsets each having at least two elements, with the property that, taking the minimum and maximum for each set, these pairs are all distinct.

It is easy to see how to inductively define a bijection between proper maximal nested sets and permutations of $1, \ldots, n$ fixing n. To see this consider $\mathbb{M}$ as a sequence $\{S_1, \ldots, S_{n-1}\}$ of subsets of $\{1, \ldots, n\}$ with the above properties. We can assume that $S_1 = (1, 2, \ldots, n)$ and have seen that $\mathbb{M}' := \mathbb{M} - \{S_1\}$ has either one or two maximal elements. If S_2 is the unique maximal element

and $1 \notin S_2$, by induction we get a permutation $p(\mathbb{M}')$ of $2, \dots, n$. We then set $p(\mathbb{M})$ equal to the permutation which fixes 1 and is equal to $p(\mathbb{M}')$ on $2, \dots, n$. If S_2 is the unique maximal element and $n \notin S_2$, we get, by induction, a permutation $p(\mathbb{M}')$ of $1, \dots, n-1$ fixing $n-1$. We then set $p(\mathbb{M})$ equal to the permutation which fixes n and is equal to $\tau p(\mathbb{M}')\tau$ on $S_2 = \{1, \dots, n-1\}$, τ being the permutation which reverses the order in S_2. If S_2 and S_3 are the two maximal elements so that $\{1, \dots, n\}$ is their disjoint union, and $1 \in S_2$, $n \in S_3$, then by induction we get two permutations p_2 and p_3 of S_2 and S_3 respectively. We then set $p(\mathbb{M})$ equal to p_3 on S_3 and equal to $\tau p_2 \tau$ on S_2 , τ being the permutation which reverses the order in S_2. In particular this shows that there are $(n-1)!$ proper maximal nested sets, which can be recursively constructed. This is the rank of the top cohomology of the complement of the corresponding hyperplane arrangement. We will see presently that this is a general phenomenon.

Remark 1.5. Notice that a proper maximal nested set inherits a total ordering from the total ordering of $\phi(\mathbb{M})$, and that this ordering is clearly a refinement of the partial ordering by reverse inclusion.

Now fix a maximal nested set $\mathbb{M}$. We clearly have:

Lemma 1.6. *Given $\alpha \in \Delta$, there exists a unique minimal irreducible $S \in \mathbb{M}$ such that $\alpha \in S$.*

This allows us to define a map $p_{\mathbb{M}} : \Delta \to \mathbb{M}$ by setting $p_{\mathbb{M}}(\alpha) := S$.

Definition 4. *If $\sigma \subset \Delta$ is a basis of V, we say that σ is adapted to $\mathbb{M}$ if the restriction of $p_{\mathbb{M}}$ to σ is a bijection.*

Notice that if $\mathbb{M}$ is proper, then the basis $\phi(\mathbb{M})$ is clearly adapted to $\mathbb{M}$.

2 A basis for homology

We have seen in Section 1 that, given a basis $\sigma = \{\gamma_1, \dots, \gamma_r\}$, we can associate to σ a maximal nested set which we now denote by $\eta(\sigma)$. $\eta(\sigma)$ is the decomposition of the flag $A_i = \Delta \cap \langle \gamma_i, \dots, \gamma_r \rangle$. Let us denote by $\mathcal{C}$ the set of no-broken-circuit bases of V, by $\mathcal{M}$ the set of proper maximal nested sets.

Lemma 2.1. *If a no-broken-circuit basis σ is adapted to a proper nested set $\mathbb{M} = \{S_1, \dots, S_r\}$, then $\sigma = \phi(\mathbb{M})$.*

Proof. Let $\sigma = \{\alpha_{i_1}, \dots, \alpha_{i_r}\}$. Clearly $i_1 = 1$, and α_{i_1} is the minimum element of Δ. Let A be the irreducible component of Δ containing α_1. We have that $A \in \mathbb{M}, \phi(A) = \alpha_1$ and so $A = S_1$. We claim that $p_{\mathbb{M}}(\alpha_1) = A$. This follows from the fact that $\mathbb{M}$ is proper so α_1 cannot be contained in two distinct elements A, B of $\mathbb{M}$, otherwise $\phi(A) = \phi(B)$. By Lemma 1.2, $\Delta' := S_2 \cup \cdots \cup S_r$

is complete. $A \not\subset \Delta'$ otherwise, still by 1.2 we would have that A is one of the S_i, $i \geq 2$. Since $\sigma' := \{\alpha_{i_2}, \ldots, \alpha_{i_r}\}$ is adapted to $\mathbb{M}' := \{S_2, \ldots, S_r\}$, we must have that the space $U_{\Delta'}$ spanned by Δ' is $r-1$ dimensional, $\{\alpha_{i_2}, \ldots, \alpha_{i_r}\}$ is a no-broken-circuit basis for $U_{\Delta'}$ relative to Δ' ordered by the total order induced from that of Δ and adapted to the proper nested set $\mathbb{M}'$. We can thus finish by induction. □

Theorem 2.2. *We have that η maps $\mathcal{C}$ to $\mathcal{M}$ and ϕ maps $\mathcal{M}$ to $\mathcal{C}$. Furthermore η and ϕ are bijections which are one the inverse of the other.*

Proof. Let $\sigma = \{\gamma_1, \ldots, \gamma_r\} \in \mathcal{C}$. By definition, for each i we have that γ_i is the minimum element in $A_i = \Delta \cap \langle \gamma_i, \ldots, \gamma_r \rangle$. Thus it is also the minimum element in one of the irreducibles decomposing A_i. It follows that $\eta(\sigma)$ is proper and that $\phi\eta(\sigma) = \sigma$.

Conversely, let $\mathbb{M} = \{S_1, \ldots, S_r\} \in \mathcal{M}$ and let $\gamma_i = \phi(S_i)$. By definition, the γ_i's are linearly independent, $\gamma_i < \gamma_{i+1}$ and $\mathbb{M}$ is the decomposition of the flag $A_i := \cup_{j \geq i} S_j$. We thus have by the definition of ϕ, that γ_i is the minimum element in A_i. Since A_i is complete we deduce that $\sigma = \{\gamma_1, \ldots, \gamma_r\} \in \mathcal{C}$. Clearly $\eta(\phi(\mathbb{M})) = \mathbb{M}$. □

Corollary 2.3. *A no-broken-circuit basis σ is adapted to a unique maximal proper nested set $\mathbb{M}$ and $\sigma = \phi(\mathbb{M})$.*

Let us now fix a basis $\sigma \subset \Delta$. Write $\sigma = \{\gamma_1, \ldots, \gamma_r\}$ and consider the r-form

$$\omega_\sigma := d\log\gamma_1 \wedge \cdots \wedge d\log\gamma_r.$$

This is a holomorphic form on the open set $\mathcal{A}_\Delta$ of U^* which is the complement of the arrangement formed by the hyperplanes whose equation is in Δ. In particular if $\mathbb{M} \in \mathcal{M}$, we shall set $\omega_{\mathbb{M}} := \omega_{\phi(\mathbb{M})}$.

Also if $\mathbb{M} \in \mathcal{M}$, we can define a homology class in $H_r(\mathcal{A}_\Delta, \mathbb{Z})$ as follows. Identify U^* with $\mathbb{A}^r$ using the coordinates $\phi(S)$, $S \in \mathbb{M}$. Consider another complex affine space $\mathbb{A}^r$ with coordinates z_S, $S \in \mathbb{M}$. In $\mathbb{A}^r$ take the small torus T of equation $|z_S| = \varepsilon$ for each $S \in \mathbb{M}$. Define a map

$$f : \mathbb{A}^r \to U^*, \quad \text{by} \quad \phi(S) := \prod_{S' \supset S} z_{S'}.$$

In [4] we have proved that this map lifts, in a neighborhood of 0, to a local system of coordinates of the model X_Δ . To be precise for a vector $\alpha \in \Delta$, set $B = p_{\mathbb{M}}(\alpha)$. In the coordinates z_S, we have that

$$\alpha = \sum_{B' \subset B} a_{B'} \prod_{S \supseteq B'} z_S = \prod_{S \supseteq B} z_S (a_B + \sum_{B' \subset B} a_{B'} \prod_{B \supsetneq S \supseteq B'} z_S) \tag{1}$$

with $a_{B'} \in \mathbb{C}$ and $a_B \neq 0$. Set $f_{M,\alpha}(z_S) := a_B + \sum_{B' \subset B} a_{B'} \prod_{B \supsetneq S \supseteq B'} z_S$ and let $A_{\mathbb{M}}$ be the complement in the affine space $\mathbb{A}^r$ of coordinates z_S of the hypersurfaces of equations $f_{M,\alpha}(z_S) = 0$. The main point is that $A_{\mathbb{M}}$ is an open set of X_Δ. The point 0 in $A_{\mathbb{M}}$ is the *point at infinity* $P_{\mathbb{M}}$. The open set $\mathcal{A}_\Delta$ is contained in $A_{\mathbb{M}}$ as the complement of the divisor with normal crossings given by the equations $z_S = 0$. From this one sees immediately that if ε is sufficiently small, f maps T homeomorphically into $\mathcal{A}_\Delta$. Let us give to T the obvious orientation coming from the total ordering of $\mathbb{M}$, so that $H_r(T, \mathbb{Z})$ is identified with $\mathbb{Z}$ and set $c_{\mathbb{M}} = f_*(1) \in H_r(\mathcal{A}_\Delta, \mathbb{Z})$.

Proposition 2.4. *Let $\sigma = \{\gamma_1, \dots, \gamma_r\} \subset \Delta$ be a basis of V. Let $\mathbb{M} \in \mathcal{M}$. Then*

1) If σ is not adapted to $\mathbb{M}$,

$$\int_{c_{\mathbb{M}}} \omega_\sigma = 0.$$

2) If σ is adapted to $\mathbb{M}$, consider the sequence $p_{\mathbb{M}}(\gamma_1), \dots, p_{\mathbb{M}}(\gamma_r)$. This is a permutation π of the totally ordered set $\mathbb{M}$ and we denote by $s(\mathbb{M}, \sigma)$ its sign. Then

$$\frac{1}{(2\pi i)^r} \int_{c_{\mathbb{M}}} \omega_\sigma = s(\mathbb{M}, \sigma).$$

Proof. Given $\alpha \in \Delta$, from equation (1) we deduce that, in the neighborhood $A_{\mathbb{M}}$, the 1-form $d \log \alpha$ equals the sum of the 1-form $\sum_{S \supseteq B} d \log z_S$ and of a 1-form $\psi_B := d \log(a_B + \sum_{B' \subset B} a_{B'} \prod_{B \supseteq S \supseteq B'} z_S)$ which is exact and holomorphic on the solid torus in $\mathbb{A}^r$ defined by $|z_S| \leq \varepsilon$.

When we substitute these expressions in the linear forms $d \log \gamma_i$ and expand the product ω_σ we obtain various terms. Some terms vanish since we repeat twice a factor $d \log z_S$, some terms contain a factor ψ_B hence they are exact. The only possible contribution which gives a non-exact form is when σ is adapted to $\mathbb{M}$, and then it is given by the term $s(\mathbb{M}, \sigma)\omega_{\mathbb{M}}$. From this observation both 1) and 2) easily follow. □

Given the class $c_{\mathbb{M}}$ and an r-dimensional differential form ψ we can compute $\int_{c_{\mathbb{M}}} \psi$. Denoting by $P_{\mathbb{M}}$ the point at infinity corresponding to 0 in the previously constructed coordinates $z_i := z_{S_i}$ we shall say:

Definition 5. *The integral $\frac{1}{(2\pi i)^r} \int_{c_{\mathbb{M}}} \psi$ is called the* **residue** *of ψ at the point at infinity $P_{\mathbb{M}}$. We will also denote it by $res_{\mathbb{M}}(\psi)$.*

Notice that the rational forms, in a neighborhood of the point $P_{\mathbb{M}}$ and in the coordinates z_i, have the form $\psi = f(z_1, \dots, z_r) dz_1 \wedge \cdots \wedge dz_r$ with $f(z_1, \dots, z_r)$ a Laurent series which can be explicitly computed. One then gets that the residue $res_{\mathbb{M}}(\psi)$ equals the coefficient of $(z_1 \dots z_r)^{-1}$, in this series.

We can summarize this section with the main theorem.

Theorem 2.5. *The set of elements $c_{\mathbb{M}}$, $\mathbb{M} \in \mathcal{M}$ is the basis of $H_r(\mathcal{A}_\Delta, \mathbb{Z})$, dual, under the residue pairing, to the basis given by the forms $\omega_{\phi(\mathbb{M})}$: the forms associated to the no-broken-circuit bases relative to the given ordering.*

Proof. This is a consequence of Theorem 2.2, Corollary 2.3 and Proposition 2.4. □

Remarks 2.6.

1. The formulas found give us an explicit formula for the projection π of $\Omega_r(\mathcal{A}_\Delta) = H^r \oplus d\Omega_{r-1}(\mathcal{A}_\Delta)$ to H^r with kernel $d\Omega_{r-1}(\mathcal{A}_\Delta)$. We have:
$$\pi(\psi) = \sum_{\mathbb{M} \in \mathcal{M}} res_{\mathbb{M}}(\psi)\omega_{\mathbb{M}}. \tag{2}$$
2. Using the projection π any linear map on H^r, in particular the Jeffrey–Kirwan residue (see below), can be thought of as a linear map on $\Omega_r(\mathcal{A}_\Delta)$ vanishing on $d\Omega_{r-1}(\mathcal{A}_\Delta)$. Our geometric description of homology allows us to describe any such map as integration on a cycle, that is a linear combination of the cycles $c_{\mathbb{M}}$.
3. There are several possible applications of these formulas to combinatorics and counting integer points in polytopes. The reader is referred to [2],[1].
4. We have treated only top homology but all homology can be described in a similar way due to the fact that for each k the k^{th} cohomology decomposes into the contributions relative to the subspaces of codimension k and the corresponding transversal configuration.

3 The Jeffrey–Kirwan residue

In this section V is a real r-dimensional vector space and $U := V \otimes_{\mathbb{R}} \mathbb{C}$, $\Delta = \{\alpha_1, \ldots, \alpha_n\} \subset V$.

Now let us assume that we have fixed once and for all an orientation of V^* by choosing an ordered basis $\xi = (x_1, \ldots, x_r)$ of V and taking the orientation form $d\underline{x} = dx_1 \wedge dx_2 \wedge \cdots \wedge dx_r$.

This gives a canonical way of identifying the r-forms with functions on $\mathcal{A}_\Delta$. The form $\omega_\sigma = d\log\gamma_1 \wedge \cdots \wedge d\log\gamma_r$ is identified with the function $d_\sigma^{-1} \prod_i \gamma_i^{-1}$ where d_σ is the determinant of the matrix expressing the basis σ in terms of the basis ξ. Let $\tilde{H}^r$ denote the space spanned by these functions.

We now further restrict to the case in which there exists a linear function on V which is positive on Δ, i.e., that all the elements in Δ are on the same side of some hyperplane. In this case there is another interesting way of representing H^r in which the Jeffrey–Kirwan residue appears in a very natural way. This is done via the Laplace transform $\int e^{-(x,y)} f(y)dy$ which in our setting has to be understood as a transform from functions on V^* with

a prescribed invariant Lebesgue measure (the one induced by ξ) to functions on V (more intrinsically to r differential forms).

Precisely, consider the cone C spanned by the vectors in Δ. For each basis $\sigma := \{\gamma_1, \dots, \gamma_r\}$ extracted from Δ let $C(\sigma)$ be the positive cone that it generates and χ_σ its characteristic function. Let finally K^r be the vector space spanned by the functions χ_σ. From the basic formula

$$\int_0^\infty \dots \int_0^\infty e^{-\sum_{i=1}^r x_i y_i} dy_1 \dots dy_r = \frac{1}{x_1 \dots x_r}$$

and linear coordinate changes, it is easy to verify that the Laplace transform of χ_σ is $|d_\sigma|^{-1} \prod_i \gamma_i^{-1}$, as a function on the dual positive cone, consisting of all x such that $(x, y) > 0,\ \forall y \in C$. Therefore combining with the isomorphism of $\tilde{H}^r$ with H^r we have a Laplace transform $L : K^r \to H^r$ with $L(\chi_\sigma) = \nu_\sigma \omega_\sigma$, where $\nu_\sigma := d_\sigma / |d_\sigma|$ equals 1 if the ordered basis σ has the same orientation as ξ, -1 otherwise.

L is a linear isomorphism in which it is easy to reinterpret geometrically the linear relations previously described. Finally the Jeffrey–Kirwan residue is a linear function $\psi \to J\langle c \,|\, \psi \rangle$ on H^r depending on a regular vector c. It corresponds to the linear function defined on K^r which just consists in evaluating the functions f in c. In other words $J\langle c \,|\, \psi \rangle = L^{-1}(\psi)(c)$. By the definition of K^r it is clear that this linear function depends only on the chamber C in which c lies. Our final result is the description of a geometric cycle $\delta(C)$ such that $J\langle c \,|\, \psi \rangle = \frac{1}{(2\pi i)^r} \int_{\delta(C)} \psi$.

For a proper maximal nested set $\mathbb{M}$ we denote $\nu_{\phi(\mathbb{M})}$ by $\nu_{\mathbb{M}}$. For each basis $\tau \subset \Delta$, set $C(\tau) = \{x \in V \,|\, x = \sum_{\alpha \in \tau} a_\alpha \alpha, a_\alpha > 0\}$. Set for simplicity, for a proper maximal nested set $\mathbb{M}$, $C(\mathbb{M}) := C(\phi(\mathbb{M}))$.

Before giving our description of $\delta(C)$, let us recall some facts from [6]. Assume that in V we have a lattice Γ which we interpret as the character group of an r-dimensional torus T. Assume that $\Delta \subset \Gamma$ is a set of characters and that the basis ξ is a basis of Γ.

We have the following sequence of ideas. First of all we use the elements $\alpha_i,\ i = 1, \dots, n$ to construct an n-dimensional representation Z of T as the direct sum of the 1-dimensional representations with character α_i^{-1}. Call $R = \mathbb{C}[t_1, \dots, t_n]$ the ring of polynomial functions on Z. On Z we have an action of the n-dimensional torus D_n of diagonal matrices and T acts via a homomorphism into D_n. Hence the torus T acts on R and t_i has weight α_i. If $\gamma \in \Gamma$ is a character, define $R(\gamma)$ to be the subspace of R of weight γ with respect to T.

We shall denote by $\mathcal{R}$ the set of regular vectors, i.e., the set of vectors which cannot be written as a linear combination of the elements in a subset S of Δ of cardinality smaller than r.

Consider a regular vector $\zeta \in \Gamma$ and let $R_\zeta = \oplus_{k=0}^\infty R(k\zeta)$.

The following facts are well known [3]. R_ζ is a finitely generated subalgebra stable under the torus D_n. So, if we grade R_ζ so that $R(k\zeta)$ has degree k, we can consider the projective variety $T_\zeta := Proj(R_\zeta)$ with a line bundle $\mathcal{L}$ such

that $H^0(T_\zeta, \mathcal{L}^{\otimes k}) = R(k\zeta)$. T_ζ is an embedding of the $n-r$ dimensional torus D_n/T and the regularity of ζ implies that T_ζ is an orbifold. The elements α_i index the boundary divisors of this torus embedding. Thus to each α_i we can associate a degree 2 cohomology class, the Chern class of the corresponding divisor, which is still expressed with the same symbol α_i. According to the theory developed by Jeffrey and Kirwan discussed in [3], one can compute the intersection numbers $\int_{T_\zeta} P(\alpha_1, \ldots, \alpha_n)$ using the notion of Jeffrey–Kirwan residue. Denoting by C the chamber in which ζ lies, one has:

$$\int_{T_\zeta} P(\alpha_1, \ldots, \alpha_n) = J\langle c \,|\, \frac{P(\alpha_1, \ldots, \alpha_n)}{\alpha_1 \alpha_2 \ldots \alpha_n} d\mu \rangle. \tag{3}$$

We want to represent this residue by integration over a cycle:

$$J\langle c \,|\, \psi \rangle = \frac{1}{(2\pi i)^r} \int_{\delta(C)} \psi. \tag{4}$$

From the formula $L(\chi_\sigma) = \nu_\sigma \omega_\sigma$, and definition of the Jeffrey–Kirwan residue discussed in the introduction, we see that for every basis $\sigma = \{\gamma_1, \ldots, \gamma_r\} \subset \Delta$ of V, the value of $\delta(C)$ on the r-form ω_σ, must be given by

$$\frac{1}{(2\pi i)} \int_{\delta(C)} \omega_\sigma = \begin{cases} 0 \text{ if } C \cap C(\sigma) = \emptyset, \\ \nu_\sigma \text{ if } C \subset C(\sigma). \end{cases} \tag{5}$$

Using this description of $\delta(C)$ and the fact that our homology basis $c_{\mathbb{M}}$ is dual to the cohomology basis $\omega_{\phi(\mathbb{M})}$, one immediately has:

Theorem 3.1.

$$\delta(C) = \sum_{\mathbb{M} \in \mathcal{M} | C \subset C(\mathbb{M})} \nu_{\mathbb{M}} c_{\mathbb{M}}.$$

References

1. W. Baldoni-Silva, J. De Loera and M. Vergne – "Counting Integer flows in Networks", 2003, `math.CO/0303228`.
2. W. Baldoni-Silva and M. Vergne – "Residues formulae for volumes and Ehrhart polynomials of convex polytopes", 2001, `math.CO/0103097`.
3. M. Brion and M. Vergne – "Arrangement of hyperplanes. I. Rational functions and Jeffrey-Kirwan residue", *Ann. Sci. École Norm. Sup. (4)* **32** (1999), no. 5, p. 715–741.
4. C. De Concini and C. Procesi – "Wonderful models of subspace arrangements", *Selecta Math. (N.S.)* **1** (1995), no. 3, p. 459–494.
5. P. Orlik and H. Terao – *Arrangements of hyperplanes*, Grundlehren der Mathematischen Wissenschaften, vol. 300, Springer-Verlag, Berlin, 1992.
6. A. Szenes and M. Vergne – "Toric reduction and a conjecture of Batyrev and Materov", 2003, `math.AT/0306311`.
7. A. Szenes – "Iterated residues and multiple Bernoulli polynomials", *Internat. Math. Res. Notices* (1998), no. 18, p. 937–956.

Counting extensions of function fields with bounded discriminant and specified Galois group

Jordan S. Ellenberg[1] and Akshay Venkatesh[2]

[1] Department of Mathematics, Princeton University, Princeton NJ 08544, U.S.A.
ellenber@math.princeton.edu
[2] Department of Mathematics, Massachusetts Institute of Technology, Cambridge MA 02139, U.S.A.
akshayv@math.mit.edu

Summary. We discuss the enumeration of function fields and number fields by discriminant. We show that Malle's conjectures agree with heuristics arising naturally from geometric computations on Hurwitz schemes. These heuristics also suggest further questions in the number field setting.

1 Introduction

The enumeration of number fields subject to various local and global conditions is an old problem, which has in recent years been the subject of renewed interest (a sampling includes [3], [2], [5], [6], [9], [12].) For a good survey of recent work, see [1]. We begin by reprising some important conjectures.

If L/K is an extension of number fields, we denote by $\mathcal{D}_{L/K}$ the relative discriminant, an ideal of K, and by $\mathbf{N}_{\mathbb{Q}}^{K}\mathcal{D}_{L/K}$ its norm, a positive integer. For $X \in \mathbb{R}^{+}$, we set $N_{K,n}(X)$ to be the number of degree-n extensions L/K (up to K-isomorphism) such that $\mathbf{N}_{\mathbb{Q}}^{K}\mathcal{D}_{L/K} < X$. It is a classical problem to understand the asymptotics of $N_{K,n}(X)$ as X goes to infinity; in particular, we have the folk conjecture:

Conjecture 1.1. *There is a constant $c_{K,n}$ such that, as $X \to \infty$,*

$$N_{K,n}(X) \sim c_{K,n}X.$$

This conjecture is now known for $n \leq 5$.

A more general conjecture applies to enumerating extensions with specified Galois group. It is due to Malle [11] and refines a previous conjecture of Cohen. To describe Malle's conjecture, we need to introduce some notation.

Let $G \leq S_n$ be a transitive subgroup. For $g \in G$, set $ind(g) = n - r$, where r is the number of orbits of g on $\{1, 2, \ldots, n\}$. Denote by $\mathcal{C}$ the set of non-trivial

conjugacy classes of G; then ind descends to a function $ind : \mathcal{C} \to \mathbb{Z}$. The group $Gal(\bar{K}/K)$ acts on $\mathcal{C}$ via $g \cdot c = c^{\chi(g)}$, where $g \in Gal(\bar{K}/K), c \in \mathcal{C}$ and $\chi : Gal(\bar{K}/K) \to \hat{\mathbb{Z}}^*$ is the cyclotomic character. Set $a(G) = \max_{c \in \mathcal{C}}(ind(c)^{-1})$, and set $b_K(G)$ to be the number of $Gal(\bar{K}/K)$-orbits on the set $\{c \in \mathcal{C} : ind(c) = 1/a(G)\}$.

Let H be any point stabilizer in the G-action on $\{1, 2, \ldots, n\}$. For each Galois extension L/K with Galois group G, let L_0/K be the degree n subextension of L/K corresponding to the subgroup $H \leq G$. Since G acts transitively on $\{1, 2, \ldots, n\}$, the K-isomorphism class of L_0 is independent of the choice of H. We then denote by $N_{K,G}(X)$ the number of Galois G-extensions L/K such that $\mathbf{N}_{\mathbb{Q}}^{K} \mathcal{D}_{L_0/K} < X$.

Conjecture 1.2. *(Malle) There is a nonzero constant $C_K(G)$ such that*

$$N_{K,G}(X) \sim C_K(G) X^{a(G)} (\log X)^{b_K(G)-1}.$$

This conjecture is known to be correct in certain special cases, including that where $G = S_3$ or D_4 (embedded in S_3 and S_4 respectively) and that where G is abelian. In general, however, little is known about Malle's conjecture – and indeed, its difficulty is ensured by the fact that it implies a positive solution to the inverse Galois problem.

A related problem, raised for example in [8], is the question of multiplicity of a fixed discriminant.

Conjecture 1.3. *The number of number fields $K/\mathbb{Q}$ with degree n and discriminant D is $\ll_{\epsilon,n} D^{\epsilon}$.*

Conjecture 1.3 is unknown, and seems quite difficult, even for $n = 3$. In that case it is intimately related to questions about 3-torsion in class groups of quadratic fields.

The arithmetic of function fields and their covers is often much more approachable than that of number fields, since one can appeal to the geometry of varieties over finite fields. In particular, one may replace K by $\mathbb{F}_q(t)$ in the above discussion, and ask whether Conjectures 1.1 and 1.2 remain true (with evident modifications) in this setting. We note that this is known to be the case when $G = S_3$, by the work of Datskovsky and Wright [6].

We do not know how to prove Conjecture 1.2 even in the function field setting. However, we will establish in the present paper certain (weak) approximations to Conjecture 1.2. In Lemma 2.4 we show that the upper bound of Malle's conjecture is nearly valid when q is large relative to $|G|$. Moreover, we prove in Proposition 3.1 a result showing that Malle's conjecture is compatible with a heuristic arising from the geometry of Hurwitz spaces. A little more precisely, Proposition 3.1 studies Malle's conjecture using the following heuristic:

(A) If X is a geometrically irreducible d-dimensional variety over $\mathbb{F}_q$, one has $|X(\mathbb{F}_q)| = q^d$.

The heuristic **(A)** can be thought of as an assertion of extremely (indeed, implausibly) strong cancellation between Frobenius eigenvalues on the cohomology of X. Despite its crudeness, **(A)** allows one to recover, in the function field setting, the precise constants $a(G)$ and $b_K(G)$ found in Malle's conjecture.

This line of reasoning suggests further questions about the distribution of discriminants of number fields. We discuss these in Section 4. For instance, Section 4.2 gives a heuristic for the number of icosahedral modular forms of conductor $\leq N$, and Section 4.3 proposes some still more general heuristics for number fields with prescribed ramification data.

We note that the approach via **(A)** is very much in the spirit of that used by Batyrev in developing precise heuristics for the distribution of rational points on Fano varieties; we thank Yuri Tschinkel for explaining this to us.

Acknowledgments. The authors thank Karim Belabas, Manjul Bhargava, Henri Cohen, and Johan de Jong for many useful conversations about the topic of this chapter, and the organizers of the Miami Winter School in Geometric Methods in Algebra and Number Theory for inviting the first author to give the lecture on which this article is based.

The first author was partially supported by NSF Grant DMS-0401616 and the second author by NSF Grant DMS-0245606.

Notation: Throughout this paper, G will be a transitive subgroup of the permutation group S_n and q will be a prime power that is coprime to $|G|$.

2 Counting extensions of function fields

2.1 Hurwitz spaces

In this section, we recall basic facts about Hurwitz spaces, i.e., moduli spaces for covers of $\mathbb{P}^1$. We will make constant use of the fact that the category of finite extensions $L/\mathbb{F}_q(t)$, with the morphisms being field homomorphisms fixing $\mathbb{F}_q(t)$, is equivalent to the category of finite (branched) covers of smooth curves $f : Y \to \mathbb{P}^1$ defined over $\mathbb{F}_q$, the morphisms being maps of covers over $\mathbb{P}^1$. Recall that q is coprime to $|G|$, eliminating painful complications concerning the residue characteristic.

Let Y be a geometrically connected curve over $\mathbb{F}_q$ and $f : Y \to \mathbb{P}^1$ a Galois covering equipped with an isomorphism $G \to \mathrm{Aut}(Y/\mathbb{P}^1)$. We refer to such a pair (Y, f) as a *G-cover.* Let H be a point stabilizer in the G-action on $\{1, 2, \ldots, n\}$, and let $f_0 : Y_0 \to \mathbb{P}^1$ be the degree-n covering corresponding to the subgroup $H \leq G$. We then set $r(f)$ to be the degree of the ramification divisor of f_0. Call $q^{r(f)}$ the discriminant of f.

We denote by $N_{q,G}(X)$ the number of isomorphism classes of G-covers $f : Y \to \mathbb{P}^1/\mathbb{F}_q$ with $q^{r(f)} < X$. Note that, by requiring that Y be geometrically connected, we have excused ourselves from counting extensions of $\mathbb{F}_q(t)$

which contain some $\mathbb{F}_{q^f}/\mathbb{F}_q$ as a subextension. This decision will not affect the powers of X and $\log X$ in the heuristics we compute, though it may change the constant terms.

The G-covers $\mathbb{P}^1$ with discriminant q^r are parametrized by a Hurwitz variety $\mathcal{H}_r$. More precisely:

Proposition 2.1. *There is a smooth scheme $\mathcal{H}_r$ over $\mathbb{Z}[\frac{1}{|G|}]$ which is a coarse moduli space for G-covers of $\mathbb{P}^1$ with discriminant r. The natural map*

$$\{\textit{isomorphism classes of } G\textit{-covers of } \mathbb{P}^1/\mathbb{F}_q\} \to \mathcal{H}(\mathbb{F}_q) \tag{1}$$

is surjective, and the fibers have size at most $|Z|$, where Z is the center of G.

Proof. We refer to [16] for details of the construction of $\mathcal{H}_r$ in positive characteristic. Let h be an $\mathbb{F}_q$-rational point of $\mathcal{H}$. Then the obstruction to h arising from a cover $Y \to \mathbb{P}^1$ defined over $\mathbb{F}_q$ lies in $H^2(\mathbb{F}_q, Z)$ where Z is the center of G; since $Gal(\bar{\mathbb{F}}_q/\mathbb{F}_q)$ has cohomological dimension 1, this obtruction is trivial (see [7, Cor. 3.3] for more discussion of this point.) Further, the isomorphism classes of covers f parametrized by the point h are indexed by the cohomology group $H^1(\mathbb{F}_q, Z)$, which has size at most $|Z|$. □

What's more, $\mathcal{H}_r$ is the union of open and closed subschemes which parametrize G-covers with specified ramification data. In order to express this decomposition, we need a bit more notation.

We call a multiset $\underline{c} = \{c_1, \dots, c_k\}$ of conjugacy classes of G a *Nielsen class*, and denote by $r(\underline{c})$ the total index $\sum_{i=1}^k ind(c_i)$. We also write $|\underline{c}|$ for the number of branch points k. Finally, for each Nielsen class $\underline{c}$ we define $\tilde{\Sigma}_{\underline{c}}$ to be the subset of G^k consisting of all k-tuples $(g_1, \dots, g_k)$ such that

- The multisets $\underline{c}$ and $\{c(g_i), \dots, c(g_k)\}$ are equal, where $c(g)$ denotes the conjugacy class of g;
- $g_1 g_2 \dots g_k = 1$;
- the g_i generate G.

Note that $\tilde{\Sigma}_{\underline{c}}$ is preserved by the action of G given by

$$(g_1, \dots, g_k) \mapsto (gg_1g^{-1}, \dots, gg_kg^{-1}).$$

We denote by $\Sigma_{\underline{c}}$ the quotient of $\tilde{\Sigma}_{\underline{c}}$ by this action.

Let $f : Y \to \mathbb{P}^1_{\bar{\mathbb{F}}_q}$ be a G-cover whose branch locus in $\mathbb{P}^1(\bar{\mathbb{F}}_q)$ is $\{x_1, \dots, x_k\}$. By consideration of the action of tame inertia at $x_1, \dots, x_k$, we can associate a Nielsen class $\underline{c}$ to f which is fixed by $Gal(\bar{K}/K)$ and which satisfies $r(\underline{c}) = r(f)$ [4, 1.2.4]. The set of Nielsen classes inherits a $Gal(\bar{\mathbb{F}}_q/\mathbb{F}_q)$-action from the cyclotomic action on $\mathcal{C}$, as described in Section 1; we call a Nielsen class which is fixed by this action an *$\mathbb{F}_q$-rational Nielsen class.* If f descends to a G-cover $Y \to \mathbb{P}^1_{\mathbb{F}_q}$, it follows that the Nielsen class $\underline{c}$ is $\mathbb{F}_q$-rational.

Denote by C_k the configuration space of k disjoint points in $\mathbb{P}^1$. The (geometric) fundamental group of C_k is the (spherical) braid group on k-strands. We denote by $\sigma_k \in C_k$ the braid that pulls strand i past strand $i+1$.

Proposition 2.2. *For each Nielsen class $\underline{c}$, there is a Hurwitz space $\mathcal{H}_{\underline{c}}/\bar{\mathbb{F}}_q$ which is a coarse moduli space for G-covers $f : Y \to \mathbb{P}^1_{\bar{\mathbb{F}}_q}$ with Nielsen class $\underline{c}$. The action of $\sigma \in Gal(\bar{\mathbb{F}}_q/\mathbb{F}_q)$ sends $\mathcal{H}_{\underline{c}}$ to $\mathcal{H}_{\underline{c}^\sigma}$; so the $\mathbb{F}_q$-rational connected components of $\mathcal{H}_r$ are each contained in $\mathcal{H}_{\underline{c}}$ for an $\mathbb{F}_q$-rational $\underline{c}$ with $r(\underline{c}) = r$.*

The map $\pi : \mathcal{H}_{\underline{c}} \to C_{|\underline{c}|}$ that sends a cover f to its ramification divisor is étale. Moreover, the geometric points of the fiber of π above $\{x_1, \dots, x_k\} \in C_k$ are naturally identified with $\Sigma_{\underline{c}}$. The action of $\pi_1(C_k)$ on $\pi^{-1}(\{x_1, \dots, x_k\})$ is given by

$$\sigma_i(g_1, \dots, g_k) = (g_1, \dots, g_i g_{i+1} g_i^{-1}, g_i, \dots, g_k)$$

so that the connected components of $\mathcal{H}_{\underline{c}}$ are in bijection with the $\pi_1(C_k)$-orbits on $\Sigma_{\underline{c}}$.

Proof. For the existence of $\mathcal{H}_{\underline{c}}$, see [4, §1.2.4]. The description of the connected components of $\mathcal{H}_{\underline{c}}$ is due to Fried; see, e.g., [10, §1.3], and [16, Cor 4.2.3] for the extension of Fried's results to positive characteristic prime to $|G|$. □

2.2 An upper bound on the number of extensions of $\mathbb{F}_q(t)$

Proposition 2.1 shows that, up to a constant factor, one can reduce the problem of controlling $N_{\mathbb{F}_q(t),G}(X)$ to the problem of controlling the number of $\mathbb{F}_q$-rational points on the varieties $\mathcal{H}_r$, as r ranges up to $\log_q X$. Bounding the number of $\mathbb{F}_q$-points on a variety of high dimension over a small finite field is a difficult matter. In the context at hand, we may give a straightforward upper bound, but the exponent is far from the one appearing in Malle's conjecture. We carry this out below; to clarify matters, we fix q and G and consider only the dependence as $X \to \infty$.

We will use the following easy lemma to bound various sequences arising in this paper.

Lemma 2.3. *Suppose $\{a_n\}$ is a sequence of real numbers with $a_n = 0$ whenever n is not a power of q, and suppose*

$$\sum_{r=1}^{\infty} a_{q^r} q^{-rs},$$

considered as a formal power series, is a rational function $f(t)$ of $t = q^s$. Let a be a positive real number. If $f(t)$ has no poles with $|t| \geq q^a$, then:

$$\sum_{n=1}^{X} a_n \ll X^a.$$

If $f(t)$ has a pole of order b at $t = q^a$ and no other poles with $|t| \geq q^a$, then:

$$\sum_{n=1}^{X} a_n \asymp X^a (\log X)^{b-1}.$$

Here we use the notation $A(X) \asymp B(X)$ to mean that there are real constants $C_1, C_2 > 0$ such that $C_1 A(X) \leq B(X) \leq C_2 A(X)$.

Proof. It follows immediately from the decomposition of $f(t)$ in partial fractions that

$$\sum_{r=1}^{R} a_{q^r} \ll q^{aR}$$

when $f(t)$ has no poles with $|t| > q^a$. Moreover, if $f(t)$ has a pole of order b at $t = q^a$ and no other poles with $|t| \geq q^a$, then

$$\sum_{r=1}^{R} a_{q^r} \sim C q^{aR} R^{b-1}$$

for some $C \in \mathbb{R}$. Then the lemma follows, since $q^{\lfloor \log_q X \rfloor} \asymp X$. □

Lemma 2.4. *Let q and G be fixed. Denote by $E(j)$ the number of elements g of G with $ind(g) = j$, and set $e(G) = \sup_j E(j)^{1/j}$. Then*

$$\limsup_{X \to \infty} \frac{\log N_{q,G}(X)}{\log X} \leq a(G) + \frac{\log(2e(G))}{\log q}.$$

In particular

$$\limsup_{X \to \infty} \frac{\log N_{q,S_n}(X)}{\log X} \leq 1 + \frac{\log(4n^2)}{\log q}. \tag{2}$$

Note that the right-hand-side of the first inequality in Lemma 2.4 approaches Malle's constant $a(G)$ when q becomes large relative to $|G|$.

Proof. Define a sequence of integers a_n such that $a_{q^r} = |\mathcal{H}_r(\mathbb{F}_q)|$ and $a_n = 0$ if n is not a power of q. So

$$N_{q,G}(X) \asymp \sum_{n=1}^{X} a_n.$$

We have seen in Proposition 2.2 that the $\mathbb{F}_q$-rational components of $\mathcal{H}_r$ are the union of Hurwitz varieties $\mathcal{H}_{\underline{c}}/\mathbb{F}_q$. Since $\mathcal{H}_{\underline{c}}$ is a finite cover of degree $|\Sigma_{\underline{c}}|$ of $C_{|\underline{c}|} \cong \mathbb{P}^{|\underline{c}|}/\mathbb{F}_q$, we have

$$|\mathcal{H}_{\underline{c}}(\mathbb{F}_q)| \ll_{q,G} |\Sigma_{\underline{c}}| q^{|\underline{c}|}$$

and

$$a_{q^r} \ll_{q,G} \sum_{\underline{c}: r(\underline{c}) = r} |\Sigma_{\underline{c}}| q^{|\underline{c}|}.$$

Let $f(r)$ be the sum of q^k over all k-tuples $(g_1, \dots, g_k)$ in G satisfying $\sum_i ind(g_i) = r$. (Here, k is allowed to vary.) Then evidently

$$\sum_{\underline{c}: r(\underline{c})=r} |\Sigma_{\underline{c}}| q^{|\underline{c}|} \leq f(r).$$

On the other hand,

$$\sum_r f(r) q^{-rs} = (1 - \sum_{g \in G} (q^{1-ind(g)s}))^{-1}.$$

We conclude that

$$\sum_r a_{q^r} q^{-rs} \ll_{q,G} (1 - \sum_{g \in G} (q^{1-ind(g)s}))^{-1} = (1 - \sum_{j \geq a(G)^{-1}} E(j) q^{1-js})^{-1}. \quad (3)$$

It is easy to see that (3) has no poles once we have

$$|q^s| > 2q^{a(G)} E(j)^{1/j}$$

for every j. The first part of the proposition now follows from Lemma 2.3.

We now show that, when $G = S_n$, we have $E(j)^{1/j} < 2n^2$ for all j; this proves the second part of the lemma. Any $\sigma \in S_n$ with $ind(\sigma) = j$ fixes at least $n - 2j$ elements of $\{1, 2, \dots, n\}$. Enumerating such σ by their number l of fixed points, we obtain $E(j) \leq \sum_{n-2j \leq l \leq n-1} \frac{n!}{l!} < 2jn^{2j}$. Thus $E(j)^{1/j} < n^2 (2j)^{1/j} \leq 2n^2$. □

Remark 2.5. It is interesting to contrast the "trivial" upper bounds of Lemma 2.4 with what can be obtained in the number field setting.

The upper bounds of Lemma 2.4 used explicit knowledge of the fundamental group of a punctured $\mathbb{P}^1$. In the number field setting, such tools are unavailable. Nevertheless in [9] an upper bound for $N_n(X)$ was derived, similar to (2), with the exponent $\log(n)$ replaced by a quantity of the form $e^{\sqrt{\log(n)}}$. The proof was considerably more complicated, but nevertheless geometric: the key idea is to find in each number field K a small set $\{x_1, x_2, \dots, x_r\}$ of algebraic integers which are "nondegenerate" in the sense that they do not satisfy an algebraic relation of low degree, and then to show that an appropriate set of traces $\mathrm{Tr}(x_1^{g_1} \dots x_r^{g_r})$ suffice to determine K.

Further, let $N_{q,n}^{\mathrm{Gal}}(X)$ denote the number of *Galois* extensions of $\mathbb{P}^1_{\mathbb{F}_q}$ of degree n and discriminant less than X. Lemma 2.4 implies that $N_{q,n}^{\mathrm{Gal}}(X) \ll_{q,n} X^{\frac{2}{n} + \frac{\log(2n)}{\log(q)}}$. Again, a result of a similar flavor was shown in [9], where it was shown that $N_{q,n}^{\mathrm{Gal}}(X) \ll X^{3/8}$ if $n \geq 3$. Again, the proof in the number field case was more elaborate and in fact relied on the classification of finite simple groups; the main idea is to prove the theorem using a low-degree permutation representation of G when G is simple, and to proceed by induction on a composition series otherwise.

3 Counting points on Hurwitz spaces under heuristic (A)

Lemma 2.4 asserts, at least, that the upper bound of Malle's conjecture is close to valid when q is large compared to $|G|$. Beyond Lemma 2.4, we can do no more than speculate about the exact number of $\mathbb{F}_q$-points on $\mathcal{H}_r$. The situation improves somewhat if we are willing to assume the heuristic (**A**) from the introduction: that is, we suppose that a geometrically irreducible d-dimensional variety over $\mathbb{F}_q$ has q^d points. This heuristic reduces the problem of estimating $|\mathcal{H}_r(\mathbb{F}_q)|$ to the substantially simpler problem of computing the number of geometric connected components of the spaces $\mathcal{H}_r$ and their fields of definition.

Let $h(q,r)$ be the sum of $q^{\dim C}$ over all geometrically connected components C of $\mathcal{H}_r$ which are defined over $\mathbb{F}_q$. Denote by $b_{\mathbb{F}_q}(G)$ the number of $Gal(\bar{\mathbb{F}}_q/\mathbb{F}_q)$-orbits on the set $\{c \in \mathcal{C} : ind(c) = 1/a(G)\}$.

We shall prove:

Proposition 3.1.

$$\sum_{q^r \leq X} h(q,r) \asymp X^{a(G)} \log(X)^{b_{\mathbb{F}_q}(G)-1}.$$

Proposition 3.1 amounts, roughly speaking, to the assertion that Malle's conjectures are compatible with naive dimension computations for Hurwitz spaces. The proof is more difficult than that of Lemma 2.4 but is still elementary.

The problem here is that the decomposition of $\mathcal{H}_r$ into geometrically connected components is somewhat subtle. Let $h'(q,r)$ be the sum of $q^{|\underline{c}|}$ over all $\mathbb{F}_q$-rational Nielsen classes $\underline{c}$ with $r(\underline{c}) = r$. If $\mathcal{H}_{\underline{c}}$ were a nonempty geometrically connected variety for every $\mathbb{F}_q$-rational Nielsen class $\underline{c}$ with $r(\underline{c}) = r$, we would have $h'(q,r) = h(q,r)$. (We remark that, in many cases, $\mathcal{H}_{\underline{c}}$ is known to be geometrically connected by the theorem of Conway and Parker [10, Appendix].) In the following proposition we show that h' is a reasonable approximation to h, at least on average.

Proposition 3.2. *There exist constants m, C_1, C_2, depending only on G, such that*

$$C_1 \sum_{r<R-m} h'(q,r) < \sum_{r<R} h(q,r) < C_2 \sum_{r<R} h'(q,r) \tag{4}$$

for all $R \gg 0$.

Proof. Recall that $\tilde{\Sigma}_{\underline{c}}$ consists of $(g_1, \ldots, g_k) \in G^k$ such that the multiset $\{c(g_i), \ldots, c(g_k)\}$ equals $\underline{c}$; $g_1, \ldots, g_k$ generate G; and $g_1 g_2 \ldots g_k = 1$. Write $n(\underline{c})$ for the number of orbits of the braid group $\pi_1(C_{|\underline{c}|})$ on $\tilde{\Sigma}_{\underline{c}}$. The right-hand inequality above thus follows immediately from the following lemma.

Lemma 3.3. *There exists a constant C_2 such that $n(\underline{c}) < C_2$ for all $\underline{c}$.*

Proof. If $\underline{g} = (g_1, \ldots, g_k)$ and $\underline{g}' = (g'_1, \ldots, g'_k)$ are two elements of G^k, we write $\underline{g} \sim \underline{g}'$ when $\underline{g}$ and $\underline{g}'$ are in the same orbit of the action of the braid group on G^k. We shall need a simple fact about the action of the braid group on G^k: suppose $\underline{g} = (g_1, \ldots, g_k) \in G^k$ with $g_1 \ldots g_k = 1$. Then, for any $1 \leq j \leq k$, there exists $(g'_1, \ldots, g'_{k-1}) \in G^{k-1}$ such that

$$(g_1, \ldots, g_k) \sim (g'_1, \ldots, g'_{k-1}, g_j). \tag{5}$$

Moreover, one knows (see, e.g., [15, Cor. 9.4]) that

$$(gg_1g^{-1}, gg_2g^{-1}, \ldots, gg_kg^{-1}) \sim (g_1, \ldots, g_k) \tag{6}$$

whenever g belongs to the subgroup generated by $(g_1, \ldots, g_k)$.

We show that $n(\underline{c}) \leq |G|^{|G|^2}$. This is clear if $|\underline{c}| \leq |G|^2$.

Suppose $k = |\underline{c}| > |G|^2$. Then any k-tuple $(g_1, g_2, \ldots, g_k)$ in $\tilde{\Sigma}_{\underline{c}}$ contains an element $g_0 \in G$ with multiplicity at least $|G| + 1$. Let g'_0 be any element in G conjugate to g_0. Thus, applying the braid operations (5) and (6) above, we deduce

$$(g_1, g_2, \ldots, g_k) \sim (g'_1, g'_2, \ldots, g'_{k-|G|-1}, g_0, g_0, \ldots, g_0) \tag{7}$$

$$\sim (g''_1, g''_2, g''_{k-|G|-1}, g'_0, g'_0, \ldots, g'_0) \tag{8}$$

for certain $g'_j, g''_j \in G$, where both g_0 and g'_0 occur $|G| + 1$ times at the end of each expression. On the other hand $g_0'^{|G|} = 1$. Thus, if $(g_1, g_2, \ldots, g_k) \in \Sigma_{\underline{c}}$, then $(g''_1, \ldots, g''_{k-|G|-1}, g'_0)$ belongs to $\Sigma_{\underline{c}'}$ where $\underline{c}'$ is $\underline{c}$ with $|G|$ copies of the conjugacy class of g_0 removed. So $n(\underline{c}) \leq n(\underline{c}')$. If $|\underline{c}'| > |G|^2$ we may apply the procedure that led to (7) again; indeed, repeatedly applying (7) and (8) we can bring elements of $\Sigma_{\underline{c}}$ to a "standard form." We see in particular that $n(\underline{c}) \leq n(\underline{c}')$ for some $|\underline{c}'| \leq |G|^2$. The result now follows. □

We now turn to the left-hand inequality in (4). Here we will make use of the theorem of Conway and Parker [10, Appendix] in order to show that $\mathcal{H}_{\underline{c}}$ has geometric components defined over $\mathbb{F}_q$ for many choices of $\underline{c}$.

We first show that $\mathcal{H}_{\underline{c}}$ is nonempty for many choices of $\underline{c}$.

Let $N \subset G$ be the normal subgroup consisting of all products $g_1 \ldots g_k$, where the Nielsen class of $(g_1, \ldots, g_k)$ is $\mathbb{F}_q$-rational. We claim that for every element $g \in N$ there exists, for some k, a k-tuple $(g_1, \ldots, g_k)$ such that

- $g_1 \ldots g_k = g$;
- the Nielsen class of $(g_1, \ldots, g_k)$ is $\mathbb{F}_q$-rational;
- the g_i generate G.

It suffices to show that this assertion holds for $g = 1$; for if we have $(g_1, \ldots, g_k)$ satisfying the last two conditions and having product 1, we can concatenate it with $(g_{k+1}, \ldots, g_\ell)$ having product g and representing an $\mathbb{F}_q$-rational Nielsen class. To see that the assertion holds for $g = 1$, merely choose

$(g_1, g_1^{-1}, \ldots, g_k, g_k^{-1})$ where $(g_1, \ldots, g_k)$ is a generating set for G which represents an $\mathbb{F}_q$-rational Nielsen class.

Now let $\underline{d}_1, \ldots, \underline{d}_K$ be a finite set of $\mathbb{F}_q$-rational Nielsen classes such that, for each $g \in N$, there exists $(g_1, \ldots, g_k)$ representing some $\underline{d}_i$ which generates G and has product g. If $\underline{c}$ and $\underline{d}$ are two Nielsen classes, we denote their concatenation by $\underline{c} + \underline{d}$. For each $\mathbb{F}_q$-rational Nielsen class $\underline{c}$, choose a representative $(g_1, \ldots, g_k)$. By the discussion above, there exists an m-tuple $(g_1, \ldots, g_k, g_{k+1}, \ldots, g_m)$ which is contained in $\Sigma_{\underline{c}+\underline{d}_i}$ for some i. It follows that $\mathcal{H}_{\underline{c}+\underline{d}_i}$ is nonempty for some i.

We now need to show that there are many Hurwitz spaces which are not only nonempty but which possess a geometric component defined over $\mathbb{F}_q$. Our main tool is the following assertion, which follows immediately from Proposition 1 and Lemma 2 of [10]:

Lemma 3.4. *There exists a group $\tilde{G}$, a surjective homomorphism $\tilde{G} \to G$, and a constant $C_3(G)$ such that, for any Nielsen class $\underline{\tilde{c}}$ of $\tilde{G}$ which contains at least $C_3(G)$ copies of each nontrivial conjugacy class of $\tilde{G}$, the Hurwitz space $\mathcal{H}_{\underline{\tilde{c}}}$ is geometrically connected.*

By the argument prior to Lemma 3.4, applied to $\tilde{G}$ instead of G, there exists a finite set of $\mathbb{F}_q$-rational Nielsen classes $\underline{\tilde{e}}_1, \ldots, \underline{\tilde{e}}_L$ such that, for every $\mathbb{F}_q$-rational Nielsen class $\underline{\tilde{c}}$ of $\tilde{G}$, the Hurwitz scheme $\mathcal{H}_{\underline{\tilde{c}}+\underline{\tilde{e}}_i}$ is nonempty.

Now consider an $\mathbb{F}_q$-rational Nielsen class $\underline{c}$ of G. We want to find an $\mathbb{F}_q$-rational Nielsen class $\underline{\tilde{c}}$ of $\tilde{G}$ which "approximately" projects to $\underline{c}$. For each $Gal(\bar{\mathbb{F}}_q/\mathbb{F}_q)$-orbit $\mathcal{O}$ on the nontrivial conjugacy classes in $\mathcal{C}$, let $\tilde{\mathcal{O}}$ be a $Gal(\bar{\mathbb{F}}_q/\mathbb{F}_q)$-orbit of conjugacy classes in $\tilde{G}$ which projects to $\mathcal{O}$. We note that the projection of the multiset $\tilde{\mathcal{O}}$ to G will be some multiple $k_{\mathcal{O}}\mathcal{O}$ of $\mathcal{O}$, where $k_{\mathcal{O}} \geq 1$. We know that $\underline{c}$ can be expressed as

$$\sum_{\mathcal{O}} c_{\mathcal{O}} \mathcal{O}$$

for some set of integers $\{c_{\mathcal{O}}\}$. Then the Nielsen class

$$\underline{\tilde{c}} = \sum_{\mathcal{O}} \lceil \frac{c_{\mathcal{O}}}{k_{\mathcal{O}}} \rceil \tilde{\mathcal{O}}$$

is $\mathbb{F}_q$-rational; moreover, the projection of $\underline{\tilde{c}}$ to G can be written as $\underline{c} + \underline{c}'$, where $\underline{c}'$ is drawn from a finite list of $\mathbb{F}_q$-rational Nielsen classes $\underline{c}'_1, \ldots, \underline{c}'_M$.

Now we fix, once and for all, an $\mathbb{F}_q$-rational Nielsen class $\underline{\tilde{c}}''$ for $\tilde{G}$, containing at least $C_3(G)$ copies of each nontrivial conjugacy class of $\tilde{G}$. We know already that, for some i, the Hurwitz space attached to $\underline{\tilde{c}} + \underline{\tilde{c}}'' + \underline{\tilde{e}}_i$ is nonempty; what's more, it is $\mathbb{F}_q$-rational, and by Lemma 3.4 it is geometrically connected.

The projection of $\underline{\tilde{c}} + \underline{\tilde{c}}'' + \underline{\tilde{e}}_i$, under the map $\tilde{G} \to G$, can be written as $\underline{c} + \underline{d}_i + n_i \mathbf{1}$, where $\underline{d}_i$ is drawn from some finite list $\underline{d}_1, \ldots, \underline{d}_N$, and $n_i \mathbf{1}$ refers to n_i copies of the trivial conjugacy class.

We claim that $\mathcal{H}_{\underline{c}+\underline{d}_i}$ has an $\mathbb{F}_q$-rational geometrically connected component. Indeed, to any $\tilde{G}$-cover $Y \to \mathbb{P}^1$ with Nielsen class $\tilde{\underline{c}} + \tilde{\underline{c}}'' + \tilde{\underline{e}}_i$, there is canonically associated a G-cover of $\mathbb{P}^1$ with Nielsen class $\underline{c} + \underline{d}_i$; namely, take the quotient of Y by $\ker(\tilde{G} \to G)$.

The associated map $\mathcal{H}_{\tilde{\underline{c}}+\tilde{\underline{c}}''+\tilde{\underline{e}}_i} \to \mathcal{H}_{\underline{c}+\underline{d}_i}$ has as its image a geometrically connected $\mathbb{F}_q$-rational component of $\mathcal{H}_{\underline{c}+\underline{d}_i}$.

For notational convenience, define $h(q, \underline{c})$ to be the number of $\mathbb{F}_q$-rational geometric components of $\mathcal{H}_{\underline{c}}$ multiplied by $q^{|\underline{c}|}$. By the discussion above, $h(q, \underline{c} + \underline{d}_i) \geq q^{|\underline{c}+\underline{d}_i|}$ for some i.

We thus have, on the one hand,

$$\sum_{i, \underline{c}: r(\underline{c}) < R} h(q, \underline{c} + \underline{d}_i) \geq \sum_{\underline{c}: r(\underline{c}) < R} q^{|\underline{c}+\underline{d}_i|} \geq \sum_{\underline{c}: r(\underline{c}) < R} h'(q, \underline{c})$$

and on the other,

$$\sum_{i, \underline{c}: r(\underline{c}) < R} h(q, \underline{c} + \underline{d}_i) \leq N \sum_{\underline{c}: r(\underline{c}) < R + r(\underline{d}_i)} h(q, \underline{c}).$$

This finishes the proof of the proposition, taking C_1 to be $1/N$ and m to be the supremum of $r(\underline{d}_i)$. □

We are now in a position to prove Proposition 3.1:

Proof (of Proposition 3.1). By definition $\sum_{r=0}^{\infty} h'(q, r) q^{-rs} = \sum_{\underline{c}} q^{|\underline{c}|} q^{-r(\underline{c})s}$, the sum being taken over all $\mathbb{F}_q$-rational Nielsen classes $\underline{c}$. This sum factorizes as a product indexed by the $Gal(\mathbb{F}_q)$-orbits $\mathcal{O}$ of conjugacy classes of G:

$$\sum_{r=0}^{\infty} h'(q, r) q^{-rs} = \prod_{\mathcal{O}} (1 - q^{|\mathcal{O}|(1 - ind(\mathcal{O})s)})^{-1}. \tag{9}$$

Here by $ind(\mathcal{O})$ we mean the ramification index of any representative of the orbit $\mathcal{O}$, and by $|\mathcal{O}|$ the number of conjugacy classes in $\mathcal{O}$.

Now (9) implies, via Lemma 2.3, that $\sum_{q^r < X} h'(q, r) \asymp X^{a(G)} \log(X)^{b_{\mathbb{F}_q}(G) - 1}$, where $a(G), b_{\mathbb{F}_q}(G)$ are as in Malle's conjecture. The claim of Proposition 3.1 now follows at once from this and Proposition 3.2. □

4 Further conjectures

In this section, we discuss first (Section 4.1) some further questions in the function field case. The heuristics used for Proposition 3.1 also suggest certain "refined" heuristics for extensions of number fields; we discuss some of these in Section 4.2. Finally in Section 4.3 we discuss some more speculative questions about the enumeration of higher-dimensional varieties.

We note by way of caution that there is little numerical evidence to suggest that some of the questions posed below have an affirmative answer.

4.1 More questions about function fields.

The following question was raised by N. Katz and J. de Jong.

Question 4.1. Let q be fixed. Is it true that there is a constant $c := c(q)$ such that the number of isomorphism classes of genus g curves over $\mathbb{F}_q$ is less than c^g, for all $g \geq 1$?

The emphasis of this question is on the case where q is fixed and $g \to \infty$. The upper bound $c^{g \log(g)}$ was established by Katz and de Jong in unpublished work. In a certain sense this bound is analogous to Lemma 2.4. Note that this problem, again, amounts to counting the number of $\mathbb{F}_q$-points on a variety (namely the moduli space $\mathfrak{M}_g$) of high dimension. One difficulty in using, e.g., the Lefschetz fixed point formula, is that the Betti numbers of $\mathfrak{M}_g$ grow very rapidly with g.

Returning to the distribution of discriminants, one may also study the properties of certain zeta functions; this serves to smooth out the irregularity in the distribution of discriminants. For instance, consider the function $\xi_{q,G}(s) := \sum_L D_{L/\mathbb{F}_q(t)}^{-s}$, where L varies over degree n extensions of $\mathbb{F}_q(t)$ with Galois group G, and $\mathcal{D}_L$ is the discriminant of L. A "geometric" variant of $\xi_{q,G}$ is the zeta function:

$$\zeta_{q,G}(s) = \sum_{r=0}^{\infty} |\mathcal{H}_r(\mathbb{F}_q)| q^{-rs}. \tag{10}$$

Question 4.2. What are the analytic properties of $\zeta_{q,G}(s)$? In particular, is it the case that $\zeta_{q,G}(s)$ has an analytic continuation to the left of $\Re s = 1/a(G)$, with a pole of order $b_{\mathbb{F}_q}(G)$ at $s = 1/a(G)$?

4.2 Questions about number fields

The discriminant of a number field $K/\mathbb{Q}$ may be regarded as a measure of ramification, where each ramified prime is weighted according to the conjugacy class of tame inertia. Here we first discuss generalizations of Malle's conjecture that allow for varying this weighting. Then we take up the question of multiplicity of discriminants, already raised in Conjecture 1.3. As an example of these heuristics, we give heuristics for the number of icosahedral modular forms with conductor $\leq N$ (Example 4.4).

The questions proposed in this section are interrelated. In particular, the upper bounds implicit in Question 4.5, Question 4.3, and Conjecture 1.3 are close to equivalent (see Remark 4.7.) In fact, these weak upper bounds seem on considerably safer ground then the general questions, as they do not presuppose a positive solution to the inverse Galois problem.

Malle's conjecture with modified weights

Set $K = \mathbb{Q}$ and let other notation be as described prior to Conjecture 1.2. Let $f : \mathcal{C} \to \mathbb{Z}_{\geq 0}$ be invariant under the $Gal(\bar{\mathbb{Q}}/\mathbb{Q})$-action and such that

$$f(g) = 0 \Longleftrightarrow g = \{\mathrm{id}\}.$$

We call such an f a *rational class function.* Set $a(f) = \max_{c \in \mathcal{C}, c \neq \{\mathrm{id}\}} f(c)^{-1}$. Let $b_{\mathbb{Q}}(f)$ be the number of $Gal(\bar{\mathbb{Q}}/\mathbb{Q})$-orbits on the set $\{c \in \mathcal{C} : f(c) = a(f)^{-1}\}$.

If $L/\mathbb{Q}$ is a Galois extension with group G and p is a prime not dividing $|G|$, let $c_p \in \mathcal{C}$ be the image of a generator of tame inertia at p. Now we define the f-discriminant of L to be:

$$\mathcal{D}_f(L) = \prod_{p \nmid |G|} p^{f(c_p)}. \tag{11}$$

For instance, if $f = ind$, then $\mathcal{D}_f(L)$ is the prime-to-$|G|$ part of $\mathcal{D}_{L_0}$, notation being as prior to Conjecture 1.2, taking $K = \mathbb{Q}$.

Let $N_{G,f}(X)$ (or, when the group is clear from context, just $N_f(X)$) be the number of Galois extensions $L/\mathbb{Q}$ with Galois group G and $D_f(L) < X$.

Question 4.3. Is it true that $N_{G,f}(X) \sim cX^{a(f)}(\log X)^{b_{\mathbb{Q}}(f)-1}$?

We note that this type of generalization is already, in some sense, anticipated in Malle's conjecture. A given G can be equipped with many different embeddings into symmetric groups; Malle's conjecture already predicts an asymptotic for $N_f(X)$ when f is the index function corresponding to *any* such embedding.

Example 4.4. Let $\rho : G \to \mathrm{GL}(V)$ be a complex representation. Then $g \in G \mapsto \mathrm{codim} V^g$, the codimension of the invariant space, defines a rational class function. If $L/\mathbb{Q}$ has Galois group G, $\mathcal{D}_f(L)$ is the prime-to-$|G|$ part of the Artin conductor of the Galois representation associated to L.

For example, we may take G to be the finite subgroup of order 240 in $GL_2(\mathbb{C})$ whose image in $PGL_2(\mathbb{C})$ is isomorphic to A_5. For this group, there is a unique conjugacy class (the conjugacy class of non-central involutions) which has $f(c) = 1$. Subject to Artin's conjecture, the holomorphic modular forms of weight 1, conductor N, quadratic Dirichlet character, and icosahedral type are in bijection with the Galois extensions with group G and Artin conductor N such that complex conjugation is sent to a non-central involution.

Question 4.3 then suggests that, if $s(N)$ is the number of icosahedral holomorphic weight-1 modular forms with quadratic character and conductor at most N, then

$$s(N) \sim cN$$

for some constant c. The best upper bound at present is $s(N) \ll_\epsilon N^{13/7+\epsilon}$ due to Michel and the second author [13]. Serre [14] speculated that the number of such forms with conductor *exactly* N is $\ll_\epsilon N^\epsilon$.

Multidiscriminants

One can use the function field heuristics described here to produce even more refined (i.e., optimistic!) heuristics for counting number fields, in which we attach to each field not just an element of $\mathbb{R}_{\geq 0}$ but an element of $\mathbb{R}^k_{\geq 0}$ for some $k > 1$. We could call such a map a "multidiscriminant."

Let G be a finite group, and let the orbits of the nontrivial conjugacy classes under the action of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ be denoted $\mathcal{O}_1, \mathcal{O}_2, \ldots \mathcal{O}_m$. Given a Galois G-extension $L/\mathbb{Q}$, set $D_{\mathcal{O}_i}(L)$ to be the product of all primes $p \nmid G$ such that the image in G of tame inertia at p is conjugate to $\mathcal{O}_i$; the map

$$L \mapsto (D_{\mathcal{O}_i}(L))_{1 \leq i \leq m}$$

can be regarded as a multidiscriminant. Set $N_G(X_1, \ldots, X_n)$ to be the number of $L/\mathbb{Q}$ such that $D_{\mathcal{O}_i}(L) < X_i$ for all i. We can then ask:

Question 4.5. Is it true that, if $X_j \to \infty$ for all $1 \leq j \leq n$, then the ratio

$$\frac{N_G(X_1, \ldots, X_m)}{X_1 \ldots X_m} \tag{12}$$

approaches a fixed limit $c = c(G)$?

As before, (12) can be heuristically justified by dimension computations over finite fields. Indeed, let notation be as above but let $\{\mathcal{O}_i\}$ now denote the orbits of the conjugacy classes in G under the cyclotomic character of $Gal(\bar{\mathbb{F}}_q/\mathbb{F}_q)$. Let $N_{G,q}(X_1, \ldots, X_m)$ be the number of Galois G-covers $f : Y \to \mathbb{P}^1/\mathbb{F}_q$ such that the number of branch points of f in $\mathbb{P}^1$ with monodromy in $\mathcal{O}_i$ is less than $a_i = \lfloor \frac{\log(X_i)}{\log(q)} \rfloor$. Such covers are parametrized (as usual, up to uniformly bounded finite ambiguity arising from descent problems) by the $\mathbb{F}_q$-points of a variety, whose largest-dimensional connected component is a Hurwitz space of dimension about $\sum_i a_i$. So our usual heuristic suggests that this variety has about $\prod_i q^{a_i}$, or $X_1 \ldots X_m$ points.

Lemma 4.6. *An affirmative answer to Question 4.5 implies an affirmative answer to Question 4.3.*

The proof of the Lemma is straightforward but tedious.

The multiplicity of discriminants

A problem of a rather different flavor is to count the extensions $L/\mathbb{Q}$ with Galois group G whose discriminant is *exactly* X. One can show, e.g., by genus theory, that this number can grow as fast as $X^{c/\log\log(X)}$. On the other hand Conjecture 1.3 asserts that this multiplicty is $\ll_{\epsilon,G} X^{\epsilon}$.

Conjecture 1.3 implies that the l-torsion part of the class group of a number field $K/\mathbb{Q}$ is $\ll_{l,[K:\mathbb{Q}]} \mathcal{D}^{\epsilon}_{K/\mathbb{Q}}$. (This follows immediately from class field theory, as l-torsion in the class group of K would give rise to unramified extensions of degree l.)

Remark 4.7. The following conjectures are equivalent:

1. Conjecture 1.3,
2. The upper bound $N_{G,f}(X) \ll_{\epsilon,G,f} X^{a(f)+\epsilon}$ in Question 4.3,
3. The upper bound $N_G(X_1, \ldots, X_n) \ll (X_1 X_2 \ldots X_n)^{1+\epsilon}$ in Question 4.5.

The implications (1) $\implies$ (3) $\implies$ (2) are trivial. We show the remaining implication in the following proposition.

Proposition 4.8. *An affirmative answer to Question 4.3, or even the weaker assumption*

$$N_{G,f}(X) \ll_{\epsilon,G,f} X^{a(f)+\epsilon}, \tag{13}$$

implies Conjecture 1.3.

Proof. Let $a_G(X)$ be the number of extensions $L/\mathbb{Q}$ with Galois group G and with discriminant X. Clearly it will suffice to show $a_G(X) \ll_{\epsilon,G} X^{\epsilon}$. The main idea will be to apply (13) with G replaced by G^k and some of its subgroups.

Fix $k > 0$ an integer. We write an element of G^k as a k-tuple $(g_1, g_2, \ldots, g_k)$. Let F be the class function on G^k that is identically 1, i.e., $F(c_1, \ldots, c_k) = 1$ for all conjugacy classes c_j of G.

Let $\mathcal{S}$ be the class of subgroups of G^k which project surjectively onto each copy of G; for each $H \in \mathcal{S}$ we also write F for the rational class function on H that is identically 1. Then the k-tuples of G-extensions $L_1, \ldots, L_k$ are in bijection with the H-extensions $L/\mathbb{Q}$, where H ranges over $\mathcal{S}$. We denote by $[L_1, \ldots, L_k]$ the H-extension associated to a k-tuple in this way, and by $\mathcal{D}_F([L_1, \ldots, L_k])$ the F-discriminant of this extension, given by (11).

$\mathcal{D}_F([L_1, \ldots, L_k])$ is, away from primes dividing $|G|$, the square-free part of the product $\prod_{j=1}^{k} \mathcal{D}_{L_j/\mathbb{Q}}$. Thus $\mathcal{D}_F([L_1, \ldots, L_k])$ is (relatively) large whenever the $\mathcal{D}_{L_i}$ have few common factors with each other. On the other hand, if $\mathcal{D}_{L_1/\mathbb{Q}} = \mathcal{D}_{L_2/\mathbb{Q}} = \cdots = \mathcal{D}_{L_k/\mathbb{Q}} = X$, it follows that $D_F([L_1, \ldots, L_k]) \leq X$. In particular,

$$\sum_{H \in \mathcal{S}} N_{H,F}(X) \geq a_G(X)^k. \tag{14}$$

Combining (14) and (13), and noting that the exponent a of (13) equals 1 whenever (G, f) is replaced by (H, F) as above, we see that $a_G(X)^k \ll_{G,k} X^{1+\epsilon}$. The result follows, k being arbitrary. □

4.3 The scarceness of arithmetic objects with prescribed bad reduction

We have discussed in previous sections heuristics for counting function fields, number fields, and Galois representations. In a certain sense all of these can be regarded as "0-dimensional" arithmetic objects. We now briefly discuss a plausible statement in higher dimensions, at least as regards *upper bounds*.

By way of motivation, we note that Conjecture 1.3 may be regarded as saying that *there are very few number fields with very little bad reduction.* If one replaces "number field" with "proper smooth variety," very little is known; however, it is generally believed that there are "relatively few" proper smooth varieties V over $\mathbb{Z}$. There are a few evident examples: one may take for V, e.g., a flag variety associated to a Chevalley group over $\mathbb{Z}$. Further, one may blow up such a variety along an appropriate locus. However, as Jason Starr and Johan de Jong pointed out to us, all such varieties are *rational*, and there seems to be no non-rational example known. A beautiful result of Fontaine states that there exist no abelian varieties over $\mathbb{Z}$.

The question we state aims to quantify this scarceness. For a finite set S of primes set $N(S) = \prod_{p \in S} p$. For concreteness and to avoid any technical hypotheses, we have phrased the question in terms of modular Galois representations.

Question 4.9. Fix a Hodge–Tate type π (i.e., a set of Hodge–Tate weights), positive integers n, d, and a prime l. Let $\mathrm{GR}(\pi, S)$ be the set of modular Galois representations $\rho_l : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_n(\overline{\mathbb{Q}}_l)$ with Hodge–Tate weights π and good reduction outside S. Here "modular" means "attached to an automorphic form on GL_n." Let $\mathrm{GR}_d(\pi, S) \subset GR(\pi, S)$ be the subset consisting of ρ_l whose Frobenius traces lie in a field extension of $\mathbb{Q}$ with degree $\leq d$.

Is it true that $|GR_d(\pi, S)| \ll_{\epsilon,d,n,\pi} N(S)^{\epsilon}$?

We can ask a similar question with a more "motivic" flavor; of course, one may expect that under suitable modularity conjectures the questions above and below are equivalent.

Question 4.10. Fix $K \in \mathbb{N}$ and let S be a finite set of primes. Consider the set $\mathcal{V}(K, S)$ of proper smooth varieties V over $\mathrm{Spec}\mathbb{Z}[\frac{1}{N(S)}]$ such that $\dim(V) \leq K$ and $\dim H^i(V_{\mathbb{C}}, \mathbb{C}) \leq K$ for each $0 \leq i \leq 2K$. To each variety $V \in \mathcal{V}(K, S)$ associate the sequence $\#(V(\mathbb{F}_p))_{p \notin S}$, indexed by the primes not in S. Then the number of distinct such sequences is $\ll_K N(S)^{\epsilon}$.

An affirmative answer to Question 4.9 would imply Conjecture 1.3. It would also imply that the number of elliptic curves over $\mathbb{Q}$ of conductor N is $\ll_{\epsilon} N^{\epsilon}$.

Note that even the *finiteness* of $GR(\pi, S)$ would not be clear without the hypothesis of modularity! Using modularity, one may probably show that $|GR(\pi, S)|$ is bounded by a polyomial in $N(S)$. The content of the assertion is then that $|GR_d(\pi, S)|$ is much smaller. A related phenomenon is well-known in the context of holomorphic forms: fix $k \geq 2$ and consider the space $S_k(N)$ of holomorphic forms of level N and weight k. Although $\dim S_k(N) \sim \mathrm{const} \cdot N$ as $N \to \infty$, the number of Hecke eigenforms whose coefficient field has degree $\leq d$ seems to grow much more slowly with N.

One can enunciate a corresponding question in the function field case; it also seems quite difficult.

Remark 4.11. It is interesting to note the contrast between the number field and function field contexts. In the number field setting, the ability to average seems to make counting objects of conductor up to N much easier than counting objects of conductor exactly N. In the function field setting, on the other hand, counting objects of conductor up to N means counting covers of $\mathbb{P}^1$ whose ramification locus varies among all divisors of $\mathbb{P}^1$ of degree less than $\log_q N$, while counting covers with a *fixed* conductor amounts to studying the arithmetic (in the case of finite covers, the étale fundamental group) of a *single* open curve inside $\mathbb{P}^1$, which might in some ways be easier. One way to express the contrast is to observe that our understanding of the étale fundamental group of an open subset of $\mathbb{P}^1_{\mathbb{F}_q}$, though very far from complete, is much greater than our understanding of the maximal Galois extension of $\mathbb{Q}$ unramified away from a fixed finite set of primes.

References

1. K. Belabas – "Paramétrisation de structures algébriques et densité de discriminants (d'après M. Bhargava)", 2004, Seminaire Bourbaki, Exp. No. 935.
2. M. Bhargava – "Higher composition laws, IV", preprint.
3. — , "Higher composition laws", Thesis, Princeton University, 2001.
4. I. Bouw and S. Wewers – "Reduction of covers and Hurwitz spaces", to appear, *J. Reine Angew. Math.*.
5. H. Cohen, F. Diaz y Diaz and M. Olivier – "A survey of discriminant counting", Algorithmic number theory (Sydney, 2002), Lecture Notes in Comput. Sci., vol. 2369, Springer, Berlin, 2002, p. 80–94.
6. B. Datskovsky and D. J. Wright – "Density of discriminants of cubic extensions", *J. Reine Angew. Math.* **386** (1988), p. 116–138.
7. P. Dèbes and J.-C. Douai – "Algebraic covers: field of moduli versus field of definition", *Ann. Sci. École Norm. Sup. (4)* **30** (1997), no. 3, p. 303–338.
8. W. Duke – "Bounds for arithmetic multiplicities", *Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998)*, no. Extra Vol. II, 1998, p. 163–172 (electronic).
9. J. Ellenberg and A. Venkatesh – "The number of extensions of a number field with fixed degree and bounded discriminant", 2004, preprint.
10. M. D. Fried and H. Völklein – "The inverse Galois problem and rational points on moduli spaces", *Math. Ann.* **290** (1991), no. 4, p. 771–800.
11. G. Malle – "On the distribution of Galois groups, II", preprint.
12. — , "On the distribution of Galois groups", *J. Number Theory* **92** (2002), no. 2, p. 315–329.
13. P. Michel and A. Venkatesh – "On the dimension of the space of cusp forms associated to 2-dimensional complex Galois representations", *Int. Math. Res. Not.* (2002), no. 38, p. 2021–2027.
14. J.-P. Serre – "Modular forms of weight one and Galois representations", Algebraic number fields: L-functions and Galois properties (Proc. Sympos., Univ. Durham, Durham, 1975), Academic Press, London, 1977, p. 193–268.
15. H. Völklein – *Groups as Galois groups*, Cambridge Studies in Advanced Mathematics, vol. 53, Cambridge University Press, Cambridge, 1996.

16. S. Wewers – "Construction of Hurwitz spaces", Thesis, University of Essen, 1998.

Classical and minimal models of the moduli space of curves of genus two

Brendan Hassett

Rice University, MS 136, Department of Mathematics, PO Box 1892, Houston, TX 77251-1892, U.S.A.
hassett@math.rice.edu

Summary. We discuss the log minimal model program as applied to the moduli space of curves, especially in the case of curves of genus two. Log canonical models for these moduli spaces can often be constructed using the techniques of Geometric Invariant Theory. In genus two, this boils down to the invariant theory of binary sextics, which was developed systematically in the 19th century.

1 Introduction

This paper is an introduction to the minimal model program, as applied to the moduli space of curves. Our long-term goal is a geometric description of the canonical model of the moduli space when it is of general type. This entails proving that the canonical model exists and interpreting it as a parameter space in its own right.

Work of Eisenbud, Harris, and Mumford shows that M_g is of general type when $g \geq 24$ (see [8] and subsequent papers). A standard conjecture of birational geometry – the finite generation of the canonical ring – would imply the canonical model is

$$\text{Proj } \oplus_{n\geq 0} \Gamma(\overline{M}_g, nK_{\overline{M}_g}).$$

Unfortunately, this has yet to be verified in a single genus!

Another possible line of attack is to consider *log* canonical models of the moduli space. The moduli space is best regarded as a pair $(\overline{M}_g, \Delta)$, where $\overline{M}_g$ is Deligne-Mumford compactification by stable curves and Δ is its boundary. It is implicit in the work of Mumford [18] that the moduli space of stable curves is its own log canonical model (see Theorem 4.7). Our basic strategy is to interpolate between the log canonical model and the (conjectural) canonical model by considering

$$\text{Proj } \oplus_{n\geq 0} \Gamma(\overline{M}_g, n(K_{\overline{M}_g} + \alpha\Delta)),$$

where $\alpha \in \mathbb{Q} \cap [0,1]$ is chosen so that $K + \alpha\Delta$ is effective.

This program is a subject of ongoing work, inspired by correspondence with S. Keel, in collaboration with D. Hyeon. Future papers will address the stable behavior of these spaces for successively smaller values of α. It is remarkable that their behavior is largely independent of the genus (see, for example, Remark 4.9.)

However, for small values of g special complexities arise. When $g = 2$ or 3, the locus in $\overline{M}_g$ of curves with automorphism has codimension ≤ 1. To include these spaces under our general framework, we must take into account the properties of the moduli stack $\overline{\mathcal{M}}_g$. In particular, it is necessary to use the canonical divisor of the moduli stack rather than its coarse moduli space. These differ substantially, as the natural morphism $\overline{\mathcal{M}}_g \to \overline{M}_g$ is ramified at stable curves admitting automorphisms.

Luckily, we have inherited a tremendously rich literature on curves of small genus. The invariant-theoretic properties of M_2 were extensively studied by the 19th century German school [4], who realized it as an open subset of the weighted projective space $\mathbb{P}(1,2,3,5)$. Theorem 4.10 reinterprets this classical construction using the modern language of stacks and minimal models.

We work over an algebraically closed field k of characteristic zero. We use the notation $\equiv$ for $\mathbb{Q}$-linear equivalence of divisors. Throughout, a *curve* is a connected, projective, reduced scheme of dimension one. The *genus* of a curve is its arithmetic genus.

The moduli stack of smooth (resp. stable) curves of genus g is denoted $\mathcal{M}_g$ (resp. $\overline{\mathcal{M}}_g$); the corresponding coarse moduli scheme is denoted M_g (resp. $\overline{M}_g$). The boundary divisors in $\overline{M}_g$ (resp. $\overline{\mathcal{M}}_g$) are denoted

$$\Delta_0, \Delta_1, \ldots, \Delta_{\lfloor g/2 \rfloor} (\text{resp. } \delta_0, \delta_1, \ldots, \delta_{\lfloor g/2 \rfloor}).$$

Let $Q : \overline{\mathcal{M}}_g \to \overline{M}_g$ denote the natural morphism from the moduli stack to the coarse moduli space, so that

$$Q^*\Delta_i = \delta_i, i \neq 1 \qquad Q^*\Delta_1 = 2\delta_1.$$

We write $\delta = \sum_{0 \leq i \leq g/2} \delta_i$; abusing notation, we also use δ for the corresponding divisor

$$\Delta_0 + 1/2\Delta_1 + \Delta_2 + \cdots + \Delta_{\lceil g/2 \rceil}$$

on $\overline{M}_g$.

Acknowledgments. The key ideas underlying this program were worked out in correspondence with S. Keel. They have been further developed in collaboration with D. Hyeon. The author also benefited from conversations with James Spencer about rigidification, quotient stacks, and moduli of curves of genus two, and from comments on the manuscript by Michael van Opstall. Part of this paper was prepared during a visit to the Mathematisches Institut of the Georg-August-Universität, Göttingen.

Partial support was provided by National Science Foundation grants 0196187 and 0134259 and the Sloan Foundation.

2 Classical geometry

2.1 Elementary facts about curves of genus two

We recall results from standard textbooks, e.g., [9] IV Ex 2.2 and §5. Let C denote a smooth curve of genus two with sheaf of differentials ω_C. The global sections of ω_C give the canonical morphism

$$j : C \to \mathbb{P}(\Gamma(C, \omega_C)) \simeq \mathbb{P}^1$$

which is finite of degree two. The corresponding covering transformation $\iota : C \to C$ is called the hyperelliptic involution. By the Hurwitz formula, j is branched over six distinct points

$$\{b_1, \ldots, b_6\} \subset \mathbb{P}(\Gamma(C, \omega_C)).$$

On fixing an identification $\mathbb{P}(\Gamma(C, \omega_C)) \simeq \mathbb{P}^1$, we can write down a nontrivial binary sextic form vanishing at the branch points

$$F \in \Gamma(\mathbb{P}^1, \mathcal{O}_{\mathbb{P}^1}(6)),$$

determined by $\{b_1, \ldots, b_6\} \in \mathbb{P}^1$ up to a scalar.

Conversely, suppose we have a binary sextic form F with six distinct zeros $b_1, \ldots, b_6 \in \mathbb{P}^1$. Then there is a unique degree-two cover of $\mathbb{P}^1$ branched over these points. This is a smooth curve of genus two and the map to $\mathbb{P}^1$ is the canonical morphism. Moreover, the isomorphism class of C depends only on the orbit of F under the action of GL_2.

To summarize: There is a one-to-one correspondence between isomorphism classes of curves of genus two and GL_2-orbits of binary sextic forms with distinct zeros.

We will need a relative version of this dictionary, following [24]. Let $\pi : \mathcal{C} \to S$ be a smooth morphism to a scheme of finite type over k, with fibers curves of genus two. Since the relative dualizing sheaf ω_π is globally generated, there is a relative double cover

$$\begin{array}{ccc} \mathcal{C} & \xrightarrow{j} & \mathbb{P}^1(\pi_*\omega_\pi) := \mathbb{P} \\ & {\scriptstyle\pi}\searrow \quad \swarrow{\scriptstyle\psi} & \\ & S & \end{array}$$

with associated involution ι. Using the trace we decompose $j_*\mathcal{O}_\mathcal{C} \simeq \mathcal{O}_\mathbb{P} \oplus \mathcal{L}$, where $\mathcal{L}$ has relative degree -3 on the fibers of ψ. The $\mathcal{O}_\mathbb{P}$-algebra structure

on $\mathcal{O}_{\mathcal{C}}$ is thus determined by an isomorphism $\mathcal{L}^2 \to \mathcal{O}_{\mathbb{P}}$, i.e., by a nonvanishing section of $\mathcal{L}^{-2}$ with zeros along the branch locus of j. By relative duality

$$j_*\omega_\pi = \mathcal{H}om_{\mathcal{O}_{\mathbb{P}}}(j_*\mathcal{O}_{\mathcal{C}}, \omega_\psi) \simeq \omega_\psi \oplus (\omega_\psi \otimes \mathcal{L}^{-1})$$

which yields

$$\pi_*\omega_\pi \simeq \psi_*(\omega_\psi \otimes \mathcal{L}^{-1}).$$

Using the identifications

$$\psi_*\mathcal{O}_{\mathbb{P}}(+1) = \pi_*\omega_\pi, \quad \omega_\psi = (\psi^* \det \pi_*\omega_\pi)(-2), \quad \mathrm{Pic}(\mathbb{P}) = \mathrm{Pic}(S) \oplus \mathbb{Z}c_1(\mathcal{O}_{\mathbb{P}}(+1)),$$

we find

$$\mathcal{L}^{-1} \simeq \mathcal{O}_{\mathbb{P}}(+1) \otimes \omega_\psi^{-1} = \mathcal{O}_{\mathbb{P}}(3) \otimes (\psi^* \det \pi_*\omega_\pi)^{-1}.$$

The class of the branch divisor is thus

$$-2c_1(\mathcal{L}) = c_1(\mathcal{O}_{\mathbb{P}}(6)) - 2\psi^* c_1(\det \pi_*\omega_\pi). \tag{1}$$

This has practical implications: An isomorphism of C induces a linear transformation on $\Gamma(C, \omega_C)$, which respects the binary sextic F up to a scalar. Formula (1) allows us to keep track of this scalar. For each $M \in \mathrm{GL}_2$, we have the linear action

$$(x, y) \mapsto (x, y) \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$$

which induces a natural left action on binary sextic forms

$$F \mapsto F(xm_{11} + ym_{21}, xm_{12} + ym_{22}).$$

We normalize this action using formula (1)

$$F \mapsto (M, F) := (\det M)^{-2} F(xm_{11} + ym_{21}, xm_{12} + ym_{22}), \tag{2}$$

so that $(M, F) = F$ if and only M is induced from an automorphism of C.

A smooth curve is *bielliptic* if it admits a degree-two morphism $i : C \to E$ to an elliptic curve; the covering transformation is called a *bielliptic involution*. For curves of genus two any bielliptic involution commutes with the hyperelliptic involution, which yields a diagram

$$\begin{array}{ccc} C & \stackrel{i}{\to} & E \\ {\scriptstyle j}\downarrow & & \downarrow{\scriptstyle \bar{j}} \\ \mathbb{P}^1 & \stackrel{\bar{i}}{\to} & \mathbb{P}^1 \end{array}$$

where $\bar{i}$ and $\bar{j}$ are the double covers induced on quotients. The branch locus of j is preserved by the covering transformation for $\bar{i}$, which is conjugate to $[x, y] \mapsto [y, x]$. The resulting involution of the branch locus will also be called a bielliptic involution. Thus C is isomorphic to a double cover branched over

$$\{[\alpha_1, 1], [1, \alpha_1], [\alpha_2, 1,][1, \alpha_2], [\alpha_3, 1], [1, \alpha_3]\}$$

for some $\alpha_1, \alpha_2, \alpha_3 \in k$. Conversely, each such curve admits a diagram as above and thus is bielliptic.

2.2 Invariant theory of binary sextics

We observe the classical convention for normalizing the coefficients of a binary sextic

$$F = ax^6 + 6bx^5y + 15cx^4y^2 + 20dx^3y^3 + 15ex^2y^4 + 6fxy^5 + gy^6.$$

The action (2) induces an action of GL_2 on $k[a,b,c,d,e,f,g]$. Recall that a polynomial $P \in k[a,b,c,d,e,f,g]$ is SL_2-*invariant* if for each $M \in \mathrm{SL}_2$, we have

$$(M,P) = P.$$

We write

$$R := k[a,b,c,d,e,f,g]^{\mathrm{SL}_2}$$

for the ring of such invariants. If P is SL_2-invariant, then each homogeneous component of P is as well, so R is a graded ring. Every homogeneous invariant satisfies the functional relation

$$(M,P) = (\det M)^{\deg(P)} P, \quad M \in \mathrm{GL}_2; \tag{3}$$

here it is essential that the action (2) include the factor $(\det M)^2$. The transformation

$$(x,y) \to (y,x)$$

thus reverses the sign of invariants of odd degree. These are called *skew invariants* in the classical literature.

Explicit generators for R were first written down in the nineteenth century, e.g., [4], pp. 296, in symbolic notation, [3] and [21] as explicit polynomials – the second edition of Salmon's *Higher algebra* has the most detailed information, and also [6] pp. 322. A nice early twentieth-century discussion is [22] pp. 90 and a modern account invoking the representation theory of SL_2 is [23].

For our purposes, the symmetric function representation of the invariants in [12] pp. 176 and 185 is the most useful. Let $\xi_1, \ldots, \xi_6$ denote the roots of the dehomogenized form $F(x,1)$, and write (ij) as shorthand for $\xi_i - \xi_j$. We write

$$A = a^2 \sum_{\text{fifteen}} (12)^2(34)^2(56)^2,$$

$$B = a^4 \sum_{\text{ten}} (12)^2(23)^2(31)^2(45)^2(56)^2(64)^2,$$

$$C = a^6 \sum_{\text{sixty}} (12)^2(23)^2(31)^2(45)^2(56)^2(64)^2(14)^2(25)^2(36)^2,$$

$$D = a^{10} \prod_{ij} (ij)^2,$$

$$E = a^{15} \prod_{\text{fifteen}} \det \begin{pmatrix} 1 & \xi_1 + \xi_2 & \xi_1\xi_2 \\ 1 & \xi_3 + \xi_4 & \xi_3\xi_4 \\ 1 & \xi_5 + \xi_6 & \xi_5\xi_6 \end{pmatrix} = a^{15} \prod_{\text{fifteen}} \left((14)(36)(52) - (16)(32)(54) \right),$$

where the summations are chosen to make the expressions $\mathfrak{S}_6$-symmetric. Consequently, A, B, C, D, and E can all be expressed as polynomials in $\mathbb{Q}[a, b, c, d, e, f, g]$, e.g.,

$$A = -240(ag - 6bf + 15ce - 10d^2),$$
$$B = -162000 \det \begin{pmatrix} a & b & c & d \\ b & c & d & e \\ c & d & e & f \\ d & e & f & g \end{pmatrix} + 1620(ag - 6bf + 15ce - 10d^2)^2.$$

In classical terminology, $(ag - 6bf + 15ce - 10d^2)$ is the *sixth transvectant of F over itself*; transvection is one of the main operations in Gordan's proof of finiteness for invariants of binary forms. The determinantal expression is the *catalecticant* of F: It vanishes precisely when F can be expressed as a sum of three sixth powers [6] pp. 276.

The following facts will be useful for subsequent analysis:

Proposition 2.1.

1. *The expressions A, B, C, D, and E are invariant and generate R [10], pp. 100, [4], etc.*
2. *D is the discriminant and vanishes precisely when the binary form has a multiple root.*
3. *B, C, D, and E vanish whenever the binary form has a triple root; A vanishes when the form has a quadruple root.*
4. *E vanishes if and only if the form admits a bielliptic involution, as defined in §2.1 [6], pp. 327 and [4], pp. 457.*
5. *The unique irreducible relation among the invariants is*

$$E^2 = G(A, B, C, D),$$

where G is weighted-homogeneous of degree 30 [4], pp. 299.

The notation used for the generating invariants is not consistent among authors. Our notation is consistent with that of Igusa, but inconsistent with Clebsch's and Salmon's. Of course, the invariants of degree two and fifteen are unique up to scalar.

2.3 The projective invariant-theory quotient

We consider

$$X := \operatorname{Proj} R = \operatorname{Proj} \frac{k[A, B, C, D, E]}{\langle E^2 - G(A, B, C, D) \rangle}.$$

If $A = B = C = D = 0$, then $E = 0$ as well, so X is covered by the distinguished affine open subsets

$$\{A \neq 0\} \quad \{B \neq 0\} \quad \{C \neq 0\} \quad \{D \neq 0\}.$$

However, in each localization

$$(k[A,B,C,D,E][A^{-1}])_0 \quad (k[A,B,C,D,E][B^{-1}])_0$$

$$(k[A,B,C,D,E][C^{-1}])_0 \quad (k[A,B,C,D,E][D^{-1}])_0$$

only even powers of E appear, so all the functions over these distinguished open subsets can be expressed in terms of A, B, C, D. In light of Proposition 2.1, we find

Proposition 2.2.

1. $X \simeq \operatorname{Proj} k[A,B,C,D] \simeq \mathbb{P}(2,4,6,10) \simeq \mathbb{P}(1,2,3,5)$ *[12], pp. 177.*
2. *A binary sextic with a zero of multiplicity three, admitting a nonvanishing invariant of positive degree, is mapped to* $p := [1,0,0,0,0] \in X$.
3. *All positive-degree invariants vanish at binary sextics with a zero of multiplicity four; they do not yield points of* X.

Geometric Invariant Theory gives an interpretation of the points of X:

Proposition 2.3.

1. *a binary sextic is stable (resp. semistable) if and only if its zeros have multiplicity* ≤ 2 *(resp.* ≤ 3*) [17], ch. 4 §1;*
2. $X - \{[1,0,0,0,0]\}$ *is a geometric quotient for binary sextics with zeros of multiplicity* ≤ 2 *[17] 1.10.*

The 'only if' part of the first assertion can be deduced from Proposition 2.2. As $X - \{D = 0\}$ is a geometric quotient for binary sextics with distinct zeros, our analysis of genus two curves in §2.1 yields

Proposition 2.4. *The moduli scheme* M_2 *can be identified with* $X - \{D = 0\}$*, where* D *is the discriminant.*

Remark 2.5. This construction definitely fails in characteristic two. If the double cover $j : C \to \mathbb{P}^1$ is wildly ramified, the branch divisor may have multiplicities > 3. These curves correspond to unstable points under the SL_2-action, and thus are not represented in the invariant-theory quotient. [11] has a detailed account of what must be done in this case.

2.4 Invariant-theory quotient as a contraction

We sketch the relationship between the invariant-theory quotient and the moduli space of stable curves.

Definition 2.6. *A birational map of normal projective varieties*

$$\beta : Y \dashrightarrow X$$

is a contraction *if β^{-1} has no exceptional divisors, i.e., the proper transform of each codimension-one subset in X has codimension one in Y.*

Proposition 2.7. *There exists a birational contraction $\beta : \overline{M}_2 \dashrightarrow X$ restricting to the identity along the open subset M_2. β is an isomorphism over $\overline{M}_2 - \Delta_1$ and contracts Δ_1 to the point p.*

Proof. To produce the birational contraction, we exhibit a morphism

$$\beta^{-1} : U \hookrightarrow \overline{M}_2$$

where $U \subset X$ is open with complement of codimension ≥ 2 and $\beta^{-1}|_{M_2 \cap U}$ is the identity. We shall take $U = X - p$, where p corresponds to the binary forms with a triple zero (cf. Proposition 2.2.)

The universal binary sextic is a hypersurface

$$W := \{ax^6+6bx^5y+15cx^4y^2+20dx^3y^3+15ex^2y^4+6fxy^5+gy^6 = 0\} \subset \mathbb{A}^7 \times \mathbb{P}^1.$$

Its class in $\mathrm{Pic}(\mathbb{A}^7 \times \mathbb{P}^1)$ is divisible by two, so there exists a double cover $\mathcal{C}' \to \mathbb{A}^7 \times \mathbb{P}^1$ simply branched over W. Composing with the projection onto the first factor, we obtain a morphism

$$\pi' : \mathcal{C}' \to \mathbb{A}^7.$$

Let $S \subset \mathbb{A}^7$ denote the open subset corresponding to forms whose zeros all have multiplicity ≤ 2 and

$$\pi : \mathcal{C} \to S$$

the restriction of π' to S. Since π is a composition of flat morphisms, it is also flat.

Consider the fiber of π over a given binary sextic F: It is a double cover $j : \mathcal{C}_F \to \mathbb{P}^1$ branched over the zeros of F. We claim $\mathcal{C}_F$ is a stable curve of genus two, not contained in Δ_1. Evidently $\mathcal{C}_F$ is smooth and simply branched over the zeros with multiplicity one. Over the double zeros $\mathcal{C}_F$ has local equation $y^2 = x^2$, which defines a node. We have $j^*\mathcal{O}_{\mathbb{P}^1}(+1) = \omega_{\mathcal{C}_F}$, which is ample on $\mathcal{C}_F$, so $\mathcal{C}_F$ is stable. The normalization $\nu : \mathcal{C}_F^\nu \to \mathcal{C}_F$ is the double cover branched along $F' = 0$, where F' is the product of the factors of F with multiplicity one; $\mathcal{C}_F$ is obtained from $\mathcal{C}_F^\nu$ by gluing the pairs of points over the each double root of F. There are three possibilities:

1. $\deg(F') = 4$, in which case $\mathcal{C}_F^\nu$ is connected of genus one;
2. $\deg(F') = 2$, in which case $\mathcal{C}_F^\nu$ is connected of genus zero;
3. $\deg(F') = 0$, in which case $\mathcal{C}_F^\nu$ has two connected components of genus zero.

Since $\mathcal{C}_F$ cannot be expressed as the union of two subcurves of genus one meeting at a point, the resulting curve is not in Δ_1.

The classifying morphism

$$S \to \overline{\mathcal{M}}_2 - \delta_1$$

is equivariant with respect to the GL_2-action on binary sextics, and therefore descends to a morphism

$$\Phi : S/\mathrm{GL}_2 \to \overline{\mathcal{M}}_2 - \delta_1.$$

We remark that this is a morphism of stacks. Since U is a geometric quotient for binary sextics with zeros of multiplicity ≤ 2 (see Proposition 2.3), U is also the coarse moduli space for S/GL_2. We define β^{-1} to be the induced morphism on coarse moduli spaces.

It remains to show this is bijective onto its image. Suppose we are given a stable curve C of genus two not contained in Δ_1. Quite generally, ω_C is globally generated for any stable curve without disconnecting nodes; the only curves in $\overline{M}_2$ with disconnecting nodes lie in Δ_1. Thus the sections of ω_C give a double cover

$$j : C \to \mathbb{P}^1$$

branched along a sextic, with zeros of multiplicity ≤ 2 because C is nodal. The analysis above shows that every such sextic arises in this way. □

2.5 Blowing up the invariant-theory quotient

We recall the principal result of [12]. Let A_g denote the moduli space of principally polarized abelian varieties of dimension g, $\overline{A}_g$ its Satake compactification. Recall that $\overline{A}_g = \mathrm{Proj}\, S$, where S is the ring of $\mathrm{Sp}(g, \mathbb{Z})$-modular forms; we use λ to denote the resulting polarization on $\overline{A}_g$. Let $t : M_g \hookrightarrow A_g$ denote the Torelli morphism, associating to each curve its Jacobian.

Now assume $g = 2$. Regarding M_2 as an open subset of X (see Proposition 2.4), t extends to a rational map

$$\tau : X \dashrightarrow \overline{A}_2.$$

The inclusion and the Torelli morphism induce

$$M_2 \hookrightarrow \widetilde{X} := \overline{\mathrm{Graph}(\tau)} \subset X \times \overline{A}_2. \tag{4}$$

In particular, $\widetilde{X}$ compactifies M_2.

Theorem 2.8. *[12] The indeterminacy of τ is the point $p = [1, 0, 0, 0, 0] \in X$ corresponding to binary sextic forms with a zero of multiplicity three. If we choose local coordinates at p,*

$$x_1 = 2^4 3^2 B/A^2 \quad x_2 = 2^6 3^3 (3C - AB)/A^3 \quad x_3 = 2 \cdot 3^5 D/A^5,$$

then τ is resolved by a weighted blow-up centered at p,

$$b : \widetilde{X} \to X$$

with weights

$$\mathit{weight}(x_1) = 2 \quad \mathit{weight}(x_2) = 3 \quad \mathit{weight}(x_3) = 6.$$

The exceptional divisor of b is mapped isomorphically to the locus of principally polarized abelian surfaces that decompose as a product of two elliptic curves (with the induced product polarization).

Remark 2.9. Igusa's result is considerably more precise: He explicitly computes the correspondence between the ring of invariants R and the ring of modular forms S. In particular, S is a polynomial ring with generators in degrees $4, 6, 10, 12$ and the locus of products is given by the vanishing of a form of weight 10.

2.6 Comparing the blow-up with moduli space

Proposition 2.10. *The open imbedding $M_2 \hookrightarrow \widetilde{X}$ extends to a birational map*

$$\gamma : \overline{M}_2 \dashrightarrow \widetilde{X}$$

which is an isomorphism in codimension one.

In particular, γ and γ^{-1} are both birational contractions. In Proposition 4.2 we will prove that γ is an isomorphism.

Proof. The Torelli morphism admits an extension $\overline{t} : \overline{M}_g \to \overline{A}_g$ [19], Theorem 3. This is not an isomorphism for $g > 1$: The divisor $\Delta_0 \subset \overline{M}_g$ is mapped to a boundary stratum of $\overline{A}_g$, which has codimension ≥ 2. However, in genus two, $\overline{t}$ is an isomorphism at the generic point of Δ_1. Indeed, the Jacobian of a curve $[E_1 \cup_q E_2] \in \Delta_1$, with E_1 and E_2 smooth of genus one, is the abelian surface $E_1 \times E_2$.

The following diagram summarizes the various birational maps and morphisms:

$$\begin{array}{ccc} \overline{A}_2 & \stackrel{\pi}{\leftarrow} & \widetilde{X} \\ {\scriptstyle \overline{t}}\uparrow & & \downarrow{\scriptstyle b} \\ \overline{M}_2 & \stackrel{\beta}{\dashrightarrow} & X \end{array}.$$

By Theorem 2.8, π is also an isomorphism over the generic point of $\overline{t}(\Delta_1)$, so γ is an isomorphism at the generic point of Δ_1. β and b are both isomorphisms over the generic point of the divisor $\beta(\Delta_0)$ (see Proposition 2.7), so γ is an isomorphism at the generic point of Δ_0. Since γ is regular along $M_2 = \overline{M}_2 - \Delta_0 - \Delta_1$, the result follows. □

The proper transforms of Δ_0 and Δ_1 in $\widetilde{X}$ are denoted $\widetilde{\Delta}_0$ and $\widetilde{\Delta}_1$. Thus $\widetilde{\Delta}_1$ is the exceptional divisor of $b : \widetilde{X} \to X$.

Remark 2.11 (Bibliographic note). There are a number of partial desingularizations of $\overline{A}_g$ through which $\overline{t}$ factors, e.g., the 'Igusa monoidal transform' [13], [19] and the toroidal compactification associated to the 2nd Voronoi fan [20]. When $g = 2$, these approaches coincide [20] Remark 2.8 and yield a partial desingularization $\widetilde{A}_2 \to \overline{A}_2$. See [13] Theorem 5 for a blow-up representation, expressed in terms of modular forms; the center of this blow-up is in the boundary $\overline{A}_2 - A_2$. Namikawa [19] §9 has shown that the factorization $\overline{M}_2 \xrightarrow{\sim} \widetilde{A}_2$ is an isomorphism. Notwithstanding Igusa's explicit formulas for $\tau : X \dashrightarrow \overline{A}_2$ and $\widetilde{A}_2 \to \overline{A}_2$, it is not entirely obvious how to extract an isomorphism $\widetilde{A}_2 \xrightarrow{\sim} \widetilde{X}$.

3 Stack geometry

3.1 A stack-theoretic quotient

Proposition 2.2 might suggest that the invariant E is irrelevant to the geometry of the quotient. However, we have so far ignored possible *stack structures* on the quotient, which are intertwined with the geometry of E. There are a number of natural stacks to consider, including the GL_2-quotient stack. Our choice is dictated by pedagogical imperatives, i.e., to exhibit concretely the nontrivial inertia along the bielliptic locus where E vanishes.

The ring of invariants R is graded by degree, so we have a natural $\mathbb{G}_m$-action on the affine variety $Y = \operatorname{Spec} R$. Now $\mathbb{G}_m$ acts on the open subset $Y - (0,0,0,0,0)$ with finite stabilizers and closed orbits, so the quotient stack

$$\mathcal{X} := (Y - (0,0,0,0,0))\,/\mathbb{G}_m$$

is a separated Deligne-Mumford stack with coarse moduli space $q : \mathcal{X} \to X$ (see [15] 10.13.2,7.6,8.1 for more information). The points of $Y - (0,0,0,0,0)$ with nontrivial stabilizer map to the points of $\mathcal{X}$ with nontrivial inertia groups; this is the ramification locus of q.

We collect some geometric properties of $\mathcal{X}$:

Proposition 3.1.

1. *The closed imbedding* $\operatorname{Spec} R \hookrightarrow \operatorname{Spec} k[A,B,C,D,E]$ *induces a closed imbedding* $i : \mathcal{X} \hookrightarrow \mathcal{P}(2,4,6,10,15)$*, where*

$$\mathcal{P}(2,4,6,10,15) := (\operatorname{Spec} k[A,B,C,D,E] - (0,0,0,0,0))\,/\mathbb{G}_m,$$

with $\mathbb{G}_m$ *acting with weights* $(2,4,6,10,15)$*,*

$$t \cdot (A,B,C,D,E) \mapsto (t^2A, t^4B, t^6C, t^{10}D, t^{15}E);$$

2. *the dualizing sheaf of* $\mathcal{X}$ *is given via adjunction*

$$\omega_{\mathcal{X}} = i^*\omega_{\mathcal{P}(2,4,6,10,15)}(\mathcal{X}) \simeq i^*\mathcal{O}_{\mathcal{P}(2,4,6,10,15)}(-7),$$

where $\mathcal{O}_{\mathcal{P}(2,4,6,10,15)}(+1)$ *is the invertible sheaf associated with the principal* $\mathbb{G}_m$*-bundle classified by the identity character of* $\mathbb{G}_m$*;*

3. *the ramification divisor of* q *is*

$$\mathcal{E} := \{E = 0\} \subset \mathcal{X}$$

and we have

$$q^*\omega_X \simeq \omega_{\mathcal{X}}(-\mathcal{E}) \simeq i^*\mathcal{O}_{\mathcal{P}(2,4,6,10,15)}(-22).$$

Proof. The first two assertions do not require proof. As for the third, it follows from the classification of possible automorphisms of binary sextics [2], [11] §8 The only automorphism type occuring in codimension one is the bielliptic involution (cf. Proposition 2.1). □

Remark 3.2. Notwithstanding Proposition 2.4, $\mathcal{M}_2$ is *not* contained in $\mathcal{X}$ as an open substack. Using the functional relation (3), the inertia group of $[F] \in \mathcal{X} - p$ is the quotient

$$\{M \in \mathrm{GL}_2 : (M, F) = F\}/\{M \in \mathrm{SL}_2 : (M, F) = F\},$$

which is trivial for generic binary forms.

The inertia group at $[C] \in \mathcal{M}_2$ is $\mathrm{Aut}(C)$, and the presence of the hyperelliptic involution ι means this is always nontrivial. Now $\mathrm{Aut}(C)$ has a natural representation on $\Gamma(C, \omega_C)$ and an induced representation on $\wedge^2\Gamma(C, \omega_C)$ that is not faithful: We do not see elements of $\mathrm{Aut}(C)$ acting on $\Gamma(C, \omega_C)$ with determinant one, e.g., ι, which acts on $\Gamma(C, \omega_C)$ by $-I$. The corresponding quotient of $\mathrm{Aut}(C)$ is the inertia picked up by $\mathcal{X}$.

3.2 Analysis of boundary divisors in the moduli stack

Lemma 3.3. *Every stable curve of genus two admits a canonical hyperelliptic involution, which is central in its automorphism group.*

Proof. We claim that every stable curve of genus two is canonically a double cover of a nodal curve of genus zero,

$$j : C \to R, \quad R = \begin{cases} \mathbb{P}^1 & \text{if } [C] \notin \Delta_1, \\ \mathbb{P}^1 \cup_r \mathbb{P}^1 & \text{if } [C] \in \Delta_1, \end{cases}$$

branched over six smooth points $b_1, \ldots, b_6 \in R$ and the node $r \in R$. The covering transformation ι therefore commutes with each automorphism of C.

The cover is induced by

$$C \to \mathbb{P}(\Gamma(C, \omega_C^2)) \simeq \mathbb{P}^2$$

which factors

$$C \xrightarrow{j} R \subset \mathbb{P}^2,$$

where R is a plane conic and j is finite of degree two. Indeed, for curves not in Δ_1 this is the double cover $C \to \mathbb{P}^1$ discussed in the proof of Proposition 2.7. For curves $C = E_1 \cup_q E_2 \in \Delta_1$ with $q = E_1 \cap E_2$ the disconnecting node joining the genus-one components E_1 and E_2, we have a double cover

$$j : E_1 \cup_q E_2 \longrightarrow \mathbb{P}^1 \cup_r \mathbb{P}^1,$$

with $j(q) = r$ and j mapping the genus one components two-to-one onto rational components, with ramification at q along each component. □

Consider stable curves of genus two with automorphisms beyond ι. There are two possibilities: Either $(R, b_1, \ldots, b_6)$ admits automorphisms permuting the b_i or $j : C \to R$ admits covering transformations other than the canonical hyperelliptic involution. The classification of automorphism groups ([2] or [11]§8) yields the following possibilities in codimension one:

1. the curves C in Δ_1; here $j : C \to \mathbb{P}^1 \cup \mathbb{P}^1$ admits involutions fixing each component of C;
2. the closure of the locus of curves $j : C \to \mathbb{P}^1$ branched over six points admitting a bielliptic involution.

At each point $[C] \in \overline{M}_2$, the moduli space is étale-locally isomorphic to the quotient $T_{[C]}\overline{\mathcal{M}}_2/\mathrm{Aut}(C)$ at the origin, where

$$T_{[C]}\overline{\mathcal{M}}_2 = \mathrm{Ext}^1(\Omega_C^1, \mathcal{O}_C))$$

is the tangent space with the induced automorphism action. When C is smooth, Serre duality gives

$$T_{[C]}\overline{\mathcal{M}}_2 = \Gamma(C, \omega_C^2)^*;$$

equation (2) shows that the hyperelliptic involution acts trivially and a bielliptic involution acts by reflection.

This local isomorphism can be chosen so the divisors Δ_0 and Δ_1, correspond to the images of unions of distinguished hyperplanes in $T_{[C]}\overline{\mathcal{M}}_2$. The local-global spectral sequence gives an exact sequence

$$0 \to H^1(\mathcal{H}om(\Omega_C^1, \mathcal{O}_C)) \to \mathrm{Ext}^1(\Omega_C^1, \mathcal{O}_C) \to \Gamma(\mathcal{E}xt^1(\Omega_C^1, \mathcal{O}_C)) \to 0. \quad (5)$$

A local computation implies

$$\mathcal{E}xt^1(\Omega_C^1, \mathcal{O}_C) \simeq \oplus_{\text{nodes } p \in C} k.$$

If $C = E_1 \cup_q E_2$ is in Δ_1, let $\widehat{\delta_1} \subset T_{[C]}\overline{\mathcal{M}}_2$ denote the hyperplane corresponding to the node disconnecting the genus-one components of C, i.e., the kernel of the projection onto the direct summand corresponding to q. The extra covering involutions of $j : C \to \mathbb{P}^1 \cup \mathbb{P}^1$ mentioned above act on $T_{[C]}\overline{\mathcal{M}}_2$, trivially on $\widehat{\delta_1}$ and by multiplication by (-1) on $\mathcal{E}xt^1(\Omega^1_C, \mathcal{O}_C)_q$. This corresponds to reflection across $\widehat{\delta_1}$. Let $\widehat{\delta_0}$ denote the union of the hyperplanes corresponding to each of the non-disconnecting nodes of C. The surjectivity of the last arrow in (5) means that $\widehat{\delta} = \widehat{\delta_1} \cup \widehat{\delta_0}$ is normal crossings.

Definition 3.4. *Let $\xi \subset \overline{\mathcal{M}}_2$ (resp. $\Xi \subset \overline{M}_2$) denote the closure of the smooth curves admitting a bielliptic involution.*

These are irreducible of codimension one, e.g., by Proposition 2.1 and the characterization of bielliptic curves. This can also be seen infinitesimally: Under the local identification with $T_{[C]}\overline{\mathcal{M}}_2/\mathrm{Aut}(C)$, each branch of ξ is identified with the hyperplane $T_{[C]}\overline{\mathcal{M}}_2$ fixed by the corresponding bielliptic involution (acting by reflection on the tangent space). The union of these hyperplanes is denoted $\widehat{\xi}$.

The bielliptic divisor has more complicated local geometry, as Ξ may have quite a few local branches. For example,

$$F(x,y) = xy(x-\xi y)(x+\xi y)(x-\xi^{-1}y)(x+\xi^{-1}y)$$

has automorphism group isomorphic to the Klein four-group and admits two involutions

$$[x,y] \mapsto [y,x] \quad [x,y] \mapsto [-y,x].$$

The cyclotomic form

$$F(x,y) = x^6+y^6 = (x-\zeta y)(x-\zeta^3 y)(x-\zeta^5 y)(x-\zeta^7 y)(x-\zeta^9 y)(x-\zeta^{11} y) \quad \zeta^{12} = 1$$

has automorphism group isomorphic to the dihedral group with 12 elements and admits *four* distinct involutions

$$[x,y] \mapsto [y,x] \quad [x,y] \mapsto [-x,y] \quad [x,y] \mapsto [\zeta^4 y,x] \quad [x,y] \mapsto [\zeta^8 y,x].$$

In particular, $\widehat{\xi}$ is *not* normal crossings.

Proposition 3.5 (Ramification formula for $Q : \overline{\mathcal{M}}_2 \to \overline{M}_2$).

$$K_{\overline{\mathcal{M}}_2} + \alpha\delta \equiv Q^*(K_{\overline{M}_2} + \alpha\Delta_0 + \frac{1+\alpha}{2}\Delta_1 + \frac{1}{2}\Xi)$$

Proof. Lemma 3.3 allows us to consider the rigidification ([1] §5) of $\overline{\mathcal{M}}_2$ with respect to the group $\langle\iota\rangle$ generated by the canonical involution

$$r : \overline{\mathcal{M}}_2 \to \overline{\mathcal{M}}_2^{<\iota>}.$$

Given a scheme T, T-valued points of $\overline{\mathcal{M}}_2^{<\iota>}$ correspond to families of genus two stable curves over T, where two families are identified if they differ by a $\langle\iota\rangle$-valued cocycle over T. The inertia group of $\overline{\mathcal{M}}_2^{<\iota>}$ at $[C]$ is the quotient

$$\mathrm{Aut}(C)/\langle\iota\rangle\,.$$

We have a factorization

$$\overline{\mathcal{M}}_2 \xrightarrow{r} \overline{\mathcal{M}}_2^{<\iota>} \xrightarrow{\bar{q}} \overline{M}_2$$

so that r is étale of degree two and $\overline{M}_2$ is the coarse moduli space of $\overline{\mathcal{M}}_2^{<\iota>}$ [1] Theorem 5.1.5. We therefore obtain the following formula for dualizing sheaves,

$$r^*\omega_{\overline{\mathcal{M}}_2^{<\iota>}} \simeq \omega_{\overline{\mathcal{M}}_2}.$$

Let $\delta_1^{<\iota>}$ and $\xi^{<\iota>}$ denote the corresponding Cartier divisors in $\overline{\mathcal{M}}_2^{<\iota>}$. As $\bar{q}$ has simple ramification along these divisors, we obtain

$$\begin{aligned}\bar{q}^*\Delta_1 = 2\delta_1^{<\iota>}, \quad \bar{q}^*\Xi = 2\xi^{<\iota>},\\ \bar{q}^*K_{\overline{M}_2} = K_{\overline{\mathcal{M}}_2^{<\iota>}} - \delta_1^{<\iota>} - \xi^{<\iota>},\end{aligned}$$

which together imply the formula. □

3.3 Analysis of $b : \widetilde{X} \to X$ along the exceptional divisor

By Theorem 2.8, the exceptional divisor $\widetilde{\Delta}_1$ is mapped isomorphically to the locus in $\overline{A}_2$ parametrizing abelian surfaces decomposing into products of elliptic curves (as a principally polarized abelian variety); this is isomorphic to $\mathbb{P}(2,3,6)$.

Proposition 3.6. *Let $\widetilde{\Xi}$ denote the proper transform of Ξ in $\widetilde{X}$. We have the following formulas:*

$$\begin{aligned}K_{\widetilde{X}} &\equiv b^*K_X + 10\widetilde{\Delta}_1,\\ b^*\{G=0\} &\equiv \widetilde{\Xi} + 12\widetilde{\Delta}_1,\\ b^*\{D=0\} &\equiv \widetilde{\Delta}_0 + 6\widetilde{\Delta}_1,\\ \widetilde{\Xi} &\equiv 3\widetilde{\Delta}_0 + 12\widetilde{\Delta}_1.\end{aligned}$$

We pause to explore the geometry of $\widetilde{\Delta}_1$. Naively, one might expect this to be the symmetric square of the moduli space of elliptic curves. However, in taking symmetric squares we should be mindful of the stack structures. The coarse moduli space of the symmetric square need not be isomorphic to the symmetric square of the coarse moduli space.

The standard theory of modular forms implies

$$\overline{\mathcal{M}}_{1,1} \simeq \mathcal{P}(4,6) = (\operatorname{Spec} k[g_2, g_3] - (0,0))/\mathbb{G}_m,$$

with $\mathbb{G}_m$ acting with weights $(4,6)$,

$$t \cdot (g_2, g_3) \mapsto (t^4 g_2, t^6 g_3).$$

The coarse moduli space is

$$\overline{M}_{1,1} = \overline{A}_1 \simeq \operatorname{Proj} k[g_2, g_3] \simeq \mathbb{P}^1.$$

The symmetric square of the stack has the following quotient-stack presentation:

$$\begin{aligned}(\overline{\mathcal{M}}_{1,1} \times \overline{\mathcal{M}}_{1,1})/\mathfrak{S}_2 &= (\operatorname{Spec} k[g_2, g_3, h_2, h_3] - Z\}/\, H, \\ Z &= \{(g_2, g_3, h_2, h_3) : g_2 = g_3 = 0 \text{ or } h_2 = h_3 = 0\}.\end{aligned}$$

Here H is the group generated by the torus

$$(t,u) \cdot (g_2, g_3, h_2, h_3) \mapsto (t^4 g_2, t^6 g_3, u^4 h_2, u^6 h_3)$$

and the involution

$$(g_2, g_3, h_2, h_3) \mapsto (h_2, h_3, g_2, g_3),$$

i.e., $H = \mathfrak{S}_2 \ltimes \mathbb{G}_m^2$ where $\mathfrak{S}_2$ acts on $\mathbb{G}_m^2$ by interchanging the factors.

The coarse moduli space of the stack is the invariant-theory quotient for the action of H. Consider the elements $p \in k[g_2, g_3, h_2, h_3]$ with the following properties:

1. $p(g_2, g_3, h_2, h_3) = p(h_2, h_3, g_2, g_3)$;
2. $p(t^2 g_2, t^3 g_3, u^2 h_2, u^3 h_3) = (tu)^N p(g_2, g_3, h_2, h_3)$ for some N.

This ring is generated by $g_2 h_2, g_3 h_3$, and $g_2^3 h_3 + g_3^2 h_2^3$ and

$$\operatorname{Proj} k[g_2 h_2, g_3 h_3, g_2^3 h_3^2 + g_3^2 h_2^3] \simeq \mathbb{P}(4,6,12) \simeq \mathbb{P}(2,3,6),$$

which explains why the weights of b are $(2,3,6)$. See [12], Theorem 3, for a discussion in terms of the modular forms for $\operatorname{Sp}(2,\mathbb{Z})$ (see Remark 2.9).

Proof of proposition. The first equation follows because $b : \widetilde{X} \to X$ has weights $(2,3,6)$. As for the second, $\widetilde{\Xi}$ is the proper transform of the divisor $\{G = 0\} \subset X$ parametrizing forms admitting an bielliptic involution. When a bielliptic curve of genus two specializes to a stable curve in Δ_1, the bielliptic involution specializes to a morphism exchanging the elliptic components. Therefore, $\widetilde{\Xi} \cap \widetilde{\Delta}_1 \subset \widetilde{\Delta}_1$ is the diagonal in the symmetric square, which is cut out by a form of weighted-degree twelve.

For the third equation, $\widetilde{\Delta}_0$ is the proper transform of $\{D = 0\}$. The intersection $\widetilde{\Delta}_0 \cap \widetilde{\Delta}_1 \subset \widetilde{\Delta}_1$ is the locus where the discriminant $\Delta = g_2^3 - 27 g_3^2$ vanishes, and thus has weighted-degree six. The last equation follows because G has weighted degree thirty and D has weighted degree ten. □

One consequence of this analysis is worth mentioning.

Proposition 3.7. *For sufficiently small $\epsilon > 0$, the divisor*

$$\widetilde{\Delta}_0 + (6 - \epsilon)\widetilde{\Delta}_1$$

is ample on $\widetilde{X}$.

Proof. The divisor can be expressed

$$b^*(\text{ample divisor}) - \epsilon(b\text{-exceptional divisor}),$$

which yields a polarization of the blow-up $b : \widetilde{X} \to X$. □

4 Birational geometry

4.1 Divisor classes and birational contractions of $\overline{M}_2$

It is well known that the rational divisor class group of $\overline{M}_2$ is freely generated by the boundary divisors Δ_0 and Δ_1; see [7] for a nice account of divisors on $\overline{M}_g$ for arbitrary g. When $g = 2$, Proposition 2.7 gives an elementary proof of this fact: Since $X \simeq \mathbb{P}(1, 2, 3, 5)$ its divisor class group has rank one and is generated by the discriminant divisor $\{D = 0\}$; the same holds true for $X - p$. By Proposition 2.10, $\overline{M}_2$ is isomorphic to $\widetilde{X}$ up to codimension ≤ 1, so these have isomorphic class groups. It follows that the rational divisor class group of $\overline{M}_2$ is generated by Δ_1 and the proper transform of the discriminant, which is just Δ_0.

The nef cone of $\overline{M}_2$ is also well-known. We will not give a self-contained proof here, but rather rely on the general result of Cornalba-Harris [5]:

Theorem 4.1. *The line bundle $a\lambda - b\delta$ is nef on $\overline{M}_g, g \geq 2$, if and only if $a \geq 11b \geq 0$.*

Here λ is the pull-back of the polarization on $\overline{A}_g$ via the extended Torelli map $\overline{t} : \overline{M}_g \to \overline{A}_g$ (see §2.6).

To apply this in our situation, we observe that

$$\lambda \equiv \frac{1}{10}(\Delta_0 + \Delta_1)$$

over $\overline{M}_2$ (see [7] pp. 175). The factor 10 can be explained by the fact that $\overline{t}(\Delta_1)$ is defined by the vanishing of a modular form of weight ten (see Remark 2.9 and [12]). Substitution gives the first part of

Proposition 4.2. *The nef cone of $\overline{M}_2$ is generated by the divisors $\Delta_0 + \Delta_1$ and $\Delta_0 + 6\Delta_1$, respectively. These are both semiample, inducing the birational contractions*

$$\overline{t} : \overline{M}_2 \to \overline{A}_2 \quad \beta : \overline{M}_2 \to X.$$

The rational map $\gamma : \overline{M}_2 \dashrightarrow \widetilde{X}$ is an isomorphism.

Remainder of proof: Of course, $\Delta_0 + \Delta_1$ is semiample and induces the birational contraction morphism $\overline{t} : \overline{M}_2 \to \overline{A}_2$. As $\widetilde{X}$ and $\overline{M}_2$ are isomorphic in codimension one, Proposition 3.7 says that $\widetilde{\Delta}_0 + (6-\epsilon)\widetilde{\Delta}_1$ is ample on $\widetilde{X}$ and the corresponding divisor $\Delta_0 + (6-\epsilon)\Delta_1$ is ample on $\overline{M}_2$. It follows that $\widetilde{X}$ and $\overline{M}_2$ are each isomorphic to Proj of

$$\oplus_{n\geq 0}\Gamma(\mathcal{O}_{\widetilde{X}}(n(\widetilde{\Delta}_0 + (6-\epsilon)\widetilde{\Delta}_1))) \simeq \oplus_{n\geq 0}\Gamma(\mathcal{O}_{\overline{M}_2}(n(\Delta_0 + (6-\epsilon)\Delta_1))).$$

In particular, the rational map γ is an isomorphism. Thus the contractions $b : \widetilde{X} \to X$ and $\beta : \overline{M}_2 \dashrightarrow X$ coincide. □

4.2 Canonical class of $\overline{M}_2$

The canonical class $K_{\overline{M}_2}$ can also be computed by elementary methods. We know that

$$\omega_{\mathbb{P}(1,2,3,5)} \simeq \mathcal{O}_{\mathbb{P}(1,2,3,5)}(-11),$$

which also follows from the third part of Proposition 3.1. Since the discriminant has degree ten, we find

$$K_X = -\frac{11}{5}\{D = 0\}.$$

Applying the formulas of Proposition 3.6, we obtain

$$K_{\widetilde{X}} \equiv -\frac{11}{5}\widetilde{\Delta}_0 - \frac{16}{5}\widetilde{\Delta}_1,$$

$$K_{\widetilde{X}} + \frac{1}{2}\widetilde{\Delta}_1 + \frac{1}{2}\widetilde{\Xi} \equiv -\frac{7}{10}\widetilde{\Delta}_0 + \frac{3}{10}\widetilde{\Delta}_1.$$

Since $\widetilde{X}$ and $\overline{M}_2$ agree in codimension one, they have the same canonical class

$$K_{\overline{M}_2} \equiv -\frac{11}{5}\Delta_0 - \frac{16}{5}\Delta_1.$$

In particular, we obtain

$$K_{\overline{M}_2} + \alpha\Delta_0 + \frac{1+\alpha}{2}\Delta_1 + \frac{1}{2}\Xi \equiv (-\frac{7}{10} + \alpha)\Delta_0 + (\frac{3}{10} + \alpha/2)\Delta_1. \quad (6)$$

Remark 4.3. The importance of this divisor stems from the ramification equation of Proposition 3.5. This divisor class pulls back to the class

$$K_{\overline{\mathcal{M}}_2} + \alpha\delta$$

on the moduli stack. We shall interpret log canonical models of the moduli stack using this divisor.

Using the computations of Proposition 3.6, we obtain discrepancy equations for $\beta : \overline{M}_2 \to X$,

$$K_{\overline{M}_2} + \alpha\Delta_0 + \frac{1}{2}\Xi = \beta^*(K_X + \alpha\{D = 0\} + \frac{1}{2}\{G = 0\}) + (4 - 6\alpha)\Delta_1, \quad (7)$$

$$K_{\overline{M}_2} + \alpha\Delta_0 + \frac{1+\alpha}{2}\Delta_1 + \frac{1}{2}\Xi = \beta^*(K_X + \alpha\{D = 0\} + \frac{1}{2}\{G = 0\}) + \frac{9 - 11\alpha}{2}\Delta_1. \quad (8)$$

4.3 Generalities on log canonical models

See [14] §2 for definitions of the relevant terms and technical background. Let V be a normal projective variety, $D = \sum_{i=1}^{n} a_i D_i$ a $\mathbb{Q}$-divisor such that $0 \leq a_i \leq 1$ and $K_V + D$ is $\mathbb{Q}$-Cartier. Abusing notation, we write $V - D$ for $V - \cup_i D_i$.

Definition 4.4. (V, D) *is a* strict log canonical model *if* $K_V + D$ *is ample,* (V, D) *has log canonical singularities, and* $V - D$ *has canonical singularities.*

The idea here is to realize (V, D) as a log canonical model without introducing boundary divisors over $V - D$. This is natural if we want to respect the geometry of the open complement.

The following recognition criterion for strict log canonical models is based on [14] §2:

Proposition 4.5. *Consider birational projective contractions* $\rho : \widetilde{V} \to V$ *where the exceptional locus of* ρ *is divisorial,* D_i' *denotes the proper transform of* D_i*, and* E_j *(resp.* F_k*) denotes the exceptional divisors of* ρ *with* $\rho(E_j) \subset D$ *(resp.* $\rho(F_k) \not\subset D$*).*

The following are equivalent:

1. (V, D) *is a strict log canonical model.*
2. *For some resolution of singularities* $\rho : \widetilde{V} \to V$*, with the union of the exceptional locus and* $\cup_i D_i'$ *normal crossings, there exist* $b_j \in \mathbb{Q} \cap [0, 1]$ *so that*

$$\widetilde{D} = \sum_i a_i D_i' + \sum_j b_j E_j$$

satisfies the formula

$$K_{\widetilde{V}} + \widetilde{D} \equiv \rho^*(K_V + D) + \sum_j d_j E_j + \sum_k e_k F_k, \quad d_j, e_k \geq 0. \quad (9)$$

For each such choice of b_j*,*

$$K_{\widetilde{V}} + \widetilde{D} - \sum_j d_j E_j - \sum_k e_k F_k$$

is semiample and induces ρ*. Furthermore, we may take the* $b_j = 1$*.*

3. For some contraction $\rho : \widetilde{V} \to V$, there exist $b_j \in \mathbb{Q} \cap [0,1]$ so that

$$\widetilde{D} = \sum_i a_i D_i' + \sum_j b_j E_j$$

satisfies the formula

$$K_{\widetilde{V}} + \widetilde{D} \equiv \rho^*(K_V + D) + \sum_j d_j E_j + \sum_k e_k F_k, \quad d_j, e_k \geq 0,$$

$(\widetilde{V}, \widetilde{D})$ is log canonical, and $\widetilde{V} - \widetilde{D}$ has canonical singularities. The divisor

$$K_{\widetilde{V}} + \widetilde{D} - \sum_j d_j E_j - \sum_k e_k F_k$$

is semiample and induces ρ.

Some general facts are worth mentioning before we indicate the proof. First, we can decide whether a pair is canonical or log canonical by computing discrepancies on *any* resolution. Second, discrepancies increase as the coefficients of the log divisor are decreased [14] 2.17.3. Third, in situations (2) and (3) the pair (V, D) is the *log canonical model* of $(\widetilde{V}, \widetilde{D})$; such models are unique [14] 2.22.1.

Proof. It is trivial that the second statement implies the third. To see that the third implies the first, take a resolution for $(\widetilde{V}, \widetilde{D})$ so that the union of the exceptional locus and all the proper transforms of the D_i', E_j, and F_k is normal crossings. Comparing discrepancies for $(\widetilde{V}, \widetilde{D})$ and (V, D), using the fact the coefficients of components in $\widetilde{D}$ are at least as large as the coefficents of the corresponding components appearing in $\rho^*(K_V + D)$, we find that (V, D) is log-canonical and has canonical singularities along $V - D$. Since $\rho^*(K_V + D)$ induces ρ, $K_V + D$ must be ample on V.

For the remaining implication, since (V, D) has log canonical singularities and canonical singularities away from D, the discrepancy equation (9) follows. Since $K_V + D$ is ample on V, its pull-back to $\widetilde{V}$ is semiample and induces ρ. □

We shall also need the following basic fact, a special case of [14] 20.2, 20.3:

Proposition 4.6. *Let W be a smooth variety and $h : W \to V$ be a finite dominant morphism to a normal variety. Let $D = \sum_i a_i D_i, 0 \leq a_i \leq 1$ be a $\mathbb{Q}$-divisor on V containing all the divisorial components of the branch locus of h. Let $\bar{D}$ be a $\mathbb{Q}$-divisor on W so that* $\operatorname{supp}(h^{-1}(D)) = \operatorname{supp}(\bar{D})$ *and $h^*(K_V + D) = K_W + \bar{D}$. Then (V, D) has log canonical singularities along D iff $(W, \bar{D})$ has log canonical singularities along $\bar{D}$.*

If D and $\bar{D}$ have multiplicity one at each component, then $h^*(K_V + D) = K_W + \bar{D}$ follows from the other assumptions.

4.4 Example of $\overline{M}_g$

The standpoint of this section owes a great deal to Mumford [18] [16]:

Theorem 4.7. *For $g \geq 4$, the pair $(\overline{M}_g, \Delta)$ is a strict log canonical model.*

Remark 4.8. This is also the natural log canonical model from the point of view of the moduli stack. Indeed, for $g \geq 4$ the locus in M_g of curves with automorphisms has codimension ≥ 2, so the branch divisor of $Q : \overline{\mathcal{M}}_g \to \overline{M}_g$ is just Δ_1; over Δ_1, we have simple ramification. We therefore have ([8] pp. 52) $Q^* K_{\overline{M}_g} = K_{\overline{\mathcal{M}}_g} - \delta_1$ and thus

$$Q^*(K_{\overline{M}_g} + \Delta) = K_{\overline{\mathcal{M}}_g} + \delta.$$

Proof (Sketch). We first check that $K_{\overline{M}_g} + \Delta$ is ample. The formula from [8] §2 (or [7])

$$K_{\overline{M}_g} = 13\lambda - 2\Delta_0 - 3/2\Delta_1 - 2 \sum_{2 \leq i \leq g/2} \Delta_i$$

gives $K_{\overline{M}_g} + \delta = 13\lambda - \delta$, which is ample by Theorem 4.1 (see also [18]).

The singularity analysis follows [8]. $\overline{M}_g$ has canonical singularities by Theorem 1 of [8]. To show that $(\overline{M}_g, \Delta)$ has log canonical singularities, we use the fact that $\overline{M}_g$ is étale-locally a quotient of a smooth variety by a finite group. At $[C]$ it has a local presentation

$$h : T_{[C]}\overline{\mathcal{M}}_g \twoheadrightarrow T_{[C]}\overline{\mathcal{M}}_g / \mathrm{Aut}(C)$$

in terms of its tangent space

$$T_{[C]}\overline{\mathcal{M}}_g = \mathrm{Ext}^1(\Omega^1_C, \mathcal{O}_C).$$

We analyze the quotient morphism using Proposition 4.6. The preimage of the boundary divisor corresponds to a union of hyperplanes $\widehat{\delta} \subset T_{[C]}\overline{\mathcal{M}}_g$, meeting in normal crossings. The pair $(T_{[C]}\overline{\mathcal{M}}_g, \bar{\Delta})$ then has log canonical singularities. An application of Proposition 4.6, utilizing the ramification discussion in Remark 4.8, implies $(K_{\overline{\mathcal{M}}_g}, \Delta)$ is log canonical. □

Remark 4.9. Using the full force of Theorem 4.1 we get a sharper statement. Consider the pair

$$(\overline{M}_g, \alpha\Delta_0 + \frac{1+\alpha}{2}\Delta_1 + \alpha(\Delta_2 + \cdots + \Delta_{\lfloor g/2 \rfloor})),$$

with log canonical divisor pulling back to $K_{\overline{\mathcal{M}}_g} + \alpha\delta$ on the moduli stack. The $\mathbb{Q}$-divisor $K_{\overline{\mathcal{M}}_g} + \alpha\delta$ is the pull-back of an ample line bundle if and only if $9/11 < \alpha \leq 1$. Since $\overline{M}_g$ is a locally a quotient of a smooth variety by a finite group, all divisors on $\overline{M}_g$ are $\mathbb{Q}$-Cartier. An easy computation with the discrepancy equation (9) then shows that the pair remains log canonical even as the coefficients are reduced.

4.5 Application to $\overline{\mathcal{M}}_2$

The source of the special difficulties in this case is the fact that $K_{\overline{M}_2} + \Delta$ is not effective. Indeed, in §4.1 we computed

$$K_{\overline{M}_2} \equiv -\frac{11\Delta_0 + 6\Delta_1}{5}$$

so $K_{\overline{M}_2} + \Delta = -$effective divisor.

In order to recover a result analogous to Theorem 4.7, we must take the 'log canonical model of the moduli stack', as interpreted on $\overline{M}_2$ via Proposition 3.5:

Theorem 4.10. *Consider the log canonical model of $\overline{\mathcal{M}}_2$ with respect to the $K_{\overline{\mathcal{M}}_2} + \alpha\delta$, i.e., the log canonical model of $\overline{M}_2$ with respect to*

$$K_{\overline{M}_2} + \alpha\Delta_0 + \frac{1+\alpha}{2}\Delta_1 + \frac{1}{2}\Xi.$$

1. *For $9/11 < \alpha \leq 1$, we recover $\overline{M}_2$.*
2. *For $7/10 < \alpha \leq 9/11$ we recover the invariant theory quotient $X \simeq \mathbb{P}(1,2,3,5)$.*
3. *For $\alpha = 7/10$ we get a point; the log canonical divisor fails to be effective for $\alpha < 7/10$.*

Proof. The necessary ampleness results have already been stated. Proposition 4.2 and Equation (6) imply the log canonical divisor on $\overline{M}_2$ is ample if and only if $\alpha > 9/11$. Proposition 3.1 implies that

$$K_X + 1/2\{G = 0\} + \alpha\{D = 0\}$$

is positive on X if and only if $\alpha > 7/10$. When $\alpha = 7/10$ it is zero and when $\alpha < 7/10$ it is negative.

It remains to verify the singularity conditions: First, we check that $\overline{M}_2$ has canonical singularities away from Δ_0, Δ_1, and Ξ. Suppose that C is not in the boundary and does not admit admit a bielliptic involution. In Proposition 2.4 we saw

$$M_2 \simeq X - \{D = 0\} \subset \mathbb{P}(1,2,3,5),$$

so we need to analyze the singularities of $\mathbb{P}(1,2,3,5) - \{D = 0\}$. A point in weighted projective space is nonsingular when the weights corresponding to its non-vanishing coordinates are relatively prime, so the only possible singularity occurs when $A = B = C = 0$. The corresponding binary sextic form is

$$x(x^5 + y^5),$$

the unique form with an automorphism group of order five [2] pp. 51 [11] pp. 645. At this point, $\mathbb{P}(1,2,3,5)$ is locally isomorphic to the cyclic quotient singularity $\frac{1}{5}(1,2,3)$, i.e., the quotient of $\mathbb{A}^3$ under the action

$$(a, b, c) \mapsto (\zeta a, \zeta^2 b, \zeta^3 c) \quad \zeta \neq 1 \in \mu_5.$$

This is canonical by the Reid-Tai criterion; see [8] pp. 28 for a general result.

Second, we address the singularities along the boundary. For $\alpha > 9/11$ we need that

$$K_{\overline{M}_2} + \alpha\Delta_0 + \frac{1+\alpha}{2}\Delta_1 + 1/2\Xi$$

is log canonical. When $\alpha \leq 9/11$, Proposition 4.5 and the discrepancy computation (8) reduce us to showing that this is log canonical. Since $\overline{M}_2$ is $\mathbb{Q}$-factorial and discrepancies increase as coefficients of log divisors decrease [14] 2.17.3, it suffices to verify that

$$K_{\overline{M}_2} + \Delta_0 + \Delta_1 + \frac{1}{2}\Xi \tag{10}$$

is log canonical.

The proof relies on the description of the boundary divisors in terms of the local presentation

$$T_{[C]}\overline{\mathcal{M}}_2/\mathrm{Aut}(C),$$

as sketched in §3.2. The key observation is that Ξ does not play a rôle in the analysis. Each bielliptic involution acts on $T_{[C]}\overline{\mathcal{M}}_2$ by reflection across the corresponding hyperplane in $\widehat{\xi}$, so the quotient

$$h : T_{[C]}\overline{\mathcal{M}}_2 \to T_{[C]}\overline{\mathcal{M}}_2/H, \quad H = \mathrm{Aut}(C)/\langle\iota\rangle,$$

has simple ramification along $\widehat{\xi}$. Since Ξ has coefficient $1/2$ in (10), $\widehat{\xi}$ does not appear in the pull-back of the log canonical divisor to $T_{[C]}\overline{\mathcal{M}}_2$.

Thus (10) pulls back to

$$K_{T_{[C]}\overline{\mathcal{M}}_2} + \widehat{\delta},$$

and we have seen that $\widehat{\delta}$ is normal crossings. It follows that $(T_{[C]}\overline{\mathcal{M}}_2, \widehat{\delta})$ is log canonical and Proposition 4.6 gives the desired result. □

References

1. D. Abramovich, A. Corti and A. Vistoli – "Twisted bundles and admissible covers", *Comm. Algebra* **31** (2003), no. 8, p. 3547–3618, Special issue in honor of Steven L. Kleiman.
2. O. Bolza – "On binary sextics with linear transformations into themselves", *Amer. J. of Math.* **10** (1887), no. 1, p. 47–70.
3. A. Cayley – "Tables for the binary sextic", *Amer. J. of Math.* **4** (1881), no. 1/4, p. 379–384, Reprinted in: *Collected Mathematical Papers*, vol. XI, pp. 372–376, Cambridge University Press, 1896.
4. A. Clebsch – *Theorie der binären algebraischen Formen*, Verlag von B.G. Teubner, Leipzig, 1872.

5. M. Cornalba and J. Harris – "Divisor classes associated to families of stable varieties, with applications to the moduli space of curves", *Ann. Sci. École Norm. Sup. (4)* **21** (1988), no. 3, p. 455–475.
6. E. B. Elliott – *An introduction to the Algebra of Quantics*, Oxford University/Clarendon Press, 1895.
7. J. Harris and I. Morrison – *Moduli of curves*, Graduate Texts in Mathematics, vol. 187, Springer-Verlag, New York, 1998.
8. J. Harris and D. Mumford – "On the Kodaira dimension of the moduli space of curves", *Invent. Math.* **67** (1982), no. 1, p. 23–88, With an appendix by William Fulton.
9. R. Hartshorne – *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52.
10. D. Hilbert – *Theory of algebraic invariants*, Cambridge University Press, Cambridge, 1993, Edited and with an introduction by Bernd Sturmfels.
11. J.-i. Igusa – "Arithmetic variety of moduli for genus two", *Ann. of Math. (2)* **72** (1960), p. 612–649.
12. — , "On Siegel modular forms of genus two", *Amer. J. of Math.* **84** (1962), p. 175–200.
13. — , "A desingularization problem in the theory of Siegel modular functions", *Math. Ann.* **168** (1967), p. 228–260.
14. J. Kollár – *Flips and abundance for algebraic threefolds*, Société Mathématique de France, Paris, 1992, Astérisque No. 211.
15. G. Laumon and L. Moret-Bailly – *Champs algébriques*, Ergebnisse der Mathematik und ihrer Grenzgebiete, vol. 39, Springer-Verlag, Berlin, 2000.
16. D. Mumford – "Hirzebruch's proportionality theorem in the noncompact case", *Invent. Math.* **42** (1977), p. 239–272.
17. D. Mumford, J. Fogarty and F. Kirwan – *Geometric invariant theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete, vol. 34, Springer-Verlag, Berlin, 1994, 3rd edition.
18. D. Mumford – "Stability of projective varieties", *Enseignement Math. (2)* **23** (1977), no. 1-2, p. 39–110.
19. Y. Namikawa – "On the canonical holomorphic map from the moduli space of stable curves to the Igusa monoidal transform", *Nagoya Math. J.* **52** (1973), p. 197–259.
20. — , "A new compactification of the Siegel space and degeneration of Abelian varieties. I/II", *Math. Ann.* **221** (1976), no. 2/3, p. 97–141/201–241.
21. G. Salmon – *Lessons introductory to the modern higher algebra*, Hodges Smith, Dublin, 1866, 2nd edition.
22. I. Schur – *Vorlesungen über Invariantentheorie*, Bearbeitet und herausgegeben von Helmut Grunsky. Die Grundlehren der mathematischen Wissenschaften, Band 143, Springer-Verlag, Berlin, 1968.
23. T. A. Springer – *Invariant theory*, Springer-Verlag, Berlin, 1977, Lecture Notes in Mathematics, Vol. 585.
24. A. Vistoli – "The Chow ring of $\mathcal{M}_2$. Appendix to "Equivariant intersection theory" [Invent. Math. **131** (1998), no. 3, 595–634; by D. Edidin and W. Graham"], *Invent. Math.* **131** (1998), no. 3, p. 635–644.

Mirror symmetry and Langlands duality in the non-Abelian Hodge theory of a curve

Tamás Hausel

Department of Mathematics, University of Texas, Austin TX 78712, U.S.A.
hausel@math.utexas.edu

Summary. The paper surveys the mirror symmetry conjectures of Hausel–Thaddeus and Hausel–Rodriguez-Villegas concerning the equality of certain Hodge numbers of $SL(n,\mathbb{C})$ vs. $PGL(n,\mathbb{C})$ flat connections and character varieties for curves, respectively. Several new results and conjectures and their relations to works of Hitchin, Gothen, Garsia–Haiman and Earl–Kirwan are explained. These use the representation theory of finite groups of Lie-type via the arithmetic of character varieties and lead to an unexpected conjecture for a Hard Lefschetz theorem for their cohomology.

1 Introduction

Non-Abelian Hodge theory [29], [44] of a genus g smooth complex projective curve C studies three moduli spaces attached to C and a reductive complex algebraic group G, which in this paper will be either $GL(n,\mathbb{C})$ or $SL(n,\mathbb{C})$ or $PGL(n,\mathbb{C})$. They are

- $\mathcal{M}^d_{Dol}(G)$, the moduli space of semistable G-Higgs bundles on C;
- $\mathcal{M}^d_{DR}(G)$, the moduli space of flat G-connections on C and
- $\mathcal{M}^d_B(G)$ the character variety, i.e., the moduli space of representations of $\pi_1(C)$ into G modulo conjugation.

Under certain assumptions these moduli spaces are smooth varieties (or orbifolds when $G = PGL(n,\mathbb{C})$) with the underlying differentiable manifolds canonically identified and endowed with a natural hyperkähler metric.

The cohomology of this underlying manifold has been studied mostly from the perspective of $\mathcal{M}^d_{Dol}(G)$. Using a natural circle action on it [29] and [16] calculated the Poincaré polynomials for $G = SL(2,\mathbb{C})$ and $G = SL(3,\mathbb{C})$ respectively; while [25] and [39] found a simple set of generators for the cohomology ring for $G = PGL(2,\mathbb{C})$, respectively $G = PGL(n,\mathbb{C})$. The paper [26] then calculated the cohomology ring explicitly for $G = PGL(2,\mathbb{C})$. The techniques used in these papers do not seem to generalize easily to higher n.

A new point of view was introduced in [27] and [28]. It was shown that hyperkähler metrics and Hitchin systems [30] for $\mathcal{M}^d_{DR}(G)$ and $\mathcal{M}^d_{DR}(G^L)$, with $G = SL(n,\mathbb{C})$ and Langlands dual $G^L = PGL(n,\mathbb{C})$, realize the geometrical setup for mirror symmetry proposed in [49]. Based on this observation, [28] conjectured the existence of a topological version of mirror symmetry, i.e., the equality of certain Hodge numbers of $\mathcal{M}^d_{DR}(G)$ and $\mathcal{M}^d_{DR}(G^L)$. This was checked for $G = SL(2,\mathbb{C})$ and $SL(3,\mathbb{C})$ using [29] and [16].

This mirror symmetry conjecture suggests to study not only the cohomology of $\mathcal{M}^d_{DR}(G)$, $\mathcal{M}^d_{Dol}(G)$ and $\mathcal{M}^d_B(G)$ but also its mixed Hodge structure. It was shown in [28] that the mixed Hodge structures of $\mathcal{M}^d_{Dol}(G)$ and $\mathcal{M}^d_{DR}(G)$ agree, and are pure (see Theorem 2.1 or [38]). However, the mixed Hodge structure of the character variety $\mathcal{M}^d_B(G)$ has not been investigated until recently.

In this paper we study this mixed Hodge structure, more precisely,

- the *mixed Hodge polynomial* $H(x,y,t)$;
- the E-polynomial

$$E(x,y) := x^n y^n H(1/x, 1/y, -1),$$

  where $n = \dim_{\mathbb{C}} \mathcal{M}^d_B(G)$ and
- the Poincaré polynomial $H(1,1,t)$.

Here the H-polynomial encodes the dimensions of the graded pieces of the mixed Hodge structure on $\mathcal{M}^d_B(G)$ (see Section 2.2).

In [23] an arithmetic method was used to calculate the E-polynomial of $\mathcal{M}^d_B(G)$. The idea was to count the $\mathbb{F}_q$-rational points of $\mathcal{M}^d_B(G(\mathbb{F}_q))$, for the variety $\mathcal{M}^d_B(G)$ over the finite field $\mathbb{F}_q$. Using a result of [37] we found a closed formula, resembling the famous Verlinde formula [51], as a simple sum over irreducible representations of $G(\mathbb{F}_q)$. In particular, the representation theory behind the E-polynomial of the character variety is that of *finite* groups of Lie type. This could be considered as an analog of Nakajima's principle [40], stating that the representation theory of a Kac–Moody algebra is encoded in the cohomology of (hyperkähler) quiver varieties.

The shape of the E-polynomials of the various character varieties lead us to conjecture [23] that mirror symmetry also holds for the pair $\mathcal{M}^d_B(G)$ and $\mathcal{M}^d_B(G^L)$, at least for $G = SL(n,\mathbb{C})$. Calculating Hodge numbers via number theory, we were able to check this conjecture for $n = 4$ or a prime. Since the two mirror symmetry conjectures of [28] and [23] are equivalent on the level of Euler characteristics, we get a proof of the original mirror symmetry conjecture of [28] on the level of Euler characteristics as well.

More interestingly, [23] arrives at explicit formulas, in terms of a simple generating function, for E-polynomials of $\mathcal{M}_B^d(GL(n,\mathbb{C}))$. For example, the Euler characteristic of $\mathcal{M}_B^d(PGL(n,\mathbb{C}))$ equals $\mu(n)n^{2g-3}$, where μ is a basic number theoretic function: the Möbius function, i.e., the sum of all primitive nth roots of unity. This hints at an interesting link between number theory and topology of the above hyperkähler manifolds. Furthermore, the E-polynomials turn out to be palindromic, i.e., they satisfy an unexpected Poincaré duality-type symmetry. We can trace back this symmetry to the Alvis–Curtis duality [1, 5] in the representation theory of finite groups of Lie type.

We present a deformation of the E-polynomial of $\mathcal{M}_B^d(PGL(n,\mathbb{C}))$, which, conjecturally [23], should agree with the H-polynomial. Modifying this formula we obtain the conjectural H-polynomial of the Higgs moduli space $\mathcal{M}_{Dol}^d(GL(n,\mathbb{C}))$. We then explain how, using our mirror symmetry conjectures as a guide, one arrives at conjectures regarding H-polynomials of the varieties associated to $SL(n,\mathbb{C})$. These conjectures imply a conjecture on Poincaré polynomials of $\mathcal{M}_{Dol}^d(PGL(n,\mathbb{C}))$. The latter resembles Lusztig's conjecture [36] on Poincaré polynomials of Nakajima's quiver varieties, also hyperkähler manifolds, similar to the Higgs moduli space $\mathcal{M}_{Dol}^d(G)$. We should also mention Zagier's [52] formula for the Poincaré polynomial of the moduli space $\mathcal{N}^d$ of stable bundles (the "Kähler version" of $\mathcal{M}_{Dol}^d(SL(n,\mathbb{C}))$), where the formula is a similar sum, but is parametrized by ordered partitions of n.

We discuss in detail several checks of these conjectures, showing how they imply results of Hitchin [29], Gothen [16] and Earl–Kirwan [8]. The combinatorics of these formulas are quite non-trivial. Surprisingly, the calculus of Garsia–Haiman [12] is used to check the conjecture for $g = 0$.

The curious Poincaré duality satisfied by the conjectured Hodge numbers of $\mathcal{M}_B^d(PGL(n,\mathbb{C}))$, leads to a conjecture that a version of the Hard Lefschetz theorem is satisfied for the non-compact varieties under consideration. This can be thought of as a generalization of a result in [21] on the quaternionic geometry of matroids, and as an analogue of Faber's conjecture [9] on the moduli space of curves.

Acknowledgments. This survey paper is based on the author's talk at the "Geometric Methods in Algebra and Number Theory" conference at the University of Miami in December 2003. I would like to thank the organizers for the invitation and for the memorable conference. Most of the results and conjectures surveyed here have been obtained in joint work with Michael Thaddeus [27, 28] and with Fernando Rodriguez-Villegas [23]. This research was partially supported by the NSF grant DMS-0305505.

2 Abelian and non-Abelian Hodge theory

In this section we recall some basic definitions on Abelian and non-Abelian Hodge theory.

2.1 Hodge–De Rham theory

There are various cohomology theories associating a graded anti-commutative ring to a smooth complex algebraic variety M. First of all, the singular, or Betti, cohomology $H^*_B(M,\mathbb{C})$ of M with complex coefficients, defined for any reasonable topological space. The dimension $b_k(M) := \dim H^k_B(M,\mathbb{C})$ is called the k-th Betti number. The Poincaré polynomial is

$$P(t;M) := \sum_k b_k(M)t^k.$$

Next, the De Rham cohomology $H^*_{DR}(M,\mathbb{C})$, the space of closed differential forms modulo exact forms, defined for any differentiable manifold. The De Rham theorem establishes the isomorphism:

$$H^*_B(M,\mathbb{C}) \cong H^*_{DR}(M,\mathbb{C}). \tag{1}$$

For projective M we have the Dolbault cohomology

$$H^k_{Dol}(M,\mathbb{C}) = \bigoplus_{p+q=k} H^q(M,\Omega^p_M).$$

The Hodge theorem establishes a natural isomorphism

$$H^k_{DR}(M,\mathbb{C}) \cong H^k_{Dol}(M,\mathbb{C}). \tag{2}$$

The above isomorphisms imply the Hodge decomposition theorem:

$$H^k_B(M,\mathbb{C}) \cong \bigoplus_{p+q=k} H^{p,q}(M), \tag{3}$$

where $H^{p,q}(M) := H^p(M,\Omega^q_M)$. The numbers $h^{p,q}(M) := \dim H^{p,q}(M)$ are called Hodge numbers of M. The Hodge polynomial is:

$$H(x,y;M) := \sum_{p,q} h^{p,q}(M)x^p y^q.$$

For more details on these cohomology theories see [15].

2.2 Mixed Hodge structures

Deligne [7] generalized the Hodge decomposition theorem (3) to any complex variety M, not necessarily smooth or projective, by introducing a so-called *mixed Hodge structure* on $H_B^*(M, \mathbb{C})$. This implies a decomposition [1]

$$H_B^k(M, \mathbb{C}) \cong \bigoplus_{p,q} H^{p,q;k}(M),$$

where $p + q$ is called the weight of $H^{p,q;k}(M)$. For a smooth projective variety we have $H^{p,q;p+q}(M) = H^{p,q}(M)$, i.e. the weight of $H^{p,q;k}(M)$ is always k (called the *pure weight*). In general, other weights appear in the mixed Hodge structure; we will see such examples later. The dimensions $h^{p,q;k}(M) := H^{p,q;k}(M)$ are called mixed Hodge numbers of M. Form the three variable polynomial:

$$H(x, y, t; M) := \sum_{p,q,k} h^{p,q;k}(M) x^p y^q t^k. \tag{4}$$

Similarly, Deligne [7] constructs a mixed Hodge structure on the compactly supported $H_{B,cpt}^*(M, \mathbb{C})$ singular cohomology of a complex algebraic variety M. This yields the decomposition

$$H_{B,cpt}^k(M, \mathbb{C}) \cong \bigoplus_{p,q} H_{cpt}^{p,q;k}(M),$$

and compactly supported mixed Hodge numbers $h_{cpt}^{p,q;k}(M) := \dim H_{cpt}^{p,q;k}(M)$. One introduces the e-numbers

$$e^{p,q}(M) = \sum_k (-1)^k h_{cpt}^{p,q;k}(M)$$

and the E-polynomial:

$$E(x, y; M) := \sum_{p,q} e^{p,q}(M) x^p y^q. \tag{5}$$

Clearly, for M smooth projective, $E(x, y) = H(-x, -y)$. Moreover, for a smooth variety Poincaré duality implies that

$$E(x, y) = (xy)^n H(1/x, 1/y, -1),$$

where n is the complex dimension of M. The significance of the E-polynomial is that it is additive for decompositions and multiplicative for Zariski locally trivial fibrations.

For more details see [7] or [3].

[1] In fact what one gets from a mixed Hodge structure are two filtrations on the cohomology, and the decomposition in question is the associated graded.

2.3 Stringy cohomology

Let Γ be a finite group acting on M. By the naturality of the mixed Hodge structure, Γ will act on $H^{p,q,k}(M)$ and we have

$$H^{p,q;k}(M/\Gamma) \cong \left(H^{p,q;k}(M)\right)^{\Gamma}.$$

However, for a Calabi–Yau M and a Γ preserving the Calabi–Yau structure string theorists [50, 53] introduced different Hodge numbers on the Calabi–Yau orbifold M/Γ: the so-called stringy Hodge numbers, which are the "right" Hodge numbers for mirror symmetry. Their mathematical significance is highlighted by a theorem of Kontsevich [34] which says that stringy Hodge numbers agree with ordinary Hodge numbers of any crepant resolution. Following [3] we can define the stringy E-polynomials:

$$E_{\mathrm{st}}(x,y;M/\Gamma) := \sum_{[\gamma]} E(x,y;M^{\gamma})^{C(\gamma)}(xy)^{F(\gamma)},$$

where the sum runs over the conjugacy classes of Γ; $C(\gamma)$ is the centralizer of γ; M^{γ} is the subvariety fixed by γ; and $F(\gamma)$ is an integer, called the *fermionic shift*, which is defined as follows. The group element γ has finite order, so it acts on $TM|_{M^{\gamma}}$ as a linear automorphism with eigenvalues $e^{2\pi i w_1},\dots,e^{2\pi i w_n}$, where each $w_j \in [0,1)$. Let $F(\gamma) = \sum w_j$; this is an integer since, by hypothesis, γ acts trivially on the canonical bundle.

The last cohomology theory needed is the stringy cohomology of a Calabi–Yau orbifold twisted by a B-field. Following [31] we let $B \in H^2_{\Gamma}(M,U(1))$, i.e., an isomorphism class of a Γ-equivariant flat unitary gerbe. For any $\gamma \in \Gamma$ this B-field induces a $C(\gamma)$-equivariant local system $[L_{B,\gamma}] \in H^1_{C(\gamma)}(M^{\gamma},U(1))$ on the fixed point set M^{γ} and we can twist the stringy E-polynomial:

$$E^B_{\mathrm{st}}(x,y;M/\Gamma) := \sum_{[\gamma]} E(x,y;M^{\gamma};L_{B,\gamma})^{C(\gamma)}(xy)^{F(\gamma)}. \tag{6}$$

For more information about stringy cohomology see [3], for twisting with a B-field see [28].

2.4 Non-Abelian Hodge theory

The starting point of non-Abelian Hodge theory is the identification of the space $H^1_B(M,\mathbb{C}^{\times})$ with the space of homomorphisms from $\pi_1(M) \to \mathbb{C}^{\times}$; the space $H^1_{DR}(M,\mathbb{C}^{\times})$ with algebraic local systems on M and the space

$$H_{Dol}(M,\mathbb{C}^{\times}) \cong H^1(M,\mathcal{O}^{\times}) \oplus H^0(M,\Omega^1)$$

with pairs of a holomorphic line bundle and a holomorphic one-form.

This can be generalized to any non-Abelian complex reductive group G. We define $H^1_B(M,G)$ to be conjugacy classes of representations of $\pi_1(M) \to G$. I.e.,

$$H^1_B(M,G) := \mathrm{Hom}(\pi_1(M),G)//G,$$

the affine GIT quotient of the affine variety $\mathrm{Hom}(\pi_1(M),G)$ by the conjugation action of G, called the *character variety*. The space $H^1_{DR}(M,G)$ can be identified as the moduli space of algebraic G-local systems on M. Finally, $H^1_{Dol}(M,G)$ is the moduli space of certain semistable G-Higgs bundles on M. We will give a precise definition in the case of a curve below. The identification between $H^1_B(M,G)$ and $H^1_{DR}(M,G)$, which is analogous to the De Rham map (1), is given by the Riemann–Hilbert correspondence [6, 47], while the identification between $H^1_{DR}(M,G)$ and $H^1_{Dol}(M,G)$, analogous to the Hodge decomposition (2), is given in [4, 45] by the theory of harmonic bundles, the non-Abelian generalization of Hodge theory.

For an introduction to non-Abelian Hodge theory see [44], and ([33], Section 3), for more details on the construction of the spaces appearing in non-Abelian Hodge theory and the maps between them see [45, 46, 47].

2.5 The case of a curve

We fix a smooth projective complex curve C of genus g and specify our spaces in the case when $M = C$ and $G = GL(n,\mathbb{C})$. We have:

$$\begin{aligned}&\mathcal{M}_B(GL(n,\mathbb{C})) := H^1_B(C, GL(n,\mathbb{C}))\\ &= \{A_1,B_1,\ldots,A_g,B_g \in GL(n,\mathbb{C}) | [A_1,B_1]\cdot\cdots\cdot[A_g,B_g] = Id\}//GL(n,\mathbb{C}).\end{aligned}$$

There is a natural way to twist these varieties. This will be needed for $PGL(n,\mathbb{C})$ and we introduce these twists below. For $d \in \mathbb{Z}$, consider:

$$\begin{aligned}\mathcal{M}^d_B(GL(n,\mathbb{C})) := \{&A_1,B_1,\ldots,A_g,B_g \in GL(n,\mathbb{C})|\\ &[A_1,B_1]\cdot\cdots\cdot[A_g,B_g] = e^{\frac{2\pi i d}{n}} Id\}//GL(n,\mathbb{C}).\end{aligned}$$

The De Rham space looks like

$$\begin{aligned}\mathcal{M}_{DR}(GL(n,\mathbb{C})) &:= H^1_{DR}(C, GL(n,\mathbb{C}))\\ &= \{\text{moduli space of flat } GL(n,\mathbb{C})\text{-connections on } C\}\end{aligned}$$

and in the twisted case we need to fix a point $p \in C$, and define

$$\mathcal{M}^d_{DR}(GL(n,\mathbb{C})) := \left\{ \begin{array}{c} \text{moduli space of flat } GL(n,\mathbb{C})\text{-connections on } C \smallsetminus \{p\},\\ \text{with holonomy } e^{\frac{2\pi i d}{n}} Id \text{ around } p \end{array} \right\}.$$

Finally, the Dolbeault spaces are:

$$\mathcal{M}_{Dol}(GL(n,\mathbb{C})) := H^1_{Dol}(C, GL(n,\mathbb{C})) = \{\text{moduli space of semistable rank } n \text{ degree 0 Higgs bundles on } C\},$$

where a rank n Higgs bundle is a pair (E,ϕ) of a rank n algebraic vector bundle E on C, with degree 0 and Higgs field $\phi \in H^0(C, K_C \otimes \mathrm{End}(E))$. A Higgs bundle is called semistable if for any Higgs subbundle (F,ψ) (i.e., a subbundle with compatible Higgs fields) we have

$$\frac{\deg(F)}{\mathrm{rank}(F)} \leq \frac{\deg(E)}{\mathrm{rank}(E)} = 0.$$

The twisted version of $\mathcal{M}_{Dol}(GL(n,\mathbb{C}))$ is defined:

$$\mathcal{M}^d_{Dol}(GL(n,\mathbb{C})) := \{\text{moduli space of semistable rank } n \text{ degree } d \text{ Higgs bundles on } C\}.$$

The varieties above for $GL(n,\mathbb{C})$ have dimension $n^2(2g-2)+2$. The Betti space is affine, while the De Rham space is analytically (but not algebraically) isomorphic, via the Riemann–Hilbert correspondence, to the Betti space, so that the De Rham space is a Stein manifold as a complex manifold but not an affine variety as an algebraic variety. Finally, the Dolbeault space is a quasi-projective variety with large projective subvarieties.

From now on we fix a d with $(n,d) = 1$. In this case the corresponding twisted spaces are smooth, have a diffeomorphic underlying manifold $\mathcal{M}^d(GL(n,\mathbb{C}))$ which carries a complete hyperkähler metric [29]. The complex structures of $\mathcal{M}^d_{Dol}(GL(n,\mathbb{C}))$ and $\mathcal{M}^d_{DR}(GL(n,\mathbb{C}))$ appear in the hyperkähler structure.

We started this subsection by determining these spaces for $GL(1,\mathbb{C}) \cong \mathbb{C}^\times$. With the identifications explained, we see that

$$\begin{aligned} \mathcal{M}^d{}_B(GL(1,\mathbb{C})) &\cong (\mathbb{C}^\times)^{2g}, \\ \mathcal{M}^d_{Dol}(GL(1,\mathbb{C})) &\cong T^* Jac^d(C) \end{aligned} \tag{7}$$

and $\mathcal{M}^d_{DR}$ is a certain affine bundle over $Jac^d(C)$.

Interestingly, for $d = 0$ they are all algebraic groups and they act on the corresponding spaces for $GL(n,\mathbb{C})$ and any d by tensorization.

We can consider the map

$$\begin{array}{rccc} \lambda_{Dol} : & \mathcal{M}^d_{Dol}(GL(n,\mathbb{C})) & \to & \mathcal{M}^d_{Dol}(GL(1,\mathbb{C})), \\ & (E,\Phi) & \mapsto & (\det(E), \mathrm{tr}(\phi)). \end{array}$$

The fibres of this map can be shown to be isomorphic using the above tensorization action. It follows that up to isomorphism it is irrelevant which fibre we take, but we usually take a point $(\Lambda, 0) \in \mathcal{M}^d_{Dol}(GL(1,\mathbb{C}))$ and define

$$\mathcal{M}^d_{Dol}(SL(n,\mathbb{C})) := \lambda^{-1}_{Dol}((\Lambda, 0)).$$

For the other two spaces we have:

$$\mathcal{M}^d_{DR}(SL(n,\mathbb{C})) = \left\{ \begin{array}{c} \text{moduli space of flat } SL(n,\mathbb{C})\text{-connections on } C \smallsetminus \{p\} \\ \text{with holonomy } e^{\frac{2\pi i d}{n}} Id \text{ around } p \end{array} \right\},$$

and

$$\mathcal{M}^d_B(SL(n,\mathbb{C})) = \{A_1, B_1, \ldots, A_g, B_g \in SL(n,\mathbb{C}) | \\ [A_1, B_1] \cdot \cdots \cdot [A_g, B_g] = e^{\frac{2\pi i d}{n}} Id\} // SL(n,\mathbb{C}).$$

The varieties $\mathcal{M}^d_B(SL(n,\mathbb{C}))$, $\mathcal{M}^d_{DR}(SL(n,\mathbb{C}))$ and $\mathcal{M}^d_{DR}(SL(n,\mathbb{C}))$ are smooth of dimension $(n^2-1)(2g-2)$, with diffeomorphic underlying manifold $\mathcal{M}^d(SL(n,\mathbb{C}))$. The Betti space is affine, and the Betti and De Rham spaces are again analytically, but not algebraically, isomorphic.

We see that the finite subgroup $Jac[n] \cong \mathbb{Z}_n^{2g} \subset \mathcal{M}_{Dol}(GL(1,\mathbb{C}))$ preserves the fibration λ_{Dol} and thus acts on $\mathcal{M}^d_{Dol}(SL(n,\mathbb{C}))$. The quotient then is:

$$\mathcal{M}^d_{Dol}(PGL(n,\mathbb{C})) := \mathcal{M}^d_{Dol}(SL(n,\mathbb{C}))/Jac[n]$$

and similarly

$$\mathcal{M}^d_{DR}(PGL(n,\mathbb{C})) := \mathcal{M}^d_{DR}(SL(n,\mathbb{C}))/Jac[n],$$

and

$$\mathcal{M}^d_B(PGL(n,\mathbb{C})) = \mathcal{M}^d_B(SL(n,\mathbb{C}))/\mathbb{Z}_n^{2g}.$$

This shows that all the three spaces $\mathcal{M}^d_B(PGL(n,\mathbb{C}))$, $\mathcal{M}^d_{DR}(PGL(n,\mathbb{C}))$ and $\mathcal{M}^d_{Dol}(PGL(n,\mathbb{C}))$ are hyperkähler orbifolds of dimension $(n^2-1)(2g-2)$. As they are orbifolds we can talk about their stringy mixed Hodge numbers as defined above in Section 2.3. Moreover, they carry natural orbifold B-fields, constructed as follows: Consider a universal Higgs pair $(\mathbf{E}, \mathbf{\Phi})$ on $\mathcal{M}^d_{Dol}(SL(n,\mathbb{C})) \times C$; it exists because $(d,n) = 1$. Restrict $\mathbf{E}$ to $\mathcal{M}^d_{Dol} \times \{p\}$ to get the vector bundle $\mathbf{E}_p$ on $\mathcal{M}^d_{Dol}(SL(n,\mathbb{C}))$. Now we can consider the projective bundle $\mathbb{P}\mathbf{E}_p$ of $\mathbf{E}_p$ which is a $PGL(n,\mathbb{C})$-bundle. The bundle $\mathbf{E}_p$ is a $GL(n,\mathbb{C})$-bundle but not a $SL(n,\mathbb{C})$-bundle, because it has a non-trivial determinant. The obstruction class to lifting the $PGL(n,\mathbb{C})$-bundle $\mathbb{P}\mathbf{E}$ to an $SL(n,\mathbb{C})$-bundle is a class

$$B \in H^2(\mathcal{M}^d_{Dol}(SL(n,\mathbb{C}), \mathbb{Z}_n)) \subset H^2(\mathcal{M}^d_{Dol}(SL(n,\mathbb{C})), U(1)),$$

which gives us a B-field on $\mathcal{M}^d_{Dol}(SL(n,\mathbb{C}))$. By ([28], Section 3), this field has a natural equivariant extension $\hat{B} \in H^2_\Gamma(\mathcal{M}^d_{Dol}(SL(n,\mathbb{C})), U(1))$, giving a B-field on $\mathcal{M}^d_{Dol}(PGL(n,\mathbb{C}))$. This B-field will appear in our mirror symmetry discussions below.

Non-Abelian Hodge theory on a curve is explained in [29], via a gauge-theoretical approach. This yields natural hyperkähler metrics on our spaces. The case $GL(1,\mathbb{C})$ is treated in [13]. Further information about the geometry and cohomology of $\mathcal{M}^1_{Dol}(SL(2,\mathbb{C}))$ is in [20].

2.6 Mixed Hodge structure on non-Abelian Hodge cohomologies

The main subject of this paper is the mixed Hodge polynomial of the (stringy, sometimes with a B-field) cohomology of $\mathcal{M}^d_{Dol}(G)$, $\mathcal{M}^d_{DR}(G)$ and $\mathcal{M}^d_B(G)$, for $G = GL(n,\mathbb{C}), PGL(n,\mathbb{C})$ or $SL(n,\mathbb{C})$. For notational convenience, we omit G and simply write $\mathcal{M}^d_B$, $\mathcal{M}^d_{DR}$ and $\mathcal{M}^d_{Dol}$.

Consider first $G = GL(1,\mathbb{C})$. From (7) we calculate:

$$\begin{aligned} H(x,y,t;\mathcal{M}^d_B(GL(1,\mathbb{C}))) &= (1+xyt)^{2g}, \\ H(x,y,t;\mathcal{M}^d_{Dol}(GL(1,\mathbb{C}))) &= H(x,y,t;\mathcal{M}^d_{DR}(GL(1,\mathbb{C}))) \\ &= (1+xt)^g(1+yt)^g. \end{aligned}$$

It is remarkable that $H(x,y,t;\mathcal{M}^d_B(GL(1,\mathbb{C}))) \neq H(x,y,t;\mathcal{M}^d_{DR}(GL(1,\mathbb{C})))$ even though the spaces are analytically isomorphic. Furthermore, we can explicitly see that the mixed Hodge structure on $H^k(\mathcal{M}^d_{Dol}(GL(1,\mathbb{C})),\mathbb{C})$ and on $H^k(\mathcal{M}^d_{DR}(GL(1,\mathbb{C})),\mathbb{C})$ is pure, while on $H^k(\mathcal{M}^d_B(GL(1,\mathbb{C})),\mathbb{C})$ it is not.

A Künneth argument implies that:

$$\begin{aligned} &H(x,y,t;\mathcal{M}^d_{Dol}(GL(n,\mathbb{C}))) \\ &\qquad = H(x,y,t;\mathcal{M}^d_{Dol}(PGL(n,\mathbb{C})))H(x,y,t;\mathcal{M}^d_{Dol}(GL(1,\mathbb{C}))), \end{aligned}$$

and similarly for the other two spaces. Thus the calculation for $GL(n,\mathbb{C})$ is equivalent to the calculation for $PGL(n,\mathbb{C})$.

Now we list what is known about the cohomologies $H^*(\mathcal{M}^d,\mathbb{C})$. The Poincaré polynomials $P(t;\mathcal{M}^1(SL(2,\mathbb{C})))$ and $P(t;\mathcal{M}^1(PGL(2,\mathbb{C})))$ were calculated in [29], while $P(t;\mathcal{M}^1(SL(3,\mathbb{C})))$ and $P(t;\mathcal{M}^1(PGL(3,\mathbb{C})))$ have been calculated in [16]. Both papers used Morse theory for a natural $\mathbb{C}^\times$-action on $\mathcal{M}^d_{Dol}$ (acting by multiplication on the Higgs field). The idea was to calculate the Poincaré polynomial of the various fixed point components of this action, and then sum them up with a certain shift. The largest of the fixed point components, when $\phi = 0$, is the important and well-studied space:

$$\begin{aligned} \mathcal{N}^d(SL(n,\mathbb{C})) := \{&\text{the moduli space of stable vector bundles} \\ &\text{of fixed determinant bundle of degree } d\}. \end{aligned} \quad (8)$$

Its Poincaré polynomial was calculated in [18] by arithmetic and in [2] by gauge-theoretical methods, with explicit formulas given in [52]. Thus its contribution to $P(t;\mathcal{M}^d(SL(n,\mathbb{C})))$ is easy to handle. However, the other components of the fixed point set of the circle action are more cumbersome to

determine already for $n = 4$. Consequently, the Morse theory approach has not been completed for $n \geq 4$.

As a representative example we calculate from [29] the Poincaré polynomial $P(t; \mathcal{M}^1(PGL(2,\mathbb{C})))$, when $g = 3$:

$$3\,t^{12}+12\,t^{11}+18\,t^{10}+32\,t^9+18\,t^8+12\,t^7+17\,t^6+6\,t^5+2\,t^4+6\,t^3+t^2+1. \quad (9)$$

The cohomology ring of $\mathcal{M}^1_{Dol}(PGL(2,\mathbb{C}))$ has been described explicitly by generators [25] and relations [26]. This information was essential for our main Conjecture 5.1. Finally, Markman [39] showed that for $PGL(n,\mathbb{C})$ the universal cohomology classes do generate the cohomology ring.

The following result first appeared in [38] using a construction of [25]. Here we present a simple proof.

Theorem 2.1. *The Hodge structure on $H^k(\mathcal{M}^d_{Dol},\mathbb{C})$ is pure of weight k.*

Proof. The compactification $\overline{\mathcal{M}}^d_{Dol}$ of $\mathcal{M}^d_{Dol}$ constructed in [19] is a projective orbifold so that the Hodge structure on $H^k(\overline{\mathcal{M}}^d_{Dol},\mathbb{C})$ is pure of weight k. Now [19] also implies that the natural map $H^*(\overline{\mathcal{M}}^d_{Dol},\mathbb{C}) \to H^*(\mathcal{M}^d_{Dol},\mathbb{C})$ is surjective. The claim follows from the functoriality of mixed Hodge structures [7]. □

One can similarly prove the same result for $\mathcal{M}^d_{DR}$.

Theorem 2.2. *The Hodge structure on $H^k(\mathcal{M}^d_{DR},\mathbb{C})$ is pure of weight k.*

Proof. As explained in ([28], Theorem 6.2) one can deform the complex structure of $\overline{\mathcal{M}}^d_{Dol}$ to the projective orbifold $\overline{\mathcal{M}}^d_{DR}$, which is the compactification of $\mathcal{M}^d_{DR}$ given by Simpson in [48]. This way we see that the natural map $H^*(\overline{\mathcal{M}}^d_{DR},\mathbb{C}) \to H^*(\mathcal{M}^d_{DR},\mathbb{C})$ is a surjection, and we conclude as above. □

In fact the argument in ([28], Theorem 6.2) shows that

Theorem 2.3 (HT4). *The mixed Hodge structure on $H^*(\mathcal{M}^d_{Dol},\mathbb{C})$ is isomorphic to the mixed Hodge structure on $H^*(\mathcal{M}^d_{DR},\mathbb{C})$.*

However, the Hodge structure on $\mathcal{M}^d_B$ has not been studied in the literature. We will see later that it is not pure anymore. In the following section we explain our interest in the Hodge structures on $\mathcal{M}^d_{Dol}$, $\mathcal{M}^d_{DR}$ and $\mathcal{M}^d_B$. Our motivation is mirror symmetry.

3 Mirror symmetry conjectures

The starting point in [28] was the observation that the pairs $\mathcal{M}^d_{DR}(SL(n,\mathbb{C}))$ together with the B-field B^e and $\mathcal{M}^d_{DR}(PGL(n,\mathbb{C}))$ with the B-field $\hat{B}^d$ give a geometric realization for mirror symmetry as proposed in Strominger–Yau–Zaslow [49] and modified for B-fields by Hitchin in [31]. This geometric picture predicts the existence of a special Lagrangian fibration on each space, with dual fibres. In [28] it is shown that the so-called Hitchin map [30] provides the required special Lagrangian fibration on our spaces, with dual Abelian varieties as fibers. For details see ([28], Section 3).

Our focus in this survey is on the topological implications of this manifestation of mirror symmetry. The following conjecture can be called the topological mirror test for our SYZ-mirror partners.

Conjecture 3.1 ([28]). *For $d, e \in \mathbb{Z}$, with $(d,n) = (e,n) = 1$, we have*

$$E_{\mathrm{st}}^{B^e}\Big(x,y;\mathcal{M}^d_{DR}(SL(n,\mathbb{C}))\Big) = E_{\mathrm{st}}^{\hat{B}^d}\Big(x,y;\mathcal{M}^e_{DR}(PGL(n,\mathbb{C}))\Big).$$

Remark 3.2. Since $M^d_{DR}(SL(n,\mathbb{C}))$ is smooth, the left-hand side actually equals the E-polynomial $E\Big(x,y;M^d_{DR}(SL(n,\mathbb{C}))\Big)$, which is independent of e. This motivates the following:

Conjecture 3.3 (HT4). *For $d_1.d_2 \in \mathbb{Z}$ with $(d_1,n) = (d_2,n) = 1$ we have:*

$$E\Big(x,y;\mathcal{M}^{d_1}_{Dol}(SL(n,\mathbb{C}))\Big) = E\Big(x,y;\mathcal{M}^{d_2}_{Dol}(SL(n,\mathbb{C}))\Big). \tag{10}$$

This is quite interesting since the Betti numbers of $\mathcal{N}^d(SL(n,\mathbb{C}))$, the moduli space of stable vector bundles, with fixed determinant of degree d (the "Kähler version" of $\mathcal{M}^d_{Dol}(SL(n,\mathbb{C}))$), are known to depend on d. Already for $n = 5$, Zagier's explicit formulas [52] for $P(t;\mathcal{N}^1(SL(5,\mathbb{C})))$ and $P(t;\mathcal{N}^3(SL(5,\mathbb{C})))$ are different. We will see evidence for this Conjecture 3.3 in Corollary 3.11.

Remark 3.4. Conjecture 3.1 was proved for $n = 2$ and $n = 3$ in [28]. The proof proceeds by first transforming the calculation to $\mathcal{M}^d_{Dol}$ via Theorem 2.3 and then using the Morse theoretic method of [29] and [16]. It is unclear how to extend this method to $n \geq 4$.

Remark 3.5. An important ingredient of the proofs was a modification of a result of Narasimhan–Ramanan [41] to Higgs bundles. It describes the fixed points of the action of elements of $Jac[n]$ on $\mathcal{M}^d_{Dol}(SL(n,\mathbb{C}))$. The fixed point sets have lower rank $m|n$ Higgs moduli spaces $\mathcal{M}^d_{Dol}(SL(m,\mathbb{C});\tilde{C})$ for a certain covering $\tilde{C}$ of C. Their cohomology enters in the stringy contribution to the right-hand side of Conjecture 3.1 (recall (6)).

3.1 Number theory to the rescue

Although the mirror symmetry Conjecture 3.1 is still open for $n \geq 4$, recently some evidence for its validity has been achieved in form of progress on a related conjecture.

Conjecture 3.6 (HRV). *For $d, e \in \mathbb{Z}$ with $(d, n) = (e, n) = 1$, we have*

$$E_{\mathrm{st}}^{B^e}\Big(x, y, \mathcal{M}_B^d(SL(n, \mathbb{C}))\Big) = E_{\mathrm{st}}^{\hat{B}^d}\Big(x, y, \mathcal{M}_B^e(PGL(n, \mathbb{C}))\Big).$$

This conjecture has been proved [23] when n is a prime and when $n = 4$; which implies Conjecture 3.1 on the level of Euler characteristic in these cases. The method is arithmetic: one counts rational points on the variety $\mathcal{M}_B^d$ over a finite field $\mathbb{F}_q$, when n divides $q - 1$, where $q = p^r$ is a prime power. We get:

Theorem 3.7 ([23]). *The E-polynomial of $\mathcal{M}_B^d$ has only $x^k y^k$ type terms, and*

$$E(q) = \#(\mathcal{M}_B^d(G)(\mathbb{F}_q)).$$

The problem reduces to the count of solutions of the equation:

$$[A_1, B_1] \cdot \cdots \cdot [A_g, B_g] = \xi_n,$$

in the finite group of Lie type $G(\mathbb{F}_q)$, i.e., so that $A_i, B_i \in G(\mathbb{F}_q)$, where $\xi_n \in G$ is a central element of order n. A simple modification of a theorem of Mednykh [37], (which goes back to Frobenius–Schur [11], and has since been rediscovered by many authors, (see Freed–Quinn [10], (5.19)), implies:

Theorem 3.8. *Let $G = SL(n, \mathbb{C})$ or $G = GL(n, \mathbb{C})$. Then the number of rational points on $\mathcal{M}_B^d(G)$ over a finite field $\mathbb{F}_q$, where $q = p^r$ is a prime power, with $n|(q-1)$ is given by:*

$$\#(\mathcal{M}_B^d(G)(\mathbb{F}_q)) = \sum_{\chi \in \mathrm{Irr}(\mathrm{G}(\mathbb{F}_\mathrm{q}))} \frac{|G|^{2g-2}}{\chi(1)^{2g-1}} \chi(\xi_n),$$

where the sum is over all irreducible characters of $G(\mathbb{F}_q)$.

The two theorems above imply the following

Corollary 3.9 ([23]). *The E-polynomial of $\mathcal{M}_B^d(G)$ is given by:*

$$E(q) = \sum_{\chi \in \mathrm{Irr}(\mathrm{G}(\mathbb{F}_\mathrm{q}))} \frac{|G|^{2g-2}}{\chi(1)^{2g-1}} \chi(\xi_n). \tag{11}$$

Remark 3.10. An immediate consequence of this formula is the Betti analog of Conjecture 3.3. This follows from Corollary 3.9 since that character formula transforms by a Galois automorphism when one changes from d_1 to d_2. Moreover, because our $\mathcal{M}_B^{d_1}(G)$ and $\mathcal{M}_B^{d_2}(G)$ are Galois-conjugate, we can deduce that their Betti numbers agree, and presumably, that their mixed Hodge structures also agree. In summary, we have

Corollary 3.11 ([23]). *For $d_1, d_2 \in \mathbb{Z}$ with $(d_1, n) = (d_2, n) = 1$ we have*

$$E\Big(x, y; \mathcal{M}_B^{d_1}(G)\Big) = E\Big(x, y; \mathcal{M}_B^{d_2}(G)\Big) \tag{12}$$

and

$$P\Big(t; \mathcal{M}_B^{d_1}(G)\Big) = P\Big(t; \mathcal{M}_B^{d_2}(G)\Big). \tag{13}$$

This gives an affirmative answer to Conjecture 3.3 on the level of Poincaré polynomials. In general, Galois conjugate varieties tend to be (although need not be, see, e.g. [43]) homeomorphic over $\mathbb{C}$.

Problem 3.12. Are the underlying topological spaces of the varieties $\mathcal{M}_B^{d_1}(G)$ and $\mathcal{M}_B^{d_2}(G)$ homeomorphic for $(n, d_1) = (n, d_2) = 1$? Are they birationally isomorphic? Are $\mathcal{M}_{Dol}^{d_1}(G)$ and $\mathcal{M}_{Dol}^{d_2}(G)$ birationally isomorphic?

Remark 3.13. In order to calculate the character formula in Corollary 3.9, we will need to know the values of irreducible characters of G on central elements. Fortunately, for $GL(n, \mathbb{F}_q)$ this has been done by Green [14]. For $SL(n, \mathbb{F}_q)$ the required information, i.e., the value of characters on central elements, was obtained by Lehrer in [35]. In the next section we show an explicit result for the character formula for $GL(n, \mathbb{F}_q)$.

Remark 3.14. Our mirror symmetry Conjecture 3.6 can be translated to a complicated formula which is valid for the character tables of $PGL(n, \mathbb{F}_q)$ and $SL(n, \mathbb{F}_q)$. In particular, we believe that by introducing punctures for our Riemann surfaces, a similar mirror symmetry conjecture would in fact capture the exact difference between the full character tables of $PGL(n, \mathbb{F}_q)$ and $SL(n, \mathbb{F}_q)$ (not just on central elements as above). This way our mirror symmetry proposal could be phrased as follows: *the differences between the character tables of $PGL(n, \mathbb{F}_q)$ and its Langlands dual $SL(n, \mathbb{F}_q)$ are governed by mirror symmetry.* It is particularly enjoyable to see the effect of mirror symmetry on the differences between the character tables of $GL(2, \mathbb{F}_q)$ and $SL(2, \mathbb{F}_q)$, which were first calculated in 1907 by Jordan [32] and by Schur [42].

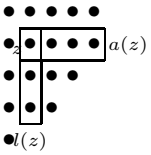
4 Explicit formulas for the E-polynomials

Here we calculate the E-polynomials of $\mathcal{M}_B^d(PGL(n, \mathbb{C}))$, which we denote by $E_n(q)$.

We start with partitions. Let $\lambda = (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_l > 0)$ be a partition of n, so that $\sum \lambda_i = n$. The *Ferrers diagram* $d(\lambda)$ of λ is the set of lattice points

$$\{(i,j) \in \mathbb{Z}_{\leq 0} \times \mathbb{N} : j < \lambda_{-i+1}\}. \tag{14}$$

The *arm length* $a(z)$ and *leg length* $l(z)$ of a point $z \in d(\lambda)$ denote the number of points strictly to the right of z and below z, respectively, as indicated in this example:



where $\lambda = (5,5,4,3,1)$, $z = (-1,1)$, $a(z) = 3$ and $l(z) = 2$. The *hook length* then is defined as

$$h(z) = l(z) + a(z) + 1.$$

Let

$$V_n(q) = E_n(q) q^{(1-g)n(n-1)} (q-1)^{2g-2},$$

and

$$Z_n(q,T) = \exp\left(\sum_{r \geq 1} V_n(q^r) \frac{T^r}{r} \right).$$

We define the Hook polynomials for a partition λ as follows :

$$\mathcal{H}^\lambda(q) = \prod_{z \in d(\lambda)} q^{-l(z)} (1 - q^{h(z)}).$$

Theorem 4.1 ([23]). *For $n = 1, 2, 3, \ldots$ one has*

$$\prod_{n=1}^{\infty} Z_n(q, T^n) = \sum_{\lambda \in \mathcal{P}} (\mathcal{H}^\lambda(q))^{2g-2} T^{|\lambda|}, \tag{15}$$

where $\mathcal{P}$ is the set of all partitions.

One simple corollary of this is a new topological result:

Corollary 4.2 ([23]). *The Euler characteristic of $\mathcal{M}^d(PGL(n,\mathbb{C}))$ equals $\mu(n) n^{2g-3}$, where μ is the Möbius function, i.e., $\mu(n)$ is the sum of primitive nth root of unities.*

Another interesting application of the theorem is the following:

Corollary 4.3. *The E-polynomial* $E_n(q) = E(q; \mathcal{M}_B^d(PGL(n,\mathbb{C})))$ *is palindromic, i.e., it satisfies, what we call, the* curious *Poincaré duality:*

$$q^{2N} E_n(1/q) = E_n(q),$$

where $2N = (n^2-1)(2g-2)$ *is the complex dimension of* $\mathcal{M}_B^d(PGL(n,\mathbb{C}))$.

Remark 4.4. This result originates in the so-called Alvis–Curtis duality [1], [5] in the character theory of $GL(n,\mathbb{F}_q)$, which is a duality between irreducible representations of $GL(n,\mathbb{F}_q)$. In particular, if $\chi, \chi' \in \mathrm{Irr}(GL(n,\mathbb{F}_q))$ are dual, then the dimension $\chi(1)$ is a polynomial in q which satisfies

$$q^{\frac{n(n-1)}{2}} \chi(1)(1/q) = \chi'(1)(q).$$

For example when $n = 2$, Theorem 4.1 gives:

$$E_2(q) = (q^2-1)^{2g-2} + q^{2g-2}(q^2-1)^{2g-2} - \frac{1}{2}q^{2g-2}(q-1)^{2g-2} - \frac{1}{2}q^{2g-2}(q+1)^{2g-2}, \quad (16)$$

when $g = 3$ this gives

$$E(x,y;\mathcal{M}_B^1(PGL(2,\mathbb{C}))) = q^{12} - 4\,q^{10} + 6\,q^8 - 14\,q^6 + 6\,q^4 - 4\,q^2 + 1, \quad (17)$$

which is a palindromic polynomial. Note also that there does not seem to be much in common with the Poincaré polynomial (9).

5 A conjectured formula for mixed Hodge polynomials

Here we present the conjecture of [23] on the H-polynomials of the spaces $\mathcal{M}_B^d(PGL(n,\mathbb{C}))$. As usual we fix the curve C and its genus g and the group $PGL(n,\mathbb{C})$ and write $\mathcal{M}_B^d$ for $\mathcal{M}_B^d(PGL(n,\mathbb{C}))$ and $H_n(x,y,t)$ for $H(x,y,t;\mathcal{M}_B^d)$.

Let

$$V_n(q,t) = H_n(q,t)\frac{(qt^2)^{(1-g)n(n-1)}(qt+1)^{2g}}{(qt^2-1)(q-1)},$$

and

$$Z_n(q,t,T) = \exp\left(\sum_{r\geq 1} V_n(q^r, -(-t)^r)\frac{T^r}{r}\right).$$

We define the t-deformed Hook polynomials for genus g and partition λ:

$$\mathcal{H}_g^\lambda(q,t) = \prod_{x\in d(\lambda)} \frac{(qt^2)^{(2-2g)l(x)}(1+q^{h(x)}t^{2l(x)+1})^{2g}}{(1-q^{h(x)}t^{2l(x)+2})(1-q^{h(x)}t^{2l(x)})}.$$

The following generating function then defines our rational functions $H_n(q,t)$:

$$\prod_{n=1}^{\infty} Z_n(q,t,T^n) = \sum_{\lambda \in \mathcal{P}} \mathcal{H}_g^{\lambda}(q,t) T^{|\lambda|}. \tag{18}$$

Because for the character variety we have that $h^{i,j;k}(\mathcal{M}_B^d) = 0$ provided that $i \neq j$, the following conjecture describes $H_n(x,y,t)$ completely.

Conjecture 5.1 ([23]). *The mixed Hodge polynomials of the character varieties $\mathcal{M}_B^d(PGL(n,\mathbb{C}))$ are given by the generating function (18):*

$$H_n(\sqrt{q},\sqrt{q},t) = H_n(q,t).$$

Thus $H_n(q,t)$, which is a priori only a rational function, is conjectured to be the H-polynomial of the character variety, so in the next conjecture we formalize our expectations from $H_n(q,t)$, with the addition of a curious, Poincaré duality-type of symmetry, which was in fact our most important guide in the derivation of these formulas:

Conjecture 5.2. *The rational functions $H_n(q,t)$ defined in the generating function of (18) satisfy the following properties:*

- *$H_n(q,t)$ is a polynomial in q and t.*
- *The q and the t degree of $H_n(q,t)$ equal $2N = 2(n^2-1)(g-1)$. The largest degree monomial in both variables is $(qt)^{2(n^2-1)(g-1)}$.*
- *All coefficients of $H_n(q,t)$ are non-negative integers.*
- *The coefficients of $H_n(q,t) = \sum h_j^i q^j t^i$ satisfy what we call the* curious Poincaré duality*:*

$$h_{N-j}^{i-j} = h_{N+j}^{i+j} \tag{19}$$

Now we list some checks and implications of the above conjectures:

Remark 5.3. Computer calculations with Maple gives $H_n(q,t)$ from the above generating function when $n = 2,3,4$. In all these cases for small g we do get a polynomial in q and t with the expected degree and positive coefficients, satisfying the curious symmetry (19).

Remark 5.4. The paper [26] contains a monomial basis, in the tautological generators, for the cohomology ring $H^*(\mathcal{M}_B^1(PGL(2,\mathbb{C})),\mathbb{C})$. Understanding the action of the Frobenius on these generators leads to a formula for the mixed Hodge polynomial of $\mathcal{M}_B^1(PGL(2,\mathbb{C}))$ of the form

$$\begin{aligned} H_2(\sqrt{q},\sqrt{q},t) = {} & \frac{(q^2t^3+1)^{2g}}{(q^2t^2-1)(q^2t^4-1)} + \frac{q^{2g-2}t^{4g-4}(q^2t+1)^{2g}}{(q^2-1)(q^2t^2-1)} \\ & - \frac{1}{2}\frac{q^{2g-2}t^{4g-4}(qt+1)^{2g}}{(qt^2-1)(q-1)} - \frac{1}{2}\frac{q^{2g-2}t^{4g-4}(qt-1)^{2g}}{(q+1)(qt^2+1)}, \end{aligned} \tag{20}$$

which agrees with the conjectured one through (18), and clearly reduces to (4.1), when $t=-1$. For example when $g=3$, this gives

$$\begin{aligned}&H(\sqrt{q},\sqrt{q},t;\mathcal{M}^d(PGL(2,\mathbb{C})))\\&=t^{12}q^{12}+t^{12}q^{10}+6\,t^{11}q^{10}+t^{12}q^8+t^{10}q^{10}+6\,t^{11}q^8+16\,t^{10}q^8+6\,t^9q^8+t^{10}q^6\\&+t^8q^8+26\,t^9q^6+16\,t^8q^6+6\,t^7q^6+t^8q^4+t^6q^6+6\,t^7q^4+16\,t^6q^4+6\,t^5q^4+t^4q^4+t^4q^2\\&+6\,t^3q^2+t^2q^2+1,\end{aligned}\tag{21}$$

which is a common refinement of (9) when $q=1$ and of (17) when $t=-1$. Note also how the curious Poincaré duality appears when one refines the Poincaré polynomial (9), which does not possess any kind of symmetry, to the mixed Hodge polynomial (21).

Remark 5.5. Note that $P_n(t)=H_n(1,t)$ should be the Poincaré polynomial of the character variety, which is the same as the Poincaré polynomial of the diffeomorphic Higgs moduli space $\mathcal{M}^d_{Dol}$. For $n=2$, Hitchin in [29] calculated the Poincaré polynomial of this latter space, and an easy calculation shows that if one substitutes $q=1$ into (20) we get $P_2(t)=H_2(1,t)$, the Poincaré polynomial of Hitchin. For $n=3$, Gothen in [16] calculated $P_3(t)$. Since it is a pleasure to work with a formula like (20), we write down what our Conjecture 5.1 gives for $n=3$:

$$\begin{aligned}&H_3(q,t)=\\&\frac{\left(q^3t^5+1\right)^{2g}\left(q^2t^3+1\right)^{2g}}{\left(q^3t^6-1\right)\left(q^3t^4-1\right)\left(q^2t^4-1\right)\left(q^2t^2-1\right)}+\frac{q^{6g-6}t^{12g-12}\left(q^3t+1\right)^{2g}\left(q^2t+1\right)^{2g}}{\left(q^3t^2-1\right)\left(q^3-1\right)\left(q^2t^2-1\right)\left(q^2-1\right)}\\&+\frac{q^{4g-4}t^{8g-8}\left(q^3t^3+1\right)^{2g}\left(qt+1\right)^{2g}}{\left(q^3t^4-1\right)\left(q^3t^2-1\right)\left(qt^2-1\right)\left(q-1\right)}+\frac{1}{3}\,\frac{q^{6g-6}t^{12g-12}\left(\left(qt+1\right)^{2g}\right)^2}{\left(qt^2-1\right)^2\left(q-1\right)^2}\\&-\frac{1}{3}\,\frac{q^{6g-6}t^{12g-12}\left(q^2t^2-qt+1\right)^{2g}}{\left(q^2t^4+qt^2+1\right)\left(q^2+q+1\right)}-\frac{q^{4g-4}t^{8g-8}\left(q^2t^3+1\right)^{2g}\left(qt+1\right)^{2g}}{\left(q^2t^4-1\right)\left(q^2t^2-1\right)\left(qt^2-1\right)\left(q-1\right)}\\&-\frac{q^{6g-6}t^{12g-12}\left(q^2t+1\right)^{2g}\left(qt+1\right)^{2g}}{\left(q^2t^2-1\right)\left(q^2-1\right)\left(qt^2-1\right)\left(q-1\right)}.\end{aligned}$$

It is a nice exercise to show that $H_3(1,t)$ does produce (the corrected version[2] of) Gothen's complicated looking formula in [16].

It is also worth noting that many terms in $H_n(q,t)$ have poles at $q=1$, which somehow cancel, according to our conjecture.

Remark 5.6. When $g=0$, we know from the definitions that $H_1(x,y,t)=1$ and $H_n(x,y,t)=0$ otherwise. One can deduce the same from Conjecture 5.1 by applying Theorem 2.10 in [12] to calculate the right-hand side of (18). Moreover Conjecture 5.2 has the same flavour as the main conjecture in [12] about q,t Catalan numbers, which was in turn proved by Haiman in [17] using subtle intersection theory on the Hilbert scheme of n points on $\mathbb{C}^2$.

[2] One accidental mistake in the calculation of [16] was pointed out in (10.3) of [28].

Apart from the fact that this Hilbert scheme is also a hyperkähler manifold, the similarities between the two conjectures are rather surprising.

Remark 5.7. When $g = 1$ we have $H_n(x, y, t) = 1$ for every n, but this we could not prove from (18) for $H_n(q, t)$.

Remark 5.8. Let us look at the conjecture (19). Recall that H^2 of our varieties are exactly one dimensional, generated by a class, call it $[\omega]$, which is the Kähler class in the complex structure of $\mathcal{M}^d_{Dol}$. This carries the weight q^2t^2 in the mixed Hodge structure. The following Hard Lefschetz type conjecture enhances the curious Poincaré duality of the conjecture (19):

Conjecture 5.9. *If L denotes the map by multiplication with $[\omega]$, then the map*

$$L^k : H^{N-k,N-k;i-k}(\mathcal{M}^d_B(PGL(n,\mathbb{C}))) \to H^{N+k,N+k;i+k}(\mathcal{M}^d_B(PGL(n,\mathbb{C})))$$

is an isomorphism.

Interestingly this conjecture implies a theorem of [21] that the Lefschetz map $L^k : H^{N-k} \to H^{N+k}$ is injective for $\mathcal{M}^d_{Dol}$, and it is explained there how this weak version of Hard Lefschetz, when applied to toric hyperkähler varieties, yields new inequalities for the h-numbers of matroids. See also [24] for the original argument on toric hyperkähler varieties. Furthermore, this conjecture can be proved when $n = 2$ using the explicit description of the cohomology ring in [26]. The general case can be thought of as an analog of Faber's conjecture [9] on the cohomology of the moduli space of curves, which is another non-compact variety whose cohomology ring is conjectured to satisfy some form of Hard Lefschetz theorem.

Remark 5.10. There are two subspaces of the cohomology $H^*(\mathcal{M}^d_B, \mathbb{C})$ which are particularly interesting. One of them is the middle dimensional cohomology $H^{2N}(\mathcal{M}^d_B, \mathbb{C})$, which is the top non-vanishing cohomology. The mixed Hodge structure breaks up into parts with respect to the q-degree. The curious Poincaré dual (19) of these spaces are also interesting: it is easy to see that they are exactly the pure part of the mixed Hodge structure, i.e., spaces of the form $H^{i,i;2i}$. (Another significance of the pure part is that if there is a smooth projective compactification of the variety, then its image is in this pure part.) Thus it would already be interesting to get the pure part of $H_n(q, t)$. It is easy to identify the pure part in our case with what we call the *Pure* ring, which is the subring of $H^*(\mathcal{M}^d, \mathbb{C})$ generated by the tautological classes $a_i \in H^{2i}(\mathcal{M}^d, \mathbb{C})$ for $i = 2, \ldots, n$ (the other tautological classes, which generate the cohomology ring, are not pure classes).

For example, when $n = 2$, it was known [29] that the middle degree cohomology of the Higgs moduli space $\mathcal{M}^d_{Dol}(PGL(2, \mathbb{C}))$ is g dimensional. The Pure ring was determined in [26], and it was found to be g dimensional due to the relation $\beta^g = 0$ (where $\beta = a_2$). Thus these two seemingly unrelated

observations are dual to each other via our curious Poincaré duality (19). To see this curious duality in action let us recall the formula (21). The terms which contain the top degree 12 in t are $t^{12}q^{12}$, $t^{12}q^{10}$ and $t^{12}q^8$, which are curious Poincaré dual via (19) to the terms 1, t^4q^2 and t^8q^4, which is exactly the ring generated by the degree-four class β, which has additive basis 1, β and β^2.

The analogous ring, generated by the corresponding classes $a_2, \ldots, a_n \in H^*(\mathcal{N}^d, \mathbb{C})$, which a priori is a quotient of our Pure ring (as $\mathcal{N}^d \subset \mathcal{M}^d_{Dol}$ naturally), was studied for the moduli space $\mathcal{N}^d$ of rank n, degree-d stable bundles (with $(n,d)=1$) in [8], where it was found that the top non-vanishing degree of this ring is $2n(n-1)(g-1)$. Computer calculations for our conjecture for $n = 2, 3, 4$ also show that our conjectured Pure ring has the same 1-dimensional top degree. This and the known situation for $n = 2$ (see [26]), yields the following

Conjecture 5.11. *The Pure rings of $\mathcal{M}^d_{Dol}$ and $\mathcal{N}^d$, i.e., the subrings of the cohomology rings generated by the classes $a_2, \ldots, a_n$ are isomorphic. In particular, unlike the whole cohomology ring of $\mathcal{N}^d$, it does not depend on d.*

Now we explain a combinatorial consequence of this conjecture. First we extract a conjectured formula for $PP_n(t)$ the Poincaré polynomial of the Pure ring. Indeed we only have to deal with monomials in Conjecture 5.1 whose t-degree is double of their q-degree.

Let

$$PV_n(t) = PP_n(t)\frac{t^{2(1-g)n(n-1)}}{(t^2-1)},$$

and

$$PZ_n(t,T) = \exp\left(\sum_{r\geq 1} PV_n(t^r)\frac{T^r}{r}\right).$$

We now define the pure part of the t-deformed Hook polynomials for genus g and partition λ as follows:

$$\mathcal{PH}^\lambda_g(t) = t^{4(1-g)n(\lambda')} \prod_{x\in d(\lambda); a(x)=0} \frac{1}{(1-t^{2h(x)})},$$

where

$$n(\lambda') := \sum_{z\in d(\lambda)} l(z).$$

We get the conjecture that $PP_n(t)$ is given by

$$\prod_{n=1}^{\infty} PZ_n(t,T^n) = \sum_{\lambda\in\mathcal{P}} \mathcal{PH}^\lambda_g(t)T^{|\lambda|}. \tag{22}$$

Combining the two conjectures above we can formulate:

Conjecture 5.12. *The rational functions $PP_n(t)$ defined in (22) satisfy*

- *$PP_n(t)$ is a polynomial in t;*
- *all coefficients of $PP_n(t)$ are non-negative integers;*
- *The degree of $PP_n(t)$ is $2n(n-1)(g-1)$, and the coefficient of the leading term is 1.*

For example, when $n = 3$ the Poincaré polynomial of the Pure ring should be:

$$PP_3(t) = \frac{1}{(t^6-1)(t^4-1)} + t^{12\,g-12} - \frac{t^{8\,g-8}}{t^2-1} + \frac{1}{3}\,\frac{t^{12\,g-12}}{(t^2-1)^2} - \frac{1}{3}\,\frac{t^{12\,g-12}}{t^4+t^2+1} - \frac{t^{8\,g-8}}{(t^4-1)(t^2-1)} + \frac{t^{12\,g-12}}{t^2-1}.$$

Remark 5.13. The formula of Conjecture 5.1 can be modified to give a conjectured formula for the mixed Hodge polynomial of $\mathcal{M}^d_{Dol}$. Recall from Theorem 2.1 that the mixed Hodge structure on $H^k(\mathcal{M}^d_{Dol}, \mathbb{C})$ is pure of weight k; thus this mixed Hodge polynomial coincides with the E-polynomial.

We now introduce polynomials $H_n(q, x, y)$ of three variables. Let

$$V_n(q, x, y) = H_n(q, x, y)\frac{(qxy)^{(1-g)n(n-1)}(qx+1)^g(qy+1)^g}{(qxy-1)(q-1)},$$

and

$$Z_n(q, x, y, T) = \exp\left(\sum_{r\geq 1} V_n(q^r, -(-x)^r, -(-y)^r)\frac{T^r}{r}\right).$$

Define the (x, y)-deformed Hook polynomials for genus g and partition λ:

$$\mathcal{H}^\lambda_g(q, x, y) = \prod_{z\in d(\lambda)} \frac{(qxy)^{(2-2g)l(z)}(1+q^{h(z)}y^{l(z)}x^{l(z)+1})^g(1+q^{h(z)}x^{l(z)}y^{l(z)+1})^g}{(1-q^{h(z)}(xy)^{l(z)+1})(1-q^{h(z)}(xy)^{l(z)})}. \quad (23)$$

The following generating function defines $H_n(q, x, y)$:

$$\prod_{n=1}^{\infty} Z_n(q, x, y, T^n) = \sum_{\lambda\in\mathcal{P}} \mathcal{H}^\lambda_g(q, x, y)T^{|\lambda|}. \quad (24)$$

Clearly we have $H_n(q, t, t) = H_n(q, t)$ which says that a specialization of $H_n(q, x, y)$ gives the mixed Hodge polynomial $H_n(q, t)$ of $\mathcal{M}^d_B$. The following conjecture says that another specialization gives the mixed Hodge polynomial of $\mathcal{M}^d_{Dol}$ and $\mathcal{M}^d_{DR}$.

Conjecture 5.14. *$H_n(q, x, y)$ is a polynomial with non-negative integer coefficients with specialization $H_n(1, x, y)$ equal to the E-polynomial of the Higgs moduli space $\mathcal{M}^d_{Dol}(PGL(n, \mathbb{C}))$.*

Thus we have a mysterious formula $H_n(q,x,y)$ which specializes, on one hand to the H-polynomial of the character variety, and on the other hand to the mixed Hodge polynomial of the Higgs (or equivalently flat connection) moduli space. It would be very interesting to find a geometrical meaning for $H_n(q,x,y)$.

Checks on this Conjecture 5.14 include a proof for $n=2$ and $n=3$, (one can easily modify Hitchin's and Gothen's argument to get the Hodge polynomial instead of the Poincaré polynomial of the Higgs moduli space) and also computer checks that the shape of the polynomial $H_n(1,x,y)$ is the expected one when $n=4$.

Consider now the specification $H_n(q,-1,y)$. Interestingly, the corresponding specification of the (x,y)-deformed Hook polynomials (23) becomes a polynomial, showing that at least $H_n(q,-1,y)$ is a polynomial. We get an even nicer formula if we make the further specification $H_n(1,-1,y)$ which by Conjecture 5.14 should be the Hirzebruch y-genus of the moduli space of Higgs bundles $\mathcal{M}^d_{Dol}$. Namely, for $g>1$, most of the (x,y) deformed Hook polynomials vanish, when one substitutes first $x=-1$ and then $q=1$. Indeed, the only partitions which will have a non-zero contribution to the y-genus are the partitions of the form $n=1+1+\cdots+1$; when $l(z)=0$ only for once. This in turn gives the following closed formula for the conjectured y-genus of $\mathcal{M}^d_{Dol}$:

Conjecture 5.15. *The Hirzebruch y-genus of $\mathcal{M}^d_{Dol}(PGL(n,\mathbb{C}))$, for $g>1$, equals*

$$(1-y+\cdots+(-y)^{n-1})^{g-1}\sum_{m|n}\frac{\mu(m)}{m}\left((-y)^{n(n-n/m)}m\prod_{i=1}^{n/m-1}(1-(-y)^{mi})^2\right)^{g-1}.$$

Note that the term corresponding to $m=1$ is exactly the known y-genus of $\mathcal{N}^d$ (see [41]). The rest should be thought of as the contribution of the other fixed point components of the circle action on $\mathcal{M}^d_{Dol}$. Of course, this conjectured formula gives the known specialization of Corollary 4.2 at $y=-1$, while the $y=1$ specialization gives $\mu(n)n^{g-2}$ when n is odd, and 0 when n is even. The specialization at $y=1$ can be thought of as the signature of the pairing on the rationalized circle equivariant cohomology of $\mathcal{M}^d_{Dol}$ as defined in [22].

Remark 5.16. Finally we discuss how to obtain a conjecture for the mixed Hodge polynomial of $\mathcal{M}^d_B(SL(n,\mathbb{C}))$. For the mixed Hodge polynomial of $\mathcal{M}^d_{Dol}(SL(n,\mathbb{C}))$ the mirror symmetry Conjecture 3.1, together with Conjecture 5.14 imply a conjecture. For $\mathcal{M}^d_B(SL(n,\mathbb{C}))$ the mixed Hodge polynomial contains more information than the E-polynomial. In order to have a conjecture on $H_n(x,y,t;\mathcal{M}^d_B(SL(n,\mathbb{C})))$ a mirror symmetry conjecture is needed on the level of the H-polynomial. We finish by formulating such a conjecture, generalizing Conjecture 3.6 for H-polynomials:

Conjecture 5.17. *For all $d, e \in \mathbb{Z}$, with $(d, n) = (e, n) = 1$ we have*

$$H_{\mathrm{st}}^{B^e}\Big(x, y, t; \mathcal{M}_B^d(SL(n, \mathbb{C}))\Big) = H_{\mathrm{st}}^{\hat{B}^d}\Big(x, y, t; \mathcal{M}_B^e(PGL(n, \mathbb{C}))\Big),$$

where H_{st}^B is the stringy mixed Hodge polynomial twisted with a B-field, which can be defined identically as E_{st}^B in (6).

References

1. D. Alvis – "The duality operation in the character ring of a finite Chevalley group", *Bull. Amer. Math. Soc. (N.S.)* **1** (1979), no. 6, 907–911.
2. M.F. Atiyah and R. Bott – "The Yang-Mills equations over Riemann surfaces", *Philos. Trans. Roy. Soc. London Ser. A* **308** (1982), 523–615.
3. V.V. Batyrev and D. Dais – "Strong McKay correspondence, string-theoretic Hodge numbers and mirror symmetry", *Topology* **35** (1996), 901–929.
4. K. Corlette – "Flat G-bundles with canonical metrics", *J. Differential Geom.* **28** (1988), no. 3, 361–382.
5. C. Curtis – "Truncation and duality in the character ring of a finite group of Lie type", *J. Algebra* **62** (1980), no. 2, 320–332.
6. P. Deligne – "Équations différentielles á points singuliers réguliers", Lecture Notes in Math., vol. 163, Springer, Berlin-New York, 1970.
7. — , "Théorie de Hodge II, *Inst. Hautes Études Sci. Publ. Math.* **40** (1971), 5–57.
8. R. Earl and F. Kirwan – "The Pontryagin rings of moduli spaces of arbitrary rank holomorphic bundles over a Riemann surface", *J. London Math. Soc.* (2) **60** (1999), no. 3, 835–846.
9. C. Faber – "A conjectural description of the tautological ring of the moduli space of curves", in "Moduli of curves and abelian varieties", 109–129, Aspects Math., E33, Vieweg, Braunschweig, 1999.
10. D. Freed and F. Quinn – "Chern-Simons theory with finite gauge group ", *Comm. Math. Phys.* **156** (1993), no. 3, 435–472.
11. G. Frobenius and I. Schur – "Über die reellen Darstellungen der endlichen Gruppen", *Sitzungsberichte der königlich preussichen Akademie der Wissenschaften* (1906), 186-208
12. A.M. Garsia and M. Haiman – "A remarkable q,t-Catalan sequence and q-Lagrange inversion", *J. Algebraic Combin.* **5** (1996) no. 3, 191-244.
13. W.M. Goldman and E.Z. Xia – "Rank One Higgs Bundles and Representations of Fundamental Groups of Riemann Surfaces", 2004, `math.DG/0402429`.
14. J.A Green – "The characters of the finite general linear groups", *Trans. Amer. Math. Soc.* **80** (1955), 402–447.
15. P. Griffiths, J. Harris – *Principles of algebraic geometry.* New York, Wiley 1978.
16. P.B. Gothen – "The Betti numbers of the moduli space of rank 3 Higgs bundles", *Internat. J. Math.* **5** (1994), 861–875.
17. M. Haiman – "t, q-Catalan numbers and the Hilbert scheme", *Discrete Math.* **193** (1998) no. 1-3, 201–224.
18. G. Harder and M.S. Narasimhan – "On the cohomology groups of moduli spaces of vector bundles over curves", *Math. Ann.* **212** (1975), 215–248.

19. T. Hausel – "Compactification of moduli of Higgs bundles", *J. Reine Angew. Math.* **503** (1998), 169–192.
20. — , *Geometry of the moduli space of Higgs bundles"*, Ph.D. thesis, University of Cambridge, 1998, `math.AG/0107040`.
21. — , "Quaternionic Geometry of Matroids", 2003, `math.AG/0308146`.
22. T. Hausel and N. Proudfoot – "Abelianization for hyperkähler quotients", (to appear in *Topology*), `math.SG/0310141`.
23. T. Hausel and F. Rodriguez-Villegas – "Mirror symmetry, Langlands duality and representations of finite groups of Lie type", (in preparation).
24. T. Hausel and B. Sturmfels – "Toric hyperkähler varieties", *Doc. Math.* **7** (2002), 495–534.
25. T. Hausel and M. Thaddeus – "Generators for the cohomology ring of the moduli space of rank 2 Higgs bundles", *Proc. London Math. Soc.* **88** (2004), 632–658.
26. — , "Relations in the cohomology ring of the moduli space of rank 2 Higgs bundles", *J. Amer. Math. Soc.*, **16** (2003), 303–329.
27. — , "Examples of mirror partners arising from integrable systems", *C. R. Acad. Sci. Paris*, **333** (4) (2001), 313–318.
28. — , "Mirror symmetry, Langlands duality and Hitchin systems", *Invent. Math.*, **153**, no. 1 (2003), 197–229.
29. N. Hitchin – "The self-duality equations on a Riemann surface", *Proc. London Math. Soc.* (3) **55** (1987), 59–126.
30. — , "Stable bundles and integrable systems", *Duke Math. J.* **54** (1987), 91–114.
31. — , "Lectures on special Lagrangian submanifolds", *Winter School on Mirror Symmetry, Vector Bundles and Lagrangian Submanifolds (Cambridge, MA, 1999)*, 151–182, Amer. Math. Soc., 2001.
32. H. Jordan – "Group characters of various types of linear groups", *Amer. J. Math*, **29** (1907), 387.
33. L. Katzarkov and T. Pantev – "Nonabelian (p, p) classes ", (in Motives, polylogarithms and Hodge theory, Part II (Irvine, CA, 1998)), *Int. Press Lect. Ser.*, 3, II, 625–715.
34. M. Kontsevich – "Motivic Integration", *Lecture at Orsay*, 1995.
35. G. Lehrer – "The characters of the finite special linear groups", *J. Algebra* **26** (1973), 564–583.
36. G. Lusztig – "Fermionic form and Betti numbers", `math.QA/0005010`.
37. A.D. Mednykh – "Determination of the number of nonequivalent coverings over a compact Riemann surface", *Soviet Mathematics Doklady* **19** (1978), 318–320.
38. M. Mehta – "Hodge structure on the cohomology of the moduli space of Higgs bundles", `math.AG/0112111`.
39. E. Markman – "Generators of the cohomology ring of moduli spaces of sheaves on symplectic surfaces", *J. Reine Angew. Math.* **544** (2002), 61–82.
40. H. Nakajima – "Instantons on ALE spaces, quiver varieties, and Kac-Moody algebras", *Duke Math. J.* **76** (1994) no. 2, 365–416.
41. M.S. Narasimhan and S. Ramanan – "Generalised Prym varieties as fixed points", *J. Indian Math. Soc.* **39** (1975), 1–19.
42. I. Schur – "Untersuchungen über die Darstellung der endlichen Gruppen durch Gebrochene Lineare Substitutionen", *J. Reine Angew. Math.* **132** (1907), 85.
43. J-P. Serre – "Exemples de variétés projectives conjuguées non homéomorphes", *C. R. Acad. Sci. Paris* **258** (1964), 4194–4196.

44. C.T. Simpson – "Nonabelian Hodge theory", *Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990)*, 747–756, Math. Soc. Japan, Tokyo, 1991.
45. — , "Higgs bundles and local systems", *Inst. Hautes Études Sci. Publ. Math.* **75** (1992), 5–95.
46. — , "Moduli of representations of the fundamental group of a smooth projective variety", I, *Inst. Hautes Études Sci. Publ. Math.* **79** (1994), 47–129.
47. — , "Moduli of representations of the fundamental group of a smooth projective variety", II, *Inst. Hautes Études Sci. Publ. Math.* **80** (1995), 5–79.
48. — , "The Hodge filtration on nonabelian cohomology", *Algebraic geometry—Santa Cruz 1995*, Proc. Symp. Pure Math. 62, ed. J. Kollár, R. Lazarsfeld, and D. Morrison, American Math. Soc., 1997.
49. A. Strominger, S.-T. Yau, and E. Zaslow – "Mirror symmetry is T-duality", *Nuclear Phys. B* **479** (1996), 243–259.
50. C. Vafa – "String vacua and orbifoldized LG models", *Modern Phys. Lett. A* **4** (1989), 1169–1185.
51. E. Verlinde – "Fusion rules and modular transformations in 2D conformal field theory", *Nucl. Phys. B* **300** (1988), 360.
52. D. Zagier – "Elementary aspects of the Verlinde formula and of the Harder-Narasimhan-Atiyah-Bott formula", (in Proceedings of the Hirzebruch 65 Conference on Algebraic Geometry (Ramat Gan, 1993), 445–462, *Israel Math. Conf. Proc.*, 9, Bar-Ilan Univ., Ramat Gan, 1996.
53. E. Zaslow – "Topological orbifold models and quantum cohomology rings", *Comm. Math. Phys.* **156** (1993), 301–331.

Mahler measure for dynamical systems on $\mathbb{P}^1$ and intersection theory on a singular arithmetic surface

Jorge Pineiro, Lucien Szpiro, and Thomas J. Tucker

Ph.D. Program in Mathematics, Graduate Center of CUNY, 365 Fifth Avenue, New York, NY 10016-4309, U.S.A.
jpineiro@gc.cuny.edu
lszpiro@gc.cuny.edu
ttucker@gc.cuny.edu

Summary. The Mahler measure formula expresses the height of an algebraic number as the integral of the log of the absolute value of its minimal polynomial on the unit circle. The height is in fact the canonical height associated to the monomial maps x^n. We show in this work that for any rational map $\varphi(x)$ the canonical height of an algebraic number with respect to φ can be expressed as the integral of the log of its equation against the invariant Brolin–Lyubich measure associated to φ, with additional adelic terms at finite places of bad reduction. We give a complete proof of this theorem using integral models for each iterate of φ. In the last chapter on equidistribution and Julia sets we give a survey of results obtained by P. Autissier, M. Baker, R. Rumely and ourselves. In particular our results, when combined with techniques of diophantine approximation, will allow us to compute the integrals in the generalized Mahler formula by averaging on periodic points.

1 Introduction

If F is the minimal polynomial over $\mathbb{Z}$ for an algebraic number x, the formula of Mahler [17] for the usual height $h(x)$ is

$$\deg(F)h(x) = \log \prod_{\text{all places } v} \sup(|x|_v, 1) = \int_0^1 \log |F(e^{2\pi i\theta})| d\theta.$$

One can notice the following facts:

1. The height satisfies the functional equation $h(x^2) = 2h(x)$.
2. $d\theta$ is supported on the unit circle, which is the closure of the set of roots of unity, each root of unity having height 0. Along with the points 0 and ∞, the roots of unity are the only points that have finite forward orbits

under iteration of the map $x \to x^2$.

We show in this article that these occurrences are general for any dynamical system on the Riemann sphere given by a rational function φ with coefficients in a number field. There is a canonical height h_φ that vanishes at precisely the points that have finite forward orbits under the iteration of φ. At each infinite place v, we have an integral $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$, where $d\mu_{v,\varphi}$ is the distribution associated to a canonical metric for φ on $\mathcal{O}_{\mathbb{P}^1}(1)$ (see Zhang [32]). At each finite v, we define an integral $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$ which is constructed via a limiting process that is analogous to Brolin's construction [6] of the φ-invariant measure at an infinite place, as we explain in Section 5. Our Theorem 6.1 asserts that

$$[K(\alpha):\mathbb{Q}]h_\varphi(\alpha) = \sum_{\text{places } v \text{ of } K} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi},$$

where α is an algebraic point and F is a minimal polynomial for α over K. We also show that for finite v we have $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} = 0$ unless φ has bad reduction at v or all the coefficients of F have nonzero v-adic valuation. Moreover, we show that $\sum_{\text{finite } v} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$ can be explicitly bounded in terms of F and polynomials P and Q for which $\varphi = P/Q$. In particular, Corollary 6.3 states that

$$\deg(F)h_\varphi(\alpha) \le \sum_{v|\infty} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$$

if P is monic and F has coprime integral coefficients.

We compute canonical heights via Arakelov intersection theory on arithmetic surfaces. The maps φ^k may not be defined on $\mathbb{P}^1_{\mathcal{O}_K}$, where $\mathcal{O}_K$ is the ring of integers of the number field K. Thus, we must work with blow-ups of $\mathbb{P}^1_{\mathcal{O}_K}$. These blow-ups may be singular arithmetic surfaces. We have found it convenient to work with Cohen–Macaulay surfaces instead of normal surfaces.

Algebraic dynamical systems have been studied by many authors; see, for example, C. T. McMullen [19] and J. Milnor [20]. Recently G. Everest and T. Ward, in their book [10], have studied algebraic dynamics on elliptic curves and on products of projective lines. They have pointed out particular cases of our main theorem (other cases have been studied by V. Maillot [18]).

In light of work of Szpiro, Ullmo, and Zhang [28], it seems natural to wonder if points with small canonical height h_φ on $\mathbb{P}^1$ are equidistributed with respect to $d\mu_{v,\varphi}$ for v an infinite place. P. Autissier [2] and M. Baker and R. Rumely [3] have recently shown that such an equidistribution result does indeed hold. One might also ask whether generalized Mahler measure can be computed via equidistribution. We will discuss conjectures and known results of this sort more precisely in Section 7.

The organization of the paper is as follows:

1. Introduction.
2. Canonical height, canonical metric, and canonical distribution associated to a dynamical system.
3. Three examples.
4. The blow-up associated to a model of a rational map from $\mathbb{P}^1$ to $\mathbb{P}^1$.
5. The integrals at finite places.
   5.1 Existence of the integrals at finite places;
   5.2 Invariance of the v-adic integral under change of variables;
   5.3 Geometry of v-adic integrals;
   5.4 A remark on the use of integral notation at finite places.
6. The Mahler formula for dynamical systems.
7. Equidistribution and the Julia set.
8. A-Appendix: Schematic intersection theory on a Cohen–Macaulay surface. (This is needed to justify the use of intersections products in Section 4.)
9. B-Appendix: Arakelov intersection with the divisor associated to a rational function. (This is used in Section 6.)

The idea of the proof of Theorem 6.1 is to follow Tate's method for computing the canonical height using the language of Arakelov theory. For a point $\alpha \in \mathbb{P}^1(\overline{K})$ and any positive integer k, we compute $h(\varphi^k(\alpha))/(\deg\varphi)^k$ using intersection theory on the (possibly singular) blow-up X_k constructed in Section 4. The height $h(\varphi^k(\alpha))/(\deg\varphi)^k$ decomposes into a sum over the places of K. The contribution at a finite place v comes from the self-intersection of components of the exceptional divisor of X_k. In Section 5, we show that this contribution at a finite place converges as k goes to infinity. We compute the contribution to $h(\varphi^k(\alpha))/(\deg\varphi)^k$ at an infinite place v in Proposition B.3.

Acknowledgments. We would like to thank D. Sullivan and S. Zhang for very interesting discussions on the subject of this paper.

The first and second authors were partially supported by NSF Grant 0071921. The first author was also supported by the NSF/AGEP MAGNET-STEM program. The third author was partially supported by NSF Grant 0101636

2 Canonical height, canonical metric, and canonical distribution associated to a dynamical system

The general global theory of canonical heights was started by J. Silverman and G. Call [8]; earlier, Néron and Tate had developed a theory of canonical heights

in the special case of abelian varieties. Let X be a variety over a number field K. Suppose φ is a finite map of X to itself. Suppose that its degree d is greater than 1 and that there is an ample line bundle L on X satisfying $\varphi^*(L) \cong L^{\otimes d}$. Tate's recipe for the definition of the Néron–Tate height on an abelian variety carries over to this general case: one has a canonical height associated to φ defined by

$$h_\varphi(\alpha) = \lim_{k\to\infty} \frac{h_L(\varphi^k(\alpha))}{d^k}.$$

In this formula, h_L is associated with any set of smooth metrics at places at infinity (cf. [26] and Appendix B).

The canonical height h_φ has the following properties:

(1) h_φ satisfies Northcott's theorem: points over $\bar{K}$ with bounded degree and bounded height are finite in number;

(2) $h_\varphi(\varphi(\alpha)) = dh_\varphi(\alpha)$;

(3) h_φ is a non-negative function;

(4) $h_\varphi(\alpha) = 0$ if and only if α has a finite forward orbit under iteration of φ;

(5) $|h_\varphi(\alpha) - h(\alpha)|$ is bounded, where h is the usual height.

The canonical height h_φ is characterized as the unique function on $\mathbb{P}^1(\bar{K})$ that satisfies (2) and (5).

Definition 2.1. *A point is called periodic if it is a fixed point of φ^k for some positive integer k. A point is called preperiodic if its image under φ^m is periodic for some integer m. Equivalently, a point is preperiodic if and only if it has a finite forward orbit under iteration of φ.*

The periodic points are separated classically into three classes important for the dynamics:

Definition 2.2. *Let f be a differentiable map $f : \mathbb{P}^1_{\mathbb{C}} \to \mathbb{P}^1_{\mathbb{C}}$. A fixed point x of f^k is called repelling (resp. attracting, resp. indifferent) if $|(f^k)'(x)| > 1$ (resp. $|(f^k)'(x)| < 1$, resp. $|(f^k)'(x)| = 1$). The closure in $\mathbb{P}^1(\mathbb{C})$ of the set of repelling periodic points is called the Julia set. The complement in $\mathbb{P}^1(\mathbb{C})$ of the Julia set is called the Fatou set of f.*

S. Zhang [30, 32] has demonstrated the interest of metrics in the study of canonical heights. In [32, Section 2], he shows (following Tate) that if a line bundle L on a projective variety W has a metric $\|\cdot\|_v$ and there is an isomorphism $\tau : L^{\otimes d} \xrightarrow{\sim} \varphi^*(L)$ for some $d > 1$, then letting $\|\cdot\|_{v,0} = \|\cdot\|_v$ and $\|\cdot\|_{v,k+1} = (\tau^*\varphi^*\|\cdot\|_{v,k})^{1/d}$, one obtains a limit metric

$$\|\cdot\|_{v,\varphi} = \lim_{k\to\infty} \|\cdot\|_{v,k}.$$

The following proposition is due to S. Zhang (see [32, Theorems 1.4 and 2.2]).

Proposition 2.3. *Let L and the metrics $\|\cdot\|_v$, $\|\cdot\|_{v,k}$, and $\|\cdot\|_{v,\varphi}$ on L be as above. Then:*

1. *The $\|\cdot\|_{v,k}$ converge uniformly to $\|\cdot\|_{v,\varphi}$.*
2. *Suppose additionally that W is a curve, v is an infinite place, and that $\|\cdot\|_v$ is smooth and semipositive. Then the (normalized) curvatures*

$$d\mu_{v,k} = -\frac{1}{(2\pi i)d^k} d\overline{d} \log \|\cdot\|_{v,k}$$

have a limit distribution $d\mu_{v,\varphi}$ such that if s is a meromorphic section of a line bundle L' with metric $\|\cdot\|'_v$ at v, then $\lim_{k\to\infty} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log \|s\|'_v d\mu_{v,k}$ exists and is equal to $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log \|s\|'_v d\mu_{v,\varphi}$.

Furthermore, neither $\|\cdot\|_{v,\varphi}$ nor $d\mu_{v,\varphi}$ depends on the choice of a smooth metric $\|\cdot\|_v$.

By convention, all of our metrics at infinity are normalized; that is, on each stalk of L on $X(\mathbb{C}_v)$ the metric $\|\cdot\|_v$ behaves like $|\cdot|_v^{[K_v:\mathbb{R}]}$ where K_v is the completion of K at an infinite place v.

We note that canonical heights are the only Arakelov type heights which are non-negative naturally. The height is an intersection number in the sense of J. Arakelov [1] as extended by P. Deligne [9] and by S. Zhang [32] to this limit situation.

3 Examples

3.1 The squaring map on the multiplicative group and the naive height

For the map $\varphi(t) = t^2$ on $\mathbb{P}^1$, the preperiodic points are zero, infinity, and the roots of unity. The unit circle is the Julia set (the closure of the repelling periodic points). The naive height satisfies the required functional equation to be the canonical height associated to φ. This can be verified via the usual definition

$$h([t_0 : t_1]) = \frac{1}{[K : \mathbb{Q}]} \log \prod_{\text{places } v \text{ of } K} \sup(|t_0|_v, |t_1|_v)^{\mathrm{N}_v},$$

where $\mathrm{N}_v = [K_v : \mathbb{Q}_v]$ and $[t_0 : t_1] \in \mathbb{P}^1(K)$. Thus, the naive height is the canonical height associated to φ.

The unit circle is the support of $d\mu_\varphi$ (in this case the Haar measure $d\theta$ on the unit circle). It is the curvature (in the sense of distributions) of the canonical metric

$$\|(\lambda T_0 + \mu T_1)([a : b])\| = \frac{|\lambda a + \mu b|}{\sup(|a|, |b|)}.$$

Note that the canonical height and curvature are the same for any map $\phi(t) = t^n$ with $n \geq 2$.

3.2 The Néron–Tate height is associated to multiplication by two on an elliptic curve

Let E be an elliptic curve with Weierstrass equation $y^2 = G(t)$. We write $E = \mathbb{C}/(\mathbb{Z}1 + \mathbb{Z}\tau)$ with τ in the upper half plane. Passing to the quotient by $[-1]$, multiplication by 2 on E gives rise to the following rational map on $\mathbb{P}^1$:

$$\varphi(t) = \frac{G'^2(t) - 8tG(t)}{4G(t)}.$$

The preperiodic points of φ are the images of the torsion points in E. To see this note first that $2^n P = \pm 2^k P$ for $n \neq k$ implies P is a torsion point. Conversely, if $nP = 0$, then writing $2^k = qn + r_k$ with $r_k < n$, we see that there must be at least two different indices k and k' with $r_k = r_{k'}$ hence $2^k P = 2^{k'} P$. The fixed points of φ are the images of the inflection points of E, i.e., of the 3-torsion points in E.

The multiplication map $[2^n]$ on E is of course étale. The derivative of this map is 2^n everywhere, so the preperiodic points are all repelling. Hence the Julia set is the entire Riemann sphere. In [21] it is established that the Haar measure on E gives the curvature of the canonical metric associated to the Néron–Tate height. Its image on $\mathbb{P}^1$ is

$$d\mu_\varphi = \frac{i}{\mathrm{Im}(\tau)} \frac{dt \wedge \overline{dt}}{|G|}.$$

The curvature and canonical height will be the same for φ and the map on $\mathbb{P}^1$ associated to any multiplication by n map on E for $n \geq 2$. In Example 6.5, we will give an explicit computation of a Mahler measure on a dynamical system associated to multiplication by 2 an elliptic curve.

3.3 Parallel projection of a conic

Consider the plane conic C over $\mathbb{Z}$ defined by the equation

$$X_0 X_1 + p X_2^2 = 0,$$

where p is an odd prime number. The reduction of C mod a prime ℓ is smooth and connected for $\ell \neq p$. The fiber over p is reduced and is the union of two lines. The arithmetic surface C is regular. The projection map from $\mathbb{P}^2$ to $\mathbb{P}^1$ defined by

$$\Phi([X_0 : X_1 : X_2]) = [X_0 + X_1 : X_2]$$

is well-defined as a map from C to $\mathbb{P}^1$. Projecting from $[0:1:0]$ yields an isomorphism between our conic and $\mathbb{P}^1$. Composing this with Φ gives rise to a map φ; the reader may check that this map is $\varphi(t) = \frac{t^2-p}{t}$. This example is an illustration of a blowing-up allowing us to define the map φ over $\mathbb{Z}$. This will be a systematic approach in the next section.

4 The Blow-up associated to a model of a rational map from $\mathbb{P}^1$ to $\mathbb{P}^1$

Let ψ be a rational map of degree d from $\mathbb{P}^1$ to $\mathbb{P}^1$ defined over the field of fractions K of a Dedekind domain B. We will assume that ∞ is a fixed point of ψ, because this simplifies the statement of our main theorem. After extending the base field K, we may assume that ψ has a fixed point α defined over K. We may then choose coordinates so that α is the point at infinity. Hence, our assumption does not restrict our generality. The definition of our integrals $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$ does, however, depend on our choice of coordinates.

Definition 4.1. *A model over B of ψ is a map of polynomial rings*

$$\psi([T_0 : T_1]) = [G(T_0, T_1) : H(T_0, T_1)],$$

where the polynomials G and H are homogeneous of degree d with coefficients in B.

The models will allow us to work on arithmetic surfaces when K is a number field. These surfaces, obtained by blowing up non-regular centers, will be singular in general.

Let Y be the closed subscheme of $X = \mathbb{P}^1_B = \mathrm{Proj}(B[T_0, T_1])$ defined by the vanishing of G and H. Let I be the sheaf of ideals in $\mathcal{O}_X$ defining Y. The choice of G and H as "generators of I" gives rise to a surjection

$$\tilde{\psi}_1 : \mathcal{O}_X^2 \twoheadrightarrow I(d).$$

The support of the scheme Y is exactly where the map ψ cannot be extended to the fibers of the model $\mathbb{P}^1_B$.

Let N denote the scheme-theoretic image of the projection from Y to $\mathrm{Spec}(B)$; we call this the *bad reduction* of the model. The scheme Y does not meet the generic fiber X_K.

Definition 4.2. *Let $\sigma : X_1 \to X$ be the blowing-up of Y in X.*

By the universal property of the blow-up, the pull-back $\sigma^* I$ is locally generated by elements which are not zero divisors, i.e., there is a positive Cartier divisor E_1 on X_1 such that

$$\sigma^* I = \mathcal{O}_{X_1}(-E_1).$$

One then has a surjective map (which is simply σ^* of $\tilde{\psi_1}$)

$$\mathcal{O}_{X_1}^2 \twoheadrightarrow \sigma^*(\mathcal{O}_X(d)) \otimes \mathcal{O}_{X_1}(-E_1).$$

By the universal property characterizing the projective line this gives rise to a map:

$$\psi_1 : X_1 \to X,$$

extending the original rational map ψ on the generic fibers.

Throughout this section we use the scheme-theoretic intersection product $(\cdot.\cdot)$, defined in Appendix A.

Proposition 4.3. *The two-dimensional scheme X_1 is reduced, irreducible and Cohen–Macaulay. The fiber of X_1 over $v \notin N$ is equal to $\mathbb{P}^1_{k_v}$. The fiber of X_1 over $v \in N$ has a finite number of components $W_v, C_{v,1}, ..., C_{v,t_v}$, where W_v is the strict transform of the fiber of X at v and the $C_{v,i}$ are components of the exceptional divisor of σ. Each $C_{v,i}$ is isomorphic to $\mathbb{P}^1_{k_{v,i}}$ where $k_{v,i}$ is the residual field of the closed point image of $C_{v,i}$ in X. The $C_{v,i}$ for $i > 0$ do not meet each other. The $C_{v,i}$ with $i > 0$ meet W_v. The exceptional divisor E_1 is a Cartier divisor equal to $\mathbb{P}^1_Y$ and, as a Weil divisor, it can be decomposed into $\sum_{v\in N, i>0} r_{v,i} C_{v,i}$ where the $r_{v,i}$ are positive integers equal to the local lengths of $\mathcal{O}_Y$ at the local rings of its support. The geometric self-intersection of $C_{v,i}$ is equal to $-[k_{v,i} : k_v]/r_{v,i}$. One has*

$$\psi_1^* \mathcal{O}_X(1) = \sigma^*(\mathcal{O}_X(d)) \otimes \mathcal{O}_{X_1}(-E_1).$$

Proof. Since the ideal I is generated by two elements, the scheme $X_1 = \mathrm{Proj}(\bigoplus I^n)$ is a closed subscheme of $\mathbb{P}^1_X = \mathrm{Proj}(\mathcal{O}_X[T_0, T_1])$. The exceptional fiber is then $\mathbb{P}^1_Y$. The surface X_1 embeds as a local complete intersection in the regular three-fold $\mathbb{P}^1_X$ and hence is Cohen–Macaulay. The $C_{v,i}$ are Weil divisors that are $\mathbb{Q}$-Cartier because E_1 and the total fibers F_v^* are Cartier divisors and the $C_{v,i}$ do not meet. Hence the intersection theory with the $C_{v,i}$ is as described in the Appendix A. One sees that the W_v are also $\mathbb{Q}$-Cartier since the total fiber of v is Cartier and the $C_{v,i}$ are $\mathbb{Q}$-Cartier. One has $\mathcal{O}_{X_1}(-E_1)|E_1 = \oplus_{i,v} \mathcal{O}_{\mathbb{P}^1_{Y_{i,v}}}(1)$. Hence

$$((-r_{v,i}C_{v,i}).(r_{v,i}C_{v,i})) = \deg(\mathcal{O}_{\mathbb{P}^1_{Y_{i,v}}}(1)) = [k_{v,i,k} : k_v] r_{v,i}.$$

This gives the value asserted for the self intersection of the components $C_{v,i}$. The formula for $\psi_1^* \mathcal{O}_X(1)$ comes from the universal property characterizing the projective line over X. □

Remark 4.4. This method of removing the indeterminacies of a map by blowing up is standard (see Hartshorne [13, p. 168] for example).

5 The integrals at finite places

Let v be a finite place of K. In this section we will:

(1) define the v-adic integral and show that it exists,

(2) show that the v-adic integral does not depend on our choice of polynomials P and Q defining a model for φ over $\mathcal{O}_K$,

(3) relate the v-adic integral to the geometry of the blow-up maps associated to the φ^k, and

(4) explain why it makes sense to think of this integral at the finite place v as a v-adic analogue of an integral at an archimedean place.

We begin by developing some terminology. Let $\varphi : \mathbb{P}^1_K \longrightarrow \mathbb{P}^1_K$ be a rational map with a fixed point at ∞. We let $\mathcal{O}_K$ denote the ring of integers of a number field K. Let

$$\varphi([T_0 : T_1]) = [P(T_0, T_1) : Q(T_0, T_1)],$$

where P and Q are homogeneous polynomials of degree d in $\mathcal{O}_K[T_0, T_1]$, be a model for φ over $\mathcal{O}_K$, in the terminology of Section 4. Letting $P_1 = P$ and $Q_1 = Q$, and recursively defining $P_{k+1} = P(P_k, Q_k)$ and $Q_{k+1} = Q(P_k, Q_k)$, we obtain models

$$\varphi^k([T_0 : T_1]) = [P_k(T_0, T_1) : Q_k(T_0, T_1)]$$

for iterates φ^k of φ. Recall that T_1 must be a factor of each Q_k since $\infty = [1 : 0]$ is a fixed point of φ_k.

Throughout this section, v will denote a finite valuation on K that has been extended to the algebraic closure $\overline{K}$ of K. We let $\mathcal{O}_v$ denote the set of all $z \in K$ for which $v(z) \geq 0$.

For $(a, b) \in \overline{K}^2 \smallsetminus (0, 0)$, we define

$$\begin{aligned} &S_{v,k}(P_k(a,b),Q_k(a,b)) \\ &\qquad := <(v(P_k(a,b)), v(Q_k(a,b))) - <(v(a^{d^k}), v(b^{d^k})). \end{aligned} \tag{1}$$

Note that $S_{v,k}(P_k(a, b), Q_k(a, b))$ is a finite number, since P_k and Q_k have no common factor, and that

$$S_{v,k}(P_k(a, b), Q_k(a, b)) = S_{v,k}(P_k(za, zb), Q_k(za, zb)) \tag{2}$$

for any nonzero $z \in \overline{K}$. It follows that $S_{v,k}(P_k(a, b), Q_k(a, b))$ is non-negative, since we may thus assume that $<(v(a), v(b)) = 0$. We also define

$$R_v(P_k, Q_k) := \sup_{(a,b)\in\overline{K}^2\smallsetminus(0,0)} (S_{v,k}(P_k(a, b), Q_k(a, b))).$$

To see that $R_v(P_k, Q_k)$ exists and is finite, we apply the Euclidean algorithm to $P_k(T_0, 1)$ and $Q_k(T_0, 1)$ to obtain

$$x_0(T_0)P_k(T_0,1) + y_0(T_0)Q_k(T_0,1) = m_0$$

with $x_0, y_0 \in \mathcal{O}_v[T_0]$ and nonzero $m_0 \in \mathcal{O}_v$. Then, for any $(a,b) \in \overline{K}^2 \smallsetminus (0,0)$ with $v(a) \geq v(b)$, it follows that $S_{v,k}(P_k(a,b), Q_k(a,b)) \leq v(m_0)$. Similarly, writing $x_1(T_1)P_k(1,T_1) + y_1(T_1)Q_k(1,T_1) = m_1$ we see that

$$S_{v,k}(P_k(a,b), Q_k(a,b)) \leq v(m_1),$$

whenever $v(b) \geq v(a)$. Thus, for any $(a,b) \in \overline{K}^2 \smallsetminus (0,0)$, we have

$$S_{v,k}(P_k(a,b), Q_k(a,b)) \leq \max(v(m_0), v(m_1)).$$

Note that

$$v(m_0) \leq v(\mathrm{Res}(P_k(T_0,1), Q_k(T_0,1))),$$

where $\mathrm{Res}(P_k(T_0,1), Q_k(T_0,1))$ is the resultant of $P_k(T_0,1)$ and $Q_k(T_0,1)$ as polynomials in T_0 (see Brieskorn and Knörrer [5, p. 279, Proposition 4], for example) and that $v(m_1) \leq v(\mathrm{Res}(P_k(1,T_1), Q_k(1,T_1)))$. If $R_v(P_k, Q_k) > 0$, then P_k and Q_k have a common root modulo the prime ideal corresponding to v and thus our model for φ^k has bad reduction at v (see Section 4).

5.1 Existence of the v-adic integral

Let $F \in K[t]$ be a nonconstant polynomial. Using the coordinates $[T_0 : T_1]$ and and letting $t = T_0/T_1$, we let $[a_1 : b_1], \ldots, [a_{\deg F} : b_{\deg F}]$ be the points in $\mathbb{P}^1(\overline{K})$ at which F vanishes.

Proposition 5.1. *The sequence* $\left(\sum_{\ell=1}^{\deg F} \frac{S_{v,k}(P_k(a_\ell,b_\ell),Q_k(a_\ell,b_\ell))}{d^k}\right)_k$ *is increasing and is bounded by* $(\deg D)(\frac{1}{d-1})R_v(P,Q)$.

This allows us to define the integrals at a finite place v as follows.

Definition 5.2. *For a finite place v,*

$$\begin{aligned}\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} := &- \lim_{k\to\infty} \sum_{\ell=1}^{\deg F} \frac{S_{v,k}(P_k(a_\ell,b_\ell), Q_k(a_\ell,b_\ell))}{d^k} \log \mathrm{N}(v) \\ &- v(F)\log \mathrm{N}(v) + (\deg F)\frac{v(A_d)}{d-1}\log \mathrm{N}(v),\end{aligned} \tag{3}$$

where $S_{v,k}$ is given by (1), $P = \sum_{i=1}^{d} A_i T_0^i T_1^{d-i}$, $\mathrm{N}(v)$ *is the cardinality of the residue field of K at v, and $v(F)$ is the v-adic valuation of the content of F, i.e., $v(F) = <_i(v(m_i))$ when $F = \sum_{i=1}^{\deg F} m_i t^i$.*

As stated, Definition 5.2 has no obvious relation to the geometry of the map φ^k. Soon, we will see that φ^k gives rise to a blow-up X_k of $\mathbb{P}^1_{\mathcal{O}_k}$ such that:

- the term $-S_{v,k}(P_k(a,b),Q_k(a,b))$ comes from intersecting the horizontal part of $\mathrm{div}(F)$ with the exceptional divisor of X_k;
- the term $-v(F)\log \mathrm{N}(v)$ represents the contribution of the vertical part of $\mathrm{div}(F)$ to an intersection product;
- the term $(\deg F)\frac{v(A_d)}{d-1}$ comes from intersecting the horizontal divisor corresponding to $[1:0]$ with the exceptional divisor of X_k and letting k go to infinity.

Call and Goldstine ([7, Theorem 3.1]) have shown that for any $\alpha \in \overline{K}$, we have

$$\lim_{k\to\infty} \frac{\log \max(|P_k(\alpha,1)|_v, |Q_k(\alpha,1)|_v)}{d^k} = \hat{\lambda}_{[1:0],\varphi,Q}(\alpha,v),$$

where $\hat{\lambda}_{[1:0],\varphi,Q}(\cdot,v)$ is the canonical local height associated to $[1:0]$, φ, and Q at the place v. Thus, Definition 5.2 is closely related to the local canonical height of the points at which F vanishes. In particular, we have

$$\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|t-\alpha|_v d\mu_{v,\varphi} = \hat{\lambda}_{[1:0],\varphi,Q}(\alpha,v) + \frac{v(A_d)}{d-1}\log \mathrm{N}(v).$$

Proposition 5.1 is a simple consequence of the following two lemmas.

Lemma 5.3. *Let $(a,b) \in \overline{K}^2 \smallsetminus (0,0)$. For all integers $k \geq 1$, we have*

$$\begin{aligned} S_{v,k}(P_{k+1}(a,b),Q_{k+1}(a,b)) &= dS_{v,k}(P_k(a,b),Q_k(a,b)) \\ &\quad + S_{v,k}(P(P_k(a,b),Q_k(a,b)),Q(P_k(a,b),Q_k(a,b))). \end{aligned}$$

Proof. By (2), we may assume that $<(v(a),v(b)) = 0$. Then

$$\begin{aligned} &S_{v,k}(P_{k+1}(a,b),Q_{k+1}(a,b)) \\ &\quad = <(v(P(P_k(a,b),Q_k(a,b))),v(Q(P_k(a,b),Q_k(a,b)))) \\ &\quad = <(v(P_k(a,b)^d),v(Q_k(a,b)^d)) - <(v(P_k(a,b)^d),v(Q_k(a,b)^d)) \\ &\quad + <(v(P(P_k(a,b),Q_k(a,b))),v(Q(P_k(a,b),Q_k(a,b)))) \\ &\quad = dS_{v,k}(P_k(a,b),Q_k(a,b)) \\ &\qquad + S_{v,k}(P(P_k(a,b),Q_k(a,b)),Q(P_k(a,b),Q_k(a,b))). \end{aligned}$$

□

Lemma 5.4. *For all integers $k \geq 1$, we have*

$$R_v(P_k,Q_k) \leq R_v(P,Q)\Big(\sum_{i=1}^{k} d^{i-1}\Big). \tag{4}$$

Proof. We proceed by induction. The case $k=1$ is obvious. Now, let $a,b \in \overline{K}^2 \smallsetminus (0,0)$; we may assume by (2) that $<(w(a),w(b)) = 1$. If (4) holds for k, then applying Lemma 5.3 to the case of $k+1$ yields

$$S_{v,k}(P_{k+1}(a,b),Q_{k+1}(a,b)) \le dR_v(P_k,Q_k) + R_v(P,Q)$$
$$= dR_v(P,Q)(\sum_{i=1}^{k} d^{i-1}) + R_v(P,Q) = R_v(P,Q)(\sum_{i=1}^{k+1} d^{i-1}).$$

□

Proof (of Proposition 5.1). It follows from Lemma 5.3 that the sequence is increasing. By Lemma 5.4, we have

$$\frac{1}{d^k}(R_v(P_k,Q_k)) \le \frac{1}{d^k}R_v(P,Q)\left(\sum_{i=1}^{k} d^{i-1}\right) = R_v(P,Q)\left(\sum_{i=1}^{k} d^{i-k-1}\right)$$
$$\le R_v(P,Q)\left(\sum_{i=1}^{\infty}\frac{1}{d^i}\right) = R_v(P,Q)\left(\frac{1}{d-1}\right),$$

which is precisely the bound given in the statement of Proposition 5.1. □

5.2 Invariance of the v-adic integral under change of variables

The quantity $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$ in Definition 5.2 involves P and Q. We can show, however, that the definition depends only on our choice of the point at infinity. Let τ be a change of variable of the form $\tau(T_0) = mU_0 + nU_1$, $\tau(T_1) = zU_1$ (so that τ fixes $[1:0]$). To get a model from this change of variables, we let $\tau^*Q = zQ(\tau(T_0),U_1)$ and let

$$\tau^*P = \frac{zP(\tau(T_0),U_1)}{m} - \frac{znQ(\tau(T_0),U_1)}{m},$$

where $z \in \mathcal{O}_K$ is chosen so that τ^*Q and τ^*P are both in $\mathcal{O}_K[U_0,U_1]$. Note that P is written as it is since $\tau^{-1}(T_0) = U_0/m - nU_1/m$. We define τ^*P_k and τ^*Q_k recursively as P_k and Q_k. Write $\tau^*F(u) = F(\tau(T_0),U_1)/U_1^d$ where $u = U_0/U_1$. Note that $\deg\tau^*F = \deg F$.

Proposition 5.5. *With τ as above, we have*

$$\lim_{k\to\infty}\sum_{\ell=1}^{\deg F}\frac{S_{v,k}(P_k(a_\ell,b_\ell),Q_k(a_\ell,b_\ell))}{d^k} - (\deg F)\frac{v(A_d)}{d-1} + v(F)$$
$$= \lim_{k\to\infty}\sum_{\ell=1}^{\deg F}\frac{S_{v,k}(\tau^*P_k(a'_\ell,b'_\ell),\tau^*Q_k(a'_\ell,b'_\ell))}{d^k} - (\deg F)\frac{v(\tau^*A_d)}{d-1} + v(\tau^*F), \tag{5}$$

where τ^*A_d is the leading coefficient of τ^*P and $\tau(a_\ell T_0 + b_\ell T_1) = a'_\ell U_0 + b'_\ell U_1$. Thus, Definition 5.2 does not depend on our choice of P and Q.

Proof. The proof is a simple computation. We compute

$$\lim_{k\to\infty}\sum_{\ell=1}^{\deg F}\frac{S_{v,k}(P_k(a_\ell,b_\ell),Q_k(a_\ell,b_\ell))}{d^k}=\lim_{k\to\infty}\sum_{\ell=1}^{\deg F}\frac{S_{v,k}(\tau^*P_k(a'_\ell,b'_\ell),\tau^*Q_k(a'_\ell,b'_\ell))}{d^k}$$
$$+\frac{(\deg F)v(z)}{d-1}+\sum_{\ell=1}^{\deg F}<(v(a_\ell),v(b_\ell))-\sum_{\ell=1}^{\deg F}<(v(a'_\ell),v(b'_\ell)). \tag{6}$$

We also see that

$$v(\tau^*A_d)=v(A_d)+v(z)-(d-1)v(m). \tag{7}$$

Now, we may choose our a_ℓ and b_ℓ so that $F(t)=\prod_{\ell=1}^{\deg F}(b_\ell t-a_\ell)$. Then $\tau^*F(u)=\prod_{\ell=1}^{\deg F}m(b'_\ell u-a'_\ell)$, so that

$$\begin{aligned}v(\tau^*F)&=(\deg F)v(m)+\sum_{\ell=1}^{\deg F}<(v(a'_\ell),v(b'_\ell))\\&=v(F)+(\deg F)v(m)+\sum_{\ell=1}^{\deg F}<(v(a'_\ell),v(b'_\ell))-\sum_{\ell=1}^{\deg F}<(v(a_\ell),v(b_\ell)).\end{aligned} \tag{8}$$

Multiplying (7) by $\deg F/(d-1)$ and subtracting it from the sum of (6) and (8) gives (5). □

5.3 Geometry of v-adic integrals

Let D be a horizontal divisor on X with support away from $[1:0]$. Let $[a_1:b_1],\dots,[a_{\deg D}:b_{\deg D}]$ denote the points in $X(\overline{K})$ corresponding to D.

Let F be a polynomial in the inhomogeneous variable $t=T_0/T_1$ obtained by taking a global section of $\mathcal{O}_{\mathbb{P}^1}(n)$ corresponding to D on the generic fiber (which will be a homogeneous polynomial of degree n) and dividing through by T_1^n. When D corresponds to a point $\alpha\in X(\overline{K})$, we will call such F a *minimal polynomial* for α over K.

Let $\sigma_k:X_k\longrightarrow X$ be the blow-up map associated to our model (P_k,Q_k) of φ^k described in Section 4. We will relate the v-adic integral of F with the geometry of σ_k^*D for D the horizontal divisor on X corresponding to F. We write

$$\sigma_k^*D=D_k+\sum_v\sum_i x_{v,i,k}C_{v,i,k},$$

where $C_{v,i,k}$ are components of the exceptional fiber of σ_k and D_k is the horizontal divisor corresponding to D on X_k. We let $k_{v,i,k}$ denote the field of definition of the closed point on X_k corresponding to the component $C_{v,i,k}$.

We begin by treating the case where D corresponds to a single point defined over K. Let $v(P_K)$ and $v(Q_k)$ be the minimum of the v-adic valuations of the coefficients of P_K and Q_k respectively.

Lemma 5.6. *If D corresponds to a point $[a : b] \in \mathbb{P}^1(K)$, then, for any nonarchimedean v, we have*

$$\sum_i x_{v,i,k} = S_{v,k}(P_k(a,b), Q_k(a,b)) - <(v(P_k), v(Q_k)).$$

Proof. We will work locally at a single nonarchimedean place v. Let $\sigma_{\mathcal{O}_v,k} : \left(X_k\right)_{\mathcal{O}_v} \longrightarrow X_{\mathcal{O}_v}$ denote σ_k with its base extended to Spec$\mathcal{O}_v$. Since D corresponds to a single rational point, there is at most one nonzero $x_{v,i,k}$ for a fixed v and k. Thus, letting D_v be the localization of D at v, we have $\sigma^*_{\mathcal{O}_v,k}(D_v) = D_{v,k} + f_{v,k}C_{v,k}$ for some horizontal divisor $D_{v,k}$, some nonnegative integer $f_{v,k}$, and some component $C_{v,k}$ of the exceptional fiber of $\sigma_{\mathcal{O}_v,k}$.

We may assume that $<(v(a), v(b)) = 0$. Let I denote the ideal sheaf of D_v in $X_{\mathcal{O}_v}$; then $\sigma^*_{\mathcal{O}_v,k}I$ will be the ideal sheaf for a subscheme of $\left(X_k\right)_{\mathcal{O}_v}$ corresponding to $D_{v,k} + f_{v,k}C_{v,k}$. Note that $\sigma^*_{\mathcal{O}_v,k}(I \otimes \mathcal{O}_X(1))$ is generated by $(bT_0 - aT_1)$. Let U be an open subset of $X_{k,\mathcal{O}_v}$ containing Supp$(D_{v,k} + C_{v,k})$. We may choose U to be the chart of $\left(X_k\right)_{\mathcal{O}_v}$ on which T_j doesn't vanish, where $j = 0$ if $v(a) = 0$ and $j = 1$ otherwise. Let i be the choice of $\{0, 1\}$ that is not equal to j. Let π be a generator for the maximal ideal in $\mathcal{O}_v$, let $\kappa = <(v(P_k), v(Q_k))$, and let $G_k = P_k/\pi^\kappa$ and $H_k = Q_k/\pi^\kappa$. Then U is isomorphic to

$$\text{Proj}\left((\mathcal{O}_v[T_i/T_j])\,[t,u]/(t\frac{G_k(T_0,T_1)}{T_j^{d^k}} - u\frac{H_k(T_0,T_1)}{T_j^{d^k}})\right).$$

Since $\mathcal{O}_v[T_i/T_j]/(b\,(T_0/T_j) - a\,(T_1/T_j)) \cong \mathcal{O}_v$, the subscheme of $\left(X_k\right)_{\mathcal{O}_v}$ determined by the vanishing of $\sigma^*_{\mathcal{O}_v,k}I$ will be isomorphic to

$$\begin{aligned}\text{Proj}(\mathcal{O}_v[t,u]/(tP_k(a,b) - uQ_k(a,b))) &\cong \text{Proj}(\mathcal{O}_v[t,u]/(\pi^r(\ell t - mu))) \\ &\cong \text{Proj}((\mathcal{O}_v/\pi^r)[t,u]) \cup \text{Spec}(\mathcal{O}_v),\end{aligned}$$

where $r = S_{v,k}(P_k(a,b), Q_k(a,b)) - \kappa$, $G_k(a,b) = \ell\pi^r$, and $H_k(a,b) = m\pi^r$. Since the divisor corresponding to $\sigma^*_{\mathcal{O}_v,k}I$ is $D_{v,k} + f_{v,k}C_{v,i,k}$ where $D_{v,k}$ is horizontal and $C_{v,i,k}$ is reduced, this means that we must have

$$S_{v,k}(P_k(a,b), Q_k(a,b)) - <(v(P_k), v(Q(P_k))) = r = f_{v,k} = \sum_i x_{v,i,k},$$

as desired. □

Since blowing up commutes with base extension from $\mathcal{O}_K$ to the ring of integers in a number field over which D splits into points, Lemma 5.6 generalizes easily to the following lemma.

Lemma 5.7. *We have*

$$\sum_i x_{v,i,k}[k_{v,i,k} : k] = \sum_{\ell=1}^{\deg F} S_{v,k}(P_k(a_\ell, b_\ell), Q_k(a_\ell, b_\ell)) - (\deg F)(<(v(P_k, Q_k))).$$

Now we show how to control the contribution of $<(v(P_k), v(Q_k))$.

Lemma 5.8. *The limit* $\lim_{k\to\infty} \frac{<(v(P_k),v(Q_k))}{d^k}$ *exists.*

Proof. We will show that the sequence $\left(\frac{<(v(P_k),v(Q_k))}{d^k}\right)$ is bounded and increasing. Since $<(v(P_k), v(Q_k)) \le R_v(P_k, Q_k)$, boundedness follows immediately from Lemma 5.4. Now, for any $k \ge 1$, we have

$$<(v(P_{k+1}), v(Q_{k+1})) \ge d<(v(P_k), v(Q_k)))$$

since $P_{k+1} = P(P_k, Q_k)$ and $Q_{k+1} = Q(P_k, Q_k)$ so the sequence is increasing. □

Let ∞ denote the horizontal divisor on X corresponding to the point $[1 : 0]$ and ∞_k the horizontal divisor on X_k corresponding to this point.

Lemma 5.9. *We have* $y_{v,i,k} = v(P_k(1,0)) - <(v(P_k), v(Q_k))$, *where*

$$\sigma_k^* \infty = \infty_k + \sum_v \sum_i y_{v,i,k} C_{v,i,k}.$$

Thus,

$$\lim_{k\to\infty} \frac{y_{v,i,k}}{d^k} = \frac{v(A_d)}{d-1} - \lim_{k\to\infty} \frac{<(v(P_k), v(Q_k))}{d^k}.$$

Proof. Since T_1 is a factor of Q, it is also a factor of Q_k for every k. Thus, for each k we have $S_{v,k}(P_k(1,0), Q_k(1,0)) = v(P_k(1,0))$, so by Lemma 5.6, we have $y_{v,i,k} = v(P_k(1,0)) - <(v(P_k), v(Q_k))$. Since $P(1,0) = A_d$ and for all $k \ge 1$ we have

$$P_{k+1}(1,0) = P(P_k(1,0), 0) = A_d P_k(1,0)^d.$$

As in the proof of Lemma 5.4, it follows that

$$\lim_{k\to\infty} \frac{v(P_k(1,0))}{d^k} = v(A_d)\left(\sum_{i=1}^{\infty} \frac{1}{d^i}\right) = \frac{v(A_d)}{d-1}.$$

□

The following is now an immediate consequence of Definition 5.2 and Lemmas 5.7, 5.8, and 5.9.

Proposition 5.10. *With notation as above, we have*

$$\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} = -\lim_{k\to\infty} \frac{x_{v,i,k}[k_{v,i,k} : k_v]}{d^k} \log \mathrm{N}(v) + (\deg F) \lim_{k\to\infty} \frac{y_{v,i,k}}{d^k} \log \mathrm{N}(v) - v(F) \log \mathrm{N}(v). \tag{9}$$

5.4 A remark on the use of integral notation at finite places

We now explain our interpretation of $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v \; d\mu_{v,\varphi}$ as an integral, when v is finite. H. Brolin [6] and M. Lyubich [16] have shown that if v is an infinite place and θ is a continuous, bounded function on $\mathbb{P}^1(\mathbb{C}_v)$, then for any $\xi \in \mathbb{C}$ with an infinite backward orbit under φ (i. e., for which the set $\cup_{k=1}^{\infty}(\varphi^k)^{-1}(\xi)$ is infinite), one has

$$\lim_{k\to\infty} \sum_{\varphi^(z)=\xi} \frac{\theta(z)}{d^k} = \int_{\mathbb{P}^1(\mathbb{C}_v)} \theta \; d\mu_{v,\varphi},$$

where $d\mu_{v,\varphi}$ is the unique φ-invariant measure supported on the Julia set of φ (see [11]) on $\mathbb{P}^1$ (coinciding with $d\mu_{v,\varphi}$, as shown in Proposition 7.2).

Our v-adic integrals can be written in a similar way. For example, suppose D is an irreducible divisor corresponding to a single point $[a : b]$ such that $P_k(a, b) \neq 0$. Then the polynomial $F(t) = bt - a$ defines D. Writing $P_k = \eta_k \prod_{j=1}^{d^k}(T_0 - u_j T_1)$, we have

$$P_k(a, b) = \eta_k \prod_{j=1}^{d^k}(b - u_j a) = \eta_k \prod_{j=1}^{d^k} F(u_j).$$

Since $\varphi^k(z) = 0$ if and only if $[z : 1] \sim [u_j : 1]$ for some j, we have

$$\frac{\log|P_k(a,b)|_v}{d^k} = \frac{\log|\eta_k|_v}{d^k} + \sum_{\varphi^k(z)=0} \frac{\log|F(z)|_v}{d^k},$$

where the z with $\varphi^k(z) = 0$ are counted with multiplicities. Similarly, when Q is not a multiple of X_1^d, we have

$$\frac{\log|Q_k(a,b)|_v}{d^k} = \frac{\log|\gamma_k|_v}{d^k} + \sum_{\varphi^k(z)=\infty} \frac{\log|F(z)|_v}{d^k},$$

where γ_k is the leading coefficient of $Q_k(T_0, 1)$. Taking limits and subtracting off $\frac{\log|A_d|_v}{d-1}$, we see that $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$ is equal to

$$\lim_{k\to\infty} \max\left(\sum_{\varphi^k(z)=0} \frac{\log|F(z)|_v}{d^k}, \sum_{\varphi^k(z)=\infty} \frac{\log|F(z)|_v}{d^k} \right).$$

With a bit of diophantine geometry, we can show that for any point $\xi \in \mathbb{C}_v$ that has an infinite backwards orbit under φ, we have

$$\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} = \lim_{k\to\infty} \sum_{\substack{\varphi^k(z)=\xi \\ F(z)\neq 0}} \frac{\log|F(z)|_v}{d^k}.$$

We will prove this in a future paper [27]. Thus, our v-adic integrals at finite places seem quite analogous to integrals at the infinite places. However, we do not know what classes of functions can be expected to be "integrable" in this way at finite places.

6 The Mahler formula for dynamical systems

The formula of Mahler for the naive height of a closed point $\alpha \neq \infty$,

$$\deg(F)h(\alpha) = \int_0^1 \log|F(\exp(2i\pi\theta))|d\theta,$$

where F is the minimal equation for the algebraic point α of the projective line over $\mathbb{Q}$ will be generalized to a dynamical system φ and its canonical height. The original Mahler formula is associated to the dynamics of the map $\varphi(t) = t^2$.

In general, the formula involves some adelic terms with support at places of bad reduction. Recall our definition of $d\mu_{v,\varphi}$ for v an infinite place from Section 2 and our definition of $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$ for v a finite place from Section 5. We prove Theorem 6.1 using integral models for each iterate of φ. It may be possible to give an independent proof using adelic metrics. Recall our definition of $d\mu_{v,\varphi}$ for v an infinite place from Section 2 and our definition of $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}$ for v a finite place from Section 5. Our main theorem is the following:

Theorem 6.1. *Let K be a number field, φ a rational map from $\mathbb{P}^1_K$ to $\mathbb{P}^1_K$. Suppose φ has a K-rational fixed point and that coordinates have been chosen so that this point has coordinates $[1:0]$. For infinite v, (resp. finite v), let $d\mu_{v,\varphi}$ for v be defined as in Proposition 2.3 (resp. Definition 5.2). Then, given any point $\alpha \in \mathbb{P}^1(\bar{K})$ with $\alpha \neq \infty$ and any minimal polynomial F for α over K one has*

$$[K(\alpha):\mathbb{Q}]h_\varphi(\alpha) = \sum_{\text{places } v \text{ of } K} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}. \tag{10}$$

For finite v, we have $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} = 0$ unless φ has bad reduction at v or all the coefficients of F have nonzero v-adic valuation. In particular, $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} = 0$ for all but finitely many v.

Proof. We will compute all heights using the Arakelov intersection product between Weil divisors and metrized line bundles (defined in Appendices A and B).

We can think of F as a rational function on X; we will then have

$$\mathrm{div}(F) = D - (\mathrm{deg}F)\infty + \sum_{\text{finite } v} v(F)X_v,$$

where D is the horizontal divisor on X corresponding to α and X_v the fiber of X at v. We let $F_k = \sigma_k^* F$ and let $X_{k,v} = \sigma_k^* X_v$ denote the fiber of X_k at the finite place v. Let D_k be the horizontal divisor on X_k corresponding to D and let ∞_k be the horizontal divisor on X_k corresponding to ∞. We have

$$\begin{aligned}\mathrm{div}(F_k) = D_k &+ \sum_{\text{finite } v} \sum_i x_{v,i,k} C_{v,i,k} - (\mathrm{deg}F)\infty_k \\ &- (\mathrm{deg}F) \sum_{\text{finite } v} \sum_i y_{v,i,k} C_{v,i,k} + \sum_{\text{finite } v} v(F) X_{k,v},\end{aligned}$$

where $\sigma_k^* D = D_k + \sum_v \sum_i x_{v,i,k} C_{v,i,k}$ and $\sigma_k^* \infty = \infty_k + \sum_v \sum_i y_{v,i,k} C_{v,i,k}$, as in Section 5.

We let L be the line bundle $\mathcal{O}_X(1)$. At each infinite place v, let $\|\cdot\|_v$ be the metric on L such that for any section $s = u_0 T_0 + u_1 T_1$, of $\mathcal{O}_{\mathbb{P}^1}(1)$, we have $\|s([T_0 : T_1])\|_v = \left(|u_0 T_0 + u_1 T_1|_v^2 / (|T_0|_v^2 + |T_1|_v^2)\right)^{[K_v:\mathbb{R}]/2}$. This metric is smooth and semipositive (see [31, Section 6]). We denote as h_L the height function given by $[K(\alpha) : \mathbb{Q}] h_L(\alpha) = (E_\alpha . L)_{\mathrm{Ar}}$, where E_α is the horizontal divisor on X corresponding to α. Note that $h_L([1 : 0]) = 0$. We denote as L_k the line bundle $\varphi_k^* L$ endowed with the metric $\varphi_k^* \|\cdot\|_v$. Then $\frac{1}{(\mathrm{deg}\varphi)^k} \frac{-1}{2\pi i} d\overline{d} \log \varphi_k^* \|\cdot\|_v$ converges to a distribution $d\mu_{v,\varphi}$ as $k \to \infty$ by Proposition 2.3.

Let $\alpha \in X(\overline{K})$ be a point on the generic fiber corresponding to D and d the degree of φ. By the definition of L_k, we have $(D_k . L_k)_{\mathrm{Ar}} = (\mathrm{deg}F) h_L(\varphi^k(\alpha))$ and $(\infty_k . L_k)_{\mathrm{Ar}} = 0$. Using Proposition B.3 and the fact that $(C_{v,i,k} . L_k) = [k_{v,i,k} : k_v]$ (by Proposition 4.3) and $(X_{k,v} . L_k) = 0$ (by Proposition A.8), we thus obtain

$$
\begin{aligned}
d^k \sum_{v|\infty} \int_{X(\mathbb{C}_v)} \log|F|_v d\mu_{v,k} &= (\mathrm{div} F_k . L_k)_{\mathrm{Ar}} \\
&= -\left((\deg F)\infty_k . L_k\right)_{\mathrm{Ar}} + (D_k . L_k)_{\mathrm{Ar}} + \sum_{\text{finite } v} (x_{v,i,k} C_{v,i,k} . L_k) \log \mathrm{N}(v) \\
&\quad + \sum_{\text{finite } v} v(F)(X_{k,v} . L_k) \log \mathrm{N}(v) - (\deg F) \sum_{\text{finite } v} (y_{v,i,k} C_{v,i,k} . L_k) \log \mathrm{N}(v) \\
&= [K(\alpha) : \mathbb{Q}] h_L(\varphi^k(\alpha)) + \sum_{\text{finite } v} x_{v,i,k} [k_{v,i,k} : k_v] \log \mathrm{N}(v) \\
&\quad + d^k \sum_{\text{finite } v} v(F) \log \mathrm{N}(v) - (\deg F) \sum_{\text{finite } v} y_{v,i,k} [k_{v,i,k} : k_v] \log \mathrm{N}(v).
\end{aligned}
$$

Now, we divide through by d^k and take limits. Since

$$
\lim_{k\to\infty} \sum_{v|\infty} \int_{X(\mathbb{C}_v)} \log|F|_v d\mu_{v,k} = \sum_{v|\infty} \int_{X(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}
$$

and $\lim_{k\to\infty} h_L(\varphi^k(\alpha))/d^k = h_\varphi(\alpha)$, applying Proposition 5.10 yields

$$
\begin{aligned}
[K(\alpha) : \mathbb{Q}] h_\varphi(\alpha) &= \sum_{v|\infty} \int_{X(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} - \sum_{\text{finite } v} v(F) \log \mathrm{N}(v) \\
&\quad - \lim_{k\to\infty} \frac{1}{d^k} \sum_{\text{finite } v} x_{v,i,k} [k_{v,i,k} : k_v] \log \mathrm{N}(v) + (\deg F) \lim_{k\to\infty} \frac{y_{v,i,k}}{d^k} \log \mathrm{N}(v) \\
&= \sum_{v|\infty} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} + \sum_{\text{finite } v} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi}.
\end{aligned}
$$

It follows from Definition 5.2 that if v is finite, then $\int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} = 0$ unless φ has bad reduction at v or all coefficients of F have nonzero v-adic valuation. □

Remark 6.2. Letting $\varphi = P/Q$ be a model for φ and applying Proposition 5.1, we have the bound

$$
\begin{aligned}
-(\deg F)\Big(\sum_{\text{finite } v} \frac{R_v(P,Q) \log \mathrm{N}(v)}{d-1} \Big) - \sum_{\text{finite } v} v(F) \log \mathrm{N}(v) & \\
\le \sum_{\text{finite } v} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log|F|_v d\mu_{v,\varphi} & \\
\le (\deg F) \sum_{\text{finite } v} \frac{v(A_d)}{d-1} \log \mathrm{N}(v) - \sum_{\text{finite } v} v(F) \log \mathrm{N}(v), &
\end{aligned}
$$

where A_d is the coefficient of the T_0^d term of P, $v(F)$ is the minimum of the v-adic valuations of the coefficients of F, and $R_v(P,Q)$ is the supremum of $< (v(P(a,b)), v(Q(a,b)))$ over all v-adic integers a and b in $\mathcal{O}_{\overline{K}}$ with $(a,b) \neq (0,0)$.

Corollary 6.3. *Suppose that φ can be written as $[P : Q]$ where $P(T_0, 1)$ is monic in T_0 and that α has a minimal polynomial F over K with coprime coefficients in $\mathcal{O}_K$ (as is always the case when $K = \mathbb{Q}$, for example). Then*

$$[K(\alpha) : \mathbb{Q}]h_\varphi(\alpha) \leq \sum_{v|\infty} \int_{X(\mathbb{C})} \log |F|_v d\mu_{v,\varphi}, \tag{11}$$

with equality if φ has good reduction everywhere.

Proof. Since $R_v(a, b) \geq 0$ for any $a, b \in \overline{K}$, as we noted in Section 5, we see that

$$\int_{\mathbb{P}^1(\mathbb{C}_v)} \log |F|_v d\mu_{v,\varphi} \leq 0$$

for any finite v when F has coprime coefficients in $\mathcal{O}_K$ and P is monic. The corollary then follows immediately from Theorem 6.1. □

Example 6.4. In general, one cannot expect equality in (11). Suppose that $P(T_0, 1)$ is monic, as in Corollary 6.3. Suppose furthermore that $\alpha \in \overline{K}$ has a minimal polynomial F with coprime coefficients in $\mathcal{O}_K$ and that $Q(\alpha, 1) = 0$. Then, we have $\int_{X(\mathbb{C})} \log |F|_v d\mu_{v,\varphi} = -\frac{v(P(\alpha,1))}{d} \log \mathrm{N}(v)$ for any finite place v. Thus, if $v(P(\alpha, 1)) > 0$ for some finite place v, then inequality (11) is strict.

Example 6.5. Let E be an elliptic curve with Weierstrass equation $y^2 = G(t)$. As in Example 2 of Section 3, multiplication by 2 on E gives rise to the rational map

$$\varphi(t) = \frac{G'^2(t) - 8tG(t)}{4G(t)}$$

on $\mathbb{P}^1$. Suppose that $G(t) = (t-a)(t-b)(t-c)$, for $a, b, c \in \mathbb{Q}$. Since a, b, and c are preperiodic points of φ, it follows from Theorem 6.1 that

$$\sum_{\text{places } v \text{ of } \mathbb{Q}} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log |G|_v d\mu_{v,\varphi} = h_\varphi(a) + h_\varphi(b) + h_\varphi(c) = 0. \tag{12}$$

Computing as in Example 6.4 and using the product formula, we obtain

$$\begin{aligned} \sum_{\text{finite } v} \int_{\mathbb{P}^1(\mathbb{C}_v)} \log |G|_v d\mu_{v,\varphi} &= \sum_{\text{finite } v} -\frac{2\, v(G'(a)G'(b)G'(c))}{4} \log \mathrm{N}(v) \\ &= \frac{-\log\left(|G'(a)||G'(b)||G'(c)|\right)}{2}, \end{aligned} \tag{13}$$

where $|\cdot|$ denotes the archimedean absolute value on $\mathbb{Q}$. It is well-known that $|G'(a)||G'(b)||G'(c)|$ is equal to the absolute value of the discriminant $\mathrm{Disc}(G)$ of G (see [15, Proposition 4.8.5]). Thus, combining equations (12) and (13) yields

$$\int_{\mathbb{P}^1(\mathbb{C})} \log |G| d\mu_{\infty,\varphi} = \frac{\log |\mathrm{Disc}(G)|}{2},$$

where ∞ is the archimedean place on $\mathbb{Q}$.

7 Equidistribution and Julia sets

Let v be an archimedean place. As we have defined it, $d\mu_{v,\varphi}$ is only a distribution; that is to say, the integral $\int_{\mathbb{P}^1(\mathbb{C})} f\, d\mu_{v,\varphi}$ is only defined for smooth functions f. We will now show that $d\mu_{v,\varphi}$ extends to a linear functional on the space of continuous functions on $\mathbb{P}^1(\mathbb{C})$ and that this linear functional is the unique φ-invariant probability measure on $\mathbb{P}^1(\mathbb{C})$ with support on the Julia set of φ.

Let $\|\cdot\|_v$ be a metric on $\mathcal{O}_{\mathbb{P}^1}(1)$. Recall the definitions of $\|\cdot\|_{v,k}$ and $d\mu_{v,k}$ from Proposition 2.3.

Following M. Lyubich [16], we define A to be the operator on the space of continuous functions of $\mathbb{P}^1(\mathbb{C})$ which sends a continuous function f on $\mathbb{P}^1(\mathbb{C})$ to

$$(Af)(z) := \frac{1}{d} \sum_{\varphi(w)=z} e_w f(w),$$

where $z \in \mathbb{P}^1(\mathbb{C})$ and e_w is the ramification index of φ at w.

Lemma 7.1. *Let U be an open set in $\mathbb{P}^1(\mathbb{C})$. Then for any continuous function f on $\mathbb{P}^1(\mathbb{C})$, we have*

$$\int_{\varphi^{-1}(U)} f\, d\mu_{v,k+1} = \int_U Af\, d\mu_{v,k} \tag{14}$$

for any $k \geq 1$.

Proof. Since the set of ramification points of φ is finite, it suffices to show that (14) holds when U contains no ramification points; since any open subset can be written as a union of simply connected open subsets, we may further assume that U is simply connected. We may then decompose $\varphi^{-1}(U)$ into d branches V_λ, $1 \leq \lambda \leq d$ such that φ is bijective on each V_λ with analytic inverse φ_λ^{-1}.

Choose a section s of $\mathcal{O}_{\mathbb{P}^1}(1)$ that does not vanish on U or $\varphi^{-1}(U)$. Let $\rho = \frac{\log \|s\|_{v,k}}{2\pi i}$. Then, on U, we have $d\overline{d}\rho = d\mu_{v,k}$ and on V_λ, we have $d\overline{d}(\rho\circ\varphi) = (\deg\varphi)(d\mu_{k+1,v})$. By change of variables, we then have

$$\begin{aligned}
\int_{V_\lambda} f\, d\mu_{v,k+1} &= \frac{1}{d} \int_{V_\lambda} f(z)\, d\overline{d}(\rho(\varphi(z))) \\
&= \frac{1}{d} \int_U f(\varphi_\lambda^{-1}(u))\, d\overline{d}(\rho(u)) \\
&= \frac{1}{d} \int_U f \circ \varphi_\lambda^{-1}\, d\mu_{v,k}.
\end{aligned}$$

Since $(Af)(u) = \frac{1}{d}\sum_{\lambda=1}^{d} f \circ \varphi_\lambda^{-1}(u)$, we thus obtain

$$\int_{\varphi^{-1}(U)} f\,d\mu_{v,k+1} = \sum_{\lambda=1}^{d} \int_{V_\lambda} f\,d\mu_{v,k+1} = \frac{1}{d}\sum_{\lambda=1}^{d} \int_U f\circ\varphi_\lambda^{-1}\,d\mu_{v,k}$$
$$= \int_U \frac{1}{d}\sum_{\lambda=1}^{d} f\circ\varphi_\lambda^{-1}\,d\mu_{v,k} = \int_U Af\,d\mu_{v,k}.$$

□

An *exceptional point* ξ for φ is a point such that $\varphi^2(\xi)=\xi$ and φ^2 ramifies completely at ξ. An exceptional point ξ is a super-attracting fixed point for φ^2 (see A. F. Beardon [4, 6.3]).

Proposition 7.2. *The measures $d\mu_{v,k}$ converge to a measure $d\mu_{v,\varphi}$ that is supported on the Julia set. Furthermore, $d\mu_{v,\varphi}$ is the unique probability measure supported on the Julia set with the property that*

$$\int_{\varphi(U)} d\mu_{v,\varphi} = \int_U d\mu_{v,\varphi}$$

for any open subset $U\subset\mathbb{P}^1(\mathbb{C})$ such that φ is injective on U.

Proof. Let f be a continuous function on $\mathbb{P}^1(\mathbb{C})$ and let $\epsilon>0$. We may choose an open set U_ϵ containing the exceptional points of φ for which

$$\int_{U_\epsilon} d\mu_{v,1} \le \epsilon/2 \sup_{z\in\mathbb{P}^1(\mathbb{C})}(|f(z)|_v).$$

Such a set exists since $d\mu_{v,1}$ is a continuous form. Let $W_\epsilon = \mathbb{P}^1(\mathbb{C})\smallsetminus U_\epsilon$. By Theorem 1 of [16], there is a constant C_f such that $(A^k f)(w)$ converges uniformly to C_f for $w\in W_\epsilon$. Thus, there is some M such that for any $k\ge M$, we have

$$|(A^k f)(w) - C_f|_v < \epsilon/2$$

for all $w\in W_\epsilon$. Using Lemma 7.1 and the fact that $\int_{W_\epsilon} d\mu_{v,1}\le 1$, we then see that for all $k\ge M$ we have

$$\left|\int_{\mathbb{P}^1(\mathbb{C})} f\,d\mu_{v,k} - C_f\right|_v = \left|\int_{\mathbb{P}^1(\mathbb{C})} (A^k f)\,d\mu_{v,1} - C_f\right|_v$$
$$\le \left|\int_{W_\epsilon} (A^k f)\,d\mu_{v,1} - C_f\right|_v + \int_{U_\epsilon} |A^k f|_v\,d\mu_{v,1}$$
$$\le \int_{W_\epsilon} \frac{\epsilon}{2}\,d\mu_{v,1} + \int_{U_\epsilon} \Big(\sup_{z\in\mathbb{P}^1(\mathbb{C})}(|f(z)|_v)\Big)\,d\mu_{v,1}$$
$$\le \epsilon/2 + \epsilon/2 = \epsilon.$$

Thus, $d\mu_{v,\varphi}$ extends to a measure such that

$$\int_{\mathbb{P}^1(\mathbb{C})} f\, d\mu_{v,\varphi} = \lim_{k\to\infty} (A^k f)(z),$$

where z is any point in V. Freire, Lopes, and Mañe [11] have shown that the map sending a continuous function f to $\lim_{k\to\infty}(A^k f)(z)$, where z is not an exceptional point of φ, is the unique φ-invariant probability measure on $\mathbb{P}^1(\mathbb{C})$ that is supported on the Julia set of φ. □

P. Autissier [2, Proposition 4.7.1] has proved the following equidistribution theorem.

Theorem 7.3. *With the same hypothesis as Theorem 6.1, for any infinite v and nonrepeating sequence of points (α_n) in $\mathbb{P}^1(\overline{K})$ with $\lim_{n\to\infty} h_\varphi(\alpha_n) = 0$, the sequence $\frac{1}{|\mathrm{Gal}(\alpha_n)|} \sum_{\sigma\in \mathrm{Gal}(\alpha_n)} \delta_{\sigma(\alpha_n)}$ of discrete measures converges weakly to $d\mu_\varphi$.*

M. Baker and R. Rumely [3] have given a new proof of Theorem 7.3, using capacity theory. Their proof also gives equidistribution results at finite places.

One might also ask whether the Mahler measure of this paper should also be computable by equidistribution; more precisely, we conjecture:

Conjecture 7.4. *With the same hypothesis as Theorem 7.3, for F the minimal equation of point α not in the Galois orbit of any α_n and v an infinite place of K, one has*

$$\lim_{n\to\infty} \frac{1}{|\mathrm{Gal}(\alpha_n)|} \sum_{\sigma\in \mathrm{Gal}(\alpha_n)} \log |F(\sigma(\alpha_n))|_v = \int_{X(\mathbb{C}_v)} \log |F|_v d\mu_{v,\varphi}.$$

We can prove Conjecture 7.4 in the case that the points α are periodic points [27]. This generalizes earlier work on "elliptic Mahler measure" by G. Everest and T. Ward in [10, Theorem 6.18].

A Schematic Intersection theory on a Cohen–Macaulay surface

The blow-up surfaces constructed in Section 4 may not be regular. In this appendix we justify our use of intersection theory on these surfaces. Intersection theory has been developed by many authors. This appendix uses material from P. Deligne [9] and W. Fulton [12, Chapter 20].

For us a surface is a noetherian, irreducible, and reduced scheme X of dimension 2. In this article, moreover, X will be Cohen–Macaulay and will be equipped with a flat and often projective, generically smooth, structural map $f : X \to \mathrm{Spec}(B)$, where B is a Dedekind domain with field of fractions K. Often K will be a number field or the field of fractions of a discrete

valuation ring V. The geometric (or schematic) intersection numbers will be rational numbers (the need for denominators for an intersection theory on a singular scheme was noted previously in D. Mumford [22] and in C. Peskine and L. Szpiro [23]).

Definition A.1. *We define the schematic (or geometric) intersection number of a Cartier divisor D with a Weil cycle C when they have no common components as*

$$(D.C) = \text{length}(\mathcal{O}_D \otimes \mathcal{O}_C) - \text{length}(\text{Tor}_1(\mathcal{O}_D, \mathcal{O}_C)).$$

Lemma A.2. *If C is of codimension 2, then $(D.C) = 0$.*

Proof. If f is locally the equation of D and I is the ideal of C in a local ring A of X, one has an exact sequence

$$0 \to \text{Tor}_1(A/I, A/f) \to A/I \xrightarrow{f} A/I \to A/(I + (f)) \to 0.$$

The lemma follows since the four modules in this sequence are of finite length. □

Proposition A.3. ***(bilinearity and symmetry)*** *The pairing $(D.C)$ is bilinear and symmetric when both sides are Cartier divisors.*

Proof. Symmetry follows immediately from Definition A.1 when C and D are both Cartier divisors. To prove linearity, we note that if C is a Weil divisor, it is a linear combination with integral coefficients of reduced and irreducible Weil divisors C_i (the coefficient of C_i is equal to $\text{length}((\mathcal{O}_C)_{\wp_i})$). Since the sheaf $\mathcal{O}_D$ has Tor dimension 1, the pairing is linear on the right by devissage. It suffices to show linearity on the left when C is reduced and irreducible. This follows from the following simple lemma, the proof of which we leave to the reader:

Lemma A.4. *Let A be a commutative ring, I be an ideal in A, and f be an element of A that is not a zero-divisor. Then one has the following exact sequence:*

$$0 \to A/I \to A/fI \to A/fA \to 0.$$

□

If C' is a $\mathbb{Q}$-Cartier divisor, i.e., a Weil divisor with an integral multiple nC' which is Cartier, we will define

$$(C'.C) = \frac{1}{n}((nC').C).$$

In this article we use only intersections between two $\mathbb{Q}$-Cartier divisors. So, up to an integral multiple, the intersection number is locally a finite sum of $\text{length}(A/(f, g))$, where A is a local ring of dimension 2 and depth 2 and (f, g) is a regular sequence. The following propositions are classical:

Proposition A.5. ***(linear equivalence)*** *If C is a Cohen–Macaulay projective curve, then*

$$(D.C) = \deg_C \mathcal{O}_X(D)_{|C}. \tag{15}$$

Proof. Note that if C is a projective curve and L is a line bundle on X a projective surface, one can then speak of $(L.C)$, since L is the difference in $\mathrm{Pic}(X)$ between two very ample line bundles each of them having sections with no common components with C. Equation (15) then follows immediately from the fact that any line bundle on a Cohen–Macaulay curve has a well-defined degree. □

Corollary A.6. *When C is a projective curve and D is a $\mathbb{Q}$-Cartier divisor, the intersection $(D.C)$ is well-defined by bilinearity and linear equivalence even when D and C have a common component.*

Corollary A.7. *Let F be a rational function on a reduced, irreducible surface X that is projective and generically smooth over B. Then for any Weil divisor C contained in a fiber over B, we have*

$$(\mathrm{div}(F).C) = 0.$$

Proof. This is clear because the line bundle $\mathcal{O}_X(\mathrm{div}(F))$ is isomorphic to $\mathcal{O}_X$ and C is a projective curve. □

Proposition A.8. ***(projection formula)*** *Let $\varphi : Y \to X$ be a map between surfaces X and Y that are projective over B. If L is a line bundle on X and C a closed subscheme of Y, one has*

$$(\varphi^*(L).C) = (L.\varphi_*(C)).$$

In particular if C is contracted by φ to a subscheme of X of codimension 2, the intersection number $(\varphi^*(L).C)$ is zero.

Proof. By additivity we can suppose C is a reduced irreducible curve in X. There are two cases: $\varphi_*(C)$ is of dimension 1 and $\varphi_*(C)$ is of dimension zero. In the first case $C \to \varphi_*(C)$ is finite and by Lemma A.4 we have

$$\begin{aligned}\mathrm{length}(\mathcal{O}_C/(f\mathcal{O}_C)) &= \mathrm{length}(\mathcal{O}_X/(f\mathcal{O}_X) \otimes \mathcal{O}_C) \\ &= \mathrm{length}(\mathcal{O}_Y/(f\mathcal{O}_Y) \otimes \mathcal{O}_C).\end{aligned}$$

In the second case L can be realized as the line bundle associated to the difference of two very ample divisors on X each of them having no intersection with $\varphi_*(C)$. The reciprocal images of these divisors in Y do not meet C, so both sides of the projection formula vanish as desired. □

The following proposition shows that intersection theory for $\mathbb{Q}$-Cartier divisors does not change when we pass to the normalization.

Proposition A.9. ***(invariance under normalization)*** *Let A be a Cohen–Macaulay integral domain of dimension 2 and let $\tilde{A}$ be its integral closure. Suppose that $\tilde{A}$ is a finitely generated A-module. Let (f, g) be a regular sequence in A. Then (f, g) is a regular sequence in $\tilde{A}$ and*

$$\text{length}(A/(f, g)) = \text{length}(\tilde{A}/(f, g)).$$

Proof. Note that $\tilde{A}$ is finitely generated over A when we are in a geometric or arithmetic situation. When (f, g) is a regular sequence in a module M the only non-vanishing Tor is the tensor product $M \otimes A/(f, g)$. One proves easily by induction that the alternating sum $\sum_i (-1)^i \text{length}(\text{Tor}_i(A/(f, g), M))$ is a non-negative additive function on the set of finitely generated A-modules. This sum is equal to zero on modules of the form A/h or $A/(h, k)$ for a system of parameters (h, k) in A and thus is zero on $A/\wp$ for any prime ideal $\wp$ containing h or (h, k) by devissage. Hence, it is equal to zero on any module of dimension less than or equal to 1. Since the module $\tilde{A}/A$ has dimension less than or equal to 1 and (f, g) remains a regular sequence in $\tilde{A}$, the proposition follows by additivity. □

The following proposition is proved in P. Deligne [9] with the additional assumption that X is normal:

Proposition A.10. *Let X be a generically smooth, reduced, irreducible and locally Cohen–Macaulay surface. The geometric intersection product on a fiber of $X \to \text{Spec}(B)$ when X is projective over B, is non-positive. Moreover, only combinations of full fibers have zero self-intersection.*

By Proposition A.9, the assumption that X is normal may be replaced with the weaker assumption that X is Cohen–Macaulay.

Thus, Proposition A.10 applies to the surfaces used in this paper.

B Arakelov intersection with the divisor associated to a rational function

Let X be as in Appendix A; in this section we specify that the Dedekind domain B we work with is the ring of integers in a number field K. Let F be a meromorphic function on X. We begin by working with "arithmetico-geometric" intersections; that is to say, if we intersect a Cartier divisor D and a Weil cycle C without common components, their arithmetico-geometric intersection $(D.C)_{\text{fin}}$ is taken to be

$$(D.C)_{\text{fin}} = \sum_{\text{finite places } v \text{ of } B} (D_v.C_v) \log \text{N}(v),$$

where $(\cdot.\cdot)$ is the intersection product from Appendix A, the D_v and C_v are the pull-backs of D and C back to the fiber X_v, and $\text{N}(v)$ is the cardinality of

the residue field k_v of our base ring B at v. These intersections $(D.C)_{\text{fin}}$ are thus sums of the geometric intersections of Appendix A "weighted" by the logarithm of the size of the residue fields k_v.

In practice, we will be computing our intersections between rational functions and reduced irreducible horizontal divisors $\text{Spec}(R)$. Thus, for R an order in a number field K and f a nonzero element of the field of fractions of R, we define $\#(R/f)$ as follows: Let S denote the primes $\mathcal{P}$ in R for which $f \in R_{\mathcal{P}}$ and let T denote the primes $\mathcal{Q}$ in R for which $f^{-1} \in R_{\mathcal{Q}}$. We let

$$\#(R/fR) = \frac{\prod_{\mathcal{P}\in S} \#(R_{\mathcal{P}}/fR_{\mathcal{P}})}{\prod_{\mathcal{Q}\in T} \#(R_{\mathcal{Q}}/f^{-1}R_{\mathcal{Q}})}.$$

Lemma B.1. *Let E_β be an irreducible horizontal divisor on X corresponding to the Galois orbit of the point $\beta \in X(\bar{\mathbb{Q}})$. Then*

$$(\text{div}F.E_\beta)_{\text{fin}} = \sum_{v|\infty} \sum_{i=1}^{\deg\beta} \text{N}_v \log |F(\beta_v^{[i]})|_v,$$

where $\beta_v^{[i]}$ are the conjugates of β in $X(\mathbb{C}_v)$, the primes $v \mid \infty$ are the set of infinite places of K, each extended to an infinite place on the field of fractions of R, and N_v is the local degree $[K_v : \mathbb{R}]$.

Proof. The divisor E_β determines a closed immersion $i_\beta : \text{Spec}R \longrightarrow X$ for an order R, so F pulls back to an element $i_\beta^* F$ of the field of fractions of R. By our definition of arithmetic intersection, we have $(\text{div}F.E_\beta)_{\text{fin}} = \log \#(R/i_\beta^* F)$. Since $\#(R/i_\beta^* F) = \text{Norm}_{K/\mathbb{Q}}(i_\beta^* F)$ (see [25, Section III.4], for example), the product formula over $\mathbb{Q}$ and the definition of the norm gives

$$\log \#(R/i_\beta^* F) - \sum_{v|\infty} \sum_{i=1}^{\deg\beta} \text{N}_v \log |F(\beta_v^{[i]})|_v = 0.$$

□

Now, let L be a line bundle on X endowed with a smooth metric $\|\cdot\|_v$ at each infinite place v of K. We will not require the metric on L to have any special properties, since we will not need the sort of adjunction formula that is used, for example, in P. Vojta [29] or S. Lang [14, Chapter IV]. Let s be a section of L such that $\text{div}(s)$ and $\text{div}(F)$ have no common horizontal components. Let v be an infinite place of K and let F_v and s_v denote the pullbacks of F and s to $X_{\mathbb{C}_v}$. We write $\text{div}(F_v) = \sum m_\alpha \alpha$ and $\text{div}(s_v) = \sum n_\beta \beta$ where all of the α and β are in $X(\mathbb{C}_v)$.

Proposition B.2. *With notation as above, we have*

$$\frac{1}{2\pi i} \int_{X(\mathbb{C}_v)} \log |F|_v \, d\overline{d} \log \|s\|_v = \sum_\alpha m_\alpha \log \|s(\alpha)\|_v - \sum_\beta \text{N}_v n_\beta \log |F(\beta)|_v.$$

Proof. We follow the proof of S. Lang [14, Lemma 2.1.1, pp. 22–23]. Since $\log|F|_v$ is harmonic away from the support of $\mathrm{div}(F_v)$, we have $d\overline{d}\log|F|_v = 0$ away from the points α such that $m_\alpha \neq 0$; thus

$$\log|F|_v d\overline{d}\log\|s\|_v = \log|F|_v\, d\overline{d}\log\|s\|_v - \log\|s\|_v d\overline{d}\log|F|_v$$

away from the α and β. Since $d\eta\wedge\overline{d}\gamma = d\gamma\wedge\overline{d}\eta$ for any smooth functions η and γ (see [14, p. 22]), we therefore have

$$\log|F|_v d\overline{d}\log\|s\|_v - \log\|s\|_v d\overline{d}\log|F|_v = d(\log|F|_v\overline{d}\|s\|_v - \log\|s\|_v\overline{d}\log|F|_v)$$

away from the α and β.

Let $Y(a)$ be the complement of the circles $C(\alpha,a)$ and $C(\beta,a)$ of radius a around all of the α and β and let

$$\omega = \log|F|_v\,\overline{d}\log\|s\|_v - \log\|s\|_v\,\overline{d}\log|F|_v.$$

By Stokes' theorem,

$$\int_{Y(a)}\log|F|_v\, d\overline{d}\|s\|_v = -\left(\sum_\alpha\int_{C(\alpha,a)}\omega + \sum_\beta\int_{C(\beta,a)}\omega\right) \tag{16}$$

(the minus sign here comes from the fact that applying Stokes' theorem to the outside of a circle gives a negative orientation). For small a,

$$\int_{C(\alpha,a)}\log|F|_v\overline{d}\log\|s\|_v = O(a\log a),$$

so

$$\lim_{a\to 0}\int_{C(\alpha,a)}\log|F|_v\overline{d}\log\|s\|_v = 0.$$

Similarly,

$$\lim_{a\to 0}\int_{C(\beta,a)}\log\|s\|_v\,\overline{d}\log|F|_v = 0.$$

Switching to polar coordinates r,θ we get

$$\overline{d}\log\|s\|_v = r\frac{\partial\log\|s\|_v}{\partial r}d\theta.$$

Since

$$\frac{\partial\log\|s\|_v}{\partial r} = \frac{\mathrm{N}_v n_\beta}{r} + C^\infty\text{-function}$$

for small r, we thus obtain

$$\lim_{a\to 0}\int_{C(\beta,a)}\log|F|_v\,\overline{d}\log\|s\|_v = (2\pi i)\mathrm{N}_v n_\beta\log|F(\beta)|_v.$$

A similar calculation shows that

$$\lim_{a\to 0}\int_{C(\alpha,a)} \log\|s\|_v \overline{d}\log|F|_v = (2\pi i)m_\alpha \log\|s(\alpha)\|_v.$$

Taking the limit of (16) as $a \to 0$, thus gives

$$\lim_{a\to 0}\frac{1}{2\pi i}\int_{Y(a)} \log|F|_v \; d\overline{d}\log\|s\|_v$$
$$= -\left(\sum_\beta \mathrm{N}_v n_\beta \log|F(\beta)|_v - \sum_\alpha m_\alpha \log\|s(\beta)\|_v\right).$$

□

Now we define the Arakelov intersection $(D.L)_{\mathrm{Ar}}$ of a metrized line bundle L with a Weil divisor. We begin by defining the Arakelov degree $\deg_{Ar} M$ of a metrized line bundle M over an order R as in J. Silverman [24]. A metrized line bundle M over an order R is a free R-module of rank 1 with a nonzero metric $\|\cdot\|_w$ on the completion M_w for each archimedean place $w \in R_\infty$ on the field of fractions of R. As with our metrized line bundles on X, the metrics at infinity are normalized so that they behave like $|\cdot|^{\mathrm{N}_w}$ where $|\cdot|$ is the usual absolute value on $\mathbb{C}$ and N_w is the local degree of R_w over $\mathbb{R}$. Let $m \neq 0$ be an element of M; then

$$\deg_{\mathrm{Ar}} M = \log\#(M/Rm) - \sum_{w\in R_\infty}\log\|m\|_w. \tag{17}$$

Note that this definition does not depend on our choice of m by the product formula.

If L is a metrized line bundle and E_β is a horizontal divisor on X, then L pulls back to a metrized line bundle $i_\beta^* L$ over the order R where E_β is $i_\beta(\mathrm{Spec}(R))$ and we define

$$(E_\beta.L)_{\mathrm{Ar}} = \deg_{\mathrm{Ar}} i_\beta^* L.$$

If s is a section of L such that Supp$s$ doesn't meet E_β on the generic fiber, this gives

$$(E_\beta.L)_{\mathrm{Ar}} = \log\#(M/R(i_\beta^* s)) - \sum_{v|\infty}\sum_{i=1}^{\deg\beta}\log\|s(\beta_v^{[i]})\|_v,$$

where $\beta_v^{[i]}$ are the conjugates of β in $X(\mathbb{C}_v)$.

If D is a reduced irreducible divisor contained in a finite fiber X_v, then we define

$$(D.L)_{\mathrm{Ar}} = (D.L)_{\mathrm{fin}} = \deg(L|_D)\log\mathrm{N}(v).$$

Write $\mathrm{div}(s) = D_{\mathrm{ver}} + D_{\mathrm{hor}}$ where D_{ver} is vertical and D_{hor} is horizontal. For each Galois orbit of points β in $X(\mathbb{C}_v)$ as above, we pick a representative β'. Then we can write $D_{\mathrm{hor}} = \sum n_{\beta'} E_{\beta'}$, where E_β is an irreducible horizontal divisor on X corresponding to the Galois orbit of the point $\beta' \in X(\bar{\mathbb{Q}})$ and $n_{\beta'} = n_\beta$ for each $\beta \in X(\mathbb{C}_v)$ in the orbit of β'. Pick representatives α' of each Galois orbit of points α and write the horizontal part of $\mathrm{div}(F)$ as $\sum m_{\alpha'} E_{\alpha'}$ where $E_{\alpha'}$ is the irreducible horizontal divisor corresponding to the Galois orbit of the point $\alpha' \in X(\bar{\mathbb{Q}})$.

We recall our definition of the curvature $d\mu_v$ of a smoothly metrized line bundle L. If s is a section of L, then away from the support of s, we have $d\mu_v = -\frac{1}{2\pi i} d\overline{d} \log \|s\|_v$.

Proposition B.3. *We have the formula*

$$(\mathrm{div}(F).L)_{\mathrm{Ar}} = \sum_{v|\infty} \int_{X(\mathbb{C}_v)} \log |F|_v d\mu_v.$$

Proof. Let s be a global meromorphic section of L. We have $(\mathrm{div}(F).D_{\mathrm{ver}}) = 0$ by Corollary A.7 (since each component of D_{ver} is projective), so letting $\beta_v^{[i]}$ denote the conjugates of β' in $X(\mathbb{C}_v)$, we have

$$(\mathrm{div}(F).L)_{\mathrm{Ar}} = (\mathrm{div}(F).E_\beta)_{\mathrm{fin}} - \sum_{v|\infty} \sum_{\alpha'} \sum_{i=1}^{\deg \alpha'} m_{\alpha'} \log \|s(\alpha_v^{[i]})\|_v$$

$$= \sum_{v|\infty} \sum_{\beta'} \sum_{i=1}^{\deg \beta'} \mathrm{N}_v n_{\beta'} \log |F(\beta_v^{[i]})|_v - \sum_{v|\infty} \sum_{\alpha'} \sum_{i=1}^{\deg \alpha'} m_{\alpha'} \log \|s(\alpha_v^{[i]})\|_v$$

by Lemma B.1. Applying Proposition B.2 at each $v \mid \infty$ and summing over v gives

$$(\mathrm{div} F.L)_{\mathrm{Ar}} = -\sum_{v|\infty} \frac{1}{2\pi i} \int_{X(\mathbb{C}_v)} \log |F|_v \; d\overline{d} \log \|s\|_v$$

$$= \sum_{v|\infty} \int_{X(\mathbb{C}_v)} \log |F|_v d\mu_v.$$

□

References

1. S. Arakelov – "Intersection theory of divisors on an arithmetic surface", *Math. USSR Izvestija* **8** (1974), p. 1167–1180.
2. P. Autissier – "Points entiers sur les surfaces arithmétiques", *J. Reine. Angew. Math.* **531** (2001), p. 201–235.

3. M. Baker and R. Rumely – "Equidistribution of small points on curves, rational dynamics, and potential theory", Preprint.
4. A. F. Beardon – *Iteration of rational functions*, Springer-Verlag, New York, 1991.
5. E. Brieskorn and H. Knörrer – *Plane algebraic curves*, Birkhäuser, Basel, 1986.
6. H. Brolin – "Invariant sets under iteration of rational functions", *Ark. Mat.* **6** (1965), p. 103–144.
7. G. S. Call and S. Goldstine – "Canonical heights on projective space", *J. Number Theory* **63** (1997), p. 211–243.
8. G. S. Call and J. Silverman – "Canonical heights on varieties with morphism", *Compositio Math.* **89** (1993), p. 163–205.
9. P. Deligne – "Le déterminant de la cohomologie", *Contemporary Mathematics* **67** (1987), p. 94–177.
10. G. Everest and T. Ward – *Heights of Polynomials and Entropy in Algebraic Dynamics*, Springer-Verlag, New York, 1999.
11. A. Freire, A. Lopes and R. Mañe – "An invariant measure for rational functions", *Boletim da Sociedade Brasileira de Matematica* **14** (1983), p. 45–62.
12. W. Fulton – *Intersection theory*, Springer-Verlag, New York, 1975.
13. R. Hartshorne – *Algebraic geometry*, Springer-Verlag, New York, 1977.
14. S. Lang – *Introduction to Arakelov theory*, Springer-Verlag, New York, 1988.
15. — , *Algebra*, third ed., Springer-Verlag, New York, 2002.
16. M. Lyubich – "Entropy properties of rational endomorphisms of the Riemann sphere", *Ergodic Theory Dynam. Systems* **3** (1983), p. 351–385.
17. K. Mahler – "An application of Jensen's formula to polynomials", *Mathematica* **7** (1960), p. 98–100.
18. V. Maillot – "Géométrie d'Arakelov des Variétés Toriques et fibrés en droites intégrables", *Mémoires de la S.M.F* **80** (2000).
19. C. T. McMullen – *Complex dynamics and renormalization*, Annals of Mathematics Studies Volume 135, Princeton, 1994.
20. J. Milnor – *Dynamics in one complex variable*, Vieweg, Braunschweig, 1999.
21. L. Moret-Bailly – "Pinceaux de variété abéliennes", *Asterisque* **129** (1985).
22. D. Mumford – "The topology of normal singularities of an algebraic surface and a criterion for simplicity", *Inst. Hautes Études Sci. Publ. Math.* **9** (1961), p. 5–22.
23. C. Peskine and L. Szpiro – "Syzygies et multiplicités", *C. R. Acad. Paris Sci. Sér A* **278** (1974), p. 1421–1424.
24. J. Silverman – "The theory of height functions", Arithmetic geometry (G. Cornell and J. Silverman, eds.), Springer-Verlag, New York, 1986, p. 151–166.
25. L. Szpiro – "Cours de géométrie arithmétique", Orsay preprint.
26. — , "Séminaire sur les pinceaux arithmétiques", *Astérisque* **127** (1985), p. 1–287.
27. L. Szpiro and T. Tucker – "Computing generalized Mahler measure via equidistribution", in preparation.
28. L. Szpiro, E. Ullmo and S. Zhang – "Equirépartition des petits points", *Invent. Math.* **127** (1997), p. 337–347.
29. P. Vojta – "A generalization of theorems of Faltings and Thue-Siegel-Roth-Wirsing", *J. Amer. Math. Soc.* **5** (1992), p. 763–804.
30. S. Zhang – "Positive line bundles on arithmetic surfaces", *Annals of Math* **136** (1992), p. 569–587.

31. — , "Positive line bundles on arithmetic varieties", *J. Amer. Math. Soc.* **8** (1995), p. 187–221.
32. — , "Small points and adelic metrics", *J. Algebraic Geometry* **4** (1995), p. 281–300.

A Combination of the Conjectures of Mordell–Lang and André–Oort

Richard Pink

Dept. of Mathematics, ETH-Zentrum, CH-8092 Zürich, Switzerland
pink@math.ethz.ch

Summary. We propose a conjecture combining the Mordell–Lang conjecture with an important special case of the André–Oort conjecture, and explain how existing results imply evidence for it.

1 Introduction

We begin with some remarks on the history of related conjectures; the reader wishing to skip them may turn directly to Conjecture 1.6. Let us start (arbitrarily) with the following theorem.

Theorem 1.1 (Mordell–Weil). *For any abelian variety A over a number field K, the group of rational points $A(K)$ is finitely generated.*

This was proved in 1922 by Mordell [31] for elliptic curves over $\mathbb{Q}$; the general case was established by Weil [48] in 1928. Mordell also posed the following statement as a question in the case $K = \mathbb{Q}$:

Conjecture 1.2 (Mordell). *For any irreducible smooth projective algebraic curve Z of genus ≥ 2 over a number field K, the set of rational points $Z(K)$ is finite.*

This conjecture was proved by Faltings [16], [17] in 1983. Later another proof was found by Vojta [46], simplified by Faltings [18], and recast in almost elementary terms by Bombieri [5]. For some accounts of these developments see Hindry [21], Vojta [47], or Wüstholz [49].

The Mordell conjecture can be translated into a statement about abelian varieties, as follows. If $Z(K)$ is empty, we are done. Otherwise we can embed Z into its Jacobian variety J, such that $Z(K) = J(K) \cap Z$. By the Mordell–Weil theorem $J(K)$ is a finitely generated group. Thus with some generalization we must prove that for any abelian variety A over a field of characteristic zero, any finitely generated subgroup $\Lambda \subset A$, and any irreducible curve $Z \subset A$ of genus

≥ 2, the intersection $Z \cap \Lambda$ is finite. Since this can be verified over the base field $\mathbb{C}$, the Mordell conjecture has thus been translated into a statement about subvarieties of complex abelian varieties, involving only complex algebraic and/or analytic geometry.

However, this point of view does not really help in proving the Mordell conjecture. It merely hides the inherent arithmetic nature of the problem that is introduced by the subgroup Λ. In fact, the only known proof of the statement consists in deducing it from the Mordell conjecture.

Nevertheless, the idea to study intersections like $Z \cap \Lambda$ led to other fruitful developments. Motivated by their attempts to prove the Mordell conjecture, Manin and Mumford independently raised the following statement as a question.

Conjecture 1.3 (Manin–Mumford). *Let A be an abelian variety over $\mathbb{C}$ and let A_{tor} denote its subgroup of all torsion points. Let $Z \subset A$ be an irreducible closed algebraic subvariety such that $Z \cap A_{\mathrm{tor}}$ is Zariski dense in Z. Then Z is a translate of an abelian subvariety of A by a torsion point.*

This conjecture has been proved in several remarkably different ways. After a partial result by Bogomolov [4], the first full proof was published in 1983 by Raynaud [36], [37]. A different proof for curves was given by Coleman [8]. In the meantime other full proofs were found by Hindry [20], by Hrushovski [23], by Ullmo [44] and Zhang [52], and by Pink–Roessler [34], see also Roessler [39]. Ullmo and Zhang actually prove a stronger statement conjectured by Bogomolov, where the torsion points are replaced by all points of sufficiently small Néron–Tate height.

Lang [24], [25] combined the preceding conjectures into a single one by starting with a finitely generated subgroup $\Lambda_0 \subset A$ and considering its division group

$$\Lambda := \big\{ a \in A \;\big|\; \exists n \in \mathbb{Z}^{>0} : na \in \Lambda_0 \big\},$$

which contains both Λ_0 and A_{tor}.

Conjecture 1.4 (Mordell–Lang). *Let A be an abelian variety over $\mathbb{C}$ and $\Lambda \subset A$ as above. Let $Z \subset A$ be an irreducible closed algebraic subvariety such that $Z \cap \Lambda$ is Zariski dense in Z. Then Z is a translate of an abelian subvariety of A.*

This conjecture, too, is now a theorem by the combination of work of Faltings [18], [19], Raynaud [35], Vojta [46], and Hindry [21]. Detailed surveys may be found in [12], [49], [47]; for a short historical overview see [22, pp.435–439]. Moreover, McQuillan [26] established the natural generalization to semi-abelian varieties.

Let us now turn to Shimura varieties. (The relevant notions and notation will be reviewed in Sections 2 through 4.) The prototype for all Shimura varieties is the Siegel moduli space of principally polarized abelian varieties of

dimension g with a level structure. The points in this moduli space corresponding to abelian varieties with complex multiplication are called *special points*, and they play a particularly important role. A major reason for this is the Shimura–Taniyama theorem [43], [9, Th. 4.19], which describes the action of the Galois group on the torsion points of any CM abelian variety. This theorem, and its generalization by Milne–Shih [11] to Galois conjugates of CM abelian varieties, determine the Galois action on special points fairly completely. The concept of special points and the results concerning the Galois action on them have also been generalized to arbitrary Shimura varieties: see Deligne [10] and Milne [27], [28].

Aside from this explicit connection between the special points on Shimura varieties and the torsion points on abelian varieties, there are several other formal analogies. For one thing, both kinds of points are those for which certain associated Galois representations have particularly small images. For another, they are invariant under an inherent additional structure, in that the torsion points on an abelian variety form a subgroup, and the set of special points on a Shimura variety is invariant under all Hecke operators. Also, both form dense subsets in the analytic topology.

This strong connection was one motivation for André [1, p.215, Problem 1] and Oort [32] around 1990 to independently pose an analogue of the Manin–Mumford conjecture, whose combination is nowadays phrased as follows. Oort [32] was motivated in part by a question of Coleman concerning the number of special points in the Torelli locus. There are also interesting connections with complex analytic and p-adic linearity properties of Shimura subvarieties: see Moonen [29, IV.1.2] and especially [30, § 6].

An irreducible component of a Shimura subvariety of a Shimura variety S, or of its image under a Hecke operator, is called a *special subvariety of* S. Moonen, Edixhoven, and Yafaev call these 'subvarieties of Hodge type', due to their description as connected components of loci of Hodge classes.

Conjecture 1.5 (André–Oort). *Let S be a Shimura variety over $\mathbb{C}$, and let $\Lambda \subset S$ denote the set of all its special points. Let $Z \subset S$ be an irreducible closed algebraic subvariety such that $Z \cap \Lambda$ is Zariski dense in Z. Then Z is a special subvariety of S.*

This conjecture remains open, although it has been established in special cases or under additional assumptions by Moonen [29], André [2], Edixhoven [13], [14], Edixhoven–Yafaev [15], and Yafaev [51], [50]. In particular it has been proved by Edixhoven–Yafaev [15] when Z is a curve and Λ is replaced by the set of special points in a single generalized Hecke orbit.

The Manin–Mumford and André–Oort conjectures are not only related by analogy, but the obvious generalization of the latter to mixed Shimura varieties includes the former for CM abelian varieties. Namely, let $A \to S$ denote the universal family of abelian varieties over some Siegel moduli space. Then A is

a mixed Shimura variety in the sense of [33], and the special points in A are precisely the torsion points in the fibers A_s over all special points $s \in S$. On the other hand an irreducible subvariety $Z \subset A_s$ is an irreducible component of a mixed Shimura subvariety of A if and only if Z is a translate of an abelian subvariety of A_s by a torsion point. Thus the Manin–Mumford conjecture for the CM abelian variety A_s becomes a special case of the generalization of the André–Oort conjecture to A.

Building on this observation, André [3, Lect. III] suggested a 'generalized André–Oort conjecture', which for mixed Shimura varieties becomes the direct generalization of the André–Oort conjecture, but which also contains the Manin–Mumford conjecture for arbitrary abelian varieties.

The aim of this article is to generalize the correspondence just explained in another direction and to propose a conjecture about subvarieties of mixed Shimura varieties which contains not only the general Manin–Mumford conjecture, but also the general Mordell–Lang conjecture, as well as an important special case of the André–Oort conjecture. The key observation is that in the above situation, the special points in A_s are all contained in a single Hecke orbit. Already in 1989 André [1, p.216, Problem 3] had posed a problem about subvarieties containing a dense subset of points from a Hecke orbit, and the conjecture below can be viewed as an attempt to give a precise answer to André's question. For this, the notion of special subvarieties must be generalized as follows.

Consider Shimura morphisms of mixed Shimura varieties $T' \xleftarrow{[\varphi]} T \xrightarrow{[i]} S$ and a point $t' \in T'$. An irreducible component of $[i]\big([\varphi]^{-1}(t')\big)$, or of its image under a Hecke operator, is called a *weakly special subvariety of* S. The proposed conjecture is this:

Conjecture 1.6. *Let S be a mixed Shimura variety over $\mathbb{C}$ and $\Lambda \subset S$ the generalized Hecke orbit of a point $s \in S$. Let $Z \subset S$ be an irreducible closed algebraic subvariety such that $Z \cap \Lambda$ is Zariski dense in Z. Then Z is a weakly special subvariety of S.*

The notions of mixed Shimura varieties, Shimura morphisms, and generalized Hecke orbits are explained in Sections 2 and 3. Special and weakly special subvarieties are studied in Section 4, and their relation with special points in Section 4. The interplay between Hodge and Galois theoretic properties of points is discussed in Section 6. On this basis we give three different kinds of evidence for the conjecture by relating it to known results:

- In Section 4 we show that, for a pure Shimura variety and the generalized Hecke orbit of a special point, the conjecture becomes a particular case of the André–Oort conjecture. If in addition Z is a curve, it is thus proved by Edixhoven and Yafaev [15].
- In Section 5 we show that for subvarieties of an abelian variety which is the fiber of a Shimura morphism to a pure Shimura variety, the conjecture is equivalent to the Mordell–Lang conjecture, which is also known.

- In Section 7 we deduce the conjecture for the Siegel moduli space and the generalized Hecke orbit of a Galois generic point from equidistribution results of Clozel, Oh, and Ullmo [6], [7].

In each case the reduction is relatively simple, in spite of the notational complexity. All the hard work is done in the cited literature. One may hope to prove the conjecture eventually by a combination of the individual approaches.

Finally, I am not sure how to generalize Conjecture 1.6 reasonably to include the full André–Oort conjecture.

The idea for this article arose in December 2003 at the workshop on "Special Points in Shimura Varieties" at the Lorentz Center in Leiden. It is my pleasure to thank the organizers for their invitation, and Y. André, D. Bertrand, L. Clozel, B. Edixhoven, B. Moonen, F. Oort, E. Ullmo, J. Wildeshaus, and A. Yafaev for interesting conversations and suggestions. I would also like to thank Y. Tschinkel for the invitation to the conference on "Diophantine Geometry" in Göttingen in June 2004, and G. Faltings, G. Harder, and N. Katz for the invitation to the conference on "Arithmetic Algebraic Geometry" in Oberwolfach in August 2004, and the opportunity to present a part of this work there.

2 Connected mixed Shimura varieties

In this section we review the necessary facts about mixed Shimura varieties over the complex numbers. In order to simplify the definitions and to avoid the language of adèles we restrict ourselves to connected mixed Shimura varieties, which are simply the connected components of usual mixed Shimura varieties. For the basic theory of pure Shimura varieties see Deligne [9], [10], for mixed Shimura varieties see Milne [28] or Pink [33].

First we recall some facts about mixed Hodge structures. Consider the real torus $\mathbb{S} := \mathcal{R}_{\mathbb{C}/\mathbb{R}}\mathbb{G}_{m,\mathbb{C}}$ defined by Weil restriction. By construction it comes with a natural isomorphism $\mathbb{S}(\mathbb{R}) \cong \mathbb{C}^*$. For any linear representation of $\mathbb{S}$ on a complex vector space $V_{\mathbb{C}}$ and any pair of integers (p,q) we let $V^{p,q}$ be the subspace of $V_{\mathbb{C}}$ on which $z \in \mathbb{S}(\mathbb{R}) \cong \mathbb{C}^*$ acts through multiplication by $z^{-p}\bar{z}^{-q}$. The induced $\mathbb{Z}$-filtrations

$$F^pV_{\mathbb{C}} := \bigoplus_{\substack{p',q\in\mathbb{Z}\\ p'\geq p}} V^{p',q} \quad \text{and} \quad W_nV_{\mathbb{C}} := \bigoplus_{\substack{p,q\in\mathbb{Z}\\ p+q\leq n}} V^{p,q}$$

are called the associated *Hodge* and *weight filtrations*. If $V_{\mathbb{C}}$ is the complexification of a $\mathbb{Q}$-vector space V, these filtrations define a *rational mixed Hodge structure on* V if and only if the subspace $W_nV_{\mathbb{C}}$ is defined over $\mathbb{Q}$ for all n, and $\overline{V^{p,q}} \equiv V^{q,p} \mathrm{mod} W_{p+q-1}V_{\mathbb{C}}$ for all (p,q). This is not the usual construction of rational mixed Hodge structures, but every rational mixed Hodge structure

arises in this way, although the representation of $\mathbb{S}$ on $V_{\mathbb{C}}$ is not uniquely determined by the filtrations. The mixed Hodge structure is called *of type* $S \subset \mathbb{Z}^2$ if $V^{p,q} = 0$ for all $(p,q) \notin S$.

Note that giving a representation of $\mathbb{S}$ on $V_{\mathbb{C}}$ is equivalent to giving a homomorphism $\mathbb{S}_{\mathbb{C}} \to \mathrm{Aut}_{\mathbb{C}}(V_{\mathbb{C}})$ of algebraic groups over $\mathbb{C}$. In the following definition we consider more generally the set of homomorphisms $\mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, P_{\mathbb{C}})$ for a linear algebraic group P over $\mathbb{Q}$. The group $P(\mathbb{C})$ acts on this set from the left by composition with the inner automorphisms $\mathrm{int}(p)\colon p' \mapsto pp'p^{-1}$ for all $p \in P(\mathbb{C})$.

Definition 2.1. *A* connected mixed Shimura datum *is a pair* (P, X^+) *where*

(a) P *is a connected linear algebraic group over* $\mathbb{Q}$*, with unipotent radical* W*, and with another algebraic subgroup* $U \subset W$ *that is normal in* P *and uniquely determined by* X^+ *using condition (iii) below, and*
(b) $X^+ \subset \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, P_{\mathbb{C}})$ *is a connected component of an orbit under the subgroup* $P(\mathbb{R}) \cdot U(\mathbb{C}) \subset P(\mathbb{C})$*,*

such that for some (or equivalently for all) $x \in X^+$*,*

(i) the composite homomorphism $\mathbb{S}_{\mathbb{C}} \xrightarrow{x} P_{\mathbb{C}} \to (P/U)_{\mathbb{C}}$ *is defined over* $\mathbb{R}$*,*
(ii) the adjoint representation induces on $\mathrm{Lie}\, P$ *a rational mixed Hodge structure of type*

$$\{(-1,1),(0,0),(1,-1)\} \cup \{(-1,0),(0,-1)\} \cup \{(-1,-1)\},$$

(iii) the weight filtration on $\mathrm{Lie}\, P$ *is given by*

$$W_n(\mathrm{Lie}\, P) = \begin{cases} 0 & \text{if } n < -2, \\ \mathrm{Lie}\, U & \text{if } n = -2, \\ \mathrm{Lie}\, W & \text{if } n = -1, \\ \mathrm{Lie}\, P & \text{if } n \geq 0, \end{cases}$$

(iv) the conjugation by $x\big(\sqrt{-1}\big)$ *induces a Cartan involution on* $(P/W)^{\mathrm{ad}}_{\mathbb{R}}$*,*
(v) P/P^{der} *is an almost direct product of a* $\mathbb{Q}$*-split torus with a torus of compact type defined over* $\mathbb{Q}$*, and*
(vi) P *possesses no proper normal subgroup* P' *defined over* $\mathbb{Q}$*, such that* x *factors through* $P'_{\mathbb{C}} \subset P_{\mathbb{C}}$*.*

If in addition P *is reductive,* (P, X^+) *is called a* connected (pure) Shimura datum.

Remark 2.2. The axioms (v) and (vi) are sometimes weakened. It suffices to assume the consequence of (v) that the center of P/W acts on $\mathrm{Lie}\, U$ and on $\mathrm{Lie}\, W/U$ through an almost direct product of a $\mathbb{Q}$-split torus with a torus of compact type defined over $\mathbb{Q}$.

Note that (v) also implies that every sufficiently small congruence subgroup of $P(\mathbb{Q})$ is contained in $P^{\mathrm{der}}(\mathbb{Q})$.

Facts 2.3. *[33, 1.18, 3.3, 9.24] Let (P, X^+) be a connected mixed Shimura datum.*

(a) The stabilizer $P(\mathbb{R})^+$ of $X^+ \subset \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, P_{\mathbb{C}})$ is open in $P(\mathbb{R})$.
(b) X^+ possesses a unique structure of complex manifold such that for every representation ρ of P on a complex vector space $V_{\mathbb{C}}$ the Hodge filtration determined by $\rho \circ x$ varies holomorphically with $x \in X^+$. In particular this complex structure is invariant under $P(\mathbb{R})^+ \cdot U(\mathbb{C})$.
(c) Set $P(\mathbb{Q})^+ := P(\mathbb{Q}) \cap P(\mathbb{R})^+$. Any congruence subgroup $\Gamma \subset P(\mathbb{Q})^+$ acts properly discontinuously on X^+, so that $\Gamma\backslash X^+$ is a complex analytic space with at most finite quotient singularities.
(d) Every sufficiently small congruence subgroup $\Gamma \subset P(\mathbb{Q})^+$ acts freely on X^+, so that $\Gamma\backslash X^+$ is a complex manifold and $X^+ \twoheadrightarrow \Gamma\backslash X^+$ an unramified covering.
(e) $\Gamma\backslash X^+$ possesses a natural structure of quasiprojective algebraic variety over $\mathbb{C}$.

Definition 2.4. *The variety $\Gamma\backslash X^+$ from 2.3 is called the* connected mixed Shimura variety *associated to (P, X^+) and Γ. It is called a* connected (pure) Shimura variety *if P reductive. The residue class of an element $x \in X^+$ is denoted $[x] \in \Gamma\backslash X^+$.*

Definition 2.5. *A* morphism *of connected mixed Shimura data $(P, X^+) \to (P', X'^+)$ is a homomorphism $\varphi\colon P \to P'$ of algebraic groups over $\mathbb{Q}$ which induces a map $X^+ \to X'^+$, $x \mapsto \varphi \circ x$.*

Facts 2.6. *[33, 3.4, 9.24] Let $\varphi : (P, X^+) \to (P', X'^+)$ be a morphism of connected mixed Shimura data and $\Gamma \subset P(\mathbb{Q})^+$ and $\Gamma' \subset P'(\mathbb{Q})^+$ congruence subgroups such that $\varphi(\Gamma) \subset \Gamma'$. Then the map*

$$[\varphi]\colon\ \Gamma\backslash X^+ \to \Gamma'\backslash X'^+,\ [x] \mapsto [\varphi \circ x]$$

is well-defined, holomorphic, and algebraic with respect to the algebraic structures from 2.3 (e). Furthermore, $[\varphi]$ is

(a) a finite morphism if the identity component of $\mathrm{Ker}(\varphi)$ is a torus,
(b) surjective if $\mathrm{Im}(\varphi)$ contains the derived group $(P')^{\mathrm{der}}$, and
(c) a (possibly ramified) covering if the conditions in (a) and (b) both hold.

Definition 2.7. *Any morphism $[\varphi]$ as in 2.6 is called a* Shimura morphism. *Moreover, it is called a*

(a) Shimura immersion *if the condition in 2.6 (a) holds,*
(b) Shimura submersion *if the condition in 2.6 (b) holds,*
(c) Shimura covering *if the condition in 2.6 (c) holds.*

Here the word 'immersion' is intended to reflect its meaning in topology, where it is a purely local condition. Any composite of an unramified covering

and a closed embedding is an immersion in this sense. The use in 2.7 (a) includes ramified coverings and is therefore somewhat abusive, but I did not find a better word.

Proposition 2.8. *Any Shimura morphism* $[\varphi]$ *possesses a factorization into Shimura morphisms* $[\varphi] = [i] \circ [\psi] \circ [\pi]$*, where*

(a) π *is surjective, and its kernel is connected and possesses no non-trivial torus quotient,*
(b) ψ *is surjective and the identity component of its kernel is a torus, and*
(c) i *is injective.*

In particular, $[\pi]$ *is a Shimura submersion,* $[\psi]$ *a Shimura covering, and* $[i]$ *a Shimura immersion.*

Proof. Let $\varphi : (P, X^+) \to (P', X'^+)$ be any morphism of connected mixed Shimura data. Let $K \lhd P$ denote the identity component of $\mathrm{Ker}(\varphi)$, and let K' be the smallest connected normal subgroup of K such that K/K' is a torus. This is a characteristic subgroup of K and therefore normal in P. Set $\bar{P} := P/K'$ and $\bar{P}' := \varphi(P) \subset P'$. Then φ is the composite of natural homomorphisms

$$P \xrightarrow{\pi} \bar{P} \xrightarrow{\psi} \bar{P}' \xrightarrow{i} P'$$

which possess the desired properties (a–c). Let

$$\bar{X}^+ \subset \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, \bar{P}_{\mathbb{C}}) \quad \text{and} \quad \bar{X}'^+ \subset \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, \bar{P}'_{\mathbb{C}})$$

denote the images of $X^+ \subset \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, P_{\mathbb{C}})$ under composition with π and $\psi \circ \pi$, respectively. Since the homomorphisms

$$P(\mathbb{R}) \xrightarrow{\pi} \bar{P}(\mathbb{R}) \xrightarrow{\psi} \bar{P}'(\mathbb{R})$$

induce surjections on the identity components, one can easily verify that $(\bar{P}, \bar{X}^+)$ and $(\bar{P}', \bar{X}'^+)$ are connected mixed Shimura data. Thus $[\varphi]$ factors through Shimura morphisms

$$\Gamma\backslash X^+ \xrightarrow{[\pi]} \bar{\Gamma}\backslash \bar{X}^+ \xrightarrow{[\psi]} \bar{\Gamma}'\backslash \bar{X}'^+ \xrightarrow{[i]} \Gamma'\backslash X'^+$$

for suitable congruence subgroups, as desired. □

Construction 2.9. Let (P, X^+) be a connected mixed Shimura datum and V a finite dimensional representation of P over $\mathbb{Q}$. Assume that for some (or equivalently for all) $x \in X^+$ the induced rational mixed Hodge structure on V has type $\{(-1,0),(0,-1)\}$. With

$$V_{\mathrm{alg}} \;:=\; V \otimes_{\mathbb{Q}} \mathbb{G}_{a,\mathbb{Q}} \;\cong\; \mathbb{G}_{a,\mathbb{Q}}^{\dim V}$$

we can define the semidirect product $P \ltimes V_{\mathrm{alg}}$ as a linear algebraic group. Set $V_{\mathbb{R}} := V \otimes_{\mathbb{Q}} \mathbb{R} \cong V_{\mathrm{alg}}(\mathbb{R})$ and let

$$X^+ \ltimes V_{\mathbb{R}} \ \subset\ \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, P_{\mathbb{C}} \ltimes V_{\mathrm{alg},\mathbb{C}})$$

denote the conjugacy class under

$$\big(P(\mathbb{R})^+ \cdot U(\mathbb{C})\big) \ltimes V_{\mathbb{R}} = (P \ltimes V_{\mathrm{alg}})(\mathbb{R})^+ \cdot U(\mathbb{C})$$

generated by $X^+ \subset \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, P_{\mathbb{C}})$. The notation $X^+ \ltimes V_{\mathbb{R}}$ is justified by the natural bijection

$$X^+ \times V_{\mathbb{R}} \xrightarrow{\sim} X^+ \ltimes V_{\mathbb{R}},\ (x, v') \mapsto \mathrm{int}(v') \circ x. \tag{2.10}$$

Under this bijection the action of $(p, v) \in (P(\mathbb{R})^+ \cdot U(\mathbb{C})) \ltimes V_{\mathbb{R}}$ corresponds to the twisted action $(p, v) \cdot (x, v') = (px, pv' + v)$. Moreover the pair $(P \ltimes V_{\mathrm{alg}}, X^+ \ltimes V_{\mathbb{R}})$ is a connected mixed Shimura datum. Let $\Gamma \subset P(\mathbb{Q})^+$ be a sufficiently small congruence subgroup and $\Gamma_V \subset V = V_{\mathrm{alg}}(\mathbb{Q})$ a Γ-invariant $\mathbb{Z}$-lattice of rank $\dim(V)$. The projection $\pi : P \ltimes V_{\mathrm{alg}} \to P$ then induces a Shimura epimorphism

$$A \ :=\ (\Gamma \ltimes \Gamma_V)\backslash(X^+ \ltimes V_{\mathbb{R}}) \xrightarrow{[\pi]} \Gamma\backslash X^+ \ =:\ S. \tag{2.11}$$

Furthermore the homomorphisms

$$\begin{array}{ccccccc} P \ltimes V_{\mathrm{alg}}^2 & \xrightarrow{\mu} & P \ltimes V_{\mathrm{alg}} & \xrightarrow{\pi} & P & \xrightarrow{\epsilon} & P \ltimes V_{\mathrm{alg}}, \\ (p, v, v') & \longmapsto & (p, v + v') & \longmapsto & p & \longmapsto & (p, 0) \end{array}$$

induce Shimura morphisms

$$\begin{array}{ccccccccc} (\Gamma \ltimes \Gamma_V^2)\backslash(X^+ \ltimes V_{\mathbb{R}}^2) & \longrightarrow & (\Gamma \ltimes \Gamma_V)\backslash(X^+ \ltimes V_{\mathbb{R}}) & \longrightarrow & \Gamma\backslash X^+ & \longrightarrow & (\Gamma \ltimes \Gamma_V)\backslash(X^+ \ltimes V_{\mathbb{R}}) \\ \wr\| & & \| & & \| & & \| \\ A \times_S A & \xrightarrow{[\mu]} & A & \xrightarrow{[\pi]} & S & \xrightarrow{[\epsilon]} & A \end{array}$$

which turn A into a family of abelian varieties over S, for which $[\mu]$ is the addition morphism and $[\epsilon]$ the zero section. All its fibers are isomorphic to the compact torus $\Gamma_V\backslash V_{\mathbb{R}} \cong \Gamma_V \otimes (\mathbb{Z}\backslash\mathbb{R})$, whose complex structure depends on the base point in S. (Compare [33, 3.13–14].)

Example 2.12. Consider an integer $g > 0$ and a non-degenerate alternating $2g \times 2g$-matrix E over $\mathbb{Q}$. The associated group of symplectic similitudes is

$$\mathrm{CSp}_{2g,\mathbb{Q}} := \big\{U \in \mathrm{GL}_{2g,\mathbb{Q}} \ \big|\ \exists \lambda \in \mathbb{G}_{m,\mathbb{Q}} \colon U^t E U = \lambda E\big\}.$$

Let $\mathcal{H}_{2g}$ be the set of homomorphisms $h \colon \mathbb{S} \to \mathrm{CSp}_{2g,\mathbb{R}}$ that induce a Hodge structure of type $\{(-1,0),(0,-1)\}$ on the tautological representation $\mathbb{Q}^{2g}$, for which the symmetric pairing

$$\mathbb{R}^{2g} \times \mathbb{R}^{2g} \to \mathbb{R},\ (v, v') \mapsto v^t E h\big(\sqrt{-1}\big) v'$$

is positive definite. It is known that the identity component of $\mathrm{CSp}_{2g}(\mathbb{R})$ acts transitively on $\mathcal{H}_{2g}$ and the pair $(\mathrm{CSp}_{2g,\mathbb{Q}}, \mathcal{H}_{2g})$ is a connected pure Shimura datum. For suitable congruence subgroups the associated connected Shimura variety is a fine moduli space for certain polarized abelian varieties of dimension g with a level structure.

Moreover, we can apply Construction 2.9 to the tautological representation of $\mathrm{CSp}_{2g,\mathbb{Q}}$ on $V_{2g} := \mathbb{Q}^{2g}$, yielding a connected mixed Shimura datum with underlying group $\mathrm{CSp}_{2g,\mathbb{Q}} \ltimes \mathbb{G}_{a,\mathbb{Q}}^{2g}$. For suitable congruence subgroups the associated connected mixed Shimura variety is the universal family of abelian varieties over the moduli space. Every abelian variety can be realized as a fiber of such a family. (See [33, 2.7, §10].)

Remark 2.13. The above construction can be generalized to semi-abelian varieties. For a brief sketch let (P, X^+) and V be as in Construction 2.9. Suppose that, as in Example 2.12, we are given a P-equivariant non-degenerate alternating form $\Psi : V \times V \to U \cong \mathbb{Q}$, where U is a representation of P of Hodge type $\{(-1,-1)\}$. For any positive integer d let $P \ltimes V_{\mathrm{alg}}^d$ act on $V \oplus U^d$ by the representation

$$(p, (v_i)) \cdot (v, (u_i)) := \big(pv, (pu_i + \Psi(v_i, v))\big).$$

Let Γ_V, $\Gamma_V^* \subset V$ and $\Gamma_U \subset U$ be Γ-invariant $\mathbb{Z}$-lattices such that Ψ induces a perfect pairing $\Gamma_V \times \Gamma_V^* \to \Gamma_U$. Then the projection homomorphisms

$$(P \ltimes V_{\mathrm{alg}}^d) \ltimes (V \oplus U^d)_{\mathrm{alg}} \twoheadrightarrow P \ltimes V_{\mathrm{alg}}^d \twoheadrightarrow P$$

and

$$(\Gamma \ltimes \Gamma_V^d) \ltimes (\Gamma_V^* \times \Gamma_U^d) \twoheadrightarrow \Gamma \ltimes \Gamma_V^d \twoheadrightarrow \Gamma$$

induce Shimura epimorphisms $B \twoheadrightarrow A^d \twoheadrightarrow S$. Here $B \twoheadrightarrow A^d$ is a family of algebraic groups, whose fibers are isomorphic to

$$\Gamma_V^* \backslash V_{\mathbb{R}} \oplus (\Gamma_U \backslash U_{\mathbb{C}})^d$$

with a twisted complex structure determined by the base point in A^d. One verifies that this is a semi-abelian variety with a torus part of dimension d. More precisely, let A_s be the fiber above $s \in S$ of the family $A \to S$ in Construction 2.9. Let A_s^* denote the abelian variety dual to A_s. Then $A_s^d \cong \mathrm{Ext}^1(A_s^*, \mathbb{G}_m^d)$, and the fiber of $B \to A^d$ above a point $a \in A_s^d$ is the associated extension

$$1 \to \mathbb{G}_m^d \to B_a \to A_s^* \to 1.$$

Every semi-abelian variety is isomorphic to such a B_a (compare [33, 3.13–14]).

3 Generalized Hecke orbits

Let $S = \Gamma\backslash X^+$ be the connected mixed Shimura variety associated to a connected mixed Shimura datum (P, X^+) and a congruence subgroup $\Gamma \subset P(\mathbb{Q})^+$.

Definition 3.1.

(a) For any automorphism φ of P that induces an automorphism of Shimura data $(P, X^+) \to (P, X^+)$, the diagram of Shimura coverings

$$S \;=\; \Gamma\backslash X^+ \xleftarrow{[\mathrm{id}]} (\Gamma \cap \varphi^{-1}(\Gamma))\backslash X^+ \xrightarrow{[\varphi]} \Gamma\backslash X^+ \;=\; S$$

is called a generalized Hecke operator on S *and is denoted by T_φ.*

(b) The generalized Hecke operator associated to an inner automorphism $\mathrm{int}(p)\colon p' \mapsto pp'p^{-1}$ for an element $p \in P(\mathbb{Q})^+$ is called a (usual) Hecke operator on S *and is denoted by T_p.*

Here $\Gamma \cap \varphi^{-1}(\Gamma)$ is again a congruence subgroup of $P(\mathbb{Q})^+$, and so both morphisms in 3.1 (a) are finite (possibly ramified) coverings. Thus Hecke operators can be viewed as finite multivalued functions from S to itself in the following sense.

Definition 3.2. *For any T_φ as in Definition 3.1 (a) and any subset $Z \subset S$, the subset*

$$T_\varphi(Z) := [\varphi]\big([\mathrm{id}]^{-1}(Z)\big)$$

is called the translate of Z under T_φ. *We also abbreviate $T_\varphi(s) := T_\varphi(\{s\})$.*

By varying the Hecke operator we obtain the following notion.

Proposition–Definition 3.3. *Fix a point $s \in S$.*

(a) The union of $T_\varphi(s)$ for all automorphisms φ of (P, X^+) is called the generalized Hecke orbit of s.

(b) The union of $T_p(s)$ for all $p \in P(\mathbb{Q})^+$ is called the (usual) Hecke orbit of s.

Remark 3.4. Using the fact that $P(\mathbb{Q})^+ \subset P(\mathbb{R})^+$ is dense, one easily shows that every Hecke orbit, and hence also every generalized Hecke orbit, is Zariski dense. It is dense for the analytic topology if $U = 1$.

Remark 3.5. Edixhoven and Yafaev [15] define generalized Hecke orbits differently. They fix a faithful representation $\rho : P \hookrightarrow \mathrm{GL}(V)$ and say that $x' \in X^+$ is equivalent to $x \in X^+$ if and only if the associated rational Hodge structures on V are isomorphic. The set of all $[x']$ thus obtained is their generalized Hecke orbit of $[x]$. (Yafaev [51] weakens this condition further by requiring only that the Mumford–Tate groups are isomorphic.)

Both definitions have their merits, but the relation between them is delicate. We restrict ourselves to the following result.

Proposition 3.6. *If P is reductive, every generalized Hecke orbit in the sense of Proposition-Definition 3.3 (a) is contained in a finite union of generalized Hecke orbits in the sense of Remark 3.5.*

Proof. Consider any automorphism φ of (P, X^+). Let $\bar{\varphi}$ denote the induced automorphism of the torus $\bar{P} := P/P^{\mathrm{der}}$, and let $\bar{x}$ denote the composite homomorphism $\mathbb{S}_{\mathbb{C}} \xrightarrow{x} P_{\mathbb{C}} \twoheadrightarrow \bar{P}_{\mathbb{C}}$ for any $x \in X^+$. Then the equation $\bar{\varphi} \circ \bar{x} = \bar{x}$ means that the image of $\bar{x}$ is contained in the kernel of the homomorphism $\bar{P} \to \bar{P}, \bar{p} \mapsto \bar{\varphi}(\bar{p})/\bar{p}$. By Axiom (vi) of Definition 2.1 this kernel must therefore be equal to $\bar{P}$. Thus $\bar{\varphi}$ is the identity, and hence the restriction of φ to the identity component of the center of P is the identity. On the other hand the outer automorphism group of P^{der} over $\bar{\mathbb{Q}}$ is finite. Thus to prove the proposition we may restrict ourselves to generalized Hecke operators coming from automorphisms φ of P which become inner automorphisms over $\bar{\mathbb{Q}}$.

Let $\rho\colon P \hookrightarrow \mathrm{GL}(V)$ be the given faithful representation. Then any automorphism φ as above fixes the isomorphism class of ρ over $\bar{\mathbb{Q}}$. Thus it also fixes the isomorphism class over $\mathbb{Q}$; hence it is induced by conjugation with an element of $\mathrm{GL}(V)$. But every such automorphism induces an isomorphism of rational Hodge structures on V. Thus it preserves the generalized Hecke orbit in the sense of 3.5, and the proposition follows. □

4 Special and weakly special subvarieties

In this section we analyze the following class of subvarieties of connected mixed Shimura varieties. Since special subvarieties can be described as connected components of loci of Hodge classes, they are also called 'subvarieties of Hodge type'.

Definition 4.1.

(a) The image of any Shimura morphism $T \xrightarrow{[i]} S$ is called a connected mixed Shimura subvariety of S, *or a* special subvariety of S.

(b) Consider any Shimura morphisms $T' \xleftarrow{[\varphi]} T \xrightarrow{[i]} S$ and any point $t' \in T'$. Then any irreducible component of $[i]\big([\varphi]^{-1}(t')\big)$ is called a weakly special subvariety of S.

Proposition 4.2. *Any special subvariety is a weakly special subvariety.*

Proof. Let $T_1 = \{t_1\}$ be the connected Shimura variety associated to the trivial algebraic group. Then for every connected mixed Shimura variety T there exists a unique Shimura morphism $[\varphi]: T \to T_1$, and the proposition follows. □

Proposition 4.3. *For any special subvariety of S the Shimura morphism in Definition 4.1 (a) can be chosen such that the underlying homomorphism of algebraic groups i is injective. Then in particular $[i]$ is a Shimura immersion.*

Proof. Direct consequence of Proposition 2.8 and Fact 2.6 (b). □

Proposition 4.4. *For any weakly special subvariety of S the Shimura morphisms in Definition 4.1 (b) can be chosen such that*

(a) the underlying homomorphism of algebraic groups φ is surjective, and its kernel is connected and possesses no non-trivial torus quotient, and
(b) the underlying homomorphism of algebraic groups i is injective.

Then in particular $[\varphi]$ is a Shimura submersion and $[i]$ a Shimura immersion.

Proof. Consider Shimura morphisms $T' \xleftarrow{[\varphi]} T \xrightarrow{[i]} S$ associated to morphisms of connected mixed Shimura data

$$(Q', Y'^{+}) \xleftarrow{\varphi} (Q, Y^{+}) \xrightarrow{i} (P, X^{+}),$$

and let Z be an irreducible component of $[i]\big([\varphi]^{-1}(t')\big)$ for some $t' \in T'$. By Proposition 2.8 we can factor $[\varphi] = [i'] \circ [\psi] \circ [\pi] = [i' \circ \psi] \circ [\pi]$, where π is surjective, and $\mathrm{Ker}(\pi)$ is connected and possesses no non-trivial torus quotient, and the identity component of $\mathrm{Ker}(i' \circ \psi)$ is a torus. The last property together with Fact 2.6 (a) implies that $[i' \circ \psi]$ is a finite morphism. It follows that every irreducible component of a fiber of $[\varphi]$ is also an irreducible component of a fiber of $[\pi]$. Since Z is the image under $[i]$ of such an irreducible component, we may replace $[\varphi]$ by $[\pi]$ without changing Z. We may therefore assume that φ already satisfies the condition (a).

Set $\bar{Q} := Q/\mathrm{Ker}(i)$ and $\bar{Q}' := Q'/\varphi(\mathrm{Ker}(i))$. Then we have a commutative diagram of homomorphisms

$$\begin{array}{ccccc} Q' & \xleftarrow{\varphi} & Q & \xrightarrow{i} & P \\ \downarrow{\scriptstyle \pi'} & & \downarrow{\scriptstyle \pi} & \nearrow{\scriptstyle \bar{i}} & \\ \bar{Q}' & \xleftarrow{\bar{\varphi}} & \bar{Q} & & \end{array}$$

where $\bar{i}$ is injective and $\bar{\varphi}$ also satisfies (a). As in the proof of Proposition 2.8 these homomorphisms induce morphisms of connected mixed Shimura data

$$\begin{array}{ccccc} (Q', Y'^{+}) & \xleftarrow{\varphi} & (Q, Y^{+}) & \xrightarrow{i} & (P, X^{+}) \\ \downarrow{\scriptstyle \pi'} & & \downarrow{\scriptstyle \pi} & \nearrow{\scriptstyle \bar{i}} & \\ (\bar{Q}', \bar{Y}'^{+}) & \xleftarrow{\bar{\varphi}} & (\bar{Q}, \bar{Y}^{+}) & & \end{array}$$

and hence Shimura morphisms

$$\begin{array}{ccccc} \Delta'\backslash Y'^{+} & \xleftarrow{[\varphi]} & \Delta\backslash Y^{+} & \xrightarrow{[i]} & \Gamma\backslash X^{+} \\ \downarrow{\scriptstyle [\pi']} & & \downarrow{\scriptstyle [\pi]} & \nearrow{\scriptstyle [\bar{i}\,]} & \\ \bar{\Delta}'\backslash \bar{Y}'^{+} & \xleftarrow{[\bar{\varphi}]} & \bar{\Delta}\backslash \bar{Y}^{+} & & \end{array}$$

for suitable congruence subgroups. The commutativity of these diagrams implies that

$$[\pi]\big([\varphi]^{-1}(t')\big) \subset [\bar{\varphi}]^{-1}\big([\pi'](t')\big).$$

On the other hand let $U \triangleleft Q$ be the subgroup of weight -2 from Definition 2.1 (a) and let d be the dimension of any orbit of $\mathrm{Ker}(\varphi)(\mathbb{R}) \cdot (U \cap \mathrm{Ker}(\varphi))(\mathbb{C})$ on $\bar{Y}^{+}$. Since $\bar{Q}' \cong \bar{Q}/\pi(\mathrm{Ker}(\varphi))$, both sides of the above inclusion are equidimensional of dimension equal to d. Thus every irreducible component of $[\pi]\big([\varphi]^{-1}(t')\big)$ is also an irreducible component of $[\bar{\varphi}]^{-1}\big([\pi'](t')\big)$. Since Z is the image under $[\,\bar{i}\,]$ of such an irreducible component, we may replace $[\varphi]$, $[i]$ by $[\bar{\varphi}]$, $[\,\bar{i}\,]$, and the proposition follows. □

Remark 4.5. There are two typical examples of weakly special subvarieties. For the first consider three connected mixed Shimura varieties T_1, T_2, T, and a Shimura immersion $[i] : T_1 \times T_2 \to T$. Then for any point $t_2 \in T_2$ the subvariety $Z := [i]\big(T_1 \times \{t_2\}\big)$ is weakly special in T. When all three Shimura varieties are pure, the subvarieties of T obtained in this way are precisely all totally geodesic irreducible subvarieties of T by a theorem of Moonen [29, II 3.1].

The other typical example is discussed in the following proposition.

Proposition 4.6. *Let $[\pi]\colon A \to S$ be a Shimura epimorphism as in (2.11), so that the fiber A_s over any point $s \in S$ is an abelian variety. Then the weakly special subvarieties of A that are contained in A_s are precisely the translates of abelian subvarieties of A_s.*

Proof. First let Z be a weakly special subvariety of A. Then by definition there exist Shimura morphisms

$$\begin{array}{ccccccc} T' & \xleftarrow{[\varphi]} & T & \xrightarrow{[i]} & A & \xrightarrow{[\pi]} & S \\ \| & & \| & & \| & & \| \\ \Gamma_{Q'}\backslash Y'^{+} & \longleftarrow & \Gamma_{Q}\backslash Y^{+} & \longrightarrow & (\Gamma \ltimes \Gamma_V)\backslash (X^{+} \ltimes V_{\mathbb{R}}) & \longrightarrow & \Gamma\backslash X^{+}, \end{array}$$

associated to morphisms of connected mixed Shimura data

$$(Q', Y'^{+}) \xleftarrow{\varphi} (Q, Y^{+}) \xrightarrow{i} \big(P \ltimes V_{\mathrm{alg}}, X^{+} \ltimes V_{\mathbb{R}}\big) \xrightarrow{\pi} (P, X^{+}),$$

such that Z is an irreducible component of $[i]\big([\varphi]^{-1}(t')\big)$ for some point $t' \in T'$. By Proposition 4.4 we may assume that φ is surjective, that $K := \mathrm{Ker}(\varphi)$ is

connected without a non-trivial torus quotient, and that i is injective. Observe that any irreducible component $Z^+ \subset X^+ \ltimes V_{\mathbb{R}}$ of the preimage of $Z \subset A$ is invariant under the action of the identity component $i(K(\mathbb{R})^\circ)$.

Assume that $Z \subset A_s$. Then $[\pi](Z) = \{s\}$ implies that π maps Z^+ to a point in X^+. Thus $\pi \circ i\,(K(\mathbb{R})^\circ)$ fixes that point. By the Axioms (ii–iv) of Definition 2.1 this implies that $\pi \circ i\,(K(\mathbb{R})^\circ)$ is compact. In particular $\pi \circ i\,(K)$ is reductive. As a quotient of K it is also connected without a non-trivial torus quotient; hence it is connected semisimple. Thus $\big(\pi \circ i\,(K)\big)^{\mathrm{ad}}$ is a connected semisimple quotient of Q of compact type, so by Definition 2.1 (vi) it must be trivial. Altogether this implies that $\pi \circ i\,(K) = 1$.

This means that $i(K) \subset \mathrm{Ker}(\pi) = V_{\mathrm{alg}}$. Since V_{alg} is of Hodge weight -1, so is K, and so all fibers of $Y^+ \to Y'^+$ are orbits under $K(\mathbb{R})$. Thus Z^+ is an orbit under $i(K(\mathbb{R}))$. Its image $Z \subset A_s \cong \Gamma_V \backslash V_{\mathbb{R}}$ is therefore a translate of a connected closed subgroup. Being a complex analytic subvariety, it must be a translate of an abelian subvariety, as desired.

Conversely, let Z be a translate of an abelian subvariety $B \subset A_s$. Recall that $A_s \cong \Gamma_V \backslash V_{\mathbb{R}}$ with the complex structure determined by any preimage $x \in X^+$ of s. Thus $B \cong (\Gamma_V \cap V') \backslash V'_{\mathbb{R}}$ for some $\mathbb{Q}$-subspace $V' \subset V$, and since B is also a complex subvariety, the action of $\mathbb{S}$ on $V_{\mathbb{R}}$ determined by x must leave $V'_{\mathbb{R}}$ invariant. In other words $x \in \mathrm{Hom}(\mathbb{S}_{\mathbb{C}}, P_{\mathbb{C}})$ factors through $P'_{\mathbb{C}}$ for $P' := \mathrm{Stab}_P(V')$.

Let $Q \subset P'$ be the smallest connected algebraic subgroup defined over $\mathbb{Q}$ such that x factors through $Q_{\mathbb{C}}$. (This is the Mumford–Tate group of x.) Using the Axioms of Definition 2.1 for (P, X^+) one easily checks that there is a unique connected mixed Shimura datum (Q, Y^+) with $x \in Y^+$, such that the inclusion defines a morphism $(Q, Y^+) \to (P, X^+)$. Consider the induced morphisms of connected mixed Shimura data

$$\big(Q \ltimes (V/V')_{\mathrm{alg}}, Y^+ \ltimes (V/V')_{\mathbb{R}}\big) \xleftarrow{\ \varphi\ } \big(Q \ltimes V_{\mathrm{alg}}, Y^+ \ltimes V_{\mathbb{R}}\big) \xrightarrow{\ i\ } \big(P \ltimes V_{\mathrm{alg}}, X^+ \ltimes V_{\mathbb{R}}\big)$$

and of the associated connected mixed Shimura varieties. In the fibers above the point $s = [x] \in \Gamma \backslash X^+$ these maps are simply

$$A_s/B \xleftarrow{\ [\varphi]\ } A_s \xrightarrow{\ \mathrm{id}\ } A_s.$$

Thus Z, being a translate of B, is also a fiber of $[\varphi]$, and therefore a weakly special subvariety according to Definition 4.1, as desired. □

The following fact shows that the conclusion in Conjecture 1.6 cannot be strengthened.

Proposition 4.7. *For any weakly special subvariety $Z \subset S$ and any point $s \in Z$, the intersection of Z with the Hecke orbit of s is Zariski dense in Z.*

Proof. Consider Shimura morphisms $T' \xleftarrow{[\varphi]} T \xrightarrow{[i]} S$ associated to morphisms of connected mixed Shimura data

$$(Q', Y'^+) \xleftarrow{\varphi} (Q, Y^+) \xrightarrow{i} (P, X^+),$$

and let Z be an irreducible component of $[i]\big([\varphi]^{-1}(t')\big)$ for a point $t' \in T'$. Then $Z = [i]\big(\tilde{Z}\big)$ for an irreducible component $\tilde{Z}$ of $[\varphi]^{-1}(t')$. For all $t \in \tilde{Z}$, the definition of Hecke operators implies that $[i]\big(T_q(t)\big) \subset T_{i(q)}\big([i](t)\big)$ for all $q \in Q(\mathbb{Q})^+$. Thus $[i]$ maps the Hecke orbit of t in T into the Hecke orbit of $[i](t)$ in S. It therefore suffices to prove that the intersection of $\tilde{Z}$ with the Hecke orbit of t is Zariski dense in $\tilde{Z}$.

Write $T = \Delta\backslash Y^+$ and $t = [y]$ for $y \in Y^+$. Set $K := \mathrm{Ker}(\varphi)$. Using the Hecke operators T_k for all $k \in K(\mathbb{Q}) \cap K(\mathbb{R})^\circ$ one finds that $\Delta\backslash\Delta\big(K(\mathbb{Q}) \cap K(\mathbb{R})^\circ\big)y$ is a subset of the Hecke orbit of t that is contained in $[\varphi]^{-1}(t')$. Its closure for the analytic topology contains the connected real analytic subset $\Delta\backslash\Delta K(\mathbb{R})^\circ y$. In particular this subset is contained in $\tilde{Z}$. Moreover, let $U \triangleleft Q$ be the subgroup of weight -2 from Definition 2.1 (a). Then any $K(\mathbb{R})^\circ$-invariant complex analytic subspace of Y^+ is also invariant under $K(\mathbb{R})^\circ \cdot (U \cap K)(\mathbb{C})$. Since this group also acts transitively on the fibers of $Y^+ \to Y'^+$, we deduce that the Zariski closure of $\Delta\backslash\Delta K(\mathbb{R})^\circ y$ is equal to $\tilde{Z}$. Altogether this shows that the Hecke orbit of t in T contains a Zariski dense subset of $\tilde{Z}$, as desired. □

Remark 4.8. For a usual, non-connected, Shimura variety $\tilde{S}$ the Definition 4.1 of special and weakly special subvarieties must be modified slightly, for a technical reason. The problem is that the images of Shimura morphisms may not meet all components of $\tilde{S}$. As a remedy one replaces $[i]$ by $T_p \circ [i]$ in Definition 4.1 for all Hecke operators T_p on $\tilde{S}$. The same applies to Definition 4.10 below.

Remark 4.9. One easily shows that any irreducible component of an intersection of special subvarieties of S is a special subvariety. By noetherian induction it follows that every irreducible subvariety of S is contained in a unique smallest special subvariety. The analogous assertions hold for weakly special subvarieties; the proofs are left to the reader.

Definition 4.10. *A point $x \in X^+$ and its image in S are called* special *if the homomorphism $x : \mathbb{S}_{\mathbb{C}} \to P_{\mathbb{C}}$ factors through $H_{\mathbb{C}}$ for a torus $H \subset P$ defined over $\mathbb{Q}$.*

Thus $x \in X^+$ is special if and only if it is the image of a morphism of connected mixed Shimura data $(H, Y^+) \to (P, X^+)$ where H is a torus. Correspondingly a point $s \in S$ is special if and only if $\{s\}$ is a special subvariety of dimension 0. Clearly the image of any special point under any Shimura morphism is again a special point.

Remark 4.11. For every point $s \in S$ the set $\{s\}$ is a fiber of the identity morphism $\mathrm{id}\colon S \to S$ and therefore a weakly special subvariety according to Definition 4.1 (b). Thus a concept of 'weakly special points' analogous to that of 'special points' would not be very useful.

Remark 4.12. A point on the connected Shimura variety associated to the pair $(\mathrm{CSp}_{2g,\mathbb{Q}}, \mathcal{H}_{2g})$ from Example 2.12 is special if and only if the associated abelian variety has complex multiplication. Here a non-simple abelian variety is said to have complex multiplication if and only if every simple constituent has complex multiplication.

Remark 4.13. Let V and $A \to S$ be as in Construction 2.9. Consider a special point $y \in X^+ \ltimes V_{\mathbb{R}}$, corresponding to a homomorphism $y\colon \mathbb{S}_{\mathbb{C}} \to H_{\mathbb{C}} \subset P_{\mathbb{C}} \ltimes V_{\mathrm{alg},\mathbb{C}}$ for a torus $H \subset P \ltimes V_{\mathrm{alg}}$ defined over $\mathbb{Q}$. Then H can be conjugated into P by an element of $V_{\mathrm{alg}}(\mathbb{Q}) = V$. The isomorphism $X^+ \times V_{\mathbb{R}} \cong X^+ \ltimes V_{\mathbb{R}}$ from (2.10) shows that this conjugation induces translation on $V_{\mathbb{R}}$. Thus it follows that the special points in A are precisely the torsion points in the fibers A_s over all special points $s \in S$. Compare also Proposition 5.1 (a).

Proposition 4.14. *Every special subvariety of S contains a Zariski dense subset of special points.*

Proof. Since any Shimura morphism maps special points to special points, it suffices to show that the set of special points of every connected mixed Shimura variety is Zariski dense. This is well-known, for example by [33, 11.7]. □

Proposition 4.15. *Every weakly special subvariety of S that contains a special point is a special subvariety.*

Proof. Consider Shimura morphisms $T' \xleftarrow{[\varphi]} T \xrightarrow{[i]} S$ associated to morphisms of connected mixed Shimura data

$$(Q', Y'^+) \xleftarrow{\varphi} (Q, Y^+) \xrightarrow{i} (P, X^+),$$

and let Z be an irreducible component of $[i]\big([\varphi]^{-1}(t')\big)$ for some $t' \in T'$. By Proposition 4.4 we may assume that i is injective. Let $y \in Y^+$ be a point whose image in S is a special point in Z. Then the homomorphism $i \circ y\colon \mathbb{S}_{\mathbb{C}} \to P_{\mathbb{C}}$ factors through a torus defined over $\mathbb{Q}$. Since i is injective, this implies that $y\colon \mathbb{S}_{\mathbb{C}} \to Q_{\mathbb{C}}$ factors through a torus defined over $\mathbb{Q}$. Let $H \subset Q$ be such a torus. Let K denote the identity component of $\mathrm{Ker}(\varphi)$ and consider the subgroup $Q_1 := K \cdot H \subset Q$. Then there exists a unique connected mixed Shimura datum (Q_1, Y_1^+) such that $y \in Y_1^+$, and the associated special subvariety $T_1 \subset T$ is an irreducible component of $[\varphi]^{-1}(t')$. It follows that Z is the image of T_1 and hence a special subvariety, as desired. □

Remark 4.16. Consider the weakly special subvariety $Z := [i]\big(T_1 \times \{t_2\}\big) \subset T$ from Remark 4.5. The preceding results imply that Z is special if and only if t_2 is a special point. (Compare Moonen [29, II 3.1].)

Remark 4.17. Let V and $A \to S$ be again as in Construction 2.9. The preceding results imply that a fiber A_s contains a special subvariety if and only if the base point $s \in S$ is special. Moreover, if s is special, then the special subvarieties contained in A_s are precisely the translates of abelian subvarieties by torsion points.

Theorem 4.18 (Edixhoven–Yafaev). *Conjecture 1.6 is true when S is a pure Shimura variety, $s \in S$ is a special point, and $Z \subset S$ is a curve.*

Proof. Under the given assumptions, if $Z \cap \Lambda$ is Zariski dense in Z for some generalized Hecke orbit Λ in the sense of Proposition-Definition 3.3 (a), Proposition 3.6 implies that the same is true for some generalized Hecke orbit in the sense of Remark 3.5. By Edixhoven and Yafaev [15] it thus follows that Z is a special subvariety of S. □

5 Relation with the Mordell–Lang conjecture

In this section we show that Conjecture 1.6 for subvarieties of an abelian fiber A_s above a pure Shimura variety is equivalent to the Mordell–Lang conjecture. The argument is based on a description of the generalized Hecke orbit of a point in A_s as a point on the ambient connected mixed Shimura variety.

We use the notation of Construction 2.9. Thus (P, X^+) is a connected Shimura datum and V is a representation of P of Hodge type $\{(-1,0),(0,-1)\}$, as in Example 2.12. To the extended connected mixed Shimura datum $(P \ltimes V_{\mathrm{alg}}, X^+ \ltimes V_{\mathbb{R}})$ and suitable congruence subgroups is associated a Shimura epimorphism

$$A \;:=\; (\Gamma \ltimes \Gamma_V)\backslash(X^+ \ltimes V_{\mathbb{R}}) \;\xrightarrow{[\pi]}\; \Gamma\backslash X^+ \;=:\; S.$$

Fix a point $s \in S$. Then the fiber $A_s := [\pi]^{-1}(s)$ is an abelian variety isomorphic to $\Gamma_V\backslash V_{\mathbb{R}}$ with a complex structure defined by s. Its torsion subgroup is $A_{s,\mathrm{tor}} = \Gamma_V\backslash V$.

Proposition 5.1. *Let ϖ denote the projection $V_{\mathbb{R}} \twoheadrightarrow \Gamma_V\backslash V_{\mathbb{R}} \cong A_s$. Let v be any point in $V_{\mathbb{R}}$, and set $a := \varpi(v)$.*

(a) The Hecke orbit of a as a point on A contains the subset

$$\varpi\big(\{v + v' \mid v' \in V\}\big) \;=\; \varpi(v) + A_{s,\mathrm{tor}} \;\subset\; A_s.$$

(b) The generalized Hecke orbit of a as a point on A contains the subset

$$\varpi\big(\{g(v) + v' \mid g \in \mathrm{Aut}_P(V),\, v' \in V\}\big) \;\subset\; A_s.$$

(c) If (P, X^+) is pure, the intersection with A_s of the generalized Hecke orbit of a as a point on A is contained in the subset

$$\big\{b \in A_s \;\big|\; \exists n \in \mathbb{Z} \smallsetminus \{0\} : \exists h \in \mathrm{End}(A_s) : nb = h(a)\big\}.$$

Proof. First consider the inner automorphisms of P induced by elements of V. Via the isomorphism $X^+ \times V_{\mathbb{R}} \cong X^+ \ltimes V_{\mathbb{R}}$ from (2.10) they act by translation on $V_{\mathbb{R}}$. Thus all translations by V on $\Gamma_V \backslash V_{\mathbb{R}}$ are induced by usual Hecke operators, proving (a).

For (b) we attach to $g \in \mathrm{Aut}_P(V)$ the automorphism $(p, v) \mapsto (p, g(v))$ of $P \ltimes V_{\mathrm{alg}}$. It is the identity on P, and therefore maps X^+ and hence $X^+ \ltimes V_{\mathbb{R}}$ to itself. Thus it defines a generalized Hecke operator on A. Via the isomorphism $X^+ \times V_{\mathbb{R}} \cong X^+ \ltimes V_{\mathbb{R}}$ from (2.10) it acts on the latter by $(x, v) \mapsto (x, g(v))$. Thus it acts by g on the fibers $V_{\mathbb{R}}$. Since $g \in \mathrm{Aut}_P(V)$ is arbitrary, and translations by V are already induced by usual Hecke operators, this proves (b).

For (c) consider any automorphism φ of the connected mixed Shimura datum $(P \ltimes V_{\mathrm{alg}}, X^+ \ltimes V_{\mathbb{R}})$. By assumption P is reductive; hence V_{alg} is the unipotent radical of $P \ltimes V_{\mathrm{alg}}$ and is therefore mapped to itself under φ. On the other hand, since all Levi decompositions of $P \ltimes V_{\mathrm{alg}}$ are conjugate under V, there exists $v' \in V$ such that $\psi := \mathrm{int}(v')^{-1} \circ \varphi$ maps P to itself. The generalized Hecke operator T_φ on A is then the composite of the generalized Hecke operator T_ψ with the translation by $\varpi(v')$ along the abelian fibers. In the fiber above s, the operator T_ψ consists of finitely many diagrams of the form

$$A_s \longleftarrow A'_{s'} \longrightarrow A_s$$

for finite morphisms between abelian varieties. By construction T_ψ preserves the zero sections; hence these morphisms are isogenies. Therefore $T_\psi(a) \cap A_s$ is contained in the indicated subset. Since this subset is invariant under translation by $\varpi(v') \in A_{s,\mathrm{tor}}$, the assertion follows. □

Theorem 5.2. *In the situation above, if (P, X^+) is pure, Conjecture 1.6 for subvarieties of A_s follows from the Mordell–Lang conjecture.*

Proof. Let $a \in A_s$ be any point and Λ its generalized Hecke orbit as a point on A. Let $Z \subset A_s$ be an irreducible closed algebraic subvariety such that $Z \cap \Lambda$ is Zariski dense in Z. Then Proposition 5.1 (c) implies that the intersection of Z with the division group of $\mathrm{End}(A_s)\, a \subset A_s$ is Zariski dense in Z. Since $\mathrm{End}(A_s)$ is a finitely generated $\mathbb{Z}$-module, the Mordell–Lang Conjecture 1.4 implies that Z is a translate of an abelian subvariety of A_s. Thus by Proposition 4.6 it is a weakly special subvariety of A, as desired. □

To prove the converse, for any point $a \in A_s$ we consider the set

$$\Lambda_a^* := \big\{b \in A_s \;\big|\; \exists m, n \in \mathbb{Z} \smallsetminus \{0\} : nb = ma\big\}.$$

If $a = \varpi(v)$ for $v \in V_{\mathbb{R}}$, we find that $\Lambda_a^* = \varpi(\mathbb{Q}^* \cdot v + V)$. Since the scalars $\mathbb{Q}^*$ are contained in $\mathrm{Aut}_P(V)$, Proposition 5.1 (b) shows that Λ_a^* lies inside in the generalized Hecke orbit of a. Using this one can already show that Conjecture 1.6 implies the Mordell–Lang conjecture for subgroups Λ of $\mathbb{Q}$-rank ≤ 1. For arbitrary rank we need one more technical lemma.

Lemma 5.3. *For any elements $a_1, \dots, a_r \in A_s$ satisfying $a_1 + \cdots + a_r = 0$, we have*

$$\big\{ b \in A_s \;\big|\; \exists n \in \mathbb{Z}^{>0} : nb \in \mathbb{Z}a_1 + \cdots + \mathbb{Z}a_r \big\} \;=\; \Lambda_{a_1}^* + \cdots + \Lambda_{a_r}^*.$$

Proof. The inclusion '$\supset$' is obvious. To prove '$\subset$' consider an element b of the left-hand side. Then there exist integers $n > 0$ and $n_1, \dots, n_r$ such that $nb = \sum_i n_i a_i$. The assumption implies that for any integer m we then also have $nb = \sum_i (n_i + m) a_i$. Since we can choose m so that all the coefficients $n_i + m$ are non-zero, this equation shows that b lies in the right-hand side, as desired. □

Theorem 5.4. *Conjecture 1.6 implies the Mordell–Lang conjecture.*

Proof. Every abelian variety over $\mathbb{C}$ is isomorphic to some A_s as above by Example 2.12. So consider a finitely generated subgroup $\Lambda_0 \subset A_s$. Fix generators $a_1, \dots, a_{r-1}$ of Λ_0 and set $a_r := -a_1 - \cdots - a_{r-1}$. Let A_s^r be the product of r copies of A_s, and let $\Sigma \colon A_s^r \to A_s$ denote the summation map $(b_1, \dots, b_r) \mapsto b_1 + \cdots + b_r$. Then Lemma 5.3 asserts that the division group Λ of Λ_0 satisfies

$$\Lambda \;=\; \Lambda_{a_1}^* + \cdots + \Lambda_{a_r}^* \;=\; \Sigma\big(\Lambda_{a_1}^* \times \cdots \times \Lambda_{a_r}^*\big).$$

Let V^r be the direct sum of r copies of V as a representation of P. Then the connected mixed Shimura variety associated to $P \ltimes V_{\mathrm{alg}}^r$,

$$A^r \;:=\; (\Gamma \ltimes \Gamma_V^r) \backslash (X^+ \ltimes V_{\mathbb{R}}^r),$$

is the r-fold fiber product of A with itself over S, and its fiber over s is A_s^r. Let tuples $(u_1, \dots, u_r)$ of elements of $\mathbb{Q}^*$ act on V^r by scalar multiplication on the respective direct summands. This defines P-equivariant automorphisms of V^r; hence Proposition 5.1 (b) implies that the generalized Hecke orbit of $(a_1, \dots, a_r) \in A_s^r$ as a point on the connected mixed Shimura variety A^r contains $\Lambda_{a_1}^* \times \cdots \times \Lambda_{a_r}^*$.

Let $Z \subset A$ be an irreducible closed algebraic subvariety such that $Z \cap \Lambda$ is Zariski dense in Z. Then with

$$\tilde{\Lambda} \;:=\; \Sigma^{-1}(Z) \cap (\Lambda_{a_1}^* \times \cdots \times \Lambda_{a_r}^*)$$

we find that

$$\Sigma(\tilde{\Lambda}) = Z \cap \Sigma(\Lambda_{a_1}^* \times \cdots \times \Lambda_{a_r}^*) = Z \cap \Lambda$$

is Zariski dense in Z. Let $\tilde{Z} \subset A_s^r$ denote the Zariski closure of $\tilde{\Lambda}$. Since taking Zariski closures commutes with taking images under proper morphisms, we deduce that $\Sigma(\tilde{Z}) = Z$. Choose an irreducible component Z' of $\tilde{Z}$ such that $\Sigma(Z') = Z$. Then by construction the intersection of Z' with the generalized Hecke orbit of $(a_1, \ldots, a_r)$ is Zariski dense in Z'. Thus Conjecture 1.6 for the connected mixed Shimura variety A^r implies that Z' is weakly special. Therefore $Z = \Sigma(Z')$ is weakly special, and so by Proposition 4.6 it is a translate of an abelian subvariety of A_s, as desired. □

Remark 5.5. The above results generalize easily to semi-abelian varieties, on the basis of Remark 2.13 and a suitable generalization of Proposition 4.6.

6 Hodge and Galois generic points

The first notion is opposite to that of special points.

Definition 6.1. *A point $x \in X^+$ and its image in S are called* Hodge generic *if the homomorphism $x\colon \mathbb{S}_{\mathbb{C}} \to P_{\mathbb{C}}$ does not factor through $P'_{\mathbb{C}}$ for any proper algebraic subgroup $P' \subset P$ defined over $\mathbb{Q}$.*

Remark 6.2. The Axiom (vi) of Definition 2.1 implies that for any fixed proper algebraic subgroup $P' \subset P$ defined over $\mathbb{Q}$, the set of $x \in X^+$ that factor through $P'_{\mathbb{C}} \subset P_{\mathbb{C}}$ is a proper analytic subspace. Since there are only countably many possibilities for P', their union is a subset of the first category of Baire. In particular its measure is zero. In this sense, 'most' points in X^+ and S are Hodge generic.

Next we look at Galois theoretic properties. To clarify the dependence on the congruence subgroup Γ, we write $S_\Gamma := \Gamma\backslash X^+$ and $[x]_\Gamma \in S_\Gamma$ from now on. We have already used the fact that S_Γ is an algebraic variety over $\mathbb{C}$. But it also possesses a model over a number field, determined by abelian class field theory. Fix any number field E that contains the so-called reflex field of (P, X^+) (compare [33, 11.1]). Then the non-connected mixed Shimura variety corresponding to (P, X^+) possesses a so-called canonical model over E by [33, 11.18]. This implies that S_Γ possesses a natural model over the maximal abelian extension E_{ab} of E. By abuse of notation we denote this model again by S_Γ. By [33, 11.10] it is functorial in (P, X^+) and Γ. In particular, for any automorphism φ of (P, X^+) and any congruence subgroups $\Gamma \subset P(\mathbb{Q})^+$ and $\Gamma' \subset \varphi^{-1}(\Gamma)$, the Shimura morphism $[\varphi]\colon S_{\Gamma'} \to S_\Gamma$ is defined over E_{ab}.

Let s be any $\mathbb{C}$-valued point on S_Γ. Let $K \subset \mathbb{C}$ be any finitely generated extension field of E_{ab}, such that s is defined over K. Let $\bar{K}$ denote the algebraic closure of K in $\mathbb{C}$. Then for any congruence subgroup $\Gamma' \subset \Gamma$ the Galois group

$\mathrm{Gal}(\bar{K}/K)$ acts on the inverse image of s under the morphism $[\mathrm{id}]\colon S_{\Gamma'} \to S_\Gamma$. When Γ is sufficiently small and Γ' is normal in Γ, this action corresponds to a continuous homomorphism

$$\mathrm{Gal}(\bar{K}/K) \longrightarrow \Gamma/\Gamma'.$$

Let $\mathbb{A}_f$ denote the ring of finite adèles of $\mathbb{Q}$ and $\bar{\Gamma}$ the closure of Γ in $P(\mathbb{A}_f)$. By going to the limit over all Γ' the above homomorphisms induce a continuous homomorphism

$$\mathrm{Gal}(\bar{K}/K) \longrightarrow \bar{\Gamma} \subset P(\mathbb{A}_f).$$

We denote its image by $\bar{\Gamma}_{s,K}$. Axiom (v) of Definition 2.1 implies that a subgroup of finite index of Γ is contained in $P^{\mathrm{der}}(\mathbb{Q})$, where P^{der} denotes the derived group of P. Thus after shrinking Γ and extending K we can always achieve $\bar{\Gamma}_{s,K} \subset \bar{\Gamma} \subset P^{\mathrm{der}}(\mathbb{A}_f)$.

Definition 6.3. *Let s and K be as above.*

(a) s is called Galois generic *if $\bar{\Gamma}_{s,K}$ is open in $\bar{\Gamma}$.*
(b) s is called strictly Galois generic with respect to K *if $\bar{\Gamma}_{s,K} = \bar{\Gamma}$.*

Proposition 6.4. *The notion 'Galois generic' is independent of the choice of K and E and invariant under taking images and preimages by morphisms induced by automorphisms of (P, X^+), and so in particular by generalized Hecke operators.*

Proof. The independence of K follows from the fact that for any finitely generated field extension K' of K, the natural homomorphism $\mathrm{Gal}(\bar{K}'/K') \to \mathrm{Gal}(\bar{K}/K)$ has an open image. The invariance under morphisms and generalized Hecke operators results from the construction. For the independence of E we must show that for any finite extension E' of E, the group $\bar{\Gamma}_{s,KE'_{\mathrm{ab}}}$ is open in $\bar{\Gamma}$, if $\bar{\Gamma}_{s,K}$ is open in $\bar{\Gamma}$. (This part of the proof could be formulated more naturally in the framework of non-connected Shimura varieties.)

For this we may assume that Γ is sufficiently small, so that it is torsion free and contained in $P^{\mathrm{der}}(\mathbb{Q})$. Next note that S_Γ already possesses a model S_{Γ,E_Γ} over some subfield $E_\Gamma \subset E_{\mathrm{ab}}$ which is finite over E. The theory of canonical models of non-connected mixed Shimura varieties yields a continuous homomorphism from the étale fundamental group $\pi_{1,\text{ét}}(S_{\Gamma,E_\Gamma}, s) \to P(\mathbb{A}_f)$. Similarly, s is defined over a subfield $L \subset K$ that is finitely generated over E_Γ. By the independence of K we may assume that $K = LE_{\mathrm{ab}}$ and $E' \subset L$. Setting $K' := KE'_{\mathrm{ab}}$, we have $\bar{K}' = \bar{K} = \bar{L}$ and hence the inclusions in the upper row of the following commutative diagram. The groups in the lower row are defined as the respective images:

$$\begin{array}{ccccccc}
\mathrm{Gal}(\bar{K}'/K') & \subset & \mathrm{Gal}(\bar{K}/K) & \subset & \mathrm{Gal}(\bar{L}/L) & \longrightarrow & \pi_{1,\text{ét}}(S_{\Gamma,E_\Gamma}, s) \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
\bar{\Gamma}_{s,K'} & \subset & \bar{\Gamma}_{s,K} & \subset & \bar{\Gamma}_{s,L} & \subset & P(\mathbb{A}_f).
\end{array}$$

Since $K' = KE'_{\mathrm{ab}} = LE'_{\mathrm{ab}}$ with $E' \subset L$, we deduce that $\bar{\Gamma}_{s,K'}$ contains the commutator subgroup of $\bar{\Gamma}_{s,L}$.

Assume now that $\bar{\Gamma}_{s,K}$ is open in $\bar{\Gamma}$. Then $\bar{\Gamma}_{s,K}$ contains an open subgroup of $W(\mathbb{A}_f)$, where W is the unipotent radical of P. We claim that the commutators of $\bar{\Gamma}_{s,L}$ with any open subgroup of $W(\mathbb{A}_f)$ generate an open subgroup of $W(\mathbb{A}_f)$. To prove this we may replace W by any simple subquotient. If P^{der} acts non-trivially on this subquotient, then so does any congruence subgroup of Γ and the assertion follows. If P^{der} acts trivially on this subquotient, the induced action of P/P^{der} must be non-trivial by the Axiom (iii) of Definition 2.1. The action of $\bar{\Gamma}_{s,L}$ on the subquotient is then determined by abelian class field theory applied to the torus P/P^{der}, and the claim follows from a straightforward exercise using the reciprocity law. Together this implies the claim for W; hence $\bar{\Gamma}_{s,K'}$ contains an open subgroup of $W(\mathbb{A}_f)$.

On the other hand, by strong approximation the image $\bar{\Gamma}_{\mathrm{ss}}$ of $\bar{\Gamma}$ in $(P^{\mathrm{der}}/W)(\mathbb{A}_f)$ is open in the image of the group of $\mathbb{A}_f$-valued points on the simply connected covering of P^{der}/W. From this one deduces that the commutator subgroup of any open subgroup of $\bar{\Gamma}_{\mathrm{ss}}$ is open. Therefore the image of $\bar{\Gamma}_{s,K'}$ in $(P^{\mathrm{der}}/W)(\mathbb{A}_f)$ is open in the image of $\bar{\Gamma}$. Altogether we deduce that $\bar{\Gamma}_{s,K'}$ is open in $\bar{\Gamma}$, as desired. □

Proposition 6.5. *For any $s \in S_\Gamma$ the following assertions are equivalent:*

(a) s is Galois generic.
(b) There exist a finitely generated extension $K \subset \mathbb{C}$ of E_{ab}, a congruence subgroup $\Gamma' \subset \Gamma$, and a point $s' \in S_{\Gamma'}$ mapping to s under $[\mathrm{id}]\colon S_{\Gamma'} \to S_\Gamma$, such that s' is strictly Galois generic with respect to K.
(c) Same as (b), but for every sufficiently large K, i.e., for every K containing some fixed finitely generated extension K_0 of E_{ab}.

Proof. The implication (c)⇒(b) is obvious, and the implication (b)⇒(a) follows from the equality $\bar{\Gamma}_{s',K} = \bar{\Gamma}_{s,K}$. For the remaining implication (a)⇒(c) assume that $\bar{\Gamma}_{s,K_0}$ is open in $\bar{\Gamma}$ for some finitely generated extension K_0 of E_{ab} over which s is defined. Then for every finitely generated extension $K \subset \mathbb{C}$ of K_0, the group $\bar{\Gamma}_{s,K}$ is open in $\bar{\Gamma}$. Set $\Gamma' := \bar{\Gamma}_{s,K} \cap \Gamma$. Then the construction of $\bar{\Gamma}_{s,K}$ shows that some point $s' \in S_{\Gamma'}$ above s is fixed by $\mathrm{Gal}(\bar{K}/K)$ and satisfies $\bar{\Gamma}_{s',K} = \bar{\Gamma}_{s,K} = \bar{\Gamma}'$, as desired. □

The following key fact translates Galois genericity into a property of Hecke orbits.

Proposition 6.6. *If s is strictly Galois generic with respect to K, then for every generalized Hecke operator T_φ on S_Γ as in Definition 3.1, the set $T_\varphi(s)$ is permuted transitively by $\mathrm{Gal}(\bar{K}/K)$.*

Proof. By definition s is strictly Galois generic with respect to K if and only if $\mathrm{Gal}(\bar{K}/K)$ acts transitively on the fiber of $[\mathrm{id}]\colon S_{\Gamma'} \to S_\Gamma$ above s for every congruence subgroup $\Gamma' \subset \Gamma$. Applying this to $\Gamma' = \Gamma \cap \varphi^{-1}(\Gamma)$, it follows that $\mathrm{Gal}(\bar{K}/K)$ also acts transitively on $T_\varphi(s) := [\varphi]\big([\mathrm{id}]^{-1}(s)\big)$, as desired. □

In the rest of this section we discuss the relation between Hodge and Galois genericity.

Proposition 6.7. *Every Galois generic point on S_Γ is Hodge generic.*

Proof. Consider any point $s = [x]_\Gamma \in \Gamma\backslash X^+ = S_\Gamma$. Let $P' \subset P$ be the smallest algebraic subgroup defined over $\mathbb{Q}$, such that $x : \mathbb{S}_\mathbb{C} \to P_\mathbb{C}$ factors through $P'_\mathbb{C}$, i.e., the Mumford–Tate group of x. One easily verifies that there exists a unique connected mixed Shimura datum (P', X'^+) with $x \in X'^+$. Thus with $\Gamma' := \Gamma \cap P'(\mathbb{Q})^+$, the given point s is the image of the point $s' := [x]_{\Gamma'} \in \Gamma'\backslash X'^+ =: S'_{\Gamma'}$ under the Shimura immersion induced by the inclusion $P' \hookrightarrow P$. By Proposition 6.4 we may choose E to contain the reflex fields of both (P, X^+) and (P', X'^+). Then both S_Γ and $S'_{\Gamma'}$ possess natural models over E_ab, and the functoriality of canonical models [33, 11.10] implies that the morphism $S'_{\Gamma'} \to S_\Gamma$ is defined over E_ab. Let $K \subset \mathbb{C}$ be any finitely generated extension of E_ab over which s' and hence s is defined. Then $\bar{\Gamma}_{s,K} \subset P(\mathbb{A}_f)$ is the image of the corresponding subgroup $\bar{\Gamma}_{s',K} \subset P'(\mathbb{A}_f)$. In particular, we have $\bar{\Gamma}_{s,K} \subset P'(\mathbb{A}_f)$.

Suppose now that s is Galois generic. By Proposition 6.5, after shrinking Γ we may assume that $\bar{\Gamma}_{s,K} = \bar{\Gamma}$. The above inclusion then implies that $\bar{\Gamma} \subset P'(\mathbb{A}_f)$, and thus $\Gamma \subset P'(\mathbb{Q})$. But the Axioms of Definition 2.1 imply that Γ contains a Zariski dense subgroup of P^der. Therefore $P^\mathrm{der} \subset P'$. In view of the Axiom 2.1 (vi) this shows that $P' = P$, and so s is Hodge generic, as desired. □

Conjecture 6.8. *Every Hodge generic point on S_Γ is Galois generic.*

Only a few cases of this conjecture are known. We begin with a tautological case, which shows in particular that Galois generic points exist. The most interesting case concerning the Siegel moduli space, Theorem 6.13, is due to Serre. Using Kummer theory we prove a relative version in Theorem 6.14.

Proposition 6.9. *Let s be any $\mathbb{C}$-valued point over the generic point of the model S_Γ over E_ab. Then s is Galois generic.*

Proof. Let K denote the function field of S_Γ as an algebraic variety over E_ab. Then s can be defined over K, and by construction we have $\bar{\Gamma}_{s,K} = \bar{\Gamma}$. Thus s is strictly Galois generic with respect to K. □

Remark 6.10. Conjecture 6.8 is related to the Mumford–Tate conjecture, as follows. Suppose that S_Γ is a Shimura subvariety of some Siegel moduli space and that $s \in S_\Gamma$ is a Hodge generic point that can be defined over a number field L. Then P is the Mumford–Tate group of the abelian variety associated to s, and the Mumford–Tate conjecture says that the image of $\mathrm{Gal}(\bar{L}/L)$ in $P(\mathbb{Q}_\ell)$ is open for any rational prime ℓ. Setting $K := LE_{\mathrm{ab}}$, this statement implies that the image of $\bar{\Gamma}_{s,K} \subset P(\mathbb{A}_f)$ under the projection to $P(\mathbb{Q}_\ell)$ is an open subgroup of $P^{\mathrm{der}}(\mathbb{Q}_\ell)$. An adèlic refinement of the Mumford–Tate conjecture would imply Conjecture 6.8 in this case.

In the following case this refinement was proved by Serre [42, Cor. de Th. 3], [41, §7, Cor. de Th. 3, Compl. 8.1]:

Theorem 6.11 (Serre). *Let A be an abelian variety over a finitely generated extension L of $\mathbb{Q}$. Assume that $\mathrm{End}(A_{\bar{L}}) = \mathbb{Z}$ and that $\dim A$ is odd or equal to 2 or 6. Then the continuous homomorphism*

$$\mathrm{Gal}(\bar{L}/L) \longrightarrow \mathrm{CSp}_{2g}(\mathbb{A}_f)$$

describing the Galois action on the adèlic Tate module of A has an open image.

Remark 6.12. The general case of Theorem 6.11 can be reduced to the number field case, as follows. Identify A with the generic fiber of an abelian scheme $\mathcal{A}$ over an integral scheme $\mathcal{X}$ of finite type over $\mathbb{Q}$ with function field L. Then the above homomorphism factors through the étale fundamental group of $\mathcal{X}$, and for any point $x \in \mathcal{X}$ with residue field L_x the homomorphism associated to the fiber $\mathcal{A}_x$ is the composite

$$\mathrm{Gal}(\bar{L}_x/L_x) \longrightarrow \pi_{1,\text{ét}}(\mathcal{X}) \longrightarrow \mathrm{CSp}_{2g}(\mathbb{A}_f).$$

In particular, the image of $\mathrm{Gal}(\bar{L}_x/L_x)$ is contained in the image of $\mathrm{Gal}(\bar{L}/L)$. By Hilbert irreducibility, as explained by Serre [40, §1], one can choose a closed point $x \in X$ such that the two images in $\mathrm{CSp}_{2g}(\mathbb{Q}_\ell)$ coincide for some prime ℓ. The Tate conjecture for endomorphisms then implies that $\mathrm{End}(\mathcal{A}_{x,\bar{L}_x}) = \mathrm{End}(A_{\bar{L}}) = \mathbb{Z}$. As L_x is a number field, Theorem 6.11 for $\mathcal{A}_x$ shows that the image of $\mathrm{Gal}(\bar{L}_x/L_x)$ is open in $\mathrm{CSp}_{2g}(\mathbb{A}_f)$; hence the same follows for $\mathrm{Gal}(\bar{L}/L)$, as desired.

Theorem 6.13. *Let S_Γ be the connected Shimura variety associated to $\mathrm{CSp}_{2g,\mathbb{Q}}$ and any Γ, as in Example 2.12. Assume that g is odd or equal to 2 or 6. Then any Hodge generic point on S_Γ is Galois generic.*

Proof. Let A_s be the abelian variety associated to a Hodge generic point $s \in S_\Gamma$. Then $\mathrm{End}(A_{s,\mathbb{C}}) = \mathbb{Z}$. Let $L \subset \mathbb{C}$ be a finitely generated extension of $\mathbb{Q}$ over which A_s is defined. Then by Theorem 6.11 the homomorphism $\mathrm{Gal}(\bar{L}/L) \to \mathrm{CSp}_{2g}(\mathbb{A}_f)$ has an open image. Since the commutator subgroup of any open subgroup of $\mathrm{CSp}_{2g}(\mathbb{A}_f)$ is an open subgroup of $\mathrm{Sp}_{2g}(\mathbb{A}_f)$, with

$K := LE_{\mathrm{ab}}$ it follows that the restriction $\mathrm{Gal}(\bar{K}/K) \to \mathrm{Sp}_{2g}(\mathbb{A}_f)$ has an open image. This image is precisely the group $\bar{\Gamma}_{s,K}$. As any sufficiently small congruence subgroup Γ of $\mathrm{CSp}_{2g}(\mathbb{Q})$ is contained in $\mathrm{Sp}_{2g}(\mathbb{Q})$, it follows that $\bar{\Gamma}_{s,K}$ is open in $\bar{\Gamma}$, as desired. □

The following result establishes a relative version of Conjecture 6.8:

Theorem 6.14. *Let $A \to S$ be as in Construction 2.9, and let $a \in A$ be a Hodge generic point. Then its image $s \in S$ is Hodge generic, and if s is Galois generic, then so is a.*

Proof. The first assertion follows from the definition of Hodge genericity. For the second assume first that $A \to S$ and a are all defined over a number field L. Then with the notation of Construction 2.9 the Galois action on the preimages of a on coverings of A corresponds to a continuous homomorphism

$$\mathrm{Gal}(\bar{L}/L) \longrightarrow P(\mathbb{A}_f) \ltimes (V \otimes \mathbb{A}_f).$$

Let Δ denote its image. With $K := LE_{\mathrm{ab}}$ we deduce that $\bar{\Gamma}_{a,K}$ is a normal subgroup of Δ containing the commutator subgroup of Δ.

Since a is Hodge generic, the subgroup $\mathbb{Z}a \subset A_s$ is Zariski dense. Thus a theorem of Ribet [38] in the form of Hindry [20, §2, Prop. 1] on the Kummer theory of A_s implies that $\Delta \cap (V \otimes \mathbb{A}_f)$ is open in $V \otimes \mathbb{A}_f$. As in the proof of Proposition 6.4 we deduce that commutators of Δ with $\Delta \cap (V \otimes \mathbb{A}_f)$ again generate an open subgroup of $V \otimes \mathbb{A}_f$. Therefore $\bar{\Gamma}_{a,K} \cap (V \otimes \mathbb{A}_f)$ is open in $V \otimes \mathbb{A}_f$. On the other hand, the Galois genericity of s means that the image $\bar{\Gamma}_{s,K}$ of $\bar{\Gamma}_{a,K}$ in $P(\mathbb{A}_f)$ is an open subgroup of $\bar{\Gamma} \subset P^{\mathrm{der}}(\mathbb{A}_f)$. Together it follows that $\bar{\Gamma}_{a,K}$ is an open subgroup of $\bar{\Gamma} \ltimes \bar{\Gamma}_V \subset P^{\mathrm{der}}(\mathbb{A}_f) \ltimes (V \otimes \mathbb{A}_f)$, and so a is Galois generic, as desired.

The case where a is defined over an arbitrary finitely generated extension of $\mathbb{Q}$ can be reduced to the number field case exactly as in Remark 6.12. □

7 Galois generic points and equidistribution

In this section we deduce Conjecture 1.6 in certain cases for the Siegel moduli space from equidistribution results by Clozel, Oh, and Ullmo. As a preparation we fix some notation. Let $P := \mathrm{CSp}_{2g,\mathbb{Q}}$ for $g \geq 1$, let $Z \cong \mathbb{R}^*$ denote the center of $P(\mathbb{R})$, and set $Y := P(\mathbb{R})/Z$. As Z is a normal subgroup of $P(\mathbb{R})$, there are natural left and right actions of $P(\mathbb{R})$ on Y. For any congruence subgroup $\Gamma \subset P(\mathbb{Q})$ and any element $p \in P(\mathbb{Q})$ we have a Hecke operator

$$\begin{array}{ccccc} \Gamma\backslash Y & \xleftarrow{[\mathrm{id}]} & (\Gamma \cap p^{-1}\Gamma p)\backslash Y & \xrightarrow{[p\cdot\,]} & \Gamma\backslash Y, \\ {[y]}_{\Gamma} & \longleftarrow\!\shortmid & {[y]}_{\Gamma\cap p^{-1}\Gamma p} & \shortmid\!\longrightarrow & {[py]}_{\Gamma}, \end{array}$$

which maps an element $[y]_\Gamma$ to the finite set

$$T_p([y]_\Gamma) \;:=\; [p\cdot]\big([\mathrm{id}]^{-1}([y]_\Gamma)\big) \;=\; \big\{[p\gamma y]_\Gamma \;\big|\; \gamma \in (\Gamma \cap p^{-1}\Gamma p)\backslash\Gamma\big\}.$$

This Hecke operator also acts on functions f on $\Gamma\backslash Y$ by

$$T_p f\big([y]_\Gamma\big) \;:=\; \frac{1}{\deg_\Gamma(p)} \cdot \sum_{\gamma\in(\Gamma\cap p^{-1}\Gamma p)\backslash\Gamma} f\big([p\gamma y]_\Gamma\big),$$

where

$$\deg_\Gamma(p) \;:=\; |\Gamma\backslash\Gamma p\Gamma| \;=\; [\Gamma : \Gamma\cap p^{-1}\Gamma p].$$

Clozel, Oh, and Ullmo [6, Thm. 1.6, Rem. (3)] have proved:

Theorem 7.1 (Clozel–Oh–Ullmo). *In the above situation assume that $\Gamma = P(\mathbb{Z})$. Let (p_n) be a sequence in $P(\mathbb{Q})$ such that $\lim_{n\to\infty}\deg_\Gamma(p_n) = \infty$. Let $d\mu$ denote the right $P(\mathbb{R})$-invariant measure on $\Gamma\backslash Y$ with total volume 1. Then for any element $[y]_\Gamma \in \Gamma\backslash Y$ the sets $T_{p_n}([y]_\Gamma)$ are equidistributed with respect to $d\mu$, in the sense that*

$$\lim_{n\to\infty} T_{p_n} f\big([y]_\Gamma\big) \;=\; \int_{\Gamma\backslash Y} f(\bar y)d\mu(\bar y)$$

for any continuous function f on Y with compact support.

Corollary 7.2. *In Theorem 7.1, the set $\bigcup_{n=1}^\infty T_{p_n}([y]_\Gamma)$ is dense in $\Gamma\backslash Y$.*

Results like these should really be true for all Γ and can presumably be proved along the lines of [6] and [7]. For our purposes a weaker consequence for all Γ will suffice, which we can deduce from the above as follows.

Let (P, X^+) denote the connected Shimura datum with $P = \mathrm{CSp}_{2g,\mathbb{Q}}$ from Example 2.12. Then $P(\mathbb{R})^+$ is the identity component of $P(\mathbb{R})$, which has index 2, and it contains the center Z. Thus $Y^+ := P(\mathbb{R})^+/Z$ is one of two connected components of Y. For any congruence subgroup $\Gamma \subset P(\mathbb{Q})^+ := P(\mathbb{Q}) \cap P(\mathbb{R})^+$ and any $p \in P(\mathbb{Q})^+$ we can define a Hecke operator T_p,

$$\Gamma\backslash Y^+ \xleftarrow{[\mathrm{id}]} (\Gamma\cap p^{-1}\Gamma p)\backslash Y^+ \xrightarrow{[p\cdot]} \Gamma\backslash Y^+$$

in the same way as above. Fix a sequence (p_n) in $P(\mathbb{Q})^+$. For any congruence subgroup $\Gamma \subset P(\mathbb{Q})^+$ we set

$$\Theta_\Gamma \;:=\; \bigcup_{n=1}^\infty T_{p_n}\big([1]_\Gamma\big) \;\subset\; \Gamma\backslash Y^+$$

and let $\bar\Theta_\Gamma$ denote its closure in $\Gamma\backslash Y^+$. The following lemma concerns its dependence on Γ.

Lemma 7.3. *For any two congruence subgroups Γ, $\Gamma' \subset P(\mathbb{Q})^+$,*

(a) Θ_Γ is finite if and only if $\Theta_{\Gamma'}$ is finite, and
(b) $\bar{\Theta}_\Gamma$ is a null set if and only if $\bar{\Theta}_{\Gamma'}$ is a null set.

Proof. It suffices to prove this when $\Gamma' \subset \Gamma$. Write $\Gamma = \bigcup_{i=1}^r \Gamma'\gamma_i$ with representatives $\gamma_i \in \Gamma$, and let $\pi\colon \Gamma'\backslash Y^+ \to \Gamma\backslash Y^+$ denote the natural projection. Then the calculation

$$T_{p_n}\big([1]_\Gamma\big) = \Gamma\backslash\Gamma p_n \Gamma Z/Z \;=\; \bigcup_{i=1}^r \Gamma\backslash\Gamma p_n \Gamma'\gamma_i Z/Z \;=\; \bigcup_{i=1}^r \pi\big(\Gamma'\backslash\Gamma' p_n \Gamma'\gamma_i Z/Z\big)$$
$$= \bigcup_{i=1}^r \pi\big(T_{p_n}([\gamma_i]_{\Gamma'})\big) \;=\; \bigcup_{i=1}^r \pi\big(T_{p_n}([1]_{\Gamma'})\cdot\gamma_i\big)$$

shows that $\Theta_\Gamma = \bigcup_{i=1}^r \pi\big(\Theta_{\Gamma'}\cdot\gamma_i\big)$. Since π is a finite (possibly ramified) covering, this implies (a). It also implies that $\bar{\Theta}_\Gamma = \bigcup_{i=1}^r \pi\big(\bar{\Theta}_{\Gamma'}\cdot\gamma_i\big)$ and thus (b). □

Using this we can deduce the following weak analogue of Corollary 7.2:

Proposition 7.4. *If $\bar{\Theta}_\Gamma$ is a null set, then it is finite.*

Proof. By Lemma 7.3 the assertion is independent of Γ. Thus it suffices to prove it for $\Gamma = P(\mathbb{Z})^+ := P(\mathbb{Z}) \cap P(\mathbb{R})^+ = \mathrm{Sp}_{2g}(\mathbb{Z})$. Here the natural map $P(\mathbb{Z})^+\backslash Y^+ \to P(\mathbb{Z})\backslash Y$ is bijective. Thus if $\lim_{n\to\infty} \deg_\Gamma(p_n) = \infty$, Corollary 7.2 shows that $\bar{\Theta}_\Gamma$ is not a null set. The same conclusion holds if $\limsup_{n\to\infty} \deg_\Gamma(p_n) = \infty$, because we may pass to a subsequence that diverges to ∞, which can only make $\bar{\Theta}_\Gamma$ smaller. If $\deg_\Gamma(p_n)$ is bounded, it is well-known that all p_n lie in finitely many cosets of ΓZ. Then the T_{p_n} form only finitely many distinct Hecke operators, and so Θ_Γ is finite. In all cases the desired implication holds. □

Theorem 7.5. *Let S_Γ be the connected Shimura variety associated to $\mathrm{CSp}_{2g,\mathbb{Q}}$ and any Γ, as in Example 2.12. Let $s \in S_\Gamma$ be a point and (p_n) a sequence in $\mathrm{CSp}_{2g}(\mathbb{Q})$. Then the subset $\bigcup_{n=1}^\infty T_{p_n}(s)$ of S_Γ is either finite or Zariski dense.*

Proof. Write $s = [x]_\Gamma$ for some $x \in X^+$, and consider the map

$$\kappa\colon\ \Gamma\backslash Y^+ \longrightarrow\!\!\!\!\!\rightarrow\ \Gamma\backslash X^+ = S_\Gamma,\ [y]_\Gamma \mapsto [yx]_\Gamma.$$

By the definition of Hecke operators it satisfies $\kappa\big(T_{p_n}([1]_\Gamma)\big) = T_{p_n}\big(\kappa([1]_\Gamma)\big) = T_{p_n}(s)$ and therefore $\kappa(\Theta_\Gamma) = \bigcup_{n=1}^\infty T_{p_n}(s)$. Consider a proper closed subvariety $Z \subset S_\Gamma$. If $\kappa(\Theta_\Gamma) \subset Z$, we have $\Theta_\Gamma \subset \kappa^{-1}(Z)$ and hence $\bar{\Theta}_\Gamma \subset \kappa^{-1}(Z)$. On the other hand, the irreducibility of S_Γ implies that Z and hence $\kappa^{-1}(Z)$ is a null set. Thus if $\kappa(\Theta_\Gamma)$ is not Zariski dense, it follows that $\bar{\Theta}_\Gamma$ is a null set. By Proposition 7.4 it is then finite, and so $\kappa(\Theta_\Gamma)$ is finite, as desired. □

Finally, we can settle Conjecture 1.6 in the following case.

Theorem 7.6. *Let S_Γ be the connected Shimura variety associated to* $\mathrm{CSp}_{2g,\mathbb{Q}}$ *and any Γ, as in Example 2.12. Let $s \in S_\Gamma$ be a Galois generic point. Then any infinite subset of the generalized Hecke orbit of s is Zariski dense in S_Γ. In particular, Conjecture 1.6 is true for $s \in S = S_\Gamma$ and any Z.*

Proof. First note that every automorphism of the connected Shimura datum (P, X^+) here is an inner automorphism. Thus every generalized Hecke operator is a usual Hecke operator, and so the generalized Hecke orbit of s coincides with the usual Hecke orbit. Consider any closed algebraic subvariety $Z \subset S_\Gamma$ containing an infinite subset of the Hecke orbit of s. We must prove that $Z = S_\Gamma$.

Let $E \subset E_{\mathrm{ab}} \subset \mathbb{C}$ be as in Section 6, and choose a finitely generated extension $K_0 \subset \mathbb{C}$ of E_{ab} such that Z is defined over K_0. By Proposition 6.5 (a)⇒(c) there exist a finitely generated extension $K \subset \mathbb{C}$ of K_0, a congruence subgroup $\Gamma' \subset \Gamma$, and a point $s' \in S_{\Gamma'}$ mapping to s under $\pi := [\mathrm{id}] \colon S_{\Gamma'} \to S_\Gamma$, such that s' is strictly Galois generic with respect to K. The inverse image under π of the Hecke orbit of s is then simply the Hecke orbit of s'. Thus $Z' := \pi^{-1}(Z)$ is a closed algebraic subvariety of $S_{\Gamma'}$ defined over K_0 that contains an infinite subset of the Hecke orbit of s'. It suffices to prove that $Z' = S_{\Gamma'}$. Thus after replacing Γ, s, Z by Γ', s', Z' we may assume that s is strictly Galois generic with respect to K. Then Proposition 6.6 says that for every Hecke operator T_p on S_Γ, the set $T_p(s)$ is permuted transitively by $\mathrm{Gal}(\bar{K}/K)$.

Since Z contains an infinite subset of the Hecke orbit of s, there exists a sequence (p_n) in $P(\mathbb{Q})$ such that Z contains a point from $T_{p_n}(s)$ for all n and that all these points are distinct. As Z is defined over K_0 and hence over K, the transitivity of $\mathrm{Gal}(\bar{K}/K)$ then implies that Z contains the whole set $\bigcup_{n=1}^{\infty} T_{p_n}(s)$. By construction this set is infinite; hence by Theorem 7.5 it is Zariski dense in S_Γ. Therefore $Z = S_\Gamma$, as desired.

Finally, in the situation of Conjecture 1.6, this shows that either $Z = S_\Gamma$, or Z is finite and hence a point. In the former case it is special by tautology, in the latter it is weakly special by Remark 4.11. □

Remarks 7.7.

(a) Proposition 6.9 and Theorem 6.13 show that Theorem 7.6 is not vacuous.

(b) In Theorem 7.5 one should expect that the subset $\bigcup_{n=1}^{\infty} T_{p_n}(s)$ of S_Γ is either finite or dense for the analytic topology. In fact, in the case $\Gamma = P(\mathbb{Z})^+$ this is a consequence of Theorem 7.1.

(c) Clozel, Oh, and Ullmo [6, Thm. 1.6] prove an analogue of Theorem 7.1 for most connected almost simple simply connected semisimple groups over $\mathbb{Q}$ with arbitrary congruence subgroups. A version for all *adjoint* groups

would apply to more Hecke operators and hence quickly yield generalizations of Theorems 7.5 and 7.6 to arbitrary connected pure Shimura varieties. Such a version can presumably be proved along the lines of [6] and [7].

(d) Clozel has pointed out to the author that the desired extension to most adjoint groups might also be deduced from [6, Cor. 1.4], using the fact that for every Hecke operator induced by an element of the adjoint group, some positive power is a linear combination of Hecke operators induced by elements of the simply connected covering.

(e) An overview of various other aspects of equidistribution is given in Ullmo [45].

References

1. Y. André – *G-functions and geometry*, Aspects of Mathematics, E13, Friedr. Vieweg & Sohn, Braunschweig, 1989.
2. — , "Finitude des couples d'invariants modulaires singuliers sur une courbe algébrique plane non modulaire", *J. Reine Angew. Math.* **505** (1998), p. 203–208.
3. — , "Shimura varieties, subvarieties, and CM points", Six lectures at the University of Hsinchu (Taiwan), August–September 2001 (with an appendix by C.-L. Chai), `http://www.math.umd.edu/~yu/notes.shtml`.
4. F. A. Bogomolov – "Points of finite order on abelian varieties", *Izv. Akad. Nauk SSSR Ser. Mat.* **44** (1980), no. 4, p. 782–804.
5. E. Bombieri – "The Mordell conjecture revisited", *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **17** (1990), no. 4, p. 615–640.
6. L. Clozel, H. Oh and E. Ullmo – "Hecke operators and equidistribution of Hecke points", *Invent. Math.* **144** (2001), no. 2, p. 327–351.
7. L. Clozel and E. Ullmo – "Équidistribution des points de Hecke", Contributions to automorphic forms, geometry, and number theory, Johns Hopkins Univ. Press, Baltimore, MD, 2004, p. 193–254.
8. R. F. Coleman – "Ramified torsion points on curves", *Duke Math. J.* **54** (1987), no. 2, p. 615–640.
9. P. Deligne – "Travaux de Shimura", Séminaire Bourbaki, 23ème année (1970/71), Exp. No. 389, Springer, Berlin, 1971, p. 123–165. Lecture Notes in Math., Vol. 244.
10. — , "Variétés de Shimura: interprétation modulaire, et techniques de construction de modèles canoniques", Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 2, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979, p. 247–289.
11. P. Deligne, J. S. Milne, A. Ogus and K.-y. Shih – *Hodge cycles, motives, and Shimura varieties*, Lecture Notes in Mathematics, vol. 900, Springer-Verlag, Berlin, 1982.
12. B. Edixhoven and J.-H. Evertse (eds.) – *Diophantine approximation and abelian varieties*, Lecture Notes in Mathematics, vol. 1566, Springer-Verlag,

Berlin, 1993, Introductory lectures, Papers from the conference held in Soesterberg, April 12–16, 1992.
13. B. Edixhoven – "Special points on the product of two modular curves", *Compositio Math.* **114** (1998), no. 3, p. 315–328.
14. — , "On the André-Oort conjecture for Hilbert modular surfaces", Moduli of abelian varieties (Texel Island, 1999), Progr. Math., vol. 195, Birkhäuser, Basel, 2001, p. 133–155.
15. B. Edixhoven and A. Yafaev – "Subvarieties of Shimura varieties", *Ann. of Math. (2)* **157** (2003), no. 2, p. 621–645.
16. G. Faltings – "Endlichkeitssätze für abelsche Varietäten über Zahlkörpern", *Invent. Math.* **73** (1983), no. 3, p. 349–366, Erratum: **75** no. 2 (1984).
17. — , "Finiteness theorems for abelian varieties over number fields", Arithmetic geometry (Storrs, Conn., 1984) (G. Cornell and J. H. Silverman, eds.), Springer, New York, 1986, p. 9–27.
18. — , "Diophantine approximation on abelian varieties", *Ann. of Math. (2)* **133** (1991), no. 3, p. 549–576.
19. — , "The general case of S. Lang's conjecture", Barsotti Symposium in Algebraic Geometry (Abano Terme, 1991), Perspect. Math., vol. 15, Academic Press, San Diego, CA, 1994, p. 175–182.
20. M. Hindry – "Autour d'une conjecture de Serge Lang", *Invent. Math.* **94** (1988), no. 3, p. 575–603.
21. — , "Sur les conjectures de Mordell et Lang (d'après Vojta, Faltings et Bombieri)", *Astérisque* (1992), no. 209, p. 11, 39–56, Journées Arithmétiques, 1991 (Geneva).
22. M. Hindry and J. H. Silverman – *Diophantine geometry*, Graduate Texts in Mathematics, vol. 201, Springer-Verlag, New York, 2000.
23. E. Hrushovski – "The Manin-Mumford conjecture and the model theory of difference fields", *Ann. Pure Appl. Logic* **112** (2001), no. 1, p. 43–115.
24. S. Lang – "Division points on curves", *Ann. Mat. Pura Appl. (4)* **70** (1965), p. 229–234.
25. — , *Fundamentals of Diophantine geometry*, Springer-Verlag, New York, 1983.
26. M. McQuillan – "Division points on semi-abelian varieties", *Invent. Math.* **120** (1995), no. 1, p. 143–159.
27. J. S. Milne – "The action of an automorphism of **C** on a Shimura variety and its special points", Arithmetic and geometry, Vol. I, Progr. Math., vol. 35, Birkhäuser Boston, Boston, MA, 1983, p. 239–265.
28. — , "Canonical models of (mixed) Shimura varieties and automorphic vector bundles", Automorphic forms, Shimura varieties, and L-functions, Vol. I (Ann Arbor, MI, 1988), Perspect. Math., vol. 10, Academic Press, Boston, MA, 1990, p. 283–414.
29. B. Moonen – "Special points and linearity properties of Shimura varieties", Thesis, University of Utrecht, 1995.
30. — , "Models of Shimura varieties in mixed characteristics", Galois representations in arithmetic algebraic geometry (Durham, 1996), London Math. Soc. Lecture Note Ser., vol. 254, Cambridge Univ. Press, Cambridge, 1998, p. 267–350.
31. L. Mordell – "On the rational solutions of the indeterminate equations of the third and fourth degrees", *Proc. Camb. Philos. Soc.* **21** (1922), p. 179–192.

32. F. Oort – "Canonical liftings and dense sets of CM-points", Arithmetic geometry (Cortona, 1994), Sympos. Math., XXXVII, Cambridge Univ. Press, Cambridge, 1997, p. 228–234.
33. R. Pink – "Arithmetical compactification of mixed shimura varieties", Thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 1989, Bonner Mathematische Schriften, **209** (1990).
34. R. Pink and D. Roessler – "On Hrushovski's proof of the Manin-Mumford conjecture", *Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002)* (Beijing), Higher Ed. Press, 2002, p. 539–546.
35. M. Raynaud – "Around the Mordell conjecture for function fields and a conjecture of Serge Lang", Algebraic geometry (Tokyo/Kyoto, 1982), Lecture Notes in Math., vol. 1016, Springer, Berlin, 1983, p. 1–19.
36. — , "Courbes sur une variété abélienne et points de torsion", *Invent. Math.* **71** (1983), no. 1, p. 207–233.
37. — , "Sous-variétés d'une variété abélienne et points de torsion", Arithmetic and geometry, Vol. I, Progr. Math., vol. 35, Birkhäuser Boston, Boston, MA, 1983, p. 327–352.
38. K. A. Ribet – "Kummer theory on extensions of abelian varieties by tori", *Duke Math. J.* **46** (1979), no. 4, p. 745–761.
39. D. Roessler – "A note on the Manin–Mumford conjecture", 2004, preprint.
40. J.-P. Serre – "Lettre à Ken Ribet du 1/1/1981 et du 29/1/1981", Oeuvres vol. IV, p. 1–20, Springer, 2000.
41. — , "Lettre à Marie-France Vignéras du 10/2/1986", Oeuvres vol. IV, p. 38–55, Springer, 2000.
42. — , "Résumé des cours de 1985–1986", Oeuvres vol. IV, p. 33–37, Springer, 2000.
43. G. Shimura and Y. Taniyama – *Complex multiplication of abelian varieties and its applications to number theory*, Publications of the Mathematical Society of Japan, vol. 6, The Mathematical Society of Japan, Tokyo, 1961.
44. E. Ullmo – "Positivité et discrétion des points algébriques des courbes", *Ann. of Math. (2)* **147** (1998), no. 1, p. 167–179.
45. — , "Théorie ergodique et géométrie arithmétique", *Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002)* (Beijing), Higher Ed. Press, 2002, p. 197–206.
46. P. Vojta – "Siegel's theorem in the compact case", *Ann. of Math. (2)* **133** (1991), no. 3, p. 509–548.
47. — , "Applications of arithmetic algebraic geometry to Diophantine approximations", Arithmetic algebraic geometry (Trento, 1991), Lecture Notes in Math., vol. 1553, Springer, Berlin, 1993, p. 164–208.
48. A. Weil – "L'arithmétique sur les courbes algébriques", *Acta Math.* **52** (1928), p. 281–315.
49. G. Wüstholz – "New developments in diophantine and arithmetic algebraic geometry", Aspects of Mathematics, E6, p. x+311, Friedr. Vieweg & Sohn, Braunschweig, third ed., 1992, Appendix to "Papers from the seminar held at the Max-Planck-Institut für Mathematik", Bonn/Wuppertal, 1983/1984.
50. A. Yafaev – "On a result of Ben Moonen on the moduli space of principally polarised abelian varieties", to appear in *Compositio Math.*
51. — , "A conjecture of Yves André", 2003, `math.NT/0302125`.
52. S.-W. Zhang – "Equidistribution of small points on abelian varieties", *Ann. of Math. (2)* **147** (1998), no. 1, p. 159–165.

Motivic approach to limit sheaves

Markus Spitzweck

Mathematisches Institut, Bunsenstr. 3-5, 37073 Göttingen, Germany
spitz@uni-math.gwdg.de

Summary. We propose a motivic analog of limit mixed Hodge structures. Working in the context of triangulated categories of motivic objects on schemes we introduce and study a limit motive functor and a motivic vanishing cycle sheaf.

1 Introduction

In this paper we explain the construction of *limit motives*, following [11]. These are motivic analogues of limit mixed Hodge structures considered by Schmid, Steenbrink, Zucker, El Zein, Saito and others. The idea of limit MHSs (coming from a geometric situation) is the following: Let $D := \{z \in \mathbb{C} \mid |z| < 1\}$, $D^\circ := D \setminus \{0\}$ and $f : X \to D$ be a proper family of complex algebraic manifolds such that $f^\circ : X^\circ := f^{-1}(D^\circ) \to D^\circ$ is smooth and $Y := f^{-1}(0)$ is a divisor with normal crossings in X. For any $t \in D^\circ$, let $X_t := f^{-1}(t)$. The cohomology groups $H^n(X_t, \mathbb{C})$ are part of variations of pure Hodge structures and there is a way to put a mixed Hodge structure on $\lim_{t\to 0} H^n(X_t, \mathbb{C})$ depending on the direction in which t moves to 0 such that the weight filtration is given in terms of the monodromy action around 0. Considerations in [3] suggest that these limit Hodge structures are fibers of a unipotent variation of mixed Hodge structures on $\mathbb{C}^*$, the pointed tangent space of D at 0.

Let us now switch to the algebraic/motivic context. Let C be a smooth curve over a field $k \subset \mathbb{C}$, $c_0 \in C(k)$ and $C^\circ := C \setminus \{c_0\}$. Let $\mathcal{F}$ be a motivic sheaf on C°, i.e., an object of a certain triangulated category with suitable Hodge realizations. To any such $\mathcal{F}$ we would like to associate a motivic sheaf $\widetilde{\mathcal{F}}$ on T°_{C,c_0}, the pointed tangent space of c_0 in C, satisfying some natural properties.

Let $f : X \to C$ be a proper morphism such that $f^\circ : X^\circ := f^{-1}(C^\circ) \to C^\circ$ is smooth and $Y := f^{-1}(c_0) \subset X$ is a divisor with normal crossings. Let $\mathcal{F} := \mathbf{R}f_*\mathbb{Z}(0)$. Then the Hodge realizations of $\widetilde{\mathcal{F}}_t$, $t \in T^\circ_{C,c_0}(\mathbb{C})$, should give the limit Hodge structures of the $H^n(X(\mathbb{C})_c, \mathbb{C})$, $c \in C^\circ(\mathbb{C})$.

In this paper we consider the unipotent case, i.e., the resulting sheaf on the pointed tangent space has unipotent monodromy around zero (appropriately formulated) and the functor defined here will only be the right one if the monodromy of the sheaf we start with is also unipotent.

One basic ingredient for the motivic construction is the usage of algebra and module structures on motivic objects in a homotopical context and functoriality of these.

We now sketch the approach in a toy model: For a manifold M let $\mathsf{DM}(M)$ be the derived category of the category of sheaves of abelian groups on M. Let $\mathsf{UDM}(M)$ be the smallest full triangulated subcategory of $\mathsf{DM}(M)$ containing the constant sheaf $\mathbb{Z}$ and closed under arbitrary sums - this will be the subcategory of generalized unipotent objects.

Let S be a 2-dimensional real manifold, $s_0 \in S$ and $S^\circ = S \setminus \{s_0\}$. Let D be a small disc around s_0 in S and $D^\circ = D \setminus \{s_0\}$. In this topological context we can easily identify $\mathsf{UDM}(D^\circ)$ and $\mathsf{UDM}(T^\circ_{S,s_0})$ via a choice of an inclusion $D \hookrightarrow T_{S,s_0}$ inducing the identity on the tangent spaces at s_0 and 0.

Let $\mathcal{F} \in \mathsf{DM}(S^\circ)$ be such that $\mathcal{F}|_{D^\circ} \in \mathsf{UDM}(D^\circ)$. We could define $\widetilde{\mathcal{F}}$ as the image of $\mathcal{F}|_{D^\circ}$ in $\mathsf{UDM}(T^\circ_{S,x})$. However, this assignment would not be algebraic if we were considering S as a complex algebraic curve. Instead we can do the following: Let $i : \{s_0\} \to S$ be the closed and $j : S^\circ \to S$ the open inclusion. Let $p : T^\circ_{S,s_0} \to \{s_0\}$ be the projection. Then there are two main observations:

- There is a canonical isomorphism $i^*\mathbf{R}j_*\mathbb{Z} \cong \mathbf{R}p_*\mathbb{Z}$ in $\mathsf{DM}(\{s_0\})$. The homotopical algebra machinery developed in [11] allows one to construct this isomorphism as an isomorphism of "commutative" algebra objects in a suitable sense. It follows that there is a canonical equivalence

$$\mathsf{D}((i^*\mathbf{R}j_*\mathbb{Z})\text{–}\mathsf{Mod}) \sim \mathsf{D}((\mathbf{R}p_*\mathbb{Z})\text{–}\mathsf{Mod})$$

  of (derived) module categories (see Section 2 for the definitions and Proposition 4.1 for the motivic versions). The argument in the algebraic context (for the "jump" from a general closed embedding to the linearized version of the zero section in the normal bundle) uses the geometric idea of the deformation to the normal cone together with sheaf theory on these spaces.

- The category $\mathsf{D}(\mathbf{R}p_*\mathbb{Z}\text{–}\mathsf{Mod})$ is equivalent to $\mathsf{UDM}(T^\circ_{S,s_0})$. This is based on the fact that the natural functor between these categories induces isomorphisms on all Ext groups between the corresponding tensor unit and itself (see Theorem 2.5). The motivic version is Corollary 3.7.

Now we can start with any $\mathcal{F} \in \mathsf{DM}(S^\circ)$ and form $i^*\mathbf{R}j_*\mathcal{F}$ as a module over $i^*\mathbf{R}j_*\mathbb{Z}$. Sending $i^*\mathbf{R}j_*\mathcal{F}$ further along the two equivalences

$$\mathsf{D}(i^*\mathbf{R}j_*\mathbb{Z}\text{–}\mathsf{Mod}) \sim \mathsf{D}(\mathbf{R}p_*\mathbb{Z}\text{–}\mathsf{Mod}) \sim \mathsf{UDM}(T^\circ_{S,s_0})$$

defines a natural functor

$$\mathsf{DM}(S^\circ) \to \mathsf{DM}(T^\circ_{S,s_0})$$

which, as we will see, has a motivic analogue.

This motivic local monodromy functor gives in particular tangential base-point functors in the motivic setting. For example, one can define the motivic fundamental group of $\mathbb{P}^1_{\mathbb{Q}} \smallsetminus \{0, 1, \infty\}$ with base point $\mathbf{01}$, as it was done for the realization categories in [3]. Another application are motivic versions of polylogarithm sheaves from [2] and of the weak Zagier conjecture. These will be addressed in a future paper.

Acknowledgments. I would like to thank Joseph Ayoub, Norbert Hoffmann, Bertrand Toen and Yuri Tschinkel for helpful discussions and suggestions.

2 Model category background

Notation 2.1.

- $\mathbb{N}$: the set of natural numbers including 0;
- $\triangle$: the category of finite ordinals $[n] = 0 < 1 < \cdots < n$ with order preserving maps;
- Set: the category of sets;
- SSet: the category of simplicial sets;
- Δ^n: the simplical set represented by $[n] \in \triangle$;
- $\partial\Delta^n$: the boundary of Δ^n;
- $\Lambda^n_k \subset \partial\Delta^n$: the cone where the k-th side is missing;
- Ab: the category of abelian groups;
- for a model category $\mathcal{C}$ (for example, SSet),
  - $\boldsymbol{Q}$ (resp., $\boldsymbol{R}$) the cofibrant (resp., fibrant) resolution functors;
  - $\mathcal{C}_c$ (resp. $\mathcal{C}_f$ resp. $\mathcal{C}_{cf}$) the full subcategory of cofibrant (resp. fibrant, resp. fibrant and cofibrant) objects of $\mathcal{C}$;
  - $\mathsf{Ho}\,\mathcal{C}$ (or $\mathsf{D}\mathcal{C}$) the corresponding homotopy category,
  - $\mathsf{Ho}^{\leq 2}\mathcal{C}$ (or $\mathsf{D}^{\leq 2}\mathcal{C}$) the homotopy 2-category, i.e., the 2-truncation of the Dwyer-Kan localisation of $\mathcal{C}_{cf}$;
  - for an object A of $\mathcal{C}$ carrying an additional structure (or an object of the corresponding homotopy category) let $A^\sharp$ be the underlying object without the structure (resp. its image in $\mathsf{Ho}\,\mathcal{C}$).
- for an abelian category $\mathcal{A}$ let $\mathsf{Cplx}(\mathcal{A})$ be the category of (homologically indexed) complexes in $\mathcal{A}$.

A model category $\mathcal{C}$ is called *stable* if the suspension functor on $\mathsf{Ho}\,\mathcal{C}$ is an equivalence (see [7, Definition 7.1.1]). In this case $\mathsf{Ho}\,\mathcal{C}$ is a (classical) triangulated category.

Let $\mathcal{D}$ be symmetric monoidal. For an operad $\mathcal{O}$ in $\mathcal{D}$ we let $\mathsf{Alg}(\mathcal{O})$ be the category of $\mathcal{O}$-algebras, and for $A \in \mathsf{Alg}(\mathcal{O})$ we let A–Mod be the category of

A-modules. For two maps $f: A \to B$ and $g: C \to D$ in $\mathcal{D}$ we denote by $f \square g$ the pushout product of f and g, i.e., the map

$$A \otimes D \sqcup_{A \otimes C} B \otimes C \to B \otimes D,$$

if it exists. If $\mathcal{D}$ is also a model category, it is called a symmetric monoidal model category if $f \square g$ is a cofibration, whenever f and g are, and, additionally, a weak equivalence if either f or g is.

For an object K of a symmetric monoidal model category $\mathcal{D}$, we denote by $\mathsf{Sp}^{\Sigma}(\mathcal{D}, K)$ the category of symmetric K-spectra as defined and studied in [8]. It is a symmetric monoidal model category (with the stable model structure) if the conditions of [8] are satisfied. Assigning to a K-spectrum X its zeroth object X_0 defines a right Quillen functor $\mathrm{Ev}_0: \mathsf{Sp}^{\Sigma}(\mathcal{D}, K) \to \mathcal{D}$ with a left adjoint Σ^{∞}_K which assigns to an object $A \in \mathcal{D}$ the free symmetric K-spectrum A.

Many constructions in homotopy theory are based on Bousfield localizations of model categories. The (left) Bousfield localization of a model category $\mathcal{C}$ with respect to a set of morphisms Φ is (if it exists) a model category $L_{\Phi}\mathcal{C}$ and a left Quillen functor $\mathcal{C} \to L_{\Phi}\mathcal{C}$ which should be thought of being universal with the property that elements of Φ are sent to weak equivalences. There are various general conditions insuring that a model category possesses a left Bousfield localization with respect to a set of morphisms. One of these conditions, used in [5], is called *cellularity*, see [5] or [8]. Cellular model categories are special cofibrantly generated model categories.

Example 2.2. If $\mathcal{A}$ is the category of modules over a unital commutative ring, then $\mathsf{Cplx}(\mathcal{A})$ carries a symmetric monoidal model structure, called the *projective* model structure. This model structure is cellular.

Assumption 2.3. Let $\mathcal{C}$ be a cofibrantly generated symmetric monoidal model category with unit $\mathbb{1}$ receiving a symmetric monoidal left Quillen functor either from SSet or $\mathsf{Cplx}(\mathsf{Ab})$. Assume that

- $\mathcal{C}$ is left proper,
- $\mathbb{1}$ and the domains of the generating cofibrations of $\mathcal{C}$ are cofibrant.

Let $\mathcal{L}$ be the image of the linear isometry operad in $\mathcal{C}$ (see [9]). Then the category of $\mathcal{L}$-algebras $\mathsf{Comm}(\mathcal{C}) := \mathsf{Alg}(\mathcal{L})$ is a semi-model category by [11, Theorem 4.7]. For $A \in \mathsf{Comm}(\mathcal{C})$ let $\mathsf{Comm}(A)$ be the category of objects in $\mathsf{Comm}(\mathcal{C})$ under A. This is a semi-model category if A is cofibrant. Thanks to the special properties of $\mathcal{L}$ the category A–Mod of A-modules carries a symmetric monoidal structure with pseudo-unit similarly defined as in [4] or [9]. If A is cofibrant, then A–Mod is a symmetric monoidal model category with weak unit (see [11, section 9]). The homotopy category $\mathsf{D}(A\text{–}\mathsf{Mod})$ is well defined for $A \in \mathsf{Ho}^{\leq 2}\mathsf{Comm}(\mathcal{C})$ and is a closed symmetric monoidal category. For $A \in \mathsf{Comm}(\mathcal{C})$ we put $\mathsf{D}(A\text{–}\mathsf{Mod}) := \mathsf{Ho}\boldsymbol{Q}A\text{–}\mathsf{Mod}$ (and similarly for $\mathsf{D}^{\leq 2}(A\text{–}\mathsf{Mod})$).

There are standard induced functors between these categories for change of algebras or change of model categories.

We now define unipotent objects relative to a base category, in the context of model categories (see the second part of [11]).

By a *homotopy colimit* in $\mathsf{Ho}\,\mathcal{C}$ we mean the image in $\mathsf{Ho}\,\mathcal{C}$ of a homotopy colimit $\mathrm{hocolim}(\mathbf{D})$ of a diagram $\mathbf{D} : \mathsf{I} \to \mathcal{C}_c$ (see [5, Definition 20.1.2] for homotopy colimits). By the homotopy colimit of a diagram $\mathbf{D} : \mathsf{I} \to \mathcal{C}$ we mean the homotopy colimit of the diagram $\boldsymbol{Q}\mathbf{D}$. For an ordinal λ a homotopy λ-sequence is a homotopy colimit of a diagram $\mathbf{D} : \lambda \to \mathcal{C}$ such that for any limit ordinal $\nu < \lambda$ the map from the homotopy colimit of $\mathrm{hocolim}(\mathbf{D}|_\nu) \to \mathbf{D}(\nu)$ is a weak equivalence.

A full subcategory $\mathcal{C}'$ of $\mathcal{C}$ is called *saturated* if $\mathcal{C}'$ is equal to its essential image.

Definition 2.4. *A saturated full subcategory $\mathcal{B}$ of $\mathsf{Ho}\,\mathcal{C}$ is called closed under homotopy colimits if every homotopy colimit in $\mathsf{Ho}\,\mathcal{C}$ whose terms map to $\mathcal{B}$ is contained in $\mathcal{B}$. Let O be a class of objects in $\mathsf{Ho}\,\mathcal{C}$. By the full subcategory of $\mathsf{Ho}\,\mathcal{C}$, O-generated by homotopy colimits, we mean the smallest full subcategory of $\mathsf{Ho}\,\mathcal{C}$ which contains O and is closed under homotopy colimits.*

In the following we will make no notational distinction between functors between model categories and the induced derived functors on the homotopy categories.

Let $F : \mathcal{C} \to \mathcal{D}$ be a symmetric monoidal left Quillen functor between cofibrantly generated symmetric monoidal model categories satisfying Assumption 2.3. Let U be the right adjoint of F. Let $\mathbb{1}_\mathcal{D} \to \boldsymbol{R}\mathbb{1}_\mathcal{D}$ be a fibrant and cofibrant replacement of the initial object in $\mathsf{Comm}(\mathcal{D})$ and set $A := \boldsymbol{Q}U(\boldsymbol{R}\mathbb{1}_\mathcal{D}) \in \mathsf{Comm}(\mathcal{C})$. There is a natural functor

$$\widetilde{U} : \mathsf{Ho}\,\mathcal{D} \to \mathsf{D}(A\text{–}\mathsf{Mod})$$

and an adjunction

$$\mathsf{D}(A\text{–}\mathsf{Mod}) \underset{\widetilde{U}}{\overset{\widetilde{F}}{\rightleftarrows}} \mathsf{Ho}\,\mathcal{D} \ ,$$

where $\widetilde{F}$ first maps to $\mathsf{D}(FA\text{–}\mathsf{Mod})$ and then to $\mathsf{Ho}\,\mathcal{D}$ via pushforward along the natural map $FA \to \boldsymbol{R}\mathbb{1}_\mathcal{D}$.

Theorem 2.5. *In this situation, assume that the following conditions are fulfilled:*

1. *Let $M \in \mathsf{Ho}\,\mathcal{C}$ be the image of a domain or codomain of a generating cofibration of $\mathcal{C}$. Then the projection morphism $M \otimes A^\sharp \to UFM$ is an isomorphism.*
2. *The functor $U : \mathsf{Ho}\,\mathcal{D} \to \mathsf{Ho}\,\mathcal{C}$ commutes with homotopy λ-sequences for all ordinals λ.*

3. $\mathcal{C}$ (and hence also $\mathcal{D}$) is a stable model category.

Then the symmetric monoidal functor $\widetilde{F}$ is fully faithful and its essential image is the full subcategory of $\mathsf{Ho}\,\mathcal{D}$ which is $F(D)$-generated by homotopy colimits, where D is the set of all domains and codomains of the generating cofibrations of $\mathcal{C}$. This subcategory is a $\otimes$-subcategory.

Proof. See [11, Theorem 13.4]. □

Remark 2.6. This Theorem also holds if we use symmetric monoidal model categories with weak unit.

Remark 2.7. The Theorem will be applied to the functor

$$f^* : \mathsf{DM}(S) \to \mathsf{DM}(X)$$

between triangulated categories of motivic sheaves (defined in the next section) induced by a morphism of schemes $f : X \to S$. The objects in the image of $(f^*)^\heartsuit$ can be viewed as generalized unipotent objects, i.e., they are constructed by some iterated homotopy colimits of objects coming from $\mathsf{DM}(S)$.

3 Motivic categories

In this section we define a class of model categories of motivic objects on schemes suitable for our applications. We essentially review the constructions of (stable) motivic homotopy categories introduced in the foundational works [10] and [12].

Let S be a separated Noetherian scheme and Sm/S the category of smooth schemes over S. Let $\mathsf{Spc}(S) := \mathsf{Shv}_{Nis}(\mathsf{Sm}/S)^{\Delta^{\mathrm{op}}}$ be the category of simplicial sheaves of sets on Sm/S for the Nisnevich topology. Let $\mathcal{M}$ be the image of a set of representatives of open inclusions $U \subset X$ in Sm/S under the composed functor $\mathsf{Sm}/S \to \mathsf{Shv}_{Nis}(\mathsf{Sm}/S) \to \mathsf{Spc}(S)$, where the first functor sends a scheme to the sheaf represented by the scheme and the second one sends a sheaf of sets to the constant simplicial object on the given sheaf. So $\mathcal{M}$ is a set of monomorphisms in $\mathsf{Spc}(S)$. Let I^s (resp. J^s) be the image of the standard set of (trivial) generating cofibrations of SSet in $\mathsf{Spc}(S)$, i.e., I^s consists of the images of the maps $\partial\Delta^n \hookrightarrow \Delta^n$ and J^s of the images of the maps $\Lambda^n_k \hookrightarrow \Delta^n$. Set $\mathsf{I}_S := \mathsf{I}^s \square \mathcal{M}$ and $\mathsf{J}_S := \mathsf{J}^s \square \mathcal{M}$.

Theorem–Definition 3.1. *The sets I_S (resp. J_S) are sets of generating (trivial) cofibrations for a symmetric monoidal (with respect to the cartesian product) finitely generated model structure on $\mathsf{Spc}(S)$ called the* cd*-model structure. This cd-model structure is left proper and cellular.*

Proof. See, e.g., [11]. □

By [5, Theorem 4.1.1] we can take the left Bousfield localization of $\mathsf{Spc}(S)$ with respect to maps $\Delta^n \times (\mathbb{A}^1_X \to X)$, $n \in \mathbb{N}$, where X runs through a set of representatives of isomorphism classes of Sm/S. We call the resulting model structure the $\mathbb{A}^1$-*local* model structure and denote the corresponding model category by $\mathsf{Spc}^{\mathbb{A}}(S)$. It is also symmetric monoidal.

We call an object of $\mathsf{Spc}(S)$ which is local with respect to the above set of morphisms $\mathbb{A}^1$-*local*, so $Z \in \mathsf{Spc}(S)$ is $\mathbb{A}^1$-local iff the maps

$$\mathrm{Hom}_{\mathsf{Ho}\,\mathsf{Spc}(S)}(X \times \Delta^n, Z) \to \mathrm{Hom}_{\mathsf{Ho}\,\mathsf{Spc}(S)}(\mathbb{A}^1_X \times \Delta^n, Z)$$

are isomorphisms for any $X \in \mathsf{Sm}/S$ and $n \in \mathbb{N}$.

Let $\mathsf{H}(S) := \mathsf{Ho}\,\mathsf{Spc}^{\mathbb{A}}(S)$. It is the motivic homotopy category defined in [10]. Let $\mathsf{Spc}^{\mathbb{A}}_\bullet(S)$ be the pointed version of $\mathsf{Spc}^{\mathbb{A}}(S)$ provided by [7, Proposition 1.1.8] and set $\mathsf{H}_\bullet(S) := \mathsf{Ho}\,\mathsf{Spc}^{\mathbb{A}}_\bullet(S)$; $\mathsf{Spc}^{\mathbb{A}}_\bullet(S)$ is a symmetric monoidal model category with respect to the smash product $\wedge$.

Let $\mathbb{T}$ be a cofibrant object in $\mathsf{Spc}^{\mathbb{A}}_\bullet(S)$ weakly equivalent to $(\mathbb{P}^1, \infty)$. Denote by $\mathsf{Sp}^\Sigma_{\mathbb{T}}(S) := \mathsf{Sp}^\Sigma(\mathsf{Spc}^{\mathbb{A}}_\bullet(S), \mathbb{T})$ the symmetric monoidal category of symmetric $\mathbb{T}$-spectra in $\mathsf{Spc}^{\mathbb{A}}_\bullet(S)$ provided by [8, Theorem 7.11]. We denote the corresponding homotopy category by $\mathsf{SH}(S)$ which is the stable motivic homotopy category. It has the property that $\Sigma^\infty_{\mathbb{T}}\mathbb{T}$ is tensor invertible. There is a factorization $\mathbb{T} \cong \mathbb{S}^1_s \wedge \mathbb{S}^1_t$ (see [10, Section 3.2.2]), where $\mathbb{S}^1_s$ is the image in $\mathsf{H}_\bullet(S)$ of the simplicial circle $\Delta^1/\partial\Delta^1$ with the obvious pointing and $\mathbb{S}^1_t$ the image of $(\mathbb{G}_m, 1)$. Hence both $\mathbb{S}^1_s$ and $\mathbb{S}^1_t$ are tensor invertible in $\mathsf{SH}(S)$. In particular $\mathsf{SH}(S)$ is a stable model category with shift functor $_ \wedge \mathbb{S}^1_s = _\,[1]$. For $M \in \mathsf{SH}(S)$ and $n \in \mathbb{Z}$ we set $M(n) := M \wedge (\mathbb{S}^1_t)^{\wedge n}[-n]$.

The model categories $\mathsf{Spc}^{\mathbb{A}}(S)$, $\mathsf{Spc}^{\mathbb{A}}_\bullet(S)$ and $\mathsf{Sp}^\Sigma_{\mathbb{T}}(S)$ depend functorially on S, i.e., they define left Quillen presheaves on the category of separated Noetherian schemes (see [6, Section 17]), since for a morphism $f : S \to T$ the isomorphism classes of the maps in $f^*(\mathsf{I}_T)$ are contained in I_S (same for the J's).

The natural Quillen functors between these model categories extend to morphisms of left Quillen presheaves.

Let $\mathcal{C}(S)$ be one of these model categories and let $f : X \to S$ in Sm/S. We have functors $M_S : \mathsf{Sm}/S \to \mathcal{C}(S)$ and $M_X : \mathsf{Sm}/X \to \mathcal{C}(X)$. In this situation f^* is also a right Quillen functor with left adjoint $f_!$ which sends $M_X(Y)$ to $M_S(Y)$ for $Y \in \mathsf{Sm}/X$.

Proposition 3.2. *Let $f : X \to S$ be in Sm/S and let $A, B \in \mathsf{Ho}\,\mathcal{C}(S)$ and $C \in \mathsf{Ho}\,\mathcal{C}(X)$. Then we have:*

1. *There is a canonical isomorphism $f_* f^* A \cong \underline{\mathrm{Hom}}_{\mathsf{Ho}\,\mathcal{C}(S)}(M_S(X), A)$.*
2. *The natural map*
$$f_!(C \otimes f^* A) \to f_! C \otimes A$$
in $\mathsf{Ho}\,\mathcal{C}(S)$ is an isomorphism.

3. The natural map

$$f^*\underline{\mathrm{Hom}}(A,B) \to \underline{\mathrm{Hom}}(f^*A, f^*B)$$

in $\mathsf{Ho}\,\mathcal{C}(X)$ *is an isomorphism.*

Proof. We prove the third point, the first two are similar but easier. Since f^* is also a right Quillen functor it commutes with homotopy limits and fiber sequences, hence we can assume that A is of the form $M_S(U)$, $U \in \mathsf{Sm}/S$. Let $B \in \mathcal{C}(S)$ be fibrant and cofibrant. The category $\mathcal{C}(S)$ consists of sheaves on Sm/S with values in some category $\mathcal{V}$, and there is a functor $v : \mathcal{V} \to \mathsf{Set}$ such that for $F \in \mathcal{C}(S)$ and $V \in \mathsf{Sm}/S$ we have $\mathrm{Hom}(M_S(V), F) = v(F(V))$. Let $V \in \mathsf{Sm}/X$. We have

$$f^*\underline{\mathrm{Hom}}_{\mathcal{C}(S)}(M_S(U), B)(V) = \underline{\mathrm{Hom}}_{\mathcal{C}(S)}(M_S(U), B)(V)$$

$$= v(B(U \times_Y V)) = v(B(U_X \times_X V)) = \underline{\mathrm{Hom}}_{\mathcal{C}(X)}(M_X(U_X), f^*B)(V)\ ,$$

which shows the claim. □

Lemma 3.3. *Let* $f : X \to S$ *be a morphism between separated Noetherian schemes. Then the map* $f_* : \mathsf{Ho}\,\mathcal{C}(X) \to \mathsf{Ho}\,\mathcal{C}(S)$ *preserves homotopy* λ*-sequences.*

Proof. We prove the case $\mathcal{C}(X) = \mathsf{Sp}^\Sigma_{\mathbb{T}}(X), \mathcal{C}(S) = \mathsf{Sp}^\Sigma_{\mathbb{T}}(S)$, the other cases are similar or easier. Consider a λ-sequence $Y : \lambda \to \mathsf{Sp}^\Sigma_{\mathbb{T}}(X)_{cf}$ with cofibrations as transition maps. Since filtered colimits in $\mathsf{Spc}(S)$ are created in presheaves, f_* commutes with λ-sequences by definition of f_*. Hence we have to check that $\mathrm{colim}_i f_* Y_i \in \mathsf{Sp}^\Sigma_{\mathbb{T}}(S)$ computes the homotopy colimit. We can find a λ-sequence $\widetilde{Y} : \lambda \to \mathsf{Sp}^\Sigma_{\mathbb{T}}(S)_{cf}$ where all transition maps are cofibrations together with an objectwise weak equivalence $\widetilde{Y} \to f_*Y$. Since these maps are weak equivalences between fibrant objects it follows that every map $\widetilde{Y}_i \to f_*Y_i$ is a level quasi-isomorphism. Hence using the injective model structure on $\mathsf{Spc}(S)$ it follows that the map $\mathrm{colim}_i \widetilde{Y}_i \to \mathrm{colim}_i f_* Y_i$ is a weak equivalence, what we wanted to show. □

The model categories $\mathsf{Spc}^{\mathbb{A}}(S)$, $\mathsf{Spc}^{\mathbb{A}}_\bullet(S)$ and $\mathsf{Sp}^\Sigma_{\mathbb{T}}(S)$ satisfy the Assumptions 2.3. Hence we have a theory of E_∞-algebras and modules in these categories. The categories of motivic objects will be categories of modules over some fixed cartesian sheaf of E_∞-algebras on some given category of schemes over some base scheme. I.e., we fix the following data:

- a separated Noetherian base scheme S_0 together with a full subcategory Sch/S_0 of the category of separated Noetherian schemes over S_0,
- a cofibrant algebra $H \in \mathsf{Comm}(\mathsf{Sp}^\Sigma_{\mathbb{T}}(S_0))$,
- for any object $f : S \to S_0$ of Sch/S_0 let $H(S) := f^*H \in \mathsf{Comm}(\mathsf{Sp}^\Sigma_{\mathbb{T}}(S))$,

- for f as above let $\mathsf{M}(S) := H(S)$–Mod be the model category of motivic objects we will consider.

The model categories $\mathsf{M}(S)$ are symmetric monoidal with weak unit. We denote the tensor product by $\otimes$. They define a left Quillen presheaf on Sch/S_0. The functoriality assertions of the last section also hold for these.

Definition 3.4. *Set* $\mathsf{DM}(S) := \mathsf{Ho}\,\mathsf{M}(S)$ *and* $\mathsf{Comm}(\mathsf{M}(S)) := \mathsf{Comm}(H(S))$.

Certainly one can use $\mathsf{M}(S) = \mathsf{Sp}^{\Sigma}_{\mathbb{T}}(S)$. As another special case, take $S_0 = Spec(\mathbb{Q})$, $\mathsf{Sch}/\mathbb{Q}$ the category of schemes of finite type over $Spec(\mathbb{Q})$ and H a cofibrant replacement of the motivic Eilenberg McLane-spectrum $(H\mathbb{Z})_{\mathbb{Q}}$ over $Spec(\mathbb{Q})$.

There is a canonical functor $M_S : \mathsf{Sm}/S \to \mathsf{DM}(S)$. We often write M instead of M_S. There is a natural extension of this functor to closed pairs $Z \subset X$ in Sm/S and we write $M(X, Z)$ for the corresponding image in $\mathsf{DM}(S)$. For $M \in \mathsf{DM}(S)$ we set $M^{\vee} := \underline{\mathrm{Hom}}_{\mathsf{DM}(S)}(M, \mathbb{1})$. For any $n \in \mathbb{Z}$ we have the Tate shift $M(n) = M \otimes M(\mathbb{G}_m, \{1\})^{\otimes n}[-n]$.

The following is an application of [10, Theorem 3.2.21]:

Proposition 3.5. *Let* $i : Z \subset X$ *be a closed embedding and* $j : U \subset X$ *the complementary open embedding in* Sch/S_0. *For every* $\mathcal{F} \in \mathsf{DM}(X)$ *there is an exact triangle*

$$j_! j^* \mathcal{F} \to \mathcal{F} \to i_* i^* \mathcal{F} \to j_! j^* \mathcal{F}[1]$$

in $\mathsf{DM}(X)$.

Proof. See [11]. □

Let $f : X \to S$ be a morphism in Sch/S_0. We wish to apply Theorem 2.5 with $\mathcal{C} = \mathsf{M}(S)$, $\mathcal{D} = \mathsf{M}(X)$ and $F = f^*$.

We have to check conditions 1-3 of Theorem 2.5.

(1) Let $M \in \mathsf{DM}(S)$. If f is smooth we have $f_* f^* B \cong \underline{\mathrm{Hom}}_{\mathsf{DM}(S)}(M(X), M)$ by Proposition 3.2, (1). Hence this condition is in this case equivalent to the rigidity statement that the canonical map

$$M(X)^{\vee} \otimes M \to \underline{\mathrm{Hom}}(M(X), M)$$

is an isomorphism.

(2) This condition follows from Lemma 3.3.

(3) This condition is clear.

Lemma 3.6. *Condition (1) is satisfied if* f *is smooth and one of the following conditions are satisfied:*

- S *is the spectrum of a field of characteristic* 0.

- *f is projective or more generally the complement of a relative normal crossing divisor in a smooth and projective morphism.*

Proof (Proof in a special case.). We prove the case where X is a vector bundle over S or the complement of the zero section of a vector bundle. We are allowed to Zariski-localize, hence in the case of a vector bundle the assertion follows from $\mathbb{A}^1$-locality. In the second case we know that Zariski-locally we have $M(X) \cong \mathbb{1} \oplus \mathbb{1}(n)[2n-1]$ which is also enough to conclude. □

We denote the full subcategory of $\mathsf{DM}(X)$ which is $f^*(\mathsf{DM}(S))$-generated by homotopy colimits by $\mathsf{UDM}(X/S)$ (see Definition 2.4).

Let $A(X) := f_*\mathbb{1} \in \mathsf{D}^{\leq 2}\mathsf{Comm}(\mathsf{M}(S))$ be the motivic cohomology algebra of X relative to S.

Then Theorem 2.5 implies

Corollary 3.7. *Suppose f satisfies one of the conditions of the above Lemma. Then there is a natural equivalence of tensor triangulated categories*

$$(f^*)\tilde{}: \mathsf{D}(A(X)\text{–}\mathsf{Mod}) \to \mathsf{UDM}(X/S).$$

4 Limit Motives (unipotent case)

Let $S \in \mathsf{Sch}/S_0$ such that any smooth scheme over S belongs to Sch/S_0.

Proposition 4.1. *Let*

$$Z \overset{i}{\hookrightarrow} X \overset{j}{\hookleftarrow} U,$$

where i (resp. j) is a closed (resp. complementary open) embedding in Sm/S. Let $p : N \to Z$ be the normal bundle of Z in X and $p^\circ : N^\circ \to Z$ the complement of the zero section. There is a natural isomorphism $i^*j_*\mathbb{1} \cong p^\circ_*\mathbb{1}$ in $\mathsf{DComm}(\mathsf{M}(Z))$.

Proof. This combines the next two Lemmas. □

Lemma 4.2. *For $Z \in \mathsf{Sm}/S$ let $p : N \to Z$ be a (geometric) vector bundle with zero section $i : Z \subset N$. Let $j : N \smallsetminus i(Z) \subset N$ be the open inclusion and let $p^\circ : N \smallsetminus i(Z) \to Z$ be the projection. Then there is a natural isomorphism $i^*j_*\mathbb{1} \cong p^\circ_*\mathbb{1}$ in $\mathsf{DComm}(\mathsf{M}(Z))$.*

Proof. The unit for the adjunction

$$\mathsf{DComm}(\mathsf{M}(Z)) \underset{i^*}{\overset{i_*}{\rightleftarrows}} \mathsf{DComm}(\mathsf{M}(N)) \ ,$$

applied to the algebra $j_*\mathbb{1}$ yields a map $j_*\mathbb{1} \to i_*i^*j_*\mathbb{1}$. Further applying p_* yields the desired map $p^\circ_*\mathbb{1} \to i^*j_*\mathbb{1}$.

We can now forget the algebra structures since all functors involved commute with them. Proposition 3.5 applied to $j_*\mathbb{1}$ yields an exact triangle

$$j_!j^*j_*\mathbb{1} \to j_*\mathbb{1} \to i_*i^*j_*\mathbb{1} \to j_!j^*j_*\mathbb{1}[1]\ ,$$

and our map comes from applying p_* to the second map of this triangle. There is an isomorphism $j_!j^*j_*\mathbb{1} \cong j_!\mathbb{1}$. Hence we are done if we show that $p_*j_!\mathbb{1} = 0$. We apply p_* to the triangle

$$j_!\mathbb{1} \to \mathbb{1} \to i_*\mathbb{1} \to j_!\mathbb{1}[1]\ .$$

The second map is mapped to an isomorphism since p is an $\mathbb{A}^1$-weak equivalence, hence $p_*j_!\mathbb{1} = 0$. □

Lemma 4.3. *Let the situation be as in Proposition 4.1 and let $i' : Z \subset N$ and $j' : N^\circ \subset N$ be the zero section and its complement. Then there is a natural isomorphism $i^*j_*\mathbb{1} \cong i'^*j'_*\mathbb{1}$ in $\mathsf{DComm}(\mathsf{M}(Z))$.*

Proof. We use a similar construction as in the proof of [10, Theorem 3.2.23]. Let $\pi : B \to X \times \mathbb{A}^1$ be the blow-up of $Z \times \{0\}$ in $X \times \mathbb{A}^1$, $f : Z \times \mathbb{A}^1 \to B$ the canonical closed embedding which splits $i(Z) \times \mathbb{A}^1$ and $g : X \to B$ the closed embedding which splits $X \times \{1\}$. We have $P := \pi^{-1}(Z \times \{0\}) \cong \mathbb{P}(N \oplus \mathcal{O}) \supset \mathbb{P}(N)$. Let E be the strict transform of $X \times \{0\}$ in B and set $\widetilde{B} := B \smallsetminus E$. We have $E \cap P = \mathbb{P}(N)$, so we have a closed embedding $h : N \to \widetilde{B}$. The maps f and g factor through $\widetilde{B}$, and we denote the factor maps also by f and g.

Let $\widetilde{B}^\circ := \widetilde{B} \smallsetminus f(Z \times \mathbb{A}^1)$ and $j'' : \widetilde{B}^\circ \subset \widetilde{B}$ the open inclusion. We have pullback squares

$$\begin{array}{ccc} U & \longrightarrow & \widetilde{B}^\circ \\ \downarrow{\scriptstyle j} & & \downarrow{\scriptstyle j''} \\ X & \xrightarrow{g} & \widetilde{B} \end{array} \quad \text{and} \quad \begin{array}{ccc} N^\circ & \xrightarrow{h^\circ} & \widetilde{B}^\circ \\ \downarrow & & \downarrow{\scriptstyle j''} \\ N & \xrightarrow{h} & \widetilde{B} \end{array}\ .$$

Claim 1: The two base change morphisms for these diagrams applied to $\mathbb{1}$ are isomorphisms, i.e., we have isomorphisms $i^*j_*\mathbb{1} \cong i_1^*f^*j''_*\mathbb{1}$ and $i'^*j'_*\mathbb{1} \cong i_0^*f^*j''_*\mathbb{1}$, where $i_r : Z \times \{r\} \to Z \times \mathbb{A}^1$, $r = 0, 1$, are the two inclusions.

Let $q : Z \times \mathbb{A}^1 \to Z$ be the projection. We have maps

$$h_r : q_*f^*j''_*\mathbb{1} \to i_r^*f^*j''_*\mathbb{1}, \quad r = 0, 1.$$

Claim 2: The maps h_r, $r = 0, 1$, are isomorphisms.

The claims imply the desired isomorphism and we turn to their proofs. After Zariski localization on X we can assume that there is an etale map $e : X \to \mathbb{A}^n_S$

such that $Z = e^{-1}(\mathbb{A}^r_S)$. By Proposition 3.2 (1) and (3), we are reduced to the case $X = \mathbb{A}^n_S$, $Z = \mathbb{A}^r_S$. Changing S to $\mathbb{A}^r_S$ we can assume that $r = 0$. Let $x_1, \dots, x_{n+1}$ be coordinates on $\mathbb{A}^{n+1}_S$ and $y_1, \dots, y_{n+1}$ homogeneous coordinates on $\mathbb{P}^n_S$. Then

$$B = \{(x, y) \in (\mathbb{A}^{n+1} \times \mathbb{P}^n)_S \mid x_i y_j = x_j y_i,\ i, j = 1, \dots, n+1\}.$$

Let $W \subset B$ be defined by the equations $x_{n+1} = y_{n+1} = 0$. Then $\widetilde{B} = B \setminus W$. We consider $\widetilde{B}$ and $\mathbb{A}^{n+1}_S$ as schemes over $\mathbb{A}^1_S$ via the last projection $\mathbb{A}^{n+1}_S \to \mathbb{A}^1_S$. The assignment $x_{n+1} \mapsto x_{n+1}$, $x_i \mapsto \frac{y_i}{y_{n+1}}$, $i = 1, \dots, n$, yields an isomorphism $\varphi : \widetilde{B} \to \mathbb{A}^{n+1}_S$ over $\mathbb{A}^1_S$ such that $(\varphi \circ f)(\mathbb{A}^1_S)$ is the closed subscheme defined by $x_1 = \dots = x_n = 0$. Hence by Lemma 4.2 and Proposition 3.2 (1) we have

$$f^* j''_* \mathbb{1} = M((\mathbb{A}^n \setminus \{0\})_{\mathbb{A}^1_S})^\vee \cong \mathbb{1} \oplus \mathbb{1}(-n)[-2n+1].$$

We have an analogous descriptions of $i^* j_* \mathbb{1}$ and ${i'}^* {j'}_* \mathbb{1}$, so Claim 1 follows. Claim 2 follows since $f^* j''_* \mathbb{1}$ is a pullback from S and $\mathbb{A}^1_S \to S$ is an $\mathbb{A}^1$-weak equivalence. □

Let now

$$D \overset{i}{\hookrightarrow} X \overset{j}{\hookleftarrow} X^\circ,$$

where D is a divisor, i (resp. j) a closed (resp. complementary open) embedding in Sm/S and let $p^\circ : N^\circ \to D$ be the pointed normal bundle of D in X. The morphism p° satisfies the conditions of Corollary 3.7, hence we have an equivalence $\mathsf{UDM}(N^\circ/D) \sim \mathsf{D}(p^\circ_* \mathbb{1}\text{–}\mathsf{Mod})$. The functor $i^* j_* : \mathsf{DM}(X^\circ) \to \mathsf{DM}(D)$ factors through $\mathsf{D}(i^* j_* \mathbb{1}\text{–}\mathsf{Mod})$. Proposition 4.1 suggests that we should have a natural equivalence $\mathsf{D}(i^* j_* \mathbb{1}\text{–}\mathsf{Mod}) \sim \mathsf{D}(p^\circ_* \mathbb{1}\text{–}\mathsf{Mod})$, but we get such an equivalence only if we have a morphism from $i^* j_* \mathbb{1}$ to $p^\circ_* \mathbb{1}$ in the 2-category $\mathsf{D}^{\leq 2}\mathsf{Comm}(\mathsf{M}(D))$. Reexamining the proofs of Lemmas 4.2 and 4.3 we find that we have a chain of weak isomorphisms $i^* j_* \mathbb{1} \leftarrow B \to B' \leftarrow p^\circ_* \mathbb{1}$ in $\mathsf{D}^{\leq 2}\mathsf{Comm}(\mathsf{M}(D))$, where all maps are unique up to a unique 2-isomorphism. Since there is no way in a 2-category to find an inverse of a weak isomorphism which is unique up to a unique 2-isomorphism this chain of weak isomorphisms is the only thing we get. Nevertheless it follows that there is a natural equivalence $\mathsf{D}(i^* j_* \mathbb{1}\text{–}\mathsf{Mod}) \sim \mathsf{D}(p^\circ_* \mathbb{1}\text{–}\mathsf{Mod})$, unique up to a unique natural isomorphism, by composing the functors induced by the maps in this chain or their adjoints. Now we can define the functor

$$\mathcal{L}_{X,D} : \mathsf{DM}(X^\circ) \to \mathsf{DM}(N^\circ)$$

to be the composition

$$\mathsf{DM}(X^\circ) \to \mathsf{D}(i^* j_* \mathbb{1}\text{–}\mathsf{Mod}) \sim \mathsf{D}(p^\circ_* \mathbb{1}\text{–}\mathsf{Mod}) \sim \mathsf{UDM}(N^\circ/D) \to \mathsf{DM}(N^\circ)\,.$$

Remark 4.4. Intuitively, the functor does the following: We first restrict a given motivic sheaf on X° to a tubular neighborhood of D (which of course does not exist). Then we identify this tubular neighborhood with a tubular neighborhood of the zero section in N (the normal bundle of D in X) and carry over the restricted sheaf. This we finally extend to the whole of N°. As long as we believe that some sort of monodromy action around D is unipotent our above definition gives a perfectly sensible simulation of this intuitive description.

Consider the more general situation: Let $X \in \mathsf{Sm}/S$ and $D \subset X$ be a divisor with normal crossings relative to S, i.e., $D = \bigcup_{i\in I} D_i$ with $D_i \in \mathsf{Sm}/S$ and locally in the etale topology the intersections of the D_i look like intersections of coordinate hyperplanes in some $\mathbb{A}^n_S$. Let $X^\circ := X \setminus D$. For $J \subset I$ let $D_J := \bigcap_{i\in J} D_i$ and $D^\circ_J := D_J \setminus \bigcup_{i\in I\setminus J} D_i$. Let N_i be the normal bundle of D_i in X and N_i° the complement of the zero section. Let N_J be the fiber product of the N_i over D_J for $i \in J$ and N°_J the corresponding product of the N_i°. Finally, let $N^{\circ\circ}_J$ be the restriction of N°_J to D°_J. Our goal is to construct functors

$$\mathcal{L}_{X,J} : \mathsf{DM}(X^\circ) \to \mathsf{DM}(N^{\circ\circ}_J),$$

for $J \subset I$, subject to certain compatibilities.

Assume that $J, J' \subset I$ are disjoint and consider the functor $\mathcal{L}_{X,J\cup J'}$. For $i \in I \setminus J$ denote by $\widetilde{D}_i$ the restriction of D_i to D_J and by D'_i the preimage of $\widetilde{D}_i$ in N°_J with respect to the natural projection. The D'_i, $i \in I \setminus J$, are again divisors with normal crossings in N°_J relative to D_J with complement $N^{\circ\circ}_J$. Applying the definitions above we get corresponding objects $D'_{J'}$, $D'^\circ_{J'}$, $N'^\circ_{J'}$ and $N'^{\circ\circ}_{J'}$ and canonical isomorphisms $D'^\circ_{J'} \cong D^\circ_{J\cup J'}$ and $N'^{\circ\circ}_{J'} \cong N^{\circ\circ}_{J\cup J'}$.

On the other hand, we have a composition of functors

$$\mathsf{DM}(X^\circ) \xrightarrow{\mathcal{L}_{X,J}} \mathsf{DM}(N^{\circ\circ}_J) \xrightarrow{\mathcal{L}_{N^\circ_J,J'}} \mathsf{DM}(N'^{\circ\circ}_{J'}) \sim \mathsf{DM}(N^{\circ\circ}_{J\cup J'}) \ . \tag{1}$$

We require the following compatibilities:

- there is a natural isomorphism $\varphi_{J',J}$ between the composition (1) and $\mathcal{L}_{X,J\cup J'}$;
- for any three disjoint subsets $J, J'J'' \subset I$,

$$\varphi_{J''\cup J',J} \circ (\varphi_{J'',J'} \underset{0}{\circ} \mathrm{Id}_{\mathcal{L}_J}) = \varphi_{J'',J'\cup J} \circ (\mathrm{Id}_{\mathcal{L}_{J''}} \underset{0}{\circ} \varphi_{J',J}) \ .$$

Below we sketch the construction of the $\varphi_{J,J'}$ and indicate the main ideas in the proof of compatibilities.

We introduce some further notation: Let $j : X^\circ \to X$, $X^J := X \setminus \bigcup_{i\in J} D_i$ and $j^J : X^J \to X$ be the open inclusion. Let $i_J : D_J \to X$ and $i^\circ_J : D^\circ_J \to X$ be the closed, respectively locally closed, embedding. Furthermore, let $p^\circ_J : N^\circ_J \to D_J$ and $p^{\circ\circ}_J : N^{\circ\circ}_J \to D^\circ_J$ be the projections.

Proposition 4.5. *Let $A \in \mathsf{DComm}(\mathsf{M}(X))$ such that Zariski-locally on X $A^\sharp$ is a pullback of an object from $\mathsf{DM}(S)$. Then there is a canonical isomorphism*

$$i_J^* j_{J*}^J j^{J*} A \cong p_{J*}^\circ p_J^{\circ *} i_J^* A$$

in $\mathsf{DComm}(\mathsf{M}(D_J))$. Furthermore, if A is an object in $\mathsf{D}^{\leq 2}\mathsf{Comm}(\mathsf{M}(X))$ there is a natural chain of weak isomorphisms in $\mathsf{D}^{\leq 2}\mathsf{Comm}(\mathsf{M}(X))$ connecting $i_J^* j_{J*}^J j^{J*} A$ and $p_{J*}^\circ p_J^{\circ *} i_J^* A$.

Proof. Analogous to Lemmas 4.2 and 4.3. The analogue of Lemma 4.2 has the same proof and states in our case that there is a natural isomorphism

$$i'^* j'_* p_J^{\circ *} i_J^* A \cong p_{J*}^\circ p_J^{\circ *} i_J^* A,$$

where $j' : N_J^\circ \to N_J$ is the open inclusion and $i' : D_J \to N_J$ the zero section. We prove the analogue of Lemma 4.3: For every $i \in J$ let π_i, B_i, f_i, g_i, $\widetilde{B}_i$ and $\widetilde{B}_i^\circ$ be as in the proof of Lemma 4.3 for $Z = D_i$. Let $\pi_J : \widetilde{B}_J \to X \times \mathbb{A}^1$ be the fiber product of the $\widetilde{B}_i$ over $X \times \mathbb{A}^1$ for $i \in J$ and $\widetilde{B}_J^\circ$ the corresponding fiber product of the $\widetilde{B}_i^\circ$. We have a closed embedding $h_J : N_J \to \widetilde{B}_J$. Let $j_J'' : \widetilde{B}_J^\circ \subset \widetilde{B}_J$ be the open inclusion and $f_J : D_J \times \mathbb{A}^1 \to \widetilde{B}_J$ the intersection of the divisors which build the complement of j_J''. Let $g_J : X \to \widetilde{B}_J$ be the product of the g_i. Then we have again pullback squares

$$\begin{array}{ccc} X^J & \longrightarrow & \widetilde{B}_J^\circ \\ \downarrow{\scriptstyle j^J} & & \downarrow{\scriptstyle j_J''} \\ X & \xrightarrow{g} & \widetilde{B} \end{array} \quad \text{and} \quad \begin{array}{ccc} N_J^\circ & \xrightarrow{h_J^\circ} & \widetilde{B}_J^\circ \\ \downarrow & & \downarrow{\scriptstyle j_J''} \\ N_J & \xrightarrow{h_J} & \widetilde{B}_J \end{array} .$$

Similar to the proof of Lemma 4.3 one shows that the base change morphisms of these diagrams yield isomorphisms

$$i_J^* j_{J*}^J j^{J*} A \cong i_1^* f_J^* j_{J*}'' \pi_J^* A \quad \text{and} \quad i'^* j'_* p_J^{\circ *} i_J^* A \cong i_0^* f_J^* j_{J*}'' \pi_J^* A,$$

where $i_r : D_J \times \{r\} \to D_J \times \mathbb{A}^1$, $r = 0, 1$, are the inclusions. Also one shows that the corresponding maps $h_r : i_r^* f_J^* j_{J*}'' \pi_J^* A \to q_* f_J^* j_{J*}'' \pi_J^* A$, as well as $q : D_J \times \mathbb{A}^1 \to D_J$, the projection, are isomorphisms. □

The functor

$$\mathcal{L}_{X,J} : \mathsf{DM}(X^\circ) \to \mathsf{DM}(N_J^{\circ\circ})$$

is defined as the composition

$$\mathsf{DM}(X^\circ) \to \mathsf{D}(i_J^{\circ *} j_* \mathbb{1}\text{–}\mathsf{Mod}) \sim D(p_{J*}^{\circ\circ} \mathbb{1}\text{–}\mathsf{Mod}) \sim \mathsf{UDM}(N_J^{\circ\circ}/D_J^\circ) \to \mathsf{DM}(N_J^{\circ\circ}) .$$

We turn to the construction of the natural isomorphisms $\varphi_{J',J}$. We may assume that $J \cup J' = I$. Consider the cartesian squares

$$\begin{array}{ccccc} D'_{J'} & \xrightarrow{i'} & N_J^\circ & \xleftarrow{j'} & N_J^{\circ\circ} = (N_J^\circ)^{J'} \\ \downarrow & & \downarrow{\scriptstyle p_J^\circ} & & \downarrow \\ D_I & \xrightarrow{\tilde{i}} & D_J & \xleftarrow{\tilde{j}} & D_J^\circ \end{array} .$$

Let $\tilde{p}^\circ : N_{J'}^\circ|_{D_I} \to D_I$ and $p'^\circ : N'^\circ_{J'} \to D'_{J'}$ be the projections. Proposition 4.5, applied to the algebra $A := p_{J*}^\circ \mathbb{1}$, gives an isomorphism

$$\tilde{p}_*^\circ \tilde{p}^{\circ *} \tilde{i}^* A \cong \tilde{i}^* \tilde{j}_* \tilde{j}^* A \tag{2}$$

and a connecting chain of weak isomorphisms in $\mathsf{D}^{\leq 2}\mathsf{Comm}(\mathsf{M}(D_I))$. The same relation for the algebra $\mathbb{1}$ on N_J° yields the functor $\mathcal{L}_{N_J^\circ;J'}$. We get a diagram of functors with natural isomorphisms in the squares

$$\begin{array}{ccccc} \mathsf{DM}((N_J^\circ)^{J'}) & \xrightarrow{i'^* j'_*} & \mathsf{D}(i'^* j'_* \mathbb{1}\text{–}\mathsf{Mod}) & \xrightarrow{\sim} & \mathsf{D}(p'^\circ_* \mathbb{1}\text{–}\mathsf{Mod}) \\ \uparrow & & \uparrow & & \uparrow \\ \mathsf{D}(\tilde{j}^* A\text{–}\mathsf{Mod}) & \xrightarrow{\tilde{i}^* \tilde{j}_*} & \mathsf{D}(\tilde{i}^* \tilde{j}_* \tilde{j}^* A\text{–}\mathsf{Mod}) & \xrightarrow{\sim} & \mathsf{D}(\tilde{p}_*^\circ \tilde{p}^{\circ *} \tilde{i}^* A\text{–}\mathsf{Mod}) \end{array} ,$$

where we used appropriate naturality of the equivalence in Corollary 3.7 and of the constructions in the proof of Proposition 4.5. We also have a cartesian square

$$\begin{array}{ccc} N'^\circ_{J'} = N_I^\circ & \xrightarrow{p'^\circ} & D'_{J'} = N_J^\circ|_{D_I} \\ \downarrow{\scriptstyle q'} & & \downarrow{\scriptstyle q} \\ N_{J'}^\circ|_{D_I} & \xrightarrow{\tilde{p}^\circ} & D_I \end{array} .$$

Hence base change morphisms induce isomorphisms

$$p_{I*}^\circ \mathbb{1} \cong \tilde{p}_*^\circ \circ q'_* \mathbb{1} \cong \tilde{p}_*^\circ \tilde{p}^{\circ *} q_* \mathbb{1} \cong \tilde{p}_*^\circ \tilde{p}^{\circ *} \tilde{i}^* A, \tag{3}$$

and we get an equivalence

$$\mathsf{D}(\tilde{p}_*^\circ \tilde{p}^{\circ *} \tilde{i}^* A\text{–}\mathsf{Mod}) \sim \mathsf{UDM}(N_I^\circ / D_I) .$$

Collecting, we get a naturally commutative square

$$\begin{array}{ccc} \mathsf{DM}((N_J^\circ)^{J'}) & \xrightarrow{\mathcal{L}_{N_J^\circ, J'}} & \mathsf{UDM}(N'^\circ_{J'} / D'_{J'}) \\ \uparrow & & \uparrow \\ \mathsf{D}(\tilde{j}^* A\text{–}\mathsf{Mod}) & \longrightarrow & \mathsf{UDM}(N_I^\circ / D_I) \end{array} .$$

Using the fact that $A \cong i_J^* j_*^J \mathbb{1}$ and applying base change to the cartesian square in the diagram

$$\begin{array}{ccccc} & & D_J^\circ & \longrightarrow & X^{J'} & \longleftarrow & X^\circ \\ & & \downarrow & & \downarrow & & \\ D_I & \longrightarrow & D_J & \longrightarrow & X & & \end{array}$$

yields an isomorphism $\tilde{i}^*\tilde{j}_*\tilde{j}^*A \cong i_I^* j_*^I \mathbb{1}$. Combining this with (2) and (3) we see that we obtain a natural chain of weak isomorphisms connecting $p_{I*}^\circ \mathbb{1}$ and $i_I^* j_*^I \mathbb{1}$ in $\mathsf{D}^{\leq 2}\mathsf{Comm}(\mathsf{M}(D_I))$. Another such chain is given by Proposition 4.5. We are done if the functors induced by these two chains are naturally isomorphic.

Keep notation as in the proof of Proposition 4.5. Let $\widetilde{B}'_{J'}$ etc. be the analogous objects defined for the situation on $D_J \times \mathbb{A}^1$ with divisors $(D_i \cap D_J) \times \mathbb{A}^1$, $i \in J'$. For $i \in J$ we have $\widetilde{D}_{J,i} = f_i(D_i \times \mathbb{A}^1) \times_{X \times \mathbb{A}^1} \widetilde{B}_{J \smallsetminus \{i\}}$ and otherwise $\widetilde{D}_{J,i} = \pi_J^{-1}(D_i \times \mathbb{A}^1)$. Let $\mathcal{B} := \widetilde{B}_J \times_X \widetilde{B}_{J'}$ and $\mathcal{B}^\circ := \widetilde{B}_J^\circ \times_X \widetilde{B}_{J'}^\circ$ We have morphisms

$$\widetilde{B}_I \xrightarrow{\iota} \mathcal{B} \xrightarrow{\rho} X \times \mathbb{A}^2,$$

and pullback squares

$$\begin{array}{ccc} \widetilde{B}'_{J'} & \longrightarrow & \mathcal{B} \\ \downarrow{\scriptstyle \pi'_{J'}} & & \downarrow \\ (D_J \times \mathbb{A}^1) \times \mathbb{A}^1 & \longrightarrow & \widetilde{B}_J \times \mathbb{A}^1 \end{array} \quad \text{and} \quad \begin{array}{ccc} \widetilde{B}_I & \xrightarrow{\iota} & \mathcal{B} \\ \downarrow{\scriptstyle \pi_I} & & \downarrow{\scriptstyle \rho} \\ X \times \mathbb{A}^1 & \xrightarrow{\mathrm{Id} \times \triangle} & X \times \mathbb{A}^2 \end{array} .$$

On $\mathcal{B}$ we have the divisors $\mathcal{D}_i := D_{J,i} \times_X \widetilde{B}_{J'}$ for $i \in J$ and $\mathcal{D}_i := \widetilde{B}_J \times_X D_{J',i}$ for $i \in J'$. The $D_{I,i}$ are the pullbacks of the $\mathcal{D}_i$ with respect to ι. Let

$$\tilde{q} : \ (D_J \times \mathbb{A}^1) \times \mathbb{A}^1 \to D_J \times \mathbb{A}^1$$

be the projection. Let $A := f_J^* j''_{J*} \mathbb{1}$, $\widetilde{A} := \pi'^*_{J'} \tilde{q}^* A$ and $B := f'^*_{J'} j'''_{J'*} j'''^*_{J'} \widetilde{A}$, so B is an algebra on $D_I \times \mathbb{A}^2$. We have canonical isomorphisms

$$\begin{aligned} i^*_{(0,0)} B &\cong \tilde{p}^\circ_* \tilde{p}^{\circ *} \tilde{i}^* p^\circ_{J*} \mathbb{1}, \\ i^*_{(0,1)} B &\cong \tilde{i}^* \tilde{j}_* \tilde{j}^* p^\circ_{J*} \mathbb{1}, \\ i^*_{(1,0)} B &\cong \tilde{p}^\circ_* \tilde{p}^{\circ *} \tilde{i}^* i^*_J j^J_* \mathbb{1}, \\ i^*_{(1,1)} B &\cong \tilde{i}^* \tilde{j}_* \tilde{j}^* i^*_J j^J_* \mathbb{1}, \end{aligned}$$

where $i_{(r,l)} : \ D_I \times \{(r,l)\} \to D_I \times \mathbb{A}^2$, $r, l = 0, 1$, are the inclusions. We have isomorphisms between the $i^*_{(r,l)} B$ by comparing them to $(D_I \times \mathbb{A}^2 \to D_I)_* B$ via base change morphisms; and the isomorphisms among the right hand sides above are compatible with these. Again via a base change morphism we have a canonical isomorphism $B \cong (D_I \times \mathbb{A}^2 \subset \mathcal{B})^* (\mathcal{B}^\circ \subset \mathcal{B})_* \mathbb{1}$, and the left square above shows that we also have canonical isomorphisms $i^*_{(0,0)} B \cong p^\circ_{I*} \mathbb{1}$

and $i^*_{(1,1)}B \cong i^*_I j^I_* \mathbb{1}$. Compatibility of base change morphisms shows that the two possible identifications of $p^\circ_{I*}\mathbb{1}$ and $i^*_I j^I_* \mathbb{1}$ constructed above actually coincide. Note that our arguments took place in homotopy categories and not in homotopy 2-categories. We leave it to the reader to exhibit the required 2-morphisms (a diagram connecting the many different ways of joining $p^\circ_{I*}\mathbb{1}$ and $i^*_I j^I_* \mathbb{1}$). For the compatibility of the $\varphi_{J,J'}$ one considers $\widetilde{B}_J \times_X \widetilde{B}_{J'} \times_X \widetilde{B}_{J''}$ and then compares the constructed 2-morphisms.

Remark 4.6. Instead of divisors $D_i \subset X$ we also can take closed $D_i \subset X$, $D_i \in \mathsf{Sm}/S$, such that the intersections of the D_i look etale locally like intersections of orthogonal standard affine subspaces of some $\mathbb{A}^n_S$. For these situations all constructions above work in exactly the same way. In particular, we also get the functors $\mathcal{L}_{X,J}$.

5 The "trivial" case: Normal bundles

Let $Z \in \mathsf{Sch}/S_0$ be such that any smooth scheme over Z lies in Sch/S_0. Let $\mathcal{E}_1, \ldots, \mathcal{E}_r$ be vector bundles over Z, $p_i : E_i \to Z$ the corresponding geometric vector bundles and $p_i^\circ : E_i^\circ \to Z$ the complement of the zero section. Let $X \to Z$ be the fibre product of the E_i over Z and $X^\circ \subset X$ the fibre product of the E_i°. Let $D_i \subset X$ be the subscheme corresponding to the zero section of E_i. Taking $I = \{1, \ldots, r\}$ we have a canonical isomorphism $X^\circ \cong N_I^{\circ\circ}$ and we would like to show that our previous functor

$$\mathcal{L}_{X,I} : \mathsf{DM}(X^\circ) \to \mathsf{DM}(N_I^{\circ\circ}) \simeq \mathsf{DM}(X^\circ)$$

is canonially isomorphic to the identity when restricted to $\mathsf{UDM}(X^\circ/Z)$.

Consider the case $r = 1$, $\mathcal{E}_1 = \mathcal{O}_Z$. Then

$$p : X = \mathbb{A}^1_Z \to Z,\ j : X^\circ = \mathbb{G}_{m,Z} \subset \mathbb{A}^1_Z,\ p^\circ : X^\circ \to Z,\ i : D = \{0\} \times Z \subset X.$$

Proposition 5.1. *The functors*

$$p^\circ_* : \mathsf{UDM}(X^\circ/Z) \to \mathsf{D}(p^\circ_*\mathbb{1}\text{–}\mathsf{Mod}) \quad \textit{and} \quad i^*j_* : \mathsf{UDM}(X^\circ/Z) \to \mathsf{D}(i^*j_*\mathbb{1}\text{–}\mathsf{Mod})$$

are canonically isomorphic under the identification $p^\circ_*\mathbb{1} \sim i^*j_*\mathbb{1}$.

Corollary 5.2. *The composition*

$$\mathsf{UDM}(X^\circ/Z) \subset \mathsf{DM}(X^\circ) \xrightarrow{\mathcal{L}_{X,D}} \mathsf{UDM}(X^\circ/Z)$$

is naturally isomorphic to the identity.

Proof. Immediate from the definition of $\mathcal{L}_{X,D}$, Proposition 5.1 and Corollary 3.7. □

Proof (Proof of the Proposition 5.1). Set $N := X$ and let $M \in \mathsf{UDM}(N/Z)$. As in Lemma 4.2 we get a module map $p_*^\circ M \cong p_*(j_*M) \to i^*j_*M$ along the map of algebras $p_*^\circ \mathbb{1} \to i^*j_*\mathbb{1}$ natural in M. We claim that this map is an isomorphism. Indeed, M can be written as a λ-sequence such that each transition map is the pushout by a map between objects coming from Z, hence we are reduced to the case where $M = (p^\circ)^*A$ with $A \in \mathsf{DM}(Z)$. We can forget the module structures. Proposition 3.5 applied to j_*M yields an exact triangle

$$j_!M \to j_*M \to i_*i^*j_*M \to j_!M[1] .$$

Applying p_* yields our transformation as the second map, hence we need to show that $p_*j_*j^*j_!M = p_*j_!M = 0$.

Set $M' = p^*A$, so $M = j^*M'$. Apply p_* to the triangle

$$j_!M \to M' \to i_*A \to j_!M[1] .$$

Since p is an $\mathbb{A}^1$-weak equivalence the second map is mapped to an isomorphism and $p_*j_!M = 0$. □

6 The motivic vanishing cycle functor

In this section we define the motivic vanishing cycle functor in the "unipotent" situation.

Let S be as in Section 4. Let again $X \in \mathsf{Sm}/S$, $D \subset X$ a relative NCD with smooth components D_i, $i \in I$, and D_I their intersection. Let N_I° be the product over D_I of the restricted pointed normal bundles of the D_i. For example, X could be a smooth curve over a field k and $D = \{x_0\}$, for some $x_0 \in X(k)$. Hence in this situation N_I° is isomorphic to $\mathbb{G}_{m,k}$.

Let $\pi : Y \to X$ be a morphism in Sch/S, $\pi^\circ : Y^\circ \to X^\circ$ the preimage of the open part X°, $\pi_I : Z := \pi^{-1}(D_I) \to D_I$ the special fiber. So there are the following pullback diagrams:

$$\begin{array}{ccccc} Z & \xrightarrow{\tilde{i}} & Y & \xleftarrow{\tilde{j}} & Y^\circ \\ \downarrow{\scriptstyle \pi_I} & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi^\circ} \\ D_I & \xrightarrow{i} & X & \xleftarrow{j} & X^\circ \end{array} .$$

Assume that for the categories $\mathsf{DM}(\cdot)$ the proper (or projective) base change isomorphism theorem holds (this has been proved by Voevodsky and is written up in [1]). For $\mathcal{F} \in \mathsf{DM}(Y^\circ)$ the sheaf $\tilde{i}^*\tilde{j}_*\mathcal{F}$ carries a canonical $\pi_I^*i^*j_*\mathbb{1}$-module structure. Furthermore, we have

$$\mathsf{D}(\pi_I^* i^* j_* \mathbb{1}\text{–}\mathsf{Mod}) \simeq \mathsf{UDM}(Z \times_{D_I} N_I^\circ / Z).$$

This defines the functor

$$\Psi_{\pi,I} : \mathsf{DM}(Y^\circ) \to \mathsf{UDM}(Z \times_{D_I} N_I^\circ / Z),$$

called the *generalized vanishing cycle* functor. Note that on fibers over N_I° it is the vanishing cycle sheaf (depending on the choice of tangent vector), and the whole sheaf encodes the monodromy.

Let

$$\begin{array}{c} Z \times_{D_I} N_I^\circ \\ {\scriptstyle \pi'}\big\downarrow \\ N_I^\circ \end{array}$$

be the projection. Let π be proper/projective. Assuming the proper/projective base change theorem it is clear that there is a canonical isomorphism

$$\mathcal{L}_{X,I}(\pi_*^\circ \mathcal{F}) \cong \pi'_*(\Psi_{\pi,I}(\mathcal{F})) ,$$

i.e., the limit motive is the pushforward of the vanishing cycle sheaf to the special locus in the base.

References

1. J. Ayoub – "Formalisme des 4 opérations", 2004, preprint.
2. A. Beilinson and P. Deligne – "Interprétation motivique de la conjecture de Zagier reliant polylogarithmes et régulateurs", Motives (Seattle, WA, 1991), Proc. Sympos. Pure Math., vol. 55, Amer. Math. Soc., Providence, RI, 1994, p. 97–121.
3. P. Deligne – "Le groupe fondamental de la droite projective moins trois points", Galois groups over **Q** (Berkeley, CA, 1987), Math. Sci. Res. Inst. Publ., vol. 16, Springer, New York, 1989, p. 79–297.
4. A. D. Elmendorf, I. Kriz, M. A. Mandell and J. P. May – *Rings, modules, and algebras in stable homotopy theory*, Mathematical Surveys and Monographs, vol. 47, American Mathematical Society, Providence, RI, 1997.
5. P. S. Hirschhorn – *Model categories and their localizations*, Mathematical Surveys and Monographs, vol. 99, American Mathematical Society, Providence, RI, 2003.
6. A. Hirschowitz and C. Simpson – "Descent for n-stacks", 1998, `math.AG/9807049`.
7. M. Hovey – *Model categories*, Mathematical Surveys and Monographs, vol. 63, American Mathematical Society, Providence, RI, 1999.
8. — , "Spectra and symmetric spectra in general model categories", *J. Pure Appl. Algebra* **165** (2001), no. 1, p. 63–127.
9. I. Kříž and J. P. May – "Operads, algebras, modules and motives", *Astérisque* **233** (1995), p. iv+145pp.

10. F. Morel and V. Voevodsky – "$\mathbf{A}^1$-homotopy theory of schemes", *Inst. Hautes Études Sci. Publ. Math.* **90** (1999), p. 45–143.
11. M. Spitzweck – "Operads, algebras and modules in model categories and motives", Thesis, University of Bonn, 2001.
12. V. Voevodsky – "$\mathbf{A}^1$-homotopy theory", *Doc. Math.* **Extra Vol. I** (1998), p. 579–604.

Counting points on cubic surfaces, II

Sir Peter Swinnerton-Dyer

DPMMS, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, U.K.
H.P.F.Swinnerton-Dyer@dpmms.cam.ac.uk

Let V be a non-singular surface defined over $\mathbb{Q}$ which is embedded in projective space $\mathbb{P}^n$ by means of anticanonical divisors, and let U be the open subset of V obtained by deleting the lines on V. For any point P in $U(\mathbb{Q})$ denote by $h(P)$ the height of P. In this paper h will usually be the standard height

$$h_1(P) = \max(|x_0|, \dots, |x_n|)$$

where $P = (x_0, \dots, x_n)$ for integers x_i with highest common factor 1; but this choice is in no way canonical, so that in a thorough investigation we should consider other heights also. Indeed, the eventual theory will almost certainly need the replacement of h by a generalized height — a concept which I am not able to make precise. Recall that for abelian varieties the canonical height function is actually a generalized height; and this is the only case in which a canonical height function is known with certainty, though the heights which have been defined for toric varieties and for certain toric bundles are probably also canonical.

For any integer $b > 0$ write

$$n(b) = n(b; U, h) = \#\{P \mid h(P) = b\},$$

the number of points in U of height exactly b; and for integral $B > 0$ write

$$N(B) = N(B; U, h) = n(1) + \cdots + n(B-1) + \tfrac{1}{2}n(B).$$

This can be regarded as the number of points in U of height at most B. The final factor $\frac{1}{2}$ is in accordance with convention; its effect on the fit between calculations and conjectures is small but helpful. If B is not an integer, write $N(B) = \sum_{b<B} n(b)$.

Let r be the rank of the Néron-Severi group of V over $\mathbb{Q}$. It is a conjecture of Manin (see [3]) that

$$N(B) \sim CB(\log B)^{r-1} \tag{1}$$

for some constant $C = C(U, h)$, which is non-zero unless there is a Brauer-Manin obstruction to the solubility of V; and Salberger [8], building on earlier work of Peyre [5], has proposed a formula for the value of C. Computational evidence for (1) in the special case of diagonal cubic surfaces was reported in [6] for $r = 1$ and in [7] for $r > 1$.

This paper describes the mixture of ideas and computation which has led me to formulate more precise conjectures related to this problem. The process of refining (1) is iterative. One first formulates a more detailed conjecture. This then suggests computations which will provide evidence about the plausibility or otherwise of that more detailed conjecture; and if the evidence is confirmatory, it may suggest a further refinement of the conjecture. This process is of course only available to those who think that a conjecture should be supported by evidence.

It is sensible to study $N(B)$ and the so-called height zeta-function

$$\xi(s) = \xi(s; U, h) = \sum (h(P))^{-s}$$

at the same time, where the sum is taken over all rational points P in U. This function was apparently first introduced by Arakelov. It follows from (1) that $\xi(s)$ is holomorphic in $\Re s > 1$ and

$$\xi(s) \longrightarrow \frac{C(r-1)!}{(s-1)^r} \text{ as } s \text{ tends to 1 within } \Re s > 1. \tag{2}$$

To go from (2) to (1) requires some additional conditions on $\xi(s)$; but in this context it is scarcely credible that (2) holds and (1) does not.

There are various toric bundles and toric varieties for which $\xi(s)$ has been determined explicitly, and this has also been done for abelian varieties. These examples might lead one to hazard some powerful conjectures. On the other hand, for the Veronese surface with $h = h_1$ we have

$$\xi(s) = 1 + 12\zeta(3s-2)/\zeta(3s) \tag{3}$$

where ζ is the Riemann zeta-function. For this surface one can be confident that the height function is not canonical; and in any case it appears to be in certain fundamental respects untypical. In particular, because the only values which $h(P)$ takes are cubes, we cannot hope to replace (1) by anything better than $N(B) = CB + O(B^{2/3})$; and standard methods of analytic number theory, applied to (3), do in fact yield $N(B) = 12B/\zeta(3) + O(B^{2/3+\epsilon})$. For this surface it is in any case much simpler to look at the associated affine cone — in other words, to look at $M(B)$, the total number of integral solutions of the equations of V, ignoring any questions of co-primacy. For $M(B)$ it is easy to give an explicit formula, by considering the Mellin transform of $2\xi(s)\zeta(3s) = 24\zeta(3s-2) + 2\zeta(3s)$.

For any V it is natural to hope that $\xi(s)$ can be continued into the half-plane $\Re s > c$ for some $c < 1$ as a function holomorphic except for an r-fold

pole at $s = 1$; indeed, if this did not appear likely to be true there would be little justification for introducing $\xi(s)$ at all. Computations which are directly relevant to the question of analytic continuation are difficult to implement; but it is easy to write down an assertion for $N(B)$ which has the same relation to the analytic continuability of $\xi(s)$ as (1) has to (2). This is

$$N(B) = Bf(\log B) + O(B^{c+\epsilon}) \tag{4}$$

for every $\epsilon > 0$, where f is a polynomial of degree $r - 1$. For the toric variety $X_0^3 = X_1X_2X_3$, for which $r = 6$, de la Bretèche [2] has obtained a result of this kind with $c = \frac{7}{8}$. For this variety there is a generalized height h which appears to be more canonical than h_1. This is given by $h(P) = |x_0|$ where $P = (x_0, x_1, x_2, x_3)$ is a minimal integral representation of P; and for it $\xi(s)$ can be expressed as an Euler product.

This conjecture is amenable to computational testing; but there are two practical constraints. The polynomial f contains r arbitrary constants; since the test of credibility is the goodness of fit between $N(B)/B$ and $f(\log B)$, it is preferable to make r as small as possible. Again, in general the time taken to compute $N(B)$ — or all the $N(b)$ for $b \leq B$, which is no harder — grows like B^3, and this is a serious limitation on the choice of B. (The amount of core memory needed is only $O(B)$, which is not a problem.) However, if V is a cubic surface given by an equation of the form

$$\phi(X_0, X_1) = \psi(X_2, X_3),$$

then the time needed is reduced to $O(B^2 \log B)$. The special case of diagonal cubics

$$a_0X_0^3 + a_1X_1^3 = a_2X_2^3 + a_3X_3^3 \tag{5}$$

is particularly advantageous, because we then only need $O(B)$ multiplications, all the rest of the arithmetic being done by additions or subtractions. (For a more detailed discussion see Bernstein [1]; he has also written freely available software for computing $N(B)$. But unfortunately rather little can be gained by incorporating congruence conditions into this method.) For computational purposes, I shall therefore confine myself to diagonal cubic surfaces (5) for which none of the three expressions like a_0a_1/a_2a_3 is a cube; the object of this last restriction is to ensure that $r = 1$, so that the polynomial f in (4) reduces to a constant C. The calculations described below have been carried out for a number of such surfaces, and they all support the conjectures which I shall put forward. I shall give explicit numerical results only for the surface

$$X_0^3 + 2X_1^3 = 5X_2^3 + 7X_3^3; \tag{6}$$

but this appears to be typical.

The numerical evidence strongly suggests that, at any rate for surfaces (5) with $r = 1$, we can take $c = \frac{1}{2}$ in (4). Indeed it even suggests that we can replace $O(B^{c+\epsilon})$ in (4) by $O(B^{1/2})$; but it would be rash to conjecture this on

the basis of numerical evidence alone. Write $E(B) = N(B) - CB$ when $r = 1$; then for example, for the surface (6) with $h = h_1$ we take $C = 0.3353$, which is probably correct to four places of decimals. For $1 \leq B \leq 20000$, which is the current limit of my enumeration of solutions of (6), we have

$$-0.6741B^{1/2} < E(B) < 0.7289B^{1/2}. \tag{7}$$

Here the critical case for the lower bound is $B = 1522$, and for the upper bound it is $B = 291$. If we confine ourselves to $1000 \leq B \leq 20000$, then the right-hand displayed inequality can be replaced by $E(B) < 0.4544B^{1/2}$, the critical case being when $B = 16305$.

Conjecture 1. *Let V be a nonsingular cubic surface defined over $\mathbb{Q}$, and let r be the rank of its Néron-Severi group over $\mathbb{Q}$. With the notation above, there is a polynomial $f(X)$ of degree $r-1$ with real coefficients such that for integral $B \geq 1$,*

$$N(B) = Bf(\log B) + O(B^{1/2+\epsilon})$$

for every $\epsilon > 0$. Moreover

$$\xi(s) = \xi(s; U, h) = \sum (h(P))^{-s}$$

can be analytically continued to $\Re s > \frac{1}{2}$, subject to an r-fold pole at $s = 1$.

There is a weaker form of this conjecture, in which it applies only to the case $r = 1$ and the polynomial $f(X)$ is replaced by a constant C. The numerical evidence which I have so far obtained only relates to this weaker version. In the opposite direction, it would be interesting to see whether the conjecture extends to singular cubic surfaces.

One expects $B^{-1/2}E(B)$ to be only very slowly varying. The numerical evidence suggests that it exhibits slow oscillations on which are superimposed what looks like white noise. There is now a suggestive analogy with the explicit formulae of prime number theory. (For the latter, see [4], Chapter 3.) This analogy leads one to hope that there is a formula

$$E(B) = \sum \Re c_\rho B^\rho + o(B^{1/2}) \tag{8}$$

where ρ runs through a discrete set of points on $\Re\rho = \frac{1}{2}$ in the upper half-plane and the c_ρ are complex constants. Here it is implicit that the series is convergent but not necessarily absolutely convergent; and it may be hoped that the error term has a comparatively simple form and is continuous in $B > 0$. We shall write $\rho = \frac{1}{2} + i\tau$. In contrast with what happens in prime number theory, we cannot expect any simple relationship between ρ and c_ρ; for if we replace h by nh for some positive integer n, which can be easily achieved by changing the projective embedding of V, then the ρ remain the same but each c_ρ is multiplied by $n^{-\rho}$.

In practice, in the estimation of the ρ and c_ρ one has to work with the sum of the squares of the errors, rather than with the maximum of the absolute values of the errors; but it is convenient to exhibit the latter also, as a guide to how successful the operation has been. The estimation process is somewhat analogous to that of computing the power spectrum of $E(B)$ as a function of $\log B$, though there are some additional complications: in particular, we cannot afford to use the small values of B because of the presence of the last term in (8), and it is desirable to introduce a weighting factor because the number of data points increases exponentially with $\log B$. In concrete terms, we are trying to choose a limited number of pairs ρ, c_ρ so as to make

$$S = \sum_{b=50}^{20000} b^{-2}(E'(b))^2 \tag{9}$$

small, where we have written the residual error as

$$E'(b) = N(b) - Cb - \Re \sum c_\rho b^\rho.$$

Here the lower bound in the sum (9) has hopefully been chosen large enough to render the effect of the final term in (8) unimportant. The initial value of S, before we introduce any pairs ρ, c_ρ, is 0.4796. By introducing the six pairs listed below, we reduce this to 0.1012; and we improve (7) to

$$-0.4573B^{1/2} < E'(B) < 0.3487B^{1/2}.$$

The critical case for the lower bound is $B = 62$, and for the upper bound it is $B = 8863$. If we confine ourselves to $300 \le B \le 20000$, then the left-hand displayed inequality can be replaced by $-0.3810B^{1/2} < E(B)$, the critical case being when $B = 2035$. The following table gives the six pairs used, in the order in which they are introduced. The four columns give respectively $\rho - \frac{1}{2}$, c_ρ, the reduction in S achieved by introducing this pair, and the resulting value of S; but I cannot assess how close to the true values the estimates in the first two columns are.

$1.58i$	$-0.2182 + 0.0037i$	0.1436	0.3360
$3.13i$	$0.1920 - 0.0066i$	0.1099	0.2261
$5.10i$	$-0.1257 + 0.0017i$	0.0471	0.1790
$8.85i$	$0.1068 - 0.0048i$	0.0338	0.1452
$6.86i$	$-0.0984 + 0.0026i$	0.0288	0.1164
$3.96i$	$-0.0719 - 0.0083i$	0.0153	0.1011

In my opinion, these figures strongly support Conjecture 1; and I believe that there are only six values of ρ for which $0 < \tau < 10$. It can scarcely be an accident that the values for c_ρ listed are all nearly real; but I have no idea why this occurs. It would no longer be true if h_1 was replaced by $h = 2h_1$.

Conjecture 2. *For given V and h there are sequences of complex numbers c_n and of points $\rho_n = \frac{1}{2} + i\tau_n$, with the τ_n real, positive and monotone increasing, having the following properties:*

(i) $\sum |c_n|^2$ *and* $\sum \tau_n^{-2}$ *are both convergent;*

(ii) *for some fixed* $c < \frac{1}{2}$ *and all* $\epsilon > 0$,

$$N(B) = Bf(\log B) + \Re \sum c_n B^{\rho_n} + O(B^{c+\epsilon}); \tag{10}$$

(iii) $\xi(s)$ *can be analytically continued to* $\Re s > c$ *with the same value of* c *as in* (ii), *subject to an* r*-fold pole at* $s = 1$ *and simple poles at each* ρ_n *and* $\overline{\rho_n}$.

The assertion that $\sum |c_n|^2$ converges is essentially equivalent to claiming that the weighted average of the $B^{-1}(E(B))^2$ is finite. The computational evidence is compatible with taking $c = 0$, but some intervention from the eye of faith is needed to make this conclusion convincing.

There is another and probably better way of expressing (iii). Let $\eta(s)$ be an integral function of order 1 which has simple zeros at the ρ_n and $\overline{\rho_n}$ and no zeros elsewhere. Such a function exists because $\sum \tau_n^{-2}$ converges, and it is uniquely determined up to a factor $\exp(a_1 s + a_2)$. Then (iii) is equivalent to the assertion that $\xi(s)\eta(s)$ can be analytically continued to $\Re s > c$ subject only to an r-fold pole at $s = 1$. With this definition, each of $\eta(s)$ and $\xi(s)\eta(s)$ has at least some of the expected properties of a number-theoretic function related to a Dirichlet series, and no obvious properties inconsistent with this.

By implication, this paper raises a number of questions to which I do not have even a conjectural answer. For some of them, the most interesting answer may well involve replacing the height h by a canonical generalized height function. Among the most obvious are the following:

1. Can one describe the whole polynomial f in terms of V and h, in the same way as its leading coefficient was described in [5] and [8]?
2. What is the nature of the error term in (10)? Note that the corresponding error term in the explicit formula of prime number theory is
$$-\log 2\pi - \tfrac{1}{2}\log(1 - B^{-2}).$$
3. Can $\xi(s)\eta(s)$ be analytically continued to the whole s-plane, subject to the pole at $s = 1$ and possible poles at negative integers? If so, does it satisfy a functional equation?
4. The function $\eta(s)$ has some of the properties of an L-series: it comes from number theory, and it satisfies a functional equation and the Riemann hypothesis. Can it be written as a (possibly generalized) Dirichlet series, and if so, can this series be written as an Euler product?

References

1. D. J. Bernstein – "Enumerating solutions to $p(a) + q(b) = r(c) + s(d)$", *Math. Comp.* **70** (2001), no. 233, p. 389–394.
2. R. de la Bretèche – "Sur le nombre de points de hauteur bornée d'une certaine surface cubique singulière", *Astérisque* (1998), no. 251, p. 51–77, Nombre et répartition de points de hauteur bornée (Paris, 1996).
3. J. Franke, Y. I. Manin and Y. Tschinkel – "Rational points of bounded height on Fano varieties", *Invent. Math.* **95** (1989), no. 2, p. 421–435.
4. S. J. Patterson – *An introduction to the theory of the Riemann zeta-function*, Cambridge Studies in Advanced Mathematics, vol. 14, Cambridge University Press, Cambridge, 1988.
5. E. Peyre – "Hauteurs et mesures de Tamagawa sur les variétés de Fano", *Duke Math. J.* **79** (1995), no. 1, p. 101–218.
6. E. Peyre and Y. Tschinkel – "Tamagawa numbers of diagonal cubic surfaces, numerical evidence", *Math. Comp.* **70** (2001), no. 233, p. 367–387.
7. — , "Tamagawa numbers of diagonal cubic surfaces of higher rank", Rational points on algebraic varieties, Progr. Math., vol. 199, Birkhäuser, Basel, 2001, p. 275–305.
8. P. Salberger – "Tamagawa measures on universal torsors and points of bounded height on Fano varieties", *Astérisque* (1998), no. 251, p. 91–258, Nombre et répartition de points de hauteur bornée (Paris, 1996).

Quantum cohomology of isotropic Grassmannians

Harry Tamvakis

Department of Mathematics, Brandeis University - MS 050, P. O. Box 9110, Waltham, MA 02454, U.S.A.
harryt@brandeis.edu

Summary. Let G be a classical Lie group and P a maximal parabolic subgroup. We describe a quantum Pieri rule which holds in the small quantum cohomology ring of G/P. We also give a presentation of this ring in terms of special Schubert class generators and relations. This is a survey paper which reports on joint work with Anders S. Buch and Andrew Kresch.

1 Introduction

Let G be a classical Lie group and P any maximal parabolic subgroup of G. Our aim in this paper is to discuss what is known about two questions regarding the small quantum cohomology ring of the homogeneous space $X = G/P$. The first problem is to formulate and prove a 'quantum Pieri rule' in the ring $QH^*(X)$, that is, a combinatorial rule which describes the quantum product of a general Schubert class with a 'special' Schubert class. The special classes should generate $QH^*(X)$, and we also seek a presentation for this ring in terms of these generators and relations.

In Lie type A, the answers to both of the above questions are well known, as X is an ordinary Grassmannian; see e.g., [9] for the classical story and [16], [1] for the extension to quantum cohomology. In the symplectic and orthogonal Lie types, the homogeneous space X is a Grassmannian which parametrizes isotropic subspaces in a vector space equipped with a nondegenerate skew-symmetric or symmetric form. When the subspaces in question have the maximum possible dimension, the corresponding classical analysis is contained in [6], [2], [8], while [10], [11] deal with the quantum case.

The above results about $QH^*(X)$ depend upon a study of the three point, genus zero Gromov–Witten invariants on X. The analysis in [1], [10], [11] required intersection theory on certain Quot scheme compactifications of the moduli space of degree d rational maps to X. More recently, using his idea of the kernel and span of a rational map to a Grassmannian, Buch [3] gave new proofs of the main structure theorems for the quantum cohomology ring. These

arguments require only basic algebraic geometry, assuming the associativity of the quantum product. The paper [4] used similar techniques to deal with the case of maximal isotropic Grassmannians. It emerged in each of these examples that the relevant Gromov–Witten invariants on X were equal (or related) to classical triple intersection numbers on other homogeneous spaces.

The latter ideas and methods have been extended in [5] to study the quantum cohomology ring of non-maximal isotropic Grassmannians X. In contrast, even the classical cohomology of these spaces is still rather unexplored; [2], [15], and [13], [14] are some previous works in this direction. The present paper includes the main results of [5], which give a presentation over $\mathbb{Z}$ for the quantum cohomology ring $QH^*(X)$ and a quantum Pieri rule as well. An important feature of the analysis in loc. cit. is that it does not lead to a quantum extension of the classical Pieri rules of Pragacz and Ratajski from [13], [14]. Instead, we need to use a different set of special Schubert classes, which are equal (or related) to the Chern classes of the universal *quotient* bundle over X. The resulting classical Pieri rules are simpler than the previously known ones, admit straightforward quantum extensions, and are parallel to the aforementioned examples (both type A and maximal isotropic).

This paper is organized as follows. Section 2 discusses the type A Grassmannian. The next section considers both the Lagrangian Grassmannian (in type C) and the maximal isotropic orthogonal Grassmannian (in types B and D). The last three sections study the non-maximal isotropic Grassmannians in types C, B, and D, respectively, and present the main results of [5].

Acknowledgments. Some of the results included here were announced at the Miami Winter School on 'Geometric Methods in Algebra and Number Theory' in December of 2003. It is a pleasure to thank the organizers Fedor Bogomolov, Bruno de Oliveira, Yuri Tschinkel, and Alan Zame for making this stimulating event possible. I also thank my collaborators Anders Buch and Andrew Kresch, without whom this survey paper could not have been written. This work was partially supported by NSF Grant DMS-0296023.

2 The type A Grassmannian

We first discuss the relevant facts about the classical and small quantum cohomology of type A Grassmannians, namely the Pieri rule and the presentation of the ring in terms of generators and relations. Let $V = \mathbb{C}^N$ and X be the Grassmannian which parametrizes m-dimensional complex linear subspaces Σ of V. We will use $G(m, N)$ or $G(m, V)$ to denote X, depending on the context. X is a smooth complex manifold of dimension mn, where $n = N - m$.

Let $\mathcal{R}(m, n)$ denote the set of integer partitions $\lambda = (\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m \geqslant 0)$ with $\lambda_1 \leqslant n$, so that the Young diagram of λ fits inside an $m \times n$ rectangle (see Figure 1). For every $\lambda \in \mathcal{R}(m, n)$, we have a Schubert variety $X_\lambda(F_\bullet)$ in X, which also depends on the choice of a complete flag of subspaces

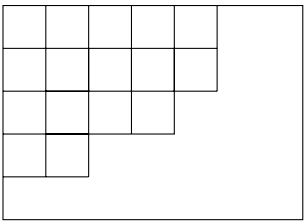


Fig. 1. The partition $(5,5,4,2)$ in $\mathcal{R}(5,7)$

$$F_\bullet \ : \ 0 = F_0 \subset F_1 \subset \cdots \subset F_N = V$$

in V. The variety $X_\lambda(F_\bullet)$ is defined as the locus of $\Sigma \in X$ such that

$$\dim(\Sigma \cap F_{n+i-\lambda_i}) \geqslant i$$

for all $i = 1, \ldots, m$, and has codimension $|\lambda| = \sum \lambda_i$ in X. We let $\sigma_\lambda = [X_\lambda(F_\bullet)]$ denote the corresponding Schubert class in $H^{2|\lambda|}(X, \mathbb{Z})$; the set of all Schubert classes σ_λ for $\lambda \in \mathcal{R}(m,n)$ forms a $\mathbb{Z}$-basis for the cohomology of X. As all cohomology classes occur in even degrees, we will adopt the convention that the degree of a class $\alpha \in H^{2k}(X, \mathbb{Z})$ is equal to k throughout this paper.

The varieties $X_p(F_\bullet)$ for $p = 1, \ldots, n$ are called *special Schubert varieties*, and the corresponding cohomology classes are the *special Schubert classes* $\sigma_1, \ldots, \sigma_n$. The Schubert variety $X_p(F_\bullet)$ may be defined by a single Schubert condition; in fact, it only depends on the subspace F_{n+1-p}:

$$X_p(F_\bullet) = \{\, \Sigma \in X \mid \Sigma \cap F_{n+1-p} \neq 0 \,\}.$$

Consider the universal short exact sequence of vector bundles over X,

$$0 \to S \to V_X \to Q \to 0, \tag{1}$$

where V_X denotes the trivial vector bundle of rank N over X, S is the tautological subbundle of rank m, and Q is the quotient bundle. The special Schubert class σ_p is equal to the p-th Chern class $c_p(Q)$, essentially by the definition of Chern classes. Indeed, let $\pi : P(S) \to X$ be the projection map and $\eta : P(S) \to P(V) = \mathbb{P}^{N-1}$ be the natural morphism induced by the inclusion $S \hookrightarrow V_X$. We then have that

$$\sigma_p = [X_p(F_\bullet)] = \pi_* \eta^* [P(F_{n+1-p})] = \pi_* \eta^* c_1(\mathcal{O}_{P(V)}(1))^{m-1+p}.$$

On the other hand, $\eta^* c_1(\mathcal{O}_{P(V)}(1)) = c_1(\eta^* \mathcal{O}_{P(V)}(1)) = c_1(\mathcal{O}_{P(S)}(1))$. Therefore

$$\pi_* \eta^* c_1(\mathcal{O}_{P(V)}(1))^{m-1+p} = \pi_* c_1(\mathcal{O}_{P(S)}(1))^{m-1+p} = s_p(S),$$

where $s_p(S)$ denotes the p-th Segre class of S (see e.g., [7, Sect. 3.1] for the definition of Segre classes). Now the Whitney sum formula applied to the

sequence (1) states that $c(S)c(Q) = 1$ in $H^*(X, \mathbb{Z})$; it follows that $s_p(S) = c_p(Q)$.

The ring structure of $H^*(X, \mathbb{Z})$ is determined by the classical *Pieri rule* [12], which gives the product of a general Schubert class with a special one. For $1 \leqslant p \leqslant n$ we have

$$\sigma_\lambda \, \sigma_p = \sum \sigma_\mu,$$

where the sum is over all $\mu \in \mathcal{R}(m, n)$ obtained from the Young diagram of λ by adding p boxes, with no two in the same column. A skew diagram μ/λ which does not contain two boxes in the same column is called a *horizontal strip*.

One can use the Pieri rule to show that the special Schubert classes generate the ring $H^*(X, \mathbb{Z})$. Moreover, the cohomology of X may be presented as a quotient of the polynomial ring $\mathbb{Z}[\sigma_1, \ldots, \sigma_n]$ by the relations

$$\det(\sigma_{1+j-i})_{1 \leqslant i,j \leqslant r} = 0, \quad m+1 \leqslant r \leqslant N.$$

The Whitney sum formula $c(S)c(Q) = 1$ can be used to show that these relations hold in $H^*(X, \mathbb{Z})$, since the coefficients of the inverse power series

$$c_t(Q)^{-1} = (1 + \sigma_1 t + \cdots + \sigma_n t^n)^{-1} = c_t(S)$$

must vanish in degrees higher than $m = \mathrm{rank}(S)$.

To extend the above picture to the quantum cohomology of X, we need to recall the enumerative definition of three point, genus zero Gromov–Witten invariants. We agree that a rational map of degree d to X is a morphism $f\colon \mathbb{P}^1 \to X$ such that

$$\int_X f_*[\mathbb{P}^1] \cdot \sigma_1 = d.$$

Given a degree $d \geqslant 0$ and partitions λ, μ, and ν such that $|\lambda| + |\mu| + |\nu| = mn + dN$, the Gromov–Witten invariant $\langle \sigma_\lambda, \sigma_\mu, \sigma_\nu \rangle_d$ is defined as the number of rational maps $f\colon \mathbb{P}^1 \to X$ of degree d such that $f(0) \in X_\lambda(F_\bullet)$, $f(1) \in X_\mu(G_\bullet)$, and $f(\infty) \in X_\nu(H_\bullet)$, for given flags $F_\bullet$, $G_\bullet$, and $H_\bullet$ in general position.

The (small) *quantum cohomology ring* $QH^*(X)$ is a $\mathbb{Z}[q]$-algebra which is isomorphic to $H^*(X, \mathbb{Z}) \otimes_{\mathbb{Z}} \mathbb{Z}[q]$ as a module over $\mathbb{Z}[q]$, where q is a formal variable. This will be the case for all the varieties considered in this paper; for the type A Grassmannian $G(m, N)$, the degree of the variable q is equal to N. The multiplicative structure of $QH^*(X)$ is defined using the relation

$$\sigma_\lambda \cdot \sigma_\mu = \sum \langle \sigma_\lambda, \sigma_\mu, \sigma_{\widehat{\nu}} \rangle_d \, \sigma_\nu \, q^d, \tag{2}$$

the sum over $d \geqslant 0$ and partitions ν with $|\nu| = |\lambda| + |\mu| - dN$. Here $\widehat{\nu}$ denotes the dual partition of ν, defined so that

$$\int_X \sigma_\nu \sigma_{\widehat{\nu}} = 1.$$

One observes that the $d = 0$ terms of the sum in (2) just give the classical cup product in the cohomology ring of X.

The ring structure of $QH^*(X)$ (and hence all three point, genus zero Gromov–Witten invariants on X) is determined by Bertram's quantum Pieri rule. To state it, we let $\mathcal{R}'(m+1,n)$ denote the set of partitions $\mu \in \mathcal{R}(m+1,n)$ with $\mu_1 = n$ and $\mu_{m+1} \geqslant 1$. For any $\mu \in \mathcal{R}'(m+1,n)$, define a partition $\widetilde{\mu} \in \mathcal{R}(m,n)$ by removing a hook of length N from μ; in other words,

$$\widetilde{\mu} = (\mu_2 - 1, \ldots, \mu_{m+1} - 1).$$

Theorem 1 ([1]). *For $1 \leqslant p \leqslant n$, we have*

$$\sigma_\lambda \cdot \sigma_p = \sum_{\mu \in \mathcal{R}(m,n)} \sigma_\mu + \sum_{\mu \in \mathcal{R}'(m+1,n)} \sigma_{\widetilde{\mu}}\, q, \tag{3}$$

where both sums are over diagrams μ obtained from λ by adding p boxes, with no two in the same column.

We remark that the partitions $\widetilde{\mu}$ which appear in the second sum in (3) are exactly those $\nu \in \mathcal{R}(m,n)$ such that $|\nu| = |\lambda| + p - N$ and

$$\lambda_1 - 1 \geqslant \nu_1 \geqslant \lambda_2 - 1 \geqslant \nu_2 \geqslant \cdots \geqslant \lambda_m - 1 \geqslant \nu_m \geqslant 0.$$

Example 1. For the Grassmannian $G(4,8)$, we have

$$\sigma_{4,3,1,1} \cdot \sigma_2 = \sigma_{4,4,2,1} + \sigma_{4,3,3,1} + \sigma_3\, q + \sigma_{2,1}\, q$$

in the quantum cohomology ring $QH^*(G(4,8))$.

In particular, it is easy to deduce Siebert and Tian's presentation of the quantum cohomology ring of X from the quantum Pieri rule; see [3] for details.

Theorem 2 ([16]). *The ring $QH^*(X)$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\sigma_1, \ldots, \sigma_n, q]$ by the relations*

$$\det(\sigma_{1+j-i})_{1 \leqslant i,j \leqslant r} = 0, \qquad m+1 \leqslant r \leqslant N-1$$

and

$$\det(\sigma_{1+j-i})_{1 \leqslant i,j \leqslant N} = (-1)^{n-1} q.$$

We will now give an approach to proving Theorem 1 using results of [4] together with the projection formula. Although one does not need all of these constructions for the proof (see the Introduction and [3]), they play an important role in the following sections. The main result of [4] equates all of the above degree d Gromov–Witten invariants on X with classical triple intersection numbers on a two-step flag variety $Y_d = F(m-d, m+d; N)$. The variety Y_d parametrizes pairs of subspaces (A, B) with $A \subset B \subset V$, $\dim A = m - d$ and $\dim B = m + d$; we agree that Y_d is empty if $d > \min(m,n)$. To each

Schubert variety $X_\lambda(F_\bullet)$ in X, we associate a Schubert variety $X_\lambda^{(d)}(F_\bullet)$ in Y_d via the prescription

$$X_\lambda^{(d)}(F_\bullet) = \{\, (A,B) \in Y_d \,|\, \exists\ \Sigma \in X_\lambda(F_\bullet) : A \subset \Sigma \subset B \,\}\,. \tag{4}$$

We let $\sigma_\lambda^{(d)}$ denote the class of $X_\lambda^{(d)}(F_\bullet)$ in $H^*(Y_d,\mathbb{Z})$. Consider three partitions λ, μ, and ν such that $|\lambda|+|\mu|+|\nu| = mn+dN$. We then have

$$\langle \sigma_\lambda, \sigma_\mu, \sigma_\nu \rangle_d = \int_{F(m-d,m+d;N)} \sigma_\lambda^{(d)} \cdot \sigma_\mu^{(d)} \cdot \sigma_\nu^{(d)}. \tag{5}$$

It is easy to justify the $d=1$ case of (5), since a straightforward argument shows that the flag variety $Y_1 = F(m-1,m+1;N)$ is exactly the parameter space of lines on the Grassmannian $G(m,N)$.

The proof of (5) in general proceeds by considering, for any morphism $f : \mathbb{P}^1 \to X$, the pair $(\mathrm{Ker}(f), \mathrm{Span}(f))$ consisting of the *kernel* and *span* of f. Here, the kernel (respectively, the span) of f is defined as the intersection (respectively, the linear span) of all the subspaces $\Sigma \subset V$ corresponding to image points of f. Setting $a = \dim \mathrm{Ker}(f)$ and $b = \dim \mathrm{Span}(f)$, a dimension count on the flag variety $F(a,b;N)$ establishes that $d \leq \min(m,n)$ whenever $\langle \sigma_\lambda, \sigma_\mu, \sigma_\nu \rangle_d \neq 0$. For such d, one then shows that the map $f \mapsto (\mathrm{Ker}(f), \mathrm{Span}(f))$ is a bijection between the set of morphisms f counted by the Gromov–Witten invariant on the left-hand side of (5) and the points (A,B) in the triple intersection $X_\lambda^{(d)}(F_\bullet) \cap X_\mu^{(d)}(G_\bullet) \cap X_\nu^{(d)}(H_\bullet)$ in Y_d, assuming the flags $F_\bullet$, $G_\bullet$, and $H_\bullet$ are in general position.

For any Young diagram λ, let $\overline{\lambda}$ denote the diagram obtained by deleting the leftmost column of λ. In terms of partitions, we have $\overline{\lambda}_i = \max\{\lambda_i - 1, 0\}$. Given any Schubert variety $X_\lambda(F_\bullet)$ in $G(m,V)$, we will consider an associated Schubert variety $X_{\overline{\lambda}}(F_\bullet)$ in $G(m+1,V)$, with cohomology class $\sigma_{\overline{\lambda}}$. It is straightforward to check that the quantum Pieri rule follows from the two relations

$$\langle \sigma_\lambda, \sigma_\mu, \sigma_p \rangle_d = 0 \quad \text{for all } d \geqslant 2 \tag{6}$$

and

$$\langle \sigma_\lambda, \sigma_\mu, \sigma_p \rangle_1 = \int_{G(m+1,N)} \sigma_{\overline{\lambda}} \cdot \sigma_{\overline{\mu}} \cdot \sigma_{\overline{p}}. \tag{7}$$

Given (5), the vanishing assertion (6) is proved by a dimension count, which shows that the sum of the codimensions of the three Schubert varieties $X_\lambda^{(d)}(F_\bullet)$, $X_\mu^{(d)}(G_\bullet)$, and $X_p^{(d)}(H_\bullet)$ is strictly greater than the dimension of Y_d.

To establish (7), one may work as follows. Consider the three-step flag variety $Z = F(m-1,m,m+1;N)$, with its natural projections $\pi_1 : Z \to X$ and $\pi_2 : Z \to Y_1$. Note that for every $\lambda \in \mathcal{R}(m,n)$, we have $X_\lambda^{(1)}(F_\bullet) = \pi_2(\pi_1^{-1}(X_\lambda(F_\bullet)))$. The morphism π_2 lies on the left-hand side of a commutative diagram

$$\begin{array}{ccccc} Z & \xrightarrow{\varphi_2} & F(m,m+1;N) & \xrightarrow{\varphi_1} & G(m,N) \\ \downarrow{\scriptstyle \pi_2} & & \downarrow{\scriptstyle \pi} & & \\ Y_1 & \xrightarrow{\psi} & G(m+1,N) & & \end{array}$$

where every arrow is a natural smooth projection map and $\pi_1 = \varphi_1\varphi_2$. We now apply (5) when $d = 1$, $\nu = (p)$ and use the projection formula repeatedly to obtain

$$\begin{aligned} \langle \sigma_\lambda, \sigma_\mu, \sigma_p \rangle_1 &= \int_{Y_1} \sigma_\lambda^{(1)} \cdot \sigma_\mu^{(1)} \cdot \sigma_p^{(1)} \\ &= \int_{Y_1} \pi_{2*}\pi_1^*\sigma_\lambda \cdot \pi_{2*}\pi_1^*\sigma_\mu \cdot \psi^*\sigma_{p-1} \\ &= \int_Z \pi_2^*\pi_{2*}\pi_1^*\sigma_\lambda \cdot \pi_1^*\sigma_\mu \cdot \varphi_2^*\pi^*\sigma_{p-1} \\ &= \int_{F(m,m+1;N)} \varphi_{2*}\pi_2^*\pi_{2*}\pi_1^*\sigma_\lambda \cdot \varphi_1^*\sigma_\mu \cdot \pi^*\sigma_{p-1} \\ &= \int_{F(m,m+1;N)} \pi^*\pi_*\varphi_1^*\sigma_\lambda \cdot \varphi_1^*\sigma_\mu \cdot \pi^*\sigma_{p-1} \\ &= \int_{G(m+1;N)} \pi_*\varphi_1^*\sigma_\lambda \cdot \pi_*\varphi_1^*\sigma_\mu \cdot \sigma_{p-1} \\ &= \int_{G(m+1,N)} \sigma_{\overline{\lambda}} \cdot \sigma_{\overline{\mu}} \cdot \sigma_{\overline{p}}. \end{aligned}$$

3 Maximal isotropic Grassmannians

In this section we will consider the Grassmannians of maximal isotropic subspaces in a vector space equipped with a nondegenerate symmetric or skew-symmetric bilinear form.

3.1 The Lagrangian Grassmannian $LG(n, 2n)$

We begin in type C with the Lagrangian Grassmannian $LG = LG(n, 2n)$ parametrizing Lagrangian subspaces in $V = \mathbb{C}^{2n}$. The variety LG has complex dimension $n(n+1)/2$. Let $\mathcal{D}_n$ denote the set of all strict partitions $\lambda = (\lambda_1 > \lambda_2 > \cdots > \lambda_\ell > 0)$ with $\lambda_1 \leqslant n$, and fix a complete isotropic flag of subspaces of V

$$F_\bullet \ : \ 0 = F_0 \subset F_1 \subset \cdots \subset F_n \subset V$$

where $\dim(F_i) = i$ for each i, and F_n is Lagrangian. For each $\lambda \in \mathcal{D}_n$, the codimension $|\lambda|$ Schubert variety $X_\lambda(F_\bullet) \subset LG$ is defined as the locus of $\Sigma \in LG$ such that

$$\dim(\Sigma \cap F_{n+1-\lambda_i}) \geqslant i, \quad \text{for} \quad i = 1, \ldots, \ell(\lambda). \tag{8}$$

Here $\ell(\lambda)$ denotes the length of λ, that is, the number of non-zero parts in λ. Let σ_λ be the class of $X_\lambda(F_\bullet)$ in the cohomology group $H^{2|\lambda|}(LG, \mathbb{Z})$; the classes σ_λ for $\lambda \in \mathcal{D}_n$ then form a $\mathbb{Z}$-basis of the cohomology of LG.

The classes $\sigma_1, \ldots, \sigma_n$ are again called special; each σ_p is the class of a special Schubert variety $X_p(F_\bullet)$, which is defined by a single Schubert condition, as in type A:

$$X_p(F_\bullet) = \{\, \Sigma \in LG \mid \Sigma \cap F_{n+1-p} \neq 0 \,\}.$$

Furthermore, if

$$0 \to S \to V_X \to Q \to 0$$

denotes the tautological short exact sequence of vector bundles over LG, then Q can be canonically identified with S^*, and we have $\sigma_p = c_p(S^*)$, for $0 \leqslant p \leqslant n$, as in Section 2.

The classical Pieri rule for LG is due to Hiller and Boe [8]. It states that for any $\lambda \in \mathcal{D}_n$ and $p = 1, \ldots, n$ we have

$$\sigma_\lambda \, \sigma_p = \sum_\mu 2^{N(\lambda,\mu)} \sigma_\mu \tag{9}$$

in $H^*(LG, \mathbb{Z})$, where the sum is over all strict partitions μ obtained from λ by adding p boxes, with no two in the same column, and $N(\lambda, \mu)$ is the number of connected components of the skew diagram μ/λ which do not meet the first column (the connected components of a skew diagram α are defined by letting two boxes in α be connected if they share a vertex or an edge).

The ring $H^*(LG, \mathbb{Z})$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\sigma_1, \ldots, \sigma_n]$ modulo the relations

$$c_t(S)c_t(S^*) = (1 - \sigma_1 t + \cdots + (-1)^n \sigma_n t^n)(1 + \sigma_1 t + \cdots + \sigma_n t^n) = 1. \tag{10}$$

By equating the coefficients of like powers of t in (10), we see that these relations are given by

$$\sigma_r^2 + 2 \sum_{i=1}^{n-r} (-1)^i \sigma_{r+i} \sigma_{r-i} = 0$$

for $1 \leqslant r \leqslant n$, where we define $\sigma_0 = 1$ and $\sigma_j = 0$ for $j < 0$.

A rational map to LG is a morphism $f\colon \mathbb{P}^1 \to LG$, and its degree is the degree of $f_*[\mathbb{P}^1] \cdot \sigma_1$. The Gromov–Witten invariant $\langle \sigma_\lambda, \sigma_\mu, \sigma_\nu \rangle_d$ is defined for $|\lambda| + |\mu| + |\nu| = \dim(LG) + d(n+1)$ and counts the number of rational maps $f\colon \mathbb{P}^1 \to LG(n, 2n)$ of degree d such that $f(0) \in X_\lambda(F_\bullet)$, $f(1) \in X_\mu(G_\bullet)$, and $f(\infty) \in X_\nu(H_\bullet)$, for given isotropic flags $F_\bullet$, $G_\bullet$, and $H_\bullet$ in general position. The quantum cohomology ring of LG is a $\mathbb{Z}[q]$-algebra isomorphic to $H^*(LG, \mathbb{Z}) \otimes_{\mathbb{Z}} \mathbb{Z}[q]$ as a module over $\mathbb{Z}[q]$, but here q is a formal variable of degree $n+1$. The product in $QH^*(LG)$ is defined by the same equation (2) as before. We can now state the quantum Pieri rule for LG, which extends the classical rule of Hiller and Boe.

Theorem 3 ([10]). *For any $\lambda \in \mathcal{D}_n$ and $p \geqslant 1$ we have*

$$\sigma_\lambda \cdot \sigma_p = \sum_\mu 2^{N(\lambda,\mu)} \sigma_\mu + \sum_\nu 2^{N'(\nu,\lambda)} \sigma_\nu \, q$$

in $QH^*(LG(n,2n))$, where the first sum is classical, as in (9), while the second is over all strict ν obtained from λ by subtracting $n+1-p$ boxes, no two in the same column, and $N'(\nu,\lambda)$ is one less than the number of connected components of λ/ν.

Example 2. For the Grassmannian $LG(4,8)$, the relations

$$\sigma_{3,2} \cdot \sigma_3 = 2\,\sigma_{4,3,1} + \sigma_3\, q + \sigma_{2,1}\, q \quad \text{and} \quad \sigma_{4,2} \cdot \sigma_3 = \sigma_{4,3,2} + \sigma_4\, q + 2\,\sigma_{3,1}\, q$$

hold in the quantum cohomology ring $QH^*(LG)$.

The proof of Theorem 3 from [4] uses a result similar to (5) which holds for LG. In this setting, the role of the two-step flag variety Y_d is played by a non-maximal isotropic Grassmannian $IG(n-d,2n)$. This is because the span of a rational map $\mathbb{P}^1 \to X$ is the orthogonal complement of its kernel, and hence is redundant. We have that for any λ, μ, $\nu \in \mathcal{D}_n$ such that $|\lambda| + |\mu| + |\nu| = n(n+1)/2 + d(n+1)$,

$$\langle \sigma_\lambda, \sigma_\mu, \sigma_\nu \rangle_d = \int_{IG(n-d,2n)} \sigma_\lambda^{(d)} \cdot \sigma_\mu^{(d)} \cdot \sigma_\nu^{(d)}.$$

Here, for each $\lambda \in D_n$, $\sigma_\lambda^{(d)}$ denotes the cohomology class of a Schubert variety $X_\lambda^{(d)}(F_\bullet)$ in $IG(n-d,2n)$, defined by an equation directly analogous to (4). To compute the line numbers $\langle \sigma_\lambda, \sigma_\mu, \sigma_\nu \rangle_1$, one shows that up to a factor of 2, they are equal to classical intersection numbers on the Lagrangian Grassmannian $LG(n+1,2n+2)$. More precisely, we have that

$$\int_{IG(n-1,2n)} \sigma_\lambda^{(1)} \cdot \sigma_\mu^{(1)} \cdot \sigma_\nu^{(1)} = \frac{1}{2} \int_{LG(n+1,2n+2)} \sigma_\lambda^{+} \cdot \sigma_\mu^{+} \cdot \sigma_\nu^{+}, \qquad (11)$$

where σ_λ^+, σ_μ^+, σ_ν^+ denote Schubert classes in $H^*(LG(n+1,2n+2),\mathbb{Z})$.

We give a brief discussion of the geometric proof of (11), since an analogous argument works to prove the quantum Pieri rule on any isotropic Grassmannian in type C. Let H be a 2-dimensional symplectic vector space and let V^+ be the orthogonal direct sum of V and H. We then consider the correspondence between $LG(n+1,2n+2)$ and $IG(n-1,2n)$ consisting of pairs (Σ^+, Σ') with Σ^+ a Lagrangian subspace of V^+ and Σ' an isotropic $(n-1)$-dimensional subspace of V, given by the condition $\Sigma' \subset \Sigma^+$. This is the correspondence induced by the rational map which sends Σ^+ to $\Sigma^+ \cap V$.

Choose general isotropic flags $E_\bullet$, $F_\bullet$, and $G_\bullet$ in V, so that the corresponding varieties $X_\lambda(E_\bullet)$, $X_\mu(F_\bullet)$, and $X_\nu(G_\bullet)$ meet transversely. We

then extend the three flags in V to a flag of subspaces in V^+ by adjoining, in each case, a general element of H. One then checks that to every point in $X^+_\lambda(E^+_\bullet) \cap X^+_\mu(F^+_\bullet) \cap X^+_\nu(G^+_\bullet)$ (intersection in $LG(n+1, 2n+2)$) there corresponds a point in $X^{(1)}_\lambda(E_\bullet) \cap X^{(1)}_\mu(F_\bullet) \cap X^{(1)}_\nu(G_\bullet)$, and conversely, each point in the latter intersection corresponds to exactly two points in $X^+_\lambda(E^+_\bullet) \cap X^+_\mu(F^+_\bullet) \cap X^+_\nu(G^+_\bullet)$, with the intersection transverse at both of these points. This is enough to prove (11).

We also have the following presentation of the quantum cohomology ring of LG.

Theorem 4 ([10]). *The ring $QH^*(LG)$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\sigma_1, \ldots, \sigma_n, q]$ by the relations*

$$\sigma_r^2 + 2\sum_{i=1}^{n-r}(-1)^i \sigma_{r+i}\sigma_{r-i} = (-1)^{n-r}\sigma_{2r-n-1}\, q$$

for $1 \leqslant r \leqslant n$.

3.2 The orthogonal Grassmannian $OG(n+1, 2n+2)$

We now turn to the even orthogonal Grassmannian $OG = OG(n+1, 2n+2)$. Here $V = \mathbb{C}^{2n+2}$ is equipped with a nondegenerate symmetric form and OG parametrizes one component of the locus of maximal isotropic subspaces in V. The variety OG is projectively equivalent to the odd orthogonal Grassmannian $OG(n, 2n+1)$, and hence the following analysis (for the maximal isotropic case) will include both of the orthogonal Lie types B and D. The dimension of OG equals $n(n+1)/2$, the same as the dimension of LG.

A good part of the classical story for OG is similar to that for $LG(n, 2n)$. The Schubert varieties $X_\lambda(F_\bullet)$ in OG are parametrized by partitions $\lambda \in \mathcal{D}_n$ and are defined by the same equations (8) as before, with respect to an isotropic flag $F_\bullet$ in V. The same is true for the special Schubert varieties $X_p(F_\bullet)$, for $p = 1, \ldots, n$. Let τ_λ be the cohomology class of $X_\lambda(F_\bullet)$; then the τ_λ for $\lambda \in \mathcal{D}_n$ form a $\mathbb{Z}$-basis for $H^*(OG, \mathbb{Z})$. Furthermore, let S (respectively, Q) denote the tautological subbundle (respectively, quotient bundle) over OG.

One important difference between the orthogonal and symplectic Grassmannians is that the natural embedding of $OG(n+1, 2n+2)$ into the type A Grassmannian $G(n+1, 2n+2)$ multiplies all degrees by a factor of 2. This occurs because the subspaces corresponding to points of OG all lie on the quadric of isotropic vectors in V. We therefore have a natural map $\theta : P(S) \to \mathcal{Q}$, where $\mathcal{Q} \hookrightarrow P(V)$ is a $2n$-dimensional quadric hypersurface.

Let us recall the structure of the cohomology ring $H^*(\mathcal{Q}, \mathbb{Z})$. If E and F are maximal isotropic subspaces in V, then $P(E)$ and $P(F)$ are subvarieties of $\mathcal{Q}$ called rulings. The two rulings represent the same class in $H^{2n}(\mathcal{Q}, \mathbb{Z})$ if and only if E and F lie in the same SO_{2n+2}-orbit. There are two families of rulings giving rise to two cohomology classes e and f, and, if $h = c_1(\mathcal{O}_{P(V)}(1)|_{\mathcal{Q}})$

denotes the hyperplane class, we have $h^n = e + f$. A $\mathbb{Z}$-basis for $H^*(\mathcal{Q}, \mathbb{Z})$ is given by $1, h, \ldots, h^n, e, eh = fh, eh^2, \ldots, eh^n$. Finally, one has that $e^2 = f^2 = 0$ and $ef = eh^n$, if n is even, while $e^2 = f^2 = eh^n$ and $ef = 0$, if n is odd.

We deduce that the special Schubert classes τ_p satisfy

$$\tau_p = \pi_* \theta^* [P(F_{n+1-p})] = \frac{1}{2}\, \pi_* \theta^* c_1(\mathcal{O}_{\mathcal{Q}}(1))^{n+p} = \frac{1}{2}\, \pi_* c_1(\mathcal{O}_{P(S)}(1))^{n+p},$$

where $\pi : P(S) \to OG$ is the projection map, and hence that $c_p(Q) = 2\tau_p$, for $p = 1, \ldots, n$. This explains the form of the Pieri rule for OG, as compared to that for $LG(n, 2n)$. We have

$$\tau_\lambda\, \tau_p = \sum_\mu 2^{N'(\lambda,\mu)} \tau_\mu$$

in $H^*(OG, \mathbb{Z})$, where the sum is over all strict partitions μ obtained from λ by adding p boxes, with no two in the same column, and $N'(\lambda, \mu)$ is one less than the number of connected components of μ/λ. The cohomology ring of OG is presented as a quotient of $\mathbb{Z}[\tau_1, \ldots, \tau_n]$ modulo the relations

$$\tau_r^2 + 2 \sum_{i=1}^{r-1} (-1)^i \tau_{r+i} \tau_{r-i} + (-1)^r \tau_{2r} = 0$$

for $1 \leqslant r \leqslant n$, where as usual $\tau_0 = 1$ and $\tau_j = 0$ for $j < 0$ or $j > n$.

The quantum cohomology of OG is isomorphic to $H^*(OG, \mathbb{Z}) \otimes \mathbb{Z}[q]$ as a module over $\mathbb{Z}[q]$, but this time the variable q has degree $2n$.

Theorem 5 ([11]). *For any $\lambda \in \mathcal{D}_n$ and $p \geqslant 1$ we have*

$$\tau_\lambda \cdot \tau_p = \sum_\mu 2^{N'(\lambda,\mu)} \tau_\mu + \sum_\nu 2^{N'(\lambda,\nu)} \tau_{\nu \smallsetminus (n,n)}\, q,$$

where the first sum is over strict μ and the second over partitions $\nu = (n, n, \overline{\nu})$ with $\overline{\nu}$ strict, such that both μ and ν are obtained from λ by adding p boxes, with no two in the same column.

Example 3. For the Grassmannian $OG(5, 10)$, we have

$$\tau_{3,2} \cdot \tau_3 = 2\, \tau_{4,3,1} \quad \text{and} \quad \tau_{4,2} \cdot \tau_3 = \tau_{4,3,2} + 2\, \sigma_1\, q$$

in the quantum cohomology ring $QH^*(OG)$. Compare this with Example 2.

The degree doubling phenomenon discussed previously allows us to conclude that for every degree d map $f : \mathbb{P}^1 \to OG$, the pullback of the quotient bundle Q has degree $2d$. It follows that the relevant parameter space of kernels of the maps counted by a Gromov–Witten invariant is the non-maximal isotropic Grassmannian $OG(n + 1 - 2d, 2n + 2)$. A dimension counting argument now implies that $\langle \tau_\lambda, \tau_\mu, \tau_p \rangle_d = 0$ for $d > 1$, as before.

One observes that the relation $\tau_n^2 = q$ holds in $QH^*(OG)$; this follows from the easy enumerative fact that there is a unique line in OG through a given point and incident to $X_n(E_\bullet)$ and $X_n(F_\bullet)$, for general complete flags $E_\bullet$ and $F_\bullet$ in V. Since the degree of q equals $2n$, this is enough to deduce the presentation which follows.

Theorem 6 ([11]). *The ring $QH^*(OG)$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\tau_1, \ldots, \tau_n, q]$ modulo the relations*

$$\tau_r^2 + 2\sum_{i=1}^{r-1}(-1)^i \tau_{r+i}\tau_{r-i} + (-1)^r \tau_{2r} = 0$$

for all $r < n$, together with the quantum relation

$$\tau_n^2 = q.$$

The remainder of the proof of the quantum Pieri rule for OG differs from those discussed in previous sections. One proves by geometric considerations that the product $\tau_\lambda \cdot \tau_p$ in $QH^*(OG)$ is classical whenever $\lambda_1 < n$. In other words, if the first row of λ is not full, then multiplying τ_λ by a special Schubert class carries no quantum correction. This implies that if λ is such that $\lambda_1 = n$, and $\lambda \smallsetminus n = (\lambda_2, \lambda_3, \ldots)$, then the equation

$$\tau_{\lambda \smallsetminus n} \cdot \tau_n = \tau_\lambda$$

holds in $QH^*(OG)$. On the other hand,

$$\tau_\lambda \cdot \tau_n = \tau_{\lambda \smallsetminus n} \cdot \tau_n^2 = \tau_{\lambda \smallsetminus n}\, q.$$

We thus have established the quantum Pieri rule for multiplication by the special Schubert class τ_n. The general case of the rule follows easily from this and the aforementioned properties.

We refer the reader to the lecture notes [17] for a more detailed discussion of the quantum cohomology of type A and maximal isotropic Grassmannians. This includes some aspects of the story which we have not touched on here, such as classical and quantum Giambelli formulas and Littlewood–Richardson rules.

4 The isotropic Grassmannian $IG(n-k, 2n)$

4.1 The classical theory

The next three sections report on joint work with Anders Buch and Andrew Kresch [5]. We return here to the symplectic vector space $V = \mathbb{C}^{2n}$ as in Section 3.1, and consider the Grassmannian $IG = IG(n-k, 2n)$ which parametrizes $(n-k)$-dimensional isotropic subspaces of V. The variety IG has

dimension $(n-k)(n+3k+1)/2$. The Schubert varieties in IG are parametrized by a certain subset of the hyperoctahedral group, which is described below.

The elements of Weyl group $W_n = S_n \ltimes \mathbb{Z}_2^n$ for the root system C_n are permutations with a sign attached to each entry; we will write these elements as barred permutations. The hyperoctahedral group W_n is an extension of the symmetric group S_n by an element s_0 which acts on the right by

$$(u_1, u_2, \ldots, u_n)s_0 = (\overline{u}_1, u_2, \ldots, u_n),$$

and is generated by the simple reflections $s_0, s_1, \ldots, s_{n-1}$, where each s_i for $i > 0$ is a simple transposition in S_n. If W_k is the parabolic subgroup of W_n generated by $\{s_i \mid i \neq k\}$, then the set $W^{(k)} \subset W_n$ of minimal length coset representatives of W_k parametrizes the Schubert varieties in $IG(n-k, 2n)$. This indexing set $W^{(k)}$ consists of barred permutations of the form

$$w = w_{u,\lambda} = (u_k, \ldots, u_1, \overline{\lambda}_1, \ldots, \overline{\lambda}_\ell, v_{n-k-\ell}, \ldots, v_1)$$

where $\lambda \in \mathcal{D}_n$ with $\ell = \ell(\lambda) \leqslant n-k$, $u_k < \cdots < u_1$, and $v_{n-k-\ell} < \cdots < v_1$.

We can define the Schubert varieties X_w in IG geometrically using a group monomorphism $\phi\colon W_n \hookrightarrow S_{2n}$ with image

$$\phi(W_n) = \{\, \sigma \in S_{2n} \mid \sigma(i) + \sigma(2n+1-i) = 2n+1, \quad \text{for all} \quad i \,\}.$$

The map ϕ is determined by setting, for each $w = (w_1, \ldots, w_n) \in W_n$ and $1 \leqslant i \leqslant n$,

$$\phi(w)(i) = \begin{cases} n + 1 - w_{n+1-i} & \text{if } w_{n+1-i} \text{ is unbarred,} \\ n + \overline{w}_{n+1-i} & \text{otherwise.} \end{cases}$$

Consider a complete isotropic flag

$$F_\bullet \;:\; 0 \subset F_1 \subset F_2 \subset \cdots \subset F_n \subset V$$

and extend it to a complete flag in V by letting $F_{n+p} = F_{n-p}^\perp$ for $1 \leqslant p \leqslant n$. For $w = w_{u,\lambda}$, the Schubert variety $X_w = X_w(F_\bullet) \subset IG$ is defined as the locus of isotropic Σ such that

$$\dim(\Sigma \cap F_i) \geqslant \#\{\, p \leqslant n-k \mid \phi(w)(p) > 2n-i \,\} \quad \text{for} \quad 1 \leqslant i \leqslant 2n.$$

Following Pragacz and Ratajski [13], each Weyl group element $w_{u,\lambda}$ corresponds to a pair of partitions $\Lambda = (\alpha \,|\, \lambda)$, where the 'top' partition $\alpha = \alpha(u, \lambda)$ is defined by

$$\alpha_i = u_i + i - k - 1 + \#\{j \mid \lambda_j > u_i\}$$

for $1 \leqslant i \leqslant k$; the 'bottom' partition is λ. The Schubert varieties X_Λ in IG are thus parametrized by the set $\mathcal{P}(k, n)$ of pairs $\Lambda = (\alpha \,|\, \lambda)$ with $\alpha \in \mathcal{R}(k, n-k)$, $\lambda \in \mathcal{D}_n$, and such that $\alpha_k \geqslant \ell(\lambda)$. Figure 2 illustrates a partition pair indexing a Schubert variety in $IG(4, 14)$, with $\alpha = (4, 3, 3)$ and $\lambda = (5, 4, 1)$.

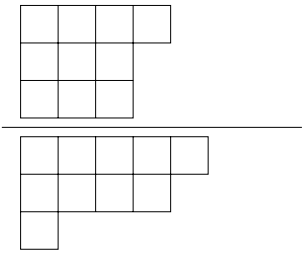


Fig. 2. The partition pair $(4,3,3\,|\,5,4,1)$

The codimension of the variety $X_\Lambda(F_\bullet)$ in IG is given by the weight $|\Lambda| = |\alpha|+|\lambda|$; we let σ_Λ denote the corresponding cohomology class in $H^{2|\Lambda|}(IG,\mathbb{Z})$.

The *special* Schubert varieties used in our analysis of IG are the varieties $X_p(F_\bullet)$ for $1 \leqslant p \leqslant n+k$ defined by a single Schubert condition, as follows:

$$X_p(F_\bullet) = \{\, \Sigma \in IG \mid \Sigma \cap F_{n+k+1-p} \neq 0 \,\}.$$

Our main reason for this choice of special classes is the application to the classical and quantum Pieri rules which follow. We note that $X_p = X_{(1^{\min(p,k)}|\max(p-k,0))}$. If σ_p denotes the cohomology class of X_p in $H^{2p}(IG,\mathbb{Z})$, and

$$0 \to S \to V \to Q \to 0$$

is the tautological exact sequence of vector bundles over IG, then σ_p is equal to the Chern class $c_p(Q)$.

We proceed to describe the classical Pieri rule for multiplying a general Schubert class σ_Λ with a special class. In contrast, Pragacz and Ratajski [13] obtain a Pieri rule for multiplication with the Chern classes of S^*.

Recall that the number of components of a skew diagram is by definition the number of connected components of its vertical projection. We adopt the following shifting conventions for the two diagrams α, λ in a pair $\Lambda = (\alpha\,|\,\lambda)$. For each i, the ith part of α is shifted to the right $k-i+1$ units, and the ith part of λ is shifted to the right $i-1$ units. Upon performing this operation, we obtain the *shifted diagram* $\mathcal{S}(\Lambda)$ of Λ. We say that a box of $(\alpha\,|\,\lambda)$ lies *shift-under* some given reference box(es) if, in $\mathcal{S}(\Lambda)$, the box lies under (at least one of) the reference box(es). In Figure 3, the boxes in $(4,3,3\,|\,5,4,1)$ which lie shift-under the box marked with an 'x' are marked with an 'o'.

Theorem 7. *For any $\Lambda = (\alpha\,|\,\lambda) \in \mathcal{P}(k,n)$ and $p \geqslant 1$ we have*

$$\sigma_\Lambda \sigma_p = \sum_M 2^{N(\Lambda,M)} \sigma_M,$$

where the sum is over all $M = (\beta\,|\,\mu)$ with $|M| = |\Lambda| + p$ such that

(i) $\alpha_{i+1} \leqslant \beta_i \leqslant \alpha_i + 1$ *for each i,*

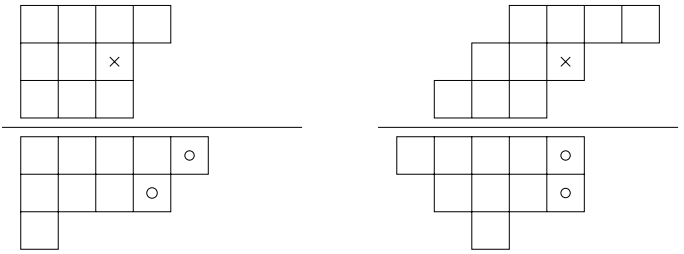


Fig. 3. $(4,3,3\,|\,5,4,1)$ and $\mathcal{S}(4,3,3\,|\,5,4,1)$

(ii) $\mu \supset \lambda$, *with* μ/λ *a horizontal strip,*

(iii) for each i *with* $\beta_i = \alpha_i$, *there is at most one box of* μ/λ *shift-under the rightmost box of* β_i,

(iv) for each i *with* $\beta_i < \alpha_i$, *there are exactly* $\alpha_i - \beta_i + 1$ *boxes of* μ/λ *shift-under the boxes between the rightmost box of* β_i *and the rightmost box of* α_i, *inclusive; furthermore, these* $\alpha_i - \beta_i + 1$ *boxes of the bottom part are contained in a single row.*

Let R_i *denote the collection of box(es) of* μ/λ *indicated in (iii) and (iv) for given* i, *with* $R_i = \emptyset$ *when* $\beta_i > \alpha_i$, *and set* $R = \bigcup_{i=1}^{k} R_i$. *Then the exponent* $N(\Lambda, M)$ *equals the number of components of* $(\mu/\lambda) \smallsetminus R$ *not meeting the first column.*

If the partition pair M with $|M| = |\Lambda| + p$ satisfies conditions (i)-(iv) of Theorem 7, then we will write $\Lambda \xrightarrow{p} M$. The argument used to prove Theorem 7 in [5] is geometric, along the lines of Hodge and Pedoe's proof of the classical Pieri rule for type A Grassmannians, via triple intersections [9, §XIV.4].

Theorem 8. *The cohomology ring* $H^*(IG, \mathbb{Z})$ *is presented as a quotient of the polynomial ring* $\mathbb{Z}[\sigma_1, \ldots, \sigma_{n+k}]$ *modulo the relations*

$$\det(\sigma_{1+j-i})_{1 \leqslant i,j \leqslant r} = 0, \qquad n-k+1 \leqslant r \leqslant n+k \tag{12}$$

and

$$\sigma_r^2 + 2 \sum_{i=1}^{n+k-r} (-1)^i \sigma_{r+i}\sigma_{r-i} = 0, \qquad k+1 \leqslant r \leqslant n. \tag{13}$$

The first set of relations (12) in Theorem 8 follow from the Whitney sum formula $c(S)c(Q) = 1$, as in Section 2. To see the relations (13), note that the symplectic form on V gives rise to a pairing $S \otimes Q \to \mathcal{O}$, and hence an injection $S \to Q^*$. The Chern classes $c_j(Q^*/S)$ vanish for $j > 2k$; multiplying with the previous relation, we deduce that the cohomology class $c(Q)c(Q^*)$ vanishes in degrees larger than $2k$. We have thus shown that the relations (12) and (13) hold in $H^*(IG, \mathbb{Z})$; more work is required to deduce that these two sets of relations suffice to obtain the presentation in the theorem.

4.2 The quantum theory

The quantum cohomology ring $QH^*(IG)$ is a $\mathbb{Z}[q]$-algebra as before, where the degree of the formal variable q is given by $\deg(q) = n + k + 1$. The ring structure on $QH^*(IG)$ is determined by a relation analogous to (2)

$$\sigma_\Lambda \cdot \sigma_M = \sum \langle \sigma_\Lambda, \sigma_M, \sigma_{\widehat{N}} \rangle_d \, \sigma_N \, q^d,$$

the sum over $d \geqslant 0$ and $N \in \mathcal{P}(k,n)$ with $|N| = |\Lambda| + |M| - (n+k+1)d$.

For each partition pair $M = (\beta \,|\, \mu)$ with $\mu_1 > 0$, define a new pair $M^* = (\beta^* \,|\, \mu^*)$ by setting $\beta_i^* = \beta_i - 1$ and $\mu_i^* = \mu_{i+1}$ for each i. In other words, M^* is obtained from M by removing one box from each row of β and the entire first row of μ. Let $\mathcal{Q}(k,n)$ denote the set of $M = (\beta \,|\, \mu) \in \mathcal{P}(k,n)$ such that $\mu_1 = n$.

Theorem 9. *For any $\Lambda \in \mathcal{P}(k,n)$ and p with $1 \leqslant p \leqslant n+k$, we have*

$$\sigma_\Lambda \cdot \sigma_p = \sum_{M \in \mathcal{P}(k,n)} 2^{N(\Lambda,M)} \, \sigma_M + \sum_{M \in \mathcal{Q}(k,n+1)} 2^{N(\Lambda,M)-1} \, \sigma_{M^*} \, q$$

in the quantum cohomology ring of $IG(n-k,2n)$, where both sums involve partition pairs M such that $\Lambda \xrightarrow{p} M$.

Example 4. In the quantum cohomology ring of $IG(4,12)$, we have

$$\sigma_5 \cdot \sigma_{(4,3|3,1)} = 2\,\sigma_{(4,4|5,3)} + 4\,\sigma_{(4,4|6,2)} + 2\,\sigma_{(4,4|5,2,1)} + \sigma_{(4,4|4,3,1)} + 2\,\sigma_{(4,3|6,2,1)}$$
$$+\sigma_{(4,2|1)}\,q + 2\,\sigma_{(3,3|1)}\,q + 4\,\sigma_{(3,2|2)}\,q + \sigma_{(2,2|2,1)}\,q.$$

The analysis used in the proof of Theorem 9 is more involved than, but analogous to the corresponding one in Section 3.1. To study the Gromov–Witten invariants $\langle \sigma_\Lambda, \sigma_M, \sigma_N \rangle_d$, we use the auxiliary variety Z_d parametrizing pairs (A,B) of subspaces of V with $A \in IG(n-k-d,2n)$, $\dim B = n-k+d$, and $A \subset B \subset A^\perp$, for each $d \leqslant n-k$. For every $\Lambda \in \mathcal{P}(k,n)$, define a subvariety $Y_\Lambda \subset Z_d$ by

$$Y_\Lambda(F_\bullet) = \{\, (A,B) \in Z_d \,|\, A \subset \Sigma \subset B \text{ for some } \Sigma \text{ with } \Sigma \in X_\Lambda(F_\bullet) \,\}.$$

One then shows that if $\Lambda, M, N \in \mathcal{P}(k,n)$ are such that $|\Lambda| + |M| + |N| = \dim(IG) + n + k + 1$, then

$$\langle \sigma_\Lambda, \sigma_M, \sigma_N \rangle_1 = \int_{Z_1} [Y_\Lambda] \cdot [Y_M] \cdot [Y_N].$$

Moreover, if $N = (1^{\min(p,k)} \mid \max(p-k,0))$ indexes a special Schubert class, then the Gromov–Witten invariant $\langle \sigma_\Lambda, \sigma_M, \sigma_p \rangle_d$ vanishes whenever $d \geqslant 2$. The key to proving these results is to associate to any rational map $f\colon \mathbb{P}^1 \to$

IG of degree d counted by $\langle \sigma_\Lambda, \sigma_M, \sigma_N \rangle_d$ the pair consisting of the kernel and the span of f.

Finally, one uses a correspondence between lines on $IG = IG(n-k, 2n)$ and points on $IG(n+1-k, 2n+2)$ to obtain an analogue of (11) for IG. For Λ, M, $N \in \mathcal{P}(k,n)$ such that the first parts of the top partitions of Λ, M, N sum to at most $2(n-k)+1$, we obtain that

$$\int_{Z_1} [Y_\Lambda] \cdot [Y_M] \cdot [Y_N] = \frac{1}{2} \int_{IG(n+1-k,2n+2)} [X^+_\Lambda] \cdot [X^+_M] \cdot [X^+_N],$$

where X^+_Λ, X^+_M, X^+_N denote Schubert varieties in $IG(n+1-k, 2n+2)$. It follows that when $\sigma_N = \sigma_p$ is a special Schubert class, then

$$\langle \sigma_\Lambda, \sigma_M, \sigma_p \rangle_1 = \frac{1}{2} \int_{IG(n+1-k,2n+2)} \sigma^+_\Lambda \cdot \sigma^+_M \cdot \sigma^+_p.$$

We deduce Theorem 9 from this, Theorem 7, and an analysis of the Poincaré duality involution on $\mathcal{P}(k,n)$.

Theorem 10. *The quantum cohomology ring $QH^*(IG, \mathbb{Z})$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\sigma_1, \ldots, \sigma_{n+k}, q]$ modulo the relations*

$$\det(\sigma_{1+j-i})_{1 \leqslant i,j \leqslant r} = 0, \qquad n-k+1 \leqslant r \leqslant n+k$$

and

$$\sigma_r^2 + 2 \sum_{i=1}^{n+k-r} (-1)^i \sigma_{r+i} \sigma_{r-i} = (-1)^{n+k-r} \sigma_{2r-n-k-1}\, q, \qquad k+1 \leqslant r \leqslant n.$$

5 The odd orthogonal Grassmannian $OG(n-k, 2n+1)$

5.1 The classical theory

In this section, we consider a vector space $V = \mathbb{C}^{2n+1}$ equipped with a nondegenerate symmetric bilinear form. The odd orthogonal Grassmannian $OG = OG(n-k, 2n+1)$ parametrizes the $(n-k)$-dimensional isotropic subspaces in V; it has the same dimension as the isotropic Grassmannian $IG(n-k, 2n)$. The Weyl group for the root system B_n is the same as that for C_n, hence most of the analysis in Section 4.1 applies here as well. To define the Schubert variety in OG indexed by $w \in W_n$, we use a monomorphism $\psi\colon W_n \hookrightarrow S_{2n+1}$ with image

$$\psi(W_n) = \{\, \sigma \in S_{2n+1} \mid \sigma(i) + \sigma(2n+2-i) = 2n+2, \ \text{ for all } \ i \,\},$$

determined by the equalities

$$\psi(w)(i) = \begin{cases} n+1-w_{n+1-i} & \text{if } w_{n+1-i} \text{ is unbarred,} \\ n+1+\overline{w}_{n+1-i} & \text{otherwise,} \end{cases}$$

for each $w = (w_1, \ldots, w_n) \in W_n$ and $1 \leqslant i \leqslant n$. Given an isotropic flag $F_\bullet$ in V, we extend it to a complete flag by setting $F_{n+p} = F_{n+1-p}^{\perp}$ for $1 \leqslant p \leqslant n+1$. For $w = w_{u,\lambda}$, the Schubert variety $X_w(F_\bullet) \subset OG$ is defined as the locus of isotropic subspaces Σ in V such that

$$\dim(\Sigma \cap F_i) \geqslant \#\{\, p \leqslant n-k \mid \psi(w)(p) > 2n+1-i \,\} \quad \text{for} \quad 1 \leqslant i \leqslant 2n.$$

To each Weyl group element $w_{u,\lambda}$ we associate a pair of partitions $\Lambda = (\alpha \,|\, \lambda)$ as in Section 4.1. The Schubert varieties in OG are indexed by the set $\mathcal{P}(k,n)$ of partition pairs; we let $\tau_\Lambda \in H^{2|\Lambda|}(OG, \mathbb{Z})$ be the cohomology class determined by the Schubert variety $X_\Lambda(F_\bullet)$.

The *special* Schubert varieties for OG are the varieties

$$X_p = X_{(1^{\min(p,k)}|\max(p-k,0))}, \qquad 1 \leqslant p \leqslant n+k.$$

These are defined by a single Schubert condition as follows: let

$$\varepsilon(p) = n+k-p+\begin{cases} 1 & \text{if } p > k, \\ 2 & \text{if } p \leqslant k. \end{cases} \tag{14}$$

Then

$$X_p(F_\bullet) = \{\, \Sigma \in OG \,|\, \Sigma \cap F_{\varepsilon(p)} \neq 0 \,\}.$$

We let τ_p denote the cohomology class of X_p, and S (respectively Q) be the tautological subbundle (respectively, quotient bundle) over $OG(n-k, 2n+1)$. As in Section 3.2, by considering the natural map $\theta : P(S) \to \mathcal{Q} \subset P(V)$, where $\mathcal{Q}$ is the $(2n-1)$-dimensional quadric of isotropic vectors, one shows that

$$c_p(Q) = \begin{cases} \tau_p & \text{if } p \leqslant k, \\ 2\tau_p & \text{if } p > k. \end{cases}$$

The classical Pieri rule for OG involves the same conditions (i)-(iv) and set $R \subset \mu/\lambda$ that appeared in Theorem 7.

Theorem 11. *For any $\Lambda = (\alpha \,|\, \lambda) \in \mathcal{P}(k,n)$ and $p \geqslant 1$ we have*

$$\tau_\Lambda \, \tau_p = \sum_M 2^{N'(\Lambda,M)} \tau_M, \tag{15}$$

where the sum is over all $M = (\beta \,|\, \mu)$ with $\Lambda \xrightarrow{p} M$. Moreover, $N'(\Lambda, M)$ equals the number (respectively, one less than the number) of components of $(\mu/\lambda) \smallsetminus R$, if $p \leqslant k$ (respectively, if $p > k$).

Let $c_p = c_p(Q)$. Using *rational* coefficients, one has a presentation of the cohomology ring of OG in terms of the c variables directly analogous to that in Theorem 8. The ring $H^*(OG, \mathbb{Q})$ is presented as a quotient of the polynomial ring $\mathbb{Q}[c_1, \ldots, c_{n+k}]$ modulo the relations

$$\det(c_{1+j-i})_{1 \leqslant i,j \leqslant r} = 0, \qquad n-k+1 \leqslant r \leqslant n+k$$

and

$$c_r^2 + 2\sum_{i=1}^{r} (-1)^i c_{r+i} c_{r-i} = 0, \qquad k+1 \leqslant r \leqslant n.$$

However, a presentation with integer coefficients using the special Schubert classes τ_p is more difficult to obtain. Let $\delta_p = 1$, if $p \leqslant k$, and $\delta_p = 2$, otherwise.

Theorem 12. *The cohomology ring $H^*(OG(n-k, 2n+1), \mathbb{Z})$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\tau_1, \ldots, \tau_{n+k}]$ modulo the relations*

$$\det(\delta_{1+j-i}\tau_{1+j-i})_{1 \leqslant i,j \leqslant r} = 0, \qquad n-k+1 \leqslant r \leqslant n,$$

$$\sum_{p=k+1}^{r} (-1)^p \tau_p \det(\delta_{1+j-i}\tau_{1+j-i})_{1 \leqslant i,j \leqslant r-p} = 0, \qquad n+1 \leqslant r \leqslant n+k,$$

and

$$\tau_r^2 + \sum_{i=1}^{r} (-1)^i \delta_{r-i} \tau_{r+i} \tau_{r-i} = 0, \qquad k+1 \leqslant r \leqslant n.$$

5.2 The quantum theory

Given a degree $d \geqslant 0$ and partition pairs $\Lambda, M, N \in \mathcal{P}(k, n)$ such that

$$|\Lambda| + |M| + |N| = \dim(OG) + d(n+k),$$

we define the Gromov–Witten invariant $\langle \tau_\Lambda, \tau_M, \tau_N \rangle_d$ as in Section 3.1. The degree of q in the quantum cohomology ring $QH^*(OG)$ is equal to $n+k$, and the quantum product is defined as usual. A notable feature of the quantum Pieri rule for OG is that it involves q^2 terms as well as q terms. To formulate it, we require some additional notation.

Let $\mathcal{P}'(k, n+1)$ be the set of $(\beta \,|\, \mu) \in \mathcal{P}(k, n+1)$ such that $\beta_1 = n+1-k$ and $\max(\beta_2 + k - 1, 1) \leqslant \mu_1 \leqslant n$. For any $M = (\beta \,|\, \mu) \in \mathcal{P}'(k, n+1)$, define a partition pair $\widetilde{M} \in \mathcal{P}(k, n)$ by

$$\widetilde{(\beta \,|\, \mu)} = (\mu_1 - k, \beta_2 - 1, \ldots, \beta_k - 1 \,|\, \mu_2, \mu_3, \ldots).$$

In other words, $\widetilde{M}$ is obtained from $(\beta \,|\, \mu)$ by removing a hook of length n from β and the entire first row of μ, then adding the part $\mu_1 - k$ to what remains in the top partition. Recall also the operation $M \mapsto M^*$ from Section 4.2.

Theorem 13. *For any $\Lambda = (\alpha \,|\, \lambda) \in \mathcal{P}(k,n)$ and p with $1 \leqslant p \leqslant n+k$, the quantum product $\tau_\Lambda \cdot \tau_p \in QH^*(OG(n-k,2n+1))$ is equal to*

$$\sum_{M \in \mathcal{P}(k,n)} 2^{N'(\Lambda,M)}\, \tau_M + \sum_{M \in \mathcal{P}'(k,n+1)} 2^{N'(\Lambda,M)}\, \tau_{\widetilde{M}}\, q + \sum_{M \in \mathcal{Q}(k,n)} 2^{N'(\Lambda^*,M)}\, \tau_{M^*}\, q^2,$$

where (i) the first sum is classical, as in (15), (ii) the second sum is over $M \in \mathcal{P}'(k,n+1)$ with $\Lambda \xrightarrow{p} M$, and (iii) the third sum is empty unless $\lambda_1 = n$, and over $M \in \mathcal{Q}(k,n)$ such that $\Lambda^* \xrightarrow{p} M$.

The q terms in Theorem 13 are explained by a phenomenon similar to what happens in type A (Section 2), while the q^2 terms are in analogy with the maximal orthogonal Grassmannian (Section 3.2); a dimension count shows that there are no higher degree contributions. For the line numbers, we observe that the parameter space of lines on OG is the orthogonal two-step flag variety $Y_1 = OF(n-k-1, n-k+1; 2n+1)$. It follows that

$$\langle \tau_\Lambda, \tau_M, \tau_p \rangle_1 = \int_{Y_1} \tau_\Lambda^{(1)} \cdot \tau_M^{(1)} \cdot \tau_p^{(1)}$$

where $\tau_\Lambda^{(1)}$, $\tau_M^{(1)}$, and $\tau_p^{(1)}$ are the associated Schubert classes in $H^*(Y_1, \mathbb{Z})$. The relevant diagram of smooth projections here is

$$\begin{array}{ccc} OF(n-k, n-k+1; 2n+1) & \xrightarrow{\varphi_1} & OG(n-k, 2n+1) \\ \Big\downarrow \pi & & \\ OG(n-k+1, 2n+1). & & \end{array}$$

We associate to any partition pair $\Lambda = (\alpha \,|\, \lambda)$ in $\mathcal{P}(k,n)$ the pair $\overline{\Lambda} \in \mathcal{P}(k-1,n)$ obtained by removing the first part α_1 of the top partition:

$$\overline{(\alpha \,|\, \lambda)} = (\alpha_2, \ldots, \alpha_k \mid \lambda).$$

One checks that $\pi(\varphi_1^{-1}(X_\Lambda(F_\bullet))) = X_{\overline{\Lambda}}(F_\bullet)$, for any $\Lambda \in \mathcal{P}(k,n)$. By arguing as in Section 2, we deduce that for $\Lambda, M \in \mathcal{P}(k,n)$ and $1 \leqslant p \leqslant n+k$, if $|\Lambda| + |M| + p = \dim OG + n + k$, then

$$\langle \tau_\Lambda, \tau_M, \tau_p \rangle_1 = \int_{OG(n-k+1,2n+1)} \tau_{\overline{\Lambda}} \cdot \tau_{\overline{M}} \cdot \tau_{p-1}. \tag{16}$$

This result characterizes the degree 1 quantum correction terms in Theorem 13.

To understand the q^2 terms appearing in the quantum Pieri rule, one uses the relation $\tau_{n+k}^2 = q^2$ and follows the model of $OG(n, 2n+1)$, which is isomorphic to $OG(n+1, 2n+2)$. Note that the correct analogy is that the

variable $q \in QH^*(OG(n,2n+1))$ corresponds to $q^2 \in QH^*(OG(n-k,2n+1))$, when $k \geqslant 1$. This is explained by degree considerations: on non-maximal orthogonal Grassmannians OG, we have $\deg(q) = n+k$, which specializes under $k=0$ to n, or half of the degree of q on $OG(n,2n+1)$. It is also related to the degree doubling phenomenon, whereby lines on the maximal orthogonal Grassmannian are mapped to conics in projective space under the Plücker embedding.

Example 5. In the ring $QH^*(OG(2,9))$, we have

$$\tau_1 \cdot \tau_{(2,2|4,3)} = \tau_{(2,1|3)}\, q + 2\, q^2.$$

The term $2\,q^2$ is explained by the enumerative fact that there is a unique conic on $OG(2,9)$ passing through two general points. This conic meets a general hyperplane section (encoded by the class τ_1) in two points, and each pair, consisting of the conic together with a point of intersection with the given hyperplane section, contributes to the coefficient of q^2.

Theorem 14. *The quantum cohomology ring $QH^*(OG,\mathbb{Z})$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\tau_1,\dots,\tau_{n+k},q]$ modulo the relations*

$$\det(\delta_{1+j-i}\tau_{1+j-i})_{1\leqslant i,j\leqslant r} = 0, \qquad n-k+1 \leqslant r \leqslant n,$$

$$\sum_{p=k+1}^{r} (-1)^p \tau_p \det(\delta_{1+j-i}\tau_{1+j-i})_{1\leqslant i,j\leqslant r-p} = 0, \qquad n+1 \leqslant r < n+k,$$

$$\sum_{p=k+1}^{n+k} (-1)^p \tau_p \det(\delta_{1+j-i}\tau_{1+j-i})_{1\leqslant i,j\leqslant n+k-p} = q$$

and

$$\tau_r^2 + \sum_{i=1}^{r} (-1)^i \delta_{r-i}\tau_{r+i}\tau_{r-i} = 0, \qquad k+1 \leqslant r \leqslant n.$$

6 The even orthogonal Grassmannian $OG(n+1-k, 2n+2)$

6.1 The classical theory

Here we take the vector space $V = \mathbb{C}^{2n+2}$ equipped with a nondegenerate symmetric form, and consider the Grassmannian $OG' = OG(n+1-k, 2n+2)$ which parametrizes the $(n+1-k)$-dimensional isotropic subspaces in V. The dimension of OG' is equal to $(n+1-k)(n+3k)/2$.

To define the Schubert varieties in OG', we require the Weyl group $\widetilde{W}_{n+1}$ of type D_{n+1}. This group is an extension of S_{n+1} by an element $s_\square$ which acts by

$$(u_1, u_2, \ldots, u_{n+1})s_\square = (\overline{u}_2, \overline{u}_1, u_3, \ldots, u_{n+1}).$$

If $\widetilde{W}_k$ is the subgroup of $\widetilde{W}_{n+1}$ generated by $\{s_i \mid i \neq k\}$, then the set $\widetilde{W}^{(k)} \subset \widetilde{W}_{n+1}$ of minimal length coset representatives of $\widetilde{W}_k$ parametrizes the Schubert varieties in $OG(n+1-k, 2n+2)$. The set $\widetilde{W}^{(k)}$ consists of barred permutations of the form

$$w = w_{u,\lambda} = (\widehat{u}_k, \ldots, u_1, \overline{z}_1, \ldots, \overline{z}_r, v_{n+1-k-r}, \ldots, v_1) \tag{17}$$

where $z_1 > \cdots > z_r$ with $r \leqslant n-k$, $u_k < \cdots < u_1$, and $v_{n+1-k-r} < \cdots < v_1$. The convention here is that $\widehat{u}_k$ is equal to u_k or $\overline{u}_k$, according to the parity of r.

The elements in $\widetilde{W}^{(k)}$ correspond to a set $\widetilde{\mathcal{P}}(k,n)$ of partition pairs, which involve a partition $\alpha \in \mathcal{R}(k, n+1-k)$ and a $\lambda \in \mathcal{D}_n$ such that $\alpha_k \geqslant \ell(\lambda)$. More precisely, $\widetilde{\mathcal{P}}(k,n)$ consists of (i) pairs $(\alpha \,|\, \lambda)$ such that $\alpha_k = \ell(\lambda)$; these correspond to elements

$$w_{(\alpha|\lambda)} = (\widehat{u}_k, \ldots, u_1, \overline{\lambda_1 + 1}, \ldots, \overline{\lambda_\ell + 1}, v_{n+1-k-\ell}, \ldots, v_1)$$

in $\widetilde{W}_{n+1}$; (ii) two types of pairs $(\alpha \,|\, \lambda)$ and $(\alpha \,|\, \lambda]$ such that $\alpha_k > \ell(\lambda)$; these correspond to the Weyl group elements

$$w_{(\alpha|\lambda)} = (\widehat{u}_k, \ldots, u_1, \overline{\lambda_1 + 1}, \ldots, \overline{\lambda_\ell + 1}, v_{n+1-k-\ell}, \ldots, v_1)$$

and

$$w_{(\alpha|\lambda]} = (\widehat{u}_k, \ldots, u_1, \overline{\lambda_1 + 1}, \ldots, \overline{\lambda_\ell + 1}, \overline{1}, v_{n-k-\ell}, \ldots, v_1),$$

respectively. We refer to the objects in this case as partition pairs of *type 1* and of *type 2*, respectively, and let $\mathrm{type}(\Lambda) \in \{1, 2\}$ denote the type of a partition pair Λ. If Λ falls under case (i), then we set $\mathrm{type}(\Lambda) = 0$. In both cases (i) and (ii), the equalities

$$\alpha_i = u_i + i - k - 1 + \#\{j \mid \lambda_j + 1 > u_i\}$$

for $1 \leqslant i \leqslant k$ determine the numbers u_i, while $v_{n+1-k-\ell} < \cdots < v_1$. The weight of a partition pair $(\alpha \,|\, \lambda)$ or $(\alpha \,|\, \lambda]$ equals $|\alpha| + |\lambda|$, as before.

The Schubert varieties in OG' are defined using a monomorphism $\varphi : \widetilde{W}_{n+1} \hookrightarrow S_{2n+2}$ whose image consists of those permutations $\sigma \in S_{2n+2}$ such that $\sigma(i) + \sigma(2n+3-i) = 2n+3$ for all i and the number of $i \leqslant n+1$ such that $\sigma(i) > n+1$ is even. The map φ is defined by the equation

$$\varphi(w)(i) = \begin{cases} n+2-w_{n+2-i} & \text{if } w_{n+2-i} \text{ is unbarred,} \\ n+1+\overline{w}_{n+2-i} & \text{otherwise,} \end{cases}$$

for $w = (w_1, \ldots, w_{n+1}) \in \widetilde{W}_{n+1}$ and $1 \leqslant i \leqslant n+1$. Choose also an isotropic flag $F_\bullet$ in V, and extend it to a complete flag of V by setting $F_{n+1+p} = F_{n+1-p}^{\perp}$ for $1 \leqslant p \leqslant n+1$.

Set $\iota(w) = 0$ if $\overline{1}$ is a part of w, and $\iota(w) = 1$ otherwise, and define $\widetilde{\varphi} : \widetilde{W}_{n+1} \hookrightarrow S_{2n+2}$ by

$$\widetilde{\varphi}(w) = s_{n+1}^{\iota(w)} \varphi(w).$$

The map $\widetilde{\varphi}$ is a modification of φ so that in the sequence of values of $\widetilde{\varphi}(w)$, $n+2$ always comes before $n+1$. We need also the alternate complete flag $\widetilde{F}_\bullet$, with $\widetilde{F}_i = F_i$ for $i \leqslant n$ but completed with a maximal isotropic subspace $\widetilde{F}_{n+1}$ in the opposite family from F_{n+1} (thus we have $\widetilde{F}_{n+1} \cap F_{n+1} = F_n$). Define

$$F_\bullet^\iota = \begin{cases} F_\bullet & \text{if } n \neq \iota \pmod 2, \\ \widetilde{F}_\bullet & \text{if } n = \iota \pmod 2. \end{cases}$$

For $w \in \widetilde{W}^{(k)}$, the Schubert variety $X_w(F_\bullet)$ is defined as the locus of isotropic Σ in OG' such that

$$\dim(\Sigma \cap F_i^{\iota(w)}) \geqslant \#\{\, p \leqslant n+1-k \mid \widetilde{\varphi}(w)(p) > 2n+2-i \,\}$$

for $1 \leqslant i \leqslant 2n+2$. We let $\tau_\Lambda \in H^{2|\Lambda|}(OG', \mathbb{Z})$ denote the cohomology class determined by the Schubert variety indexed by w_Λ for an element $\Lambda \in \widetilde{\mathcal{P}}(k,n)$.

The *special* Schubert varieties for $OG(n+1-k, 2n+2)$ are the varieties

$$X_p = X_{(1^{\min(p,k)}|\max(p-k,0))}, \quad 1 \leqslant p \leqslant n+k, \quad \text{and} \quad X'_k = X_{(1^k|0]}.$$

These are defined by a single Schubert condition as follows. For $p \neq k$, we have

$$X_p(F_\bullet) = \{\, \Sigma \in OG' \mid \Sigma \cap F_{\varepsilon(p)} \neq 0 \,\}$$

where $\varepsilon(p)$ is given by (14). If n is even, then

$$X_k(F_\bullet) = \{\, \Sigma \in OG' \mid \Sigma \cap F_{n+1} \neq 0 \,\}$$

and

$$X'_k(F_\bullet) = \{\, \Sigma \in OG' \mid \Sigma \cap \widetilde{F}_{n+1} \neq 0 \,\},$$

while the roles of F_{n+1} and $\widetilde{F}_{n+1}$ are switched if n is odd. We let τ_p denote the cohomology class of $X_p(F_\bullet)$ and τ'_k denote the cohomology class of $X'_k(F_\bullet)$.

The Pieri rule for OG' requires a slightly different shifting convention than that used for IG and OG. The shifted diagram of a partition pair $(\alpha \,|\, \lambda)$ or $(\alpha \,|\, \lambda]$ is obtained by shifting the ith part of α to the right by $k-i+1$ units, and the ith part of λ to the right by i units. Using this convention, the notation $\Lambda \xrightarrow{p} M$ and the multiplicity $N'(\Lambda, M)$ are defined in exactly the same way as in Theorem 11. Given two partition pairs Λ and M which both fall in case (ii), above, we set

$$\epsilon_{\Lambda M} = \begin{cases} 1 & \text{if type}(\Lambda) = \text{type}(M) \text{ and } \ell(\lambda) = \ell(\mu), \\ 1 & \text{if type}(\Lambda) \neq \text{type}(M) \text{ and } \ell(\lambda) \neq \ell(\mu), \\ 0 & \text{otherwise.} \end{cases}$$

In addition, let

$$h(\Lambda, M) = \#\{i \mid \beta_i \leqslant \alpha_i\} + \ell(\lambda) + \begin{cases} 1 & \text{if type}(\Lambda) = 2 \text{ or type}(M) = 2, \\ 0 & \text{otherwise.} \end{cases}$$

If $p \neq k$, then set $\delta_{\Lambda M} = 1$. If $p = k$ and $N'(\Lambda, M) > 0$, then set

$$\delta_{\Lambda M} = \delta'_{\Lambda M} = 1/2,$$

while if $N'(\Lambda, M) = 0$, define

$$\delta_{\Lambda M} = \begin{cases} 1 & \text{if } h(\Lambda, M) \text{ is even,} \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \delta'_{\Lambda M} = \begin{cases} 1 & \text{if } h(\Lambda, M) \text{ is odd,} \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 15. *Let $\Lambda = (\alpha \,|\, \lambda)$ or $\Lambda = (\alpha \,|\, \lambda]$ be an element of $\widetilde{\mathcal{P}}(k, n)$ and $p \geqslant 1$. If $\ell(\lambda) = \alpha_k$, then*

$$\tau_\Lambda \, \tau_p = \sum_M 2^{N'(\Lambda, M)} \delta_{\Lambda M} \tau_M,$$

while if $\ell(\lambda) < \alpha_k$, then

$$\tau_\Lambda \, \tau_p = \sum_{\ell(\mu) = \beta_k} 2^{N'(\Lambda, M)} \delta_{\Lambda M} \tau_M + \sum_{\ell(\mu) < \beta_k} 2^{N'(\Lambda, M)} \delta_{\Lambda M} \epsilon_{\Lambda M} \tau_M,$$

where the sums are over all $M = (\beta \,|\, \mu)$ and $M = (\beta \,|\, \mu]$ in $\widetilde{\mathcal{P}}(k, n)$ with $\Lambda \xrightarrow{p} M$ and satisfying the indicated conditions. Furthermore, the product $\tau_\Lambda \, \tau'_k$ is obtained by replacing $\delta_{\Lambda M}$ with $\delta'_{\Lambda M}$ throughout.

Let

$$0 \to S \to V_X \to Q \to 0$$

denote the tautological exact sequence of vector bundles over OG'. One checks as in Section 3.2 that

$$c_p(Q) = \begin{cases} \tau_p & \text{if } p < k, \\ \tau_k + \tau'_k & \text{if } p = k, \\ 2\tau_p & \text{if } p > k. \end{cases} \tag{18}$$

For each $r > 0$, let Δ_r denote the $r \times r$ Schur determinant

$$\Delta_r = \det(c_{1+j-i})_{1 \leqslant i, j \leqslant r},$$

where each variable c_p represents the Chern class $c_p(Q)$. Using rational coefficients, the ring $H^*(OG', \mathbb{Q})$ is presented as a quotient of the polynomial ring $\mathbb{Q}[c_1, \ldots, c_{n+k}, \xi]$ modulo the relations

$$\Delta_r = 0, \qquad n-k+1 < r \leqslant n+k, \tag{19}$$

$$\xi\, \Delta_{n+1-k} = 0, \tag{20}$$

and

$$c_k^2 + 2\sum_{i=1}^{k} (-1)^i c_{k+i} c_{k-i} = \xi^2, \tag{21}$$

$$c_r^2 + 2\sum_{i=1}^{r} (-1)^i c_{r+i} c_{r-i} = 0, \qquad k+1 \leqslant r \leqslant n. \tag{22}$$

In the above relations, the variable ξ represents the difference $\tau_k - \tau'_k$ of the two special Schubert classes in codimension k.

The new relations (20) and (21) both come from the cohomology of the quadric $\mathcal{Q} \subset P(V)$. As in Section 3.2, let $\pi : P(S) \to OG'$ and $\theta : P(S) \to \mathcal{Q}$ denote the natural morphisms. We then have

$$\xi = \tau_k - \tau'_k = \pi_*(\theta^*(e-f)),$$

where e and f are the two ruling classes in $H^{2n}(\mathcal{Q}, \mathbb{Z})$. If $\zeta = c_1(\mathcal{O}_{P(S)}(1))$, the projective bundle formula dictates

$$\zeta^{n+1-k} + c_1(\pi^* S)\zeta^{n-k} + \cdots + c_{n+1-k}(\pi^* S) = 0. \tag{23}$$

The relations in the cohomology ring of $\mathcal{Q}$ imply that $\theta^*(e-f)\zeta = \theta^*(eh - fh) = 0$, and hence $\theta^*(e-f)c_{n+1-k}(\pi^* S) = 0$ in $H^*(P(S), \mathbb{Z})$, using (23). Applying π_*, we see that $(\tau_k - \tau'_k)c_{n+1-k}(S) = 0$ in $H^*(OG', \mathbb{Z})$, which is exactly relation (20). We leave the proof of (21) as an exercise for the reader.

As in Section 5.1, the analogous presentation of $H^*(OG', \mathbb{Z})$ with integer coefficients is rather more involved.

Theorem 16. *Define polynomials c_p using the equations (18). Then the cohomology ring $H^*(OG(n+1-k, 2n+2), \mathbb{Z})$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\tau_1, \ldots, \tau_k, \tau'_k, \tau_{k+1}, \ldots, \tau_{n+k}]$ modulo the relations*

$$\Delta_r = 0, \qquad n-k+1 < r \leqslant n,$$

$$\tau_k \Delta_{n+1-k} = \tau'_k \Delta_{n+1-k} = \sum_{p=k+1}^{n+1} (-1)^{p+k+1} \tau_p \Delta_{n+1-p},$$

$$\sum_{p=k+1}^{r} (-1)^p \tau_p \Delta_{r-p} = 0, \qquad n+1 < r \leqslant n+k,$$

and

$$\tau_k \tau'_k + \sum_{i=1}^{k} (-1)^i \tau_{k+i} \tau_{k-i} = 0,$$

$$\tau_r^2 + \sum_{i=1}^{r} (-1)^i \tau_{r+i} c_{r-i} = 0, \qquad k+1 \leqslant r \leqslant n.$$

6.2 The quantum theory

The story here is rather similar to that for the odd orthogonal Grassmannian of Section 5. The degree of q in the ring $QH(OG')$ is equal to $n+k$, and the quantum Pieri rule will involve both q and q^2 terms. Since $OG(n, 2n+2)$ is not a quotient of SO_{2n+2} by a *maximal* parabolic subgroup, we will assume that $k \geqslant 2$ in this subsection.

Suppose that the classical Pieri rule for $OG(n+1-k, 2n+2)$ is given by

$$\tau_\Lambda \tau_p = \sum_M f(\Lambda, M)\tau_M, \qquad \tau_\Lambda \tau'_k = \sum_M f'(\Lambda, M)\tau_M, \tag{24}$$

where the sums are over all $M \in \widetilde{\mathcal{P}}(k,n)$, $1 \leqslant p \leqslant n+k$, and the coefficients $f(\Lambda, M)$ and $f'(\Lambda, M)$ are given as in Theorem 15.

Let $\widetilde{\mathcal{Q}}(k,n)$ denote the set of $M = (\beta \,|\, \mu)$ (or $M = (\beta \,|\, \mu]$) in $\widetilde{\mathcal{P}}(k,n)$ such that $\mu_1 = n$. For a partition pair $M \in \widetilde{\mathcal{P}}(k,n)$, the pair M^* is defined using the same prescription as in Section 4.2. Let $\widetilde{\mathcal{P}}'(k,n+1)$ be the set of pairs $(\beta \,|\, \mu)$ or $(\beta \,|\, \mu]$ in $\widetilde{\mathcal{P}}(k,n+1)$ such that $\beta_1 = n+2-k$ and $\max(\beta_2+k-2, 1) \leqslant \mu_1 \leqslant n$. For any $M \in \widetilde{\mathcal{P}}'(k,n+1)$, define a partition pair $\widetilde{M} \in \widetilde{\mathcal{P}}(k,n)$ by the equations

$$\widetilde{(\beta \,|\, \mu)} = (\mu_1 - k + 1, \beta_2 - 1, \ldots, \beta_k - 1 \mid \mu_2, \mu_3, \ldots);$$
$$\widetilde{(\beta \,|\, \mu]} = (\mu_1 - k + 1, \beta_2 - 1, \ldots, \beta_k - 1 \mid \mu_2, \mu_3, \ldots].$$

Theorem 17. *For any $\Lambda = (\alpha \,|\, \lambda)$ or $\Lambda = (\alpha \,|\, \lambda]$ in $\widetilde{\mathcal{P}}(k,n)$ and p with $1 \leqslant p \leqslant n+k$, the quantum product $\tau_\Lambda \cdot \tau_p \in QH^*(OG(n+1-k, 2n+2))$ is equal to*

$$\sum_{M \in \widetilde{\mathcal{P}}(k,n)} f(\Lambda, M)\, \tau_M + \sum_{M \in \widetilde{\mathcal{P}}'(k,n+1)} f(\Lambda, M)\, \tau_{\widetilde{M}}\, q + \sum_{M \in \widetilde{\mathcal{Q}}(k,n)} f(\Lambda^*, M)\, \tau_{M^*}\, q^2,$$

where (i) the first sum is classical, as in (24), (ii) the second sum is over $M \in \widetilde{\mathcal{P}}'(k,n+1)$ with $\Lambda \xrightarrow{p} M$, and (iii) the third sum is empty unless $\lambda_1 = n$, and over $M \in \widetilde{\mathcal{Q}}(k,n)$ such that $\Lambda^* \xrightarrow{p} M$. Furthermore, the product $\tau_\Lambda \cdot \tau'_k$ is obtained by replacing f with f' throughout.

Theorem 17 is proved using the same set of arguments used to obtain Theorem 13. For the terms which are linear in q, we consider the diagram

$$\begin{array}{ccc} OF(n+1-k, n+2-k; 2n+2) & \longrightarrow & OG(n+1-k, 2n+2) \\ \downarrow & & \\ OG(n+2-k, 2n+2) & & \end{array}$$

and apply the projection formula as in Sections 2 and 5.2. The q^2 terms in the quantum Pieri rule for OG' come from the basic relation $\tau_{n+k}^2 = q^2$.

Example 6. The quantum Pieri relations

$$\tau_2 \cdot \tau_{(2,2|1)} = \tau_2\, q \quad \text{and} \quad \tau_2' \cdot \tau_{(2,2|1)} = \tau_{(2,2|3)}$$

hold in $QH^*(OG(2,8))$. Also, in $QH^*(OG(3,10))$, we have

$$\tau_2 \cdot \tau_{(3,2|4,2)} = \tau_{(3,3|4,3]} + \tau_{(3,3|4,2,1)} + \tau_{(3,2|2)}\, q + \tau_{(3,1|3)}\, q + \tau_1\, q^2;$$
$$\tau_2' \cdot \tau_{(3,2|4,2)} = \tau_{(3,3|4,3)} + \tau_{(3,3|4,2,1)} + \tau_{(3,2|2]}\, q + \tau_{(3,1|3)}\, q + \tau_1\, q^2.$$

We conclude with a presentation of the ring $QH^*(OG(n+1-k, 2n+2))$.

Theorem 18. *The quantum cohomology ring $QH^*(OG')$ is presented as a quotient of the polynomial ring $\mathbb{Z}[\tau_1, \ldots, \tau_k, \tau_k', \tau_{k+1}, \ldots, \tau_{n+k}, q]$ modulo the relations*

$$\Delta_r = 0, \qquad n-k+1 < r \leqslant n,$$
$$\tau_k \Delta_{n+1-k} = \tau_k' \Delta_{n+1-k} = \sum_{p=k+1}^{n+1} (-1)^{p+k+1} \tau_p \Delta_{n+1-p},$$
$$\sum_{p=k+1}^{r} (-1)^p \tau_p \Delta_{r-p} = 0, \qquad n+1 < r < n+k,$$
$$\sum_{p=k+1}^{n+k} (-1)^p \tau_p \Delta_{n+k-p} = -q,$$

and

$$\tau_k \tau_k' + \sum_{i=1}^{k} (-1)^i \tau_{k+i} \tau_{k-i} = 0,$$
$$\tau_r^2 + \sum_{i=1}^{r} (-1)^i \tau_{r+i} c_{r-i} = 0, \qquad k+1 \leqslant r \leqslant n,$$

where the polynomials c_p are defined by (18).

References

1. A. Bertram – "Quantum Schubert calculus", *Adv. Math.* **128** (1997), no. 2, p. 289–305.
2. A. Borel – "Sur la cohomologie des espaces fibrés principaux et des espaces homogènes de groupes de Lie compacts", *Ann. of Math. (2)* **57** (1953), p. 115–207.
3. A. S. Buch – "Quantum cohomology of Grassmannians", *Compositio Math.* **137** (2003), no. 2, p. 227–235.

4. A. S. Buch, A. Kresch and H. Tamvakis – "Gromov-Witten invariants on Grassmannians", *J. Amer. Math. Soc.* **16** (2003), no. 4, p. 901–915.
5. — , "Quantum Pieri rules for isotropic Grassmannians", in preparation.
6. C. Ehresmann – "Sur la topologie de certains espaces homogènes", *Ann. of Math. (2)* **35** (1934), no. 2, p. 396–443.
7. W. Fulton – *Intersection Theory*, second ed., Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge, vol. 2, Springer-Verlag, Berlin, 1998.
8. H. Hiller and B. Boe – "Pieri formula for $\mathrm{SO}_{2n+1}/\mathrm{U}_n$ and $\mathrm{Sp}_n/\mathrm{U}_n$", *Adv. in Math.* **62** (1986), no. 1, p. 49–67.
9. W. V. D. Hodge and D. Pedoe – *Methods of Algebraic Geometry. Vol. II. Book III: General Theory of Algebraic Varieties in Projective Space. Book IV: Quadrics and Grassmann Varieties*, Cambridge University Press, 1952.
10. A. Kresch and H. Tamvakis – "Quantum cohomology of the Lagrangian Grassmannian", *J. Algebraic Geom.* **12** (2003), no. 4, p. 777–810.
11. — , "Quantum cohomology of orthogonal Grassmannians", *Compositio Math.* **140** (2004), no. 2, p. 482–500.
12. M. Pieri – "Sul problema degli spazi secanti. nota 1^a", *Rend. Ist. Lombardo* **26(2)** (1893), p. 534–546.
13. P. Pragacz and J. Ratajski – "A Pieri-type theorem for Lagrangian and odd orthogonal Grassmannians", *J. Reine Angew. Math.* **476** (1996), p. 143–189.
14. — , "A Pieri-type formula for even orthogonal Grassmannians", *Fund. Math.* **178** (2003), no. 1, p. 49–96.
15. S. Sertöz – "A triple intersection theorem for the varieties $\mathrm{SO}(n)/P_d$", *Fund. Math.* **142** (1993), no. 3, p. 201–220.
16. B. Siebert and G. Tian – "On quantum cohomology rings of Fano manifolds and a formula of Vafa and Intriligator", *Asian J. Math.* **1** (1997), no. 4, p. 679–695.
17. H. Tamvakis – "Gromov–Witten invariants and quantum cohomology of Grassmannians", `math.AG/0306415`.

Endomorphism algebras of superelliptic jacobians

Yuri G. Zarhin

Department of Mathematics, Pennsylvania State University, University Park, PA 16802, U.S.A.
zarhin@math.psu.edu

Summary. We describe an explicit construction that provides a plenty of complex abelian varieties whose endomorphism algebra is a product of cyclotomic fields.

1 Introduction

As usual, we write $\mathbb{Z}, \mathbb{Q}, \mathbb{F}_p, \mathbb{C}$ for the ring of integers, the field of rational numbers, the finite field with p elements and the field of complex numbers respectively. If Z is a smooth algebraic variety over an algebraically closed field, then we write $\Omega^1(Z)$ for the space of differentials of the first kind on Z. If Z is an abelian variety, then we write $\mathrm{End}(Z)$ for its ring of (absolute) endomorphisms and $\mathrm{End}^0(Z)$ for its endomorphism algebra $\mathrm{End}(Z) \otimes \mathbb{Q}$. If Z is defined over a (not necessarily algebraically closed) field K, then we write $\mathrm{End}_K(Z) \subset \mathrm{End}(Z)$ for the (sub)ring of K-endomorphisms of Z.

Let p be a prime, $q = p^r$ an integral power of p, $\zeta_q \in \mathbb{C}$ a primitive qth root of unity, $\mathbb{Q}(\zeta_q) \subset \mathbb{C}$ the qth cyclotomic field and $\mathbb{Z}[\zeta_q]$ the ring of integers in $\mathbb{Q}(\zeta_q)$. If $q = 2$ then $\mathbb{Q}(\zeta_q) = \mathbb{Q}$. It is well known that if $q > 2$, then $\mathbb{Q}(\zeta_q)$ is a CM-field of degree $(p-1)p^{r-1}$. Let us put

$$\mathcal{P}_q(t) = \frac{t^q - 1}{t - 1} = t^{q-1} + \cdots + 1 \in \mathbb{Z}[t].$$

Clearly, $\mathcal{P}_q(t) = \prod_{i=1}^r \Phi_{p^i}(t)$ where $\Phi_{p^i}(t) = t^{(p-1)p^{i-1}} + \cdots + t^{p^{i-1}} + 1 \in \mathbb{Z}[t]$ is the p^ith cyclotomic polynomial. In particular, $\mathbb{Q}[t]/\Phi_{p^i}(t)\mathbb{Q}[t] = \mathbb{Q}(\zeta_{p^i})$ and $\mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t] = \prod_{i=1}^r \mathbb{Q}(\zeta_{p^i})$.

Let $f(x) \in \mathbb{C}[x]$ be a polynomial of degree $n \geq 4$ without multiple roots. Let $C_{f,q}$ be a smooth projective model of the smooth affine curve $y^q = f(x)$. The map $(x, y) \mapsto (x, \zeta_q y)$ gives rise to a non-trivial birational automorphism $\delta_q : C_{f,q} \to C_{f,q}$ of period q. The jacobian $J(C_{f,q})$ of $C_{f,q}$ is a complex abelian variety. By Albanese functoriality, δ_q induces an automorphism of $J(C_{f,q})$ which we still denote by δ_q. One may easily check (see Lemma 4.8

below) that $\delta_q^{q-1}+\cdots+\delta_q+1=0$ in $\mathrm{End}(J(C_{f,q}))$. This implies that if $\mathbb{Q}[\delta_q]$ is the $\mathbb{Q}$-subalgebra of $\mathrm{End}^0(J(C_{f,q}))$ generated by δ_q, then there is the natural surjective homomorphism $\mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t] \twoheadrightarrow \mathbb{Q}[\delta_q]$ that sends $t+\mathcal{P}_q(t)\mathbb{Q}[t]$ to δ_q. One may check that this homomorphism is, in fact, an isomorphism (see [6, p. 149], [8, p. 458]) where the case $q=p$ was treated). This gives us an embedding $\mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t] \cong \mathbb{Q}[\delta_q] \subset \mathrm{End}^0(J(C_{f,q}))$. Our main result is the following statement.

Theorem 1.1. *Let K be a subfield of $\mathbb{C}$ such that $f(x)$ is an irreducible polynomial in $K[x]$ of degree $n \ge 5$ and its Galois group over K is either the full symmetric group $\mathbf{S}_n$ or the alternating group $\mathbf{A}_n$. In addition, assume that either p does not divide n or $q \mid n$. Then $\mathrm{End}^0(J(C_{f,q})) = \mathbb{Q}[\delta_q] \cong \mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t] = \prod_{i=1}^{r} \mathbb{Q}(\zeta_{p^i})$.*

Remark 1.2. If q is a prime (i.e., $q=p$), then $J(C_{f,p})$ is an absolutely simple abelian variety and $\mathrm{End}(J(C_{f,p})) = \mathbb{Z}[\delta_p] \cong \mathbb{Z}[\zeta_p]$ [17, 22]. In particular, if $p=2$, then $C_{f,2}$ is a hyperelliptic curve, δ_2 is multiplication by -1 and $\mathrm{End}(J(C_{f,2})) = \mathbb{Z}$. See [21, 16, 14] for a discussion of finite characteristic case.

Example 1.3. Let $n \ge 5$ be an integer, p a prime, r a positive integer, $q=p^r$. Assume also that either n is not divisible by p or $q \mid n$.

1. The polynomial $x^n-x-1 \in \mathbb{Q}[x]$ has Galois group $\mathbf{S}_n$ over $\mathbb{Q}$ ([10, p. 42]). Therefore the endomorphism algebra (over $\mathbb{C}$) of the jacobian $J(C)$ of the curve $C: y^q = x^n - x - 1$ is $\mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t]$.
2. The Galois group of the "truncated exponential"

$$\exp_n(x) := 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots + \frac{x^n}{n!} \in \mathbb{Q}[x]$$

is either $\mathbf{S}_n$ or $\mathbf{A}_n$ [9]. Therefore the endomorphism algebra (over $\mathbb{C}$) of the jacobian $J(C)$ of the curve $C: y^q = \exp_n(x)$ is $\mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t]$.

Remark 1.4. If $f(x) \in K[x]$, then the curve $C_{f,q}$ and its jacobian $J(C_{f,q})$ are defined over K. Let $K_a \subset \mathbb{C}$ be the algebraic closure of K. Clearly, all endomorphisms of $J(C_{f,q})$ are defined over K_a. This implies that in order to prove Theorem 1.1, it suffices to check that $\mathbb{Q}[\delta_q]$ coincides with the $\mathbb{Q}$-algebra of K_a-endomorphisms of $J(C_{f,q})$.

Our main technical tool used in the proof of Theorem 1.1 is a certain modular representation $V_{f,p}$ of the Galois group of f [3], [20] arising from its action on the roots of f. In the case of $q=p$ the Galois module $V_{f,p}$ is canonically isomorphic to the subgroup of δ_p-invariants in $J(C_{f,p})$ (if $\zeta_p \in K$) [6], [8]. In the present paper we construct (assuming that $\zeta_q \in K$ and p does not divide n) an abelian subvariety $J^{f,q} \subset J(C_{f,q})$ with multiplication by $\mathbb{Z}[\zeta_q]$ and prove that $V_{f,p}$ is canonically isomorphic to the subgroup of ζ_q-invariants

in $J^{f,q}$ (Lemma 4.11). (It turns out that if $q = p^r$, then $J(C_{f,q})$ is isogenous to a product of all J^{f,p^i} with $1 \le i \le r$.)

The paper is organized as follows. In Section 2 we obtain conditions that guarantee that the center of the endomorphism algebra of a complex abelian variety is a cyclotomic field (Corollary 2.2). In Section 3 we study abelian varieties X over arbitrary fields, whose endomorphism ring contains a subring isomorphic to the ring $\mathcal{O}$ of integers in a given number field E. We study the Galois action on the λ-torsion X_λ of X where λ is a maximal ideal in $\mathcal{O}$. We prove (Theorem 3.8) that if the Galois module X_λ is *very simple* in the sense of [18], [23], then the centralizer of E in the algebra $\mathrm{End}^0(X)$ of all (absolute) endomorphisms of X either coincides with E or is "very big". In Section 4 we study endomorphism algebras of $J^{(f,q)}$, using the very simplicity of the Galois module $V_{f,p}$ when $\deg(f) \ge 5$ and the Galois group of f is either the full symmetric or the alternating group. Theorem 3.8 helps us to prove that, in characteristic zero, $\mathbb{Q}(\zeta_q)$ is a maximal commutative subalgebra in $\mathrm{End}^0(J^{f,q})$. Using Corollary 2.2 and computations with differentials of the first kind (Theorem 3.10 and Remark 4.2), we prove (Theorem 4.16) that the center of $\mathrm{End}^0(J^{f,q})$ coincides with $\mathbb{Q}(\zeta_q)$ and therefore $\mathrm{End}^0(J^{f,q}) = \mathbb{Q}(\zeta_q)$. We finish the proof of Theorem 1.1 in Section 5.

2 Complex abelian varieties

Let Z be a complex abelian variety of positive dimension. We write $\mathfrak{C}_Z$ for the center of the semisimple finite-dimensional $\mathbb{Q}$-algebra $\mathrm{End}^0(Z)$.

Let E be a subfield of $\mathrm{End}^0(Z)$ that contains the identity map. Let Σ_E be the set of all field embeddings $\sigma : E \hookrightarrow \mathbb{C}$. It is well known that

$$\mathbb{C}_\sigma := E \otimes_{E,\sigma} \mathbb{C} = \mathbb{C}, \quad E_{\mathbb{C}} = E \otimes_{\mathbb{Q}} \mathbb{C} = \prod_{\sigma \in \Sigma_E} E \otimes_{E,\sigma} \mathbb{C} = \prod_{\sigma \in \Sigma_E} \mathbb{C}_\sigma.$$

Let $\mathrm{Lie}(Z)$ be the tangent space to the origin of Z; it is a $\dim(Z)$-dimensional $\mathbb{C}$-vector space. By functoriality, $\mathrm{End}^0(Z)$ and therefore E act on $\mathrm{Lie}(Z)$ and therefore provide $\mathrm{Lie}(Z)$ with a natural structure of $E \otimes_{\mathbb{Q}} \mathbb{C}$-module. Clearly,

$$\mathrm{Lie}(Z) = \bigoplus_{\sigma \in \Sigma_E} \mathbb{C}_\sigma \mathrm{Lie}(Z) = \oplus_{\sigma \in \Sigma_E} \mathrm{Lie}(Z)_\sigma$$

where $\mathrm{Lie}(Z)_\sigma := \mathbb{C}_\sigma \mathrm{Lie}(Z) = \{x \in \mathrm{Lie}(Z) \mid ex = \sigma(e)x \quad \forall e \in E\}$. Let us put $n_\sigma = n_\sigma(Z, E) = \dim_{\mathbb{C}_\sigma} \mathrm{Lie}(Z)_\sigma = \dim_{\mathbb{C}} \mathrm{Lie}(Z)_\sigma$. It is well known that the natural map $\Omega^1(Z) \to \mathrm{Hom}_{\mathbb{C}}(\mathrm{Lie}(Z), \mathbb{C})$ is an isomorphism. This allows us to define via duality the natural homomorphism $E \to \mathrm{End}_{\mathbb{C}}(\mathrm{Hom}_{\mathbb{C}}(\mathrm{Lie}(Z), \mathbb{C})) = \mathrm{End}_{\mathbb{C}}(\Omega^1(Z))$. This provides $\Omega^1(Z)$ with a natural structure of $E \otimes_{\mathbb{Q}} \mathbb{C}$-module in such a way that $\Omega^1(Z)_\sigma := \mathbb{C}_\sigma \Omega^1(Z) \cong \mathrm{Hom}_{\mathbb{C}}(\mathrm{Lie}(Z)_\sigma, \mathbb{C})$. In particular,

$$n_\sigma = \dim_{\mathbb{C}}(\mathrm{Lie}(Z)_\sigma) = \dim_{\mathbb{C}}(\Omega^1(Z)_\sigma). \tag{1}$$

The following statement is contained in [22, Th. 2.3].

Theorem 2.1. *If $E/\mathbb{Q}$ is Galois, E contains $\mathfrak{C}_Z$ and $\mathfrak{C}_Z \neq E$, then there exists a nontrivial automorphism $\kappa : E \to E$ such that $n_\sigma = n_{\sigma\kappa}$ for all $\sigma \in \Sigma_E$.*

The following assertion will be used in the proof of Theorem 4.16.

Corollary 2.2. *Suppose that there exist a prime p, a positive integer r, the prime power $q = p^r$ and an integer $n \geq 4$ such that:*

(i) $E = \mathbb{Q}(\zeta_q) \subset \mathbb{C}$, where $\zeta_q \in \mathbb{C}$ is a primitive qth root of unity;
(ii) n is not divisible by p, i.e., n and q are relatively prime;
(iii) Let $i < q$ be a positive integer not divisible by p and $\sigma_i : E = \mathbb{Q}(\zeta_q) \hookrightarrow \mathbb{C}$ the embedding that sends ζ_q to ζ_q^{-i}. Then $n_{\sigma_i} = \left[\frac{ni}{q}\right]$.

Then $\mathfrak{C}_Z = \mathbb{Q}(\zeta_q)$.

Proof. If $q = 2$, then $E = \mathbb{Q}(\zeta_2) = \mathbb{Q}$. Since $\mathfrak{C}_Z$ is a subfield of $E = \mathbb{Q}$, we conclude that $\mathfrak{C}_Z = \mathbb{Q} = \mathbb{Q}(\zeta_2)$. So, further we assume that $q > 2$.

Clearly, $\{\sigma_i\}$ is the collection Σ of all embeddings $\mathbb{Q}(\zeta_q) \hookrightarrow \mathbb{C}$. By (iii), $n_{\sigma_i} = 0$ if and only if $1 \leq i \leq [\frac{q}{n}]$. Suppose that $\mathfrak{C}_Z \neq \mathbb{Q}(\zeta_q)$. It follows from Theorem 2.1 that there exists a non-trivial field automorphism $\kappa : \mathbb{Q}[\zeta_q] \to \mathbb{Q}[\zeta_q]$ such that for all $\sigma \in \Sigma$ we have $n_\sigma = n_{\sigma\kappa}$. Clearly, there exists an integer m such that p does *not* divide m, $1 < m < q$ and $\kappa(\zeta_q) = \zeta_q^m$.

Assume that $q < n$. In this case the function $i \mapsto n_{\sigma_i} = [\frac{ni}{q}]$ is strictly increasing and therefore $n_{\sigma_i} \neq n_{\sigma_j}$ while $i \neq j$. This implies that $\sigma_i = \sigma_i\kappa$, i.e., κ is the identity map, which is not the case. The obtained contradiction implies that $n < q$. Since $n \geq 4$, we have $q \geq 5$.

If i is an integer, then we write $\bar{i} \in \mathbb{Z}/q\mathbb{Z}$ for its residue modulo q. Clearly, $n_\sigma = 0$ if and only if $\sigma = \sigma_i$ with $1 \leq i \leq [\frac{q}{n}]$. Since n and q are relatively prime, $[\frac{q}{n}] = [\frac{q-1}{n}]$. It follows that $n_{\sigma_i} = 0$ if and only if $1 \leq i \leq [\frac{q-1}{n}]$. Clearly, the map $\sigma \mapsto \sigma\kappa$ permutes the set

$$\{\sigma_i \mid 1 \leq i \leq \left[\frac{q-1}{n}\right], p \text{ does not divide } i\}.$$

Since $\kappa(\zeta_q) = \zeta_q^m$ and $\sigma_i\kappa(\zeta_q) = \zeta_q^{-im}$, it follows that if

$$A := \left\{ i \in \mathbb{Z} \mid 1 \leq i \leq \left[\frac{q-1}{n}\right] < q,\ p \text{ does not divide } i \right\},$$

then the multiplication by m in $(\mathbb{Z}/q\mathbb{Z})^* = \mathrm{Gal}(\mathbb{Q}(\zeta_q)/\mathbb{Q})$ leaves invariant the set $\bar{A} := \{\bar{i} \in \mathbb{Z}/q\mathbb{Z} \mid i \in A\}$. Clearly, A contains 1 and therefore $\bar{m} = m \cdot \bar{1} \in \bar{A}$. Since $1 < m < q$,

$$m = m \cdot 1 \leq \left[\frac{q-1}{n}\right]. \tag{2}$$

Let us consider the arithmetic progression consisting of $2m$ integers

$$[\frac{q-1}{n}] + 1, \ldots, [\frac{q-1}{n}] + 2m$$

with difference 1. All its elements lie between $[\frac{q-1}{n}] + 1$ and

$$\left[\frac{q-1}{n}\right] + 2m \leq 3\left[\frac{q-1}{n}\right] \leq 3\frac{q-1}{4} < q - 1.$$

There exist exactly two elements of A say, mc_1 and $mc_1 + m$ that are divisible by m. Clearly, c_1 is a positive integer and either c_1 or $c_1 + 1$ is not divisible by p; we put $c = c_1$ in the former case and $c = c_1 + 1$ in the latter case. However, c is not divisible by p and

$$\left[\frac{q-1}{n}\right] < mc \leq \left[\frac{q-1}{n}\right] + 2m < q - 1. \tag{3}$$

It follows that mc does not lie in A and therefore $\overline{mc}$ does not lie in $\bar{A}$. This implies that $\bar{c}$ also does not lie in $\bar{A}$ and therefore $c > \left[\frac{q-1}{n}\right]$. Using (3), we conclude that

$$(m-1)\left[\frac{q-1}{n}\right] < 2m$$

and therefore

$$\left[\frac{q-1}{n}\right] < \frac{2m}{m-1} = 2 + \frac{2}{m-1}.$$

If $m > 2$, then $m \geq 3$ and using (2), we conclude that

$$3 \leq m \leq \left[\frac{q-1}{n}\right] < 2 + \frac{2}{m-1} \leq 3$$

and therefore $3 < 3$, which is not true. Hence $m = 2$ and

$$2 = m \leq \left[\frac{q-1}{n}\right] < 2 + \frac{2}{m-1} = 4$$

and therefore $\left[\frac{q-1}{n}\right] = 2$ or 3. It follows that $q \geq 1 + 2n \geq 1 + 2 \cdot 4 = 9$. Since $m = 2$ is not divisible by p, we conclude that $p \geq 3$ and either $\bar{A} = \{\bar{1}, \bar{2}\}$ or $p > 3$ and $A = \{\bar{1}, \bar{2}, \bar{3}\}$. In both cases $\bar{4} = 2 \cdot \bar{2} = m \cdot \bar{2}$ must lie in $\bar{A}$. Contradiction. □

3 Abelian varieties over arbitrary fields

Let K be a field. Let us fix its algebraic closure K_a and denote by $\mathrm{Gal}(K)$ the absolute Galois group $\mathrm{Aut}(K_a/K)$ of K. If X is an abelian variety of positive dimension over K_a, then we write 1_X (or even just 1) for the identity automorphism of X. If Y is an (maybe another) abelian variety of positive dimension over K_a, then we write $\mathrm{Hom}(X,Y)$ for the group of all K_a-homomorphisms from X to Y. We write $\mathrm{Hom}^0(X,Y)$ for the finite-dimensional $\mathbb{Q}$-vector space $\mathrm{Hom}(X,Y)\otimes\mathbb{Q}$. Clearly, $\mathrm{End}(X)=\mathrm{Hom}(X,X)$ and $\mathrm{End}^0(X)=\mathrm{End}(X)\otimes\mathbb{Q}=\mathrm{Hom}^0(X,X)$. It is well known that $\mathrm{End}^0(X)$ is a finite-dimensional semisimple $\mathbb{Q}$-algebra and $\dim_{\mathbb{Q}}(\mathrm{End}^0(X))$ does not exceed $4\dim(X)^2$ [4, §19, Corollary 1 to Theorem 3]; the equality holds if and only if $\mathrm{char}(K)>0$ and X is a supersingular abelian variety [17, Lemma 3.1].

Let E be a number field and $\mathcal{O}\subset E$ be the ring of all its algebraic integers. Let (X,i) be a pair consisting of an abelian variety X over K_a and an embedding

$$i: E \hookrightarrow \mathrm{End}^0(X)$$

with $i(1)=1_X$. It is well known [12, Proposition 2 on p. 36] that $[E:\mathbb{Q}]$ divides $2\dim(X)$, i.e., $r=r_X:=2\dim(X)/E:\mathbb{Q}]$ is a positive integer.

Let us denote by $\mathrm{End}^0(X,i)$ the centralizer of $i(E)$ in $\mathrm{End}^0(X)$. Clearly, $i(E)$ lies in the center of the finite-dimensional $\mathbb{Q}$-algebra $\mathrm{End}^0(X,i)$. It follows that $\mathrm{End}^0(X,i)$ carries a natural structure of a finite-dimensional E-algebra. If Y is (possibly) another abelian variety over K_a and $j: E\hookrightarrow \mathrm{End}^0(Y)$ is an embedding that sends 1 to the identity automorphism of Y, then we write

$$\mathrm{Hom}^0((X,i),(Y,j))=\{u\in \mathrm{Hom}^0(X,Y)\mid ui(c)=j(c)u \quad \forall c\in E\}.$$

We have $\mathrm{End}^0(X,i)=\mathrm{Hom}^0((X,i),(X,i))$. By abuse of language, we call elements of $\mathrm{Hom}^0((X,i),(Y,j))$ *E-equivariant* homomorphisms from X to Y.

Recall that if $\psi: X\to Y$ is an isogeny, then there exist an isogeny $\phi: Y\to X$ and a positive integer N such that $\phi\psi=N1_X, \psi\phi=N1_Y$. One may easily check that if ψ is E-equivariant, then ϕ is also E-equivariant.

If d is a positive integer, then we write $i^{(d)}$ for the composition

$$E\hookrightarrow \mathrm{End}^0(X)\subset \mathrm{End}^0(X^d)$$

of i and the diagonal inclusion $\mathrm{End}^0(X)\subset \mathrm{End}^0(X^d)$.

It is known that the E-algebra $\mathrm{End}^0(X,i)$ is semisimple [15, Remark 4.1]. The following assertion is contained in [15, Theorem 4.2].

Theorem 3.1. *(i) We always have*

$$\dim_E(\mathrm{End}^0(X,i))\le \frac{4\cdot\dim(X)^2}{[E:\mathbb{Q}]^2}.$$

(ii) Suppose that

$$\dim_E(\mathrm{End}^0(X,i)) = \frac{4\cdot\dim(X)^2}{[E:\mathbb{Q}]^2}.$$

Then X is an abelian variety of CM-type isogenous to a self-product of an (absolutely) simple abelian variety. Also $\mathrm{End}^0(X,i)$ is a central simple E-algebra, i.e., E coincides with the center of $\mathrm{End}^0(X,i)$.
Furthermore, if $\mathrm{char}(K_a) = 0$, then $[E:\mathbb{Q}]$ is even and there exist an $\frac{[E:\mathbb{Q}]}{2}$-dimensional abelian variety Z, an isogeny $\psi : Z^r \to X$ and an embedding $k : E \hookrightarrow \mathrm{End}^0(Z)$ that send 1 to 1_Z and such that $\psi \in \mathrm{Hom}^0((Z^r, k^{(r)}), (X,i))$.

Remark 3.2. Suppose that

$$\dim_E(\mathrm{End}^0(X,i)) = \frac{4\cdot\dim(X)^2}{[E:\mathbb{Q}]^2}.$$

By Theorem 3.1(ii), X is isogenous to a self-product of an absolutely simple abelian variety B. It is proven in [15, §4, Proof of Theorem 4.2] that B is an abelian variety of CM-type. Recall [12, Prop. 26 on p. 96] that in characteristic zero every absolutely simple abelian variety of CM type is defined over a number field; in positive characteristic such a variety is isogenous to an abelian variety defined over a finite field (a theorem of Grothendieck [5, Th. 1.1]). It follows easily that:

1. If $\mathrm{char}(K) = 0$, then X is defined over a number field;
2. If $\mathrm{char}(K) > 0$, then X is isogenous to an abelian variety defined over a finite field.

Let d be a positive integer that is not divisible by $\mathrm{char}(K)$. Suppose that X is defined over K. We write X_d for the kernel of multiplication by d in $X(K_a)$. It is known [4, Proposition on p. 64] that the commutative group X_d is a free $\mathbb{Z}/d\mathbb{Z}$-module of rank $2\dim(X)$. Clearly, X_d is a Galois submodule in $X(K_a)$. We write $\tilde{\rho}_{d,X} : \mathrm{Gal}(K) \to \mathrm{Aut}_{\mathbb{Z}/d\mathbb{Z}}(X_d) \cong \mathrm{GL}(2\dim(X), \mathbb{Z}/d\mathbb{Z})$ for the corresponding (continuous) homomorphism defining the Galois action on X_d. Let us put

$$\tilde{G}_{d,X} = \tilde{\rho}_{d,X}(\mathrm{Gal}(K)) \subset \mathrm{Aut}_{\mathbb{Z}/d\mathbb{Z}}(X_d).$$

Clearly, $\tilde{G}_{d,X}$ coincides with the Galois group of the field extension $K(X_d)/K$, where $K(X_d)$ is the field of definition of all points on X of order dividing d. In particular, if a prime $\ell \neq \mathrm{char}(K)$, then X_ℓ is a $2\dim(X)$-dimensional vector space over the prime field $\mathbb{F}_\ell = \mathbb{Z}/\ell\mathbb{Z}$ and the inclusion $\tilde{G}_{\ell,X} \subset \mathrm{Aut}_{\mathbb{F}_\ell}(X_\ell)$ defines a faithful linear representation of $\tilde{G}_{\ell,X}$ in the vector space X_ℓ.

Now let us assume that

$$i(\mathcal{O}) \subset \mathrm{End}_K(X).$$

Let λ be a maximal ideal in $\mathcal{O}$. We write $k(\lambda)$ for the corresponding (finite) residue field. Let us put

$$X_\lambda := \{x \in X(K_a) \mid i(e)x = 0 \quad \forall e \in \lambda\}.$$

Clearly, if $\mathrm{char}(k(\lambda)) = \ell$, then $\lambda \supset \ell \cdot \mathcal{O}$ and therefore $X_\lambda \subset X_\ell$. Clearly, X_λ is a Galois submodule of X_ℓ. It is also clear that X_λ carries a natural structure of $\mathcal{O}/\lambda = k(\lambda)$-vector space. We write

$$\tilde{\rho}_{\lambda,X} : \mathrm{Gal}(K) \to \mathrm{Aut}_{k(\lambda)}(X_\lambda)$$

for the corresponding (continuous) homomorphism defining the Galois action on X_λ. Let us put

$$\tilde{G}_{\lambda,X} = \tilde{G}_{\lambda,i,X} := \tilde{\rho}_{\lambda,X}(\mathrm{Gal}(K)) \subset \mathrm{Aut}_{k(\lambda)}(X_\lambda).$$

Clearly, $\tilde{G}_{\lambda,X}$ coincides with the Galois group of the field extension $K(X_\lambda)/K$ where $K(X_\lambda) = K(X_{\lambda,i})$ is the field of definition of all points in X_λ.

In order to describe $\tilde{\rho}_{\lambda,X}$ explicitly, let us assume for the sake of simplicity that λ is the only maximal ideal of $\mathcal{O}$ dividing ℓ, i.e., $\ell \cdot \mathcal{O} = \lambda^b$ where the positive integer b satisfies $[E : \mathbb{Q}] = b \cdot [k(\lambda) : \mathbb{F}_\ell]$. Then $\mathcal{O} \otimes \mathbb{Z}_\ell = \mathcal{O}_\lambda$ where $\mathcal{O}_\lambda$ is the completion of $\mathcal{O}$ with respect to the λ-adic topology. It is well-known that $\mathcal{O}_\lambda$ is a local principal ideal domain and its only maximal ideal is $\lambda\mathcal{O}_\lambda$. One may easily check that $\ell \cdot \mathcal{O}_\lambda = (\lambda\mathcal{O}_\lambda)^b$.

Let us choose an element $c \in \lambda$ that does not lie in λ^2. Clearly, $\lambda\mathcal{O}_\lambda = c \cdot \mathcal{O}_\lambda$. This implies that there exists a unit $u \in \mathcal{O}_\lambda^*$ such that $\ell = uc^b$. It follows from the unique factorization of ideals in $\mathcal{O}$ that $\lambda = \ell \cdot \mathcal{O} + c \cdot \mathcal{O}$. It follows readily that

$$X_\lambda = \{x \in X_\ell \mid cx = 0\} \subset X_\ell.$$

Let $T_\ell(X)$ be the ℓ-adic Tate module of X defined as the projective limit of Galois modules X_{ℓ^m} [4, §18]. Recall that $T_\ell(X)$ is a free $\mathbb{Z}_\ell$-module of rank $2\dim(X)$ provided with the continuous action $\rho_{\ell,X} : \mathrm{Gal}(K) \to \mathrm{Aut}_{\mathbb{Z}_\ell}(T_\ell(X))$ and the natural embedding [4, §19, Theorem 3]

$$\mathrm{End}_K(X) \otimes \mathbb{Z}_\ell \subset \mathrm{End}(X) \otimes \mathbb{Z}_\ell \hookrightarrow \mathrm{End}_{\mathbb{Z}_\ell}(T_\ell(X)). \tag{4}$$

Clearly, the image of $\mathrm{End}_K(X) \otimes \mathbb{Z}_\ell$ commutes with $\rho_{\ell,X}(\mathrm{Gal}(K))$. In particular, $T_\ell(X)$ carries the natural structure of $\mathcal{O} \otimes \mathbb{Z}_\ell = \mathcal{O}_\lambda$-module. The following assertion is a special case of Proposition 2.2.1 on p. 769 in [7].

Lemma 3.3. *The $\mathcal{O}_\lambda$-module $T_\ell(X)$ is free of rank r_X.*

There is also the natural isomorphism of Galois modules $X_\ell = T_\ell(X)/\ell T_\ell(X)$, which is also an isomorphism of $\mathrm{End}_K(X) \supset \mathcal{O}$-modules. This implies that the $\mathcal{O}[\mathrm{Gal}(K)]$-module X_λ coincides with

$$\begin{aligned} c^{-1}\ell\, T_\ell(X)/\ell\, T_\ell(X) &= c^{b-1}T_\ell(X)/c^b T_\ell(X) = T_\ell(X)/cT_\ell(X) \\ &= T_\ell(X)/\lambda T_\ell(X) = T_\ell(X)/(\lambda\mathcal{O}_\lambda)T_\ell(X). \end{aligned}$$

Hence

$$X_\lambda = T_\ell(X)/(\lambda\mathcal{O}_\lambda)T_\ell(X) = T_\ell(X)\otimes_{\mathcal{O}_\lambda} k(\lambda),\ \dim_{k(\lambda)}X_\lambda = \frac{2\dim(X)}{[E:\mathbb{Q}]} = r_X. \tag{5}$$

Consider the $2\dim(X)$-dimensional $\mathbb{Q}_\ell$-vector space

$$V_\ell(X) = T_\ell(X)\otimes_{\mathbb{Z}_\ell}\mathbb{Q}_\ell,$$

which carries a natural structure of r_X-dimensional E_λ-vector space. Extending the embedding (4) by $\mathbb{Q}_\ell$-linearity, we get the natural embedding

$$E\otimes_{\mathbb{Q}}\mathbb{Q}_\ell = \mathcal{O}\otimes\mathbb{Q}_\ell \overset{i}{\hookrightarrow} \mathrm{End}_K(X)\otimes\mathbb{Q}_\ell \subset \mathrm{End}^0(X)\otimes_{\mathbb{Q}}\mathbb{Q}_\ell \hookrightarrow \mathrm{End}_{\mathbb{Q}_\ell}(V_\ell(X)).$$

Further we will identify $\mathrm{End}^0(X)\otimes_{\mathbb{Q}}\mathbb{Q}_\ell$ with its image in $\mathrm{End}_{\mathbb{Q}_\ell}(V_\ell(X))$.

Remark 3.4.

1. The center $\mathfrak{C}_X$ of $\mathrm{End}^0(X)$ commutes with $i(E)$ and therefore lies in $\mathrm{End}^0(X,i)$. Since $\mathfrak{C}_X$ also commutes with $\mathrm{End}^0(X,i)$, it lies in the center of $\mathrm{End}^0(X,i)$;
2. Note that $E_\lambda = E\otimes_{\mathbb{Q}}\mathbb{Q}_\ell = \mathcal{O}\otimes\mathbb{Q}_\ell = \mathcal{O}_\lambda\otimes_{\mathbb{Z}_\ell}\mathbb{Q}_\ell$ is the field coinciding with the completion of E with respect to λ-adic topology. Clearly, $V_\ell(X)$ carries a natural structure of r_X-dimensional E_λ-vector space and $\dim_{E_\lambda}(\mathrm{End}_{E_\lambda}(V_\ell(X))) = r_X^2$.
3. One may easily check that $\mathrm{End}^0(X,i)\otimes_{\mathbb{Q}}\mathbb{Q}_\ell$ is a $E\otimes_{\mathbb{Q}}\mathbb{Q}_\ell = E_\lambda$-vector subspace (even subalgebra) in $\mathrm{End}_{E_\lambda}(V_\ell(X))$. Clearly,
$$\dim_{E_\lambda}(\mathrm{End}^0(X,i)\otimes_{\mathbb{Q}}\mathbb{Q}_\ell) = \dim_E(\mathrm{End}^0(X,i)).$$
4. If $\mathrm{End}^0(X,i)\otimes_{\mathbb{Q}}\mathbb{Q}_\ell = E_\lambda\mathrm{Id}$, then $\dim_E(\mathrm{End}^0(X,i)) = 1$ and, in light of the inclusion $E\cong i(E)\subset\mathrm{End}^0(X,i)$, we obtain that $\mathrm{End}^0(X,i) = i(E)$, i.e., $i(E)\cong E$ is a maximal commutative subalgebra in $\mathrm{End}^0(X)$ and $i(\mathcal{O})\cong\mathcal{O}$ is a maximal commutative subring in $\mathrm{End}(X)$. It follows that $\mathfrak{C}_X\subset i(E)$ and therefore is isomorphic to a subfield of E. In particular, $\mathfrak{C}_X$ is a field, i.e., $\mathrm{End}^0(X)$ is a simple $\mathbb{Q}$-algebra. This means that X is isogenous to a self-product of an absolutely simple abelian variety;
5. Suppose that $\mathrm{End}^0(X,i)\otimes_{\mathbb{Q}}\mathbb{Q}_\ell = \mathrm{End}_{E_\lambda}(V_\ell(X))$. This implies that
$$\dim_E(\mathrm{End}^0(X,i)) = r_X^2.$$
Applying Theorem 3.1, we conclude that X is an abelian variety of CM-type isogenous to a self-product of an (absolutely) simple abelian variety. Also $\mathrm{End}^0(X,i)$ is a central simple E-algebra, i.e., E coincides with the center of $\mathrm{End}^0(X,i)$. Moreover, if $\mathrm{char}(K_a) = 0$, then $[E:\mathbb{Q}]$ is even and there exist an $\frac{[E:\mathbb{Q}]}{2}$-dimensional abelian variety Z, an isogeny $\psi: Z^r\to X$ and an embedding $k: E\hookrightarrow\mathrm{End}^0(Z)$ that send 1 to 1_Z and such that $\psi\in\mathrm{Hom}^0((Z^r,k^{(r)}),(X,i))$.

Using the inclusion $\mathrm{Aut}_{\mathbb{Z}_\ell}(T_\ell(X)) \subset \mathrm{Aut}_{\mathbb{Q}_\ell}(V_\ell(X))$, one may view $\rho_{\ell,X}$ as the ℓ-adic representation $\rho_{\ell,X} : \mathrm{Gal}(K) \to \mathrm{Aut}_{\mathbb{Z}_\ell}(T_\ell(X)) \subset \mathrm{Aut}_{\mathbb{Q}_\ell}(V_\ell(X))$.

Since X is defined over K, one may associate with every $u \in \mathrm{End}(X)$ and $\sigma \in \mathrm{Gal}(K)$ an endomorphism ${}^\sigma u \in \mathrm{End}(X)$ such that ${}^\sigma u(x) = \sigma u(\sigma^{-1}x)$ for all $x \in X(K_a)$. Clearly, ${}^\sigma u = u$ if $u \in \mathrm{End}_K(X)$. In particular, ${}^\sigma e = e$ if $e \in \mathcal{O}$ (here we identify $\mathcal{O}$ with $i(\mathcal{O})$). It follows easily that for each $\sigma \in \mathrm{Gal}(K)$ the map $u \mapsto {}^\sigma u$ extends by $\mathbb{Q}$-linearity to a certain automorphism of $\mathrm{End}^0(X)$. Clearly, ${}^\sigma e = e$ for each $e \in E$ and ${}^\sigma u \in \mathrm{End}^0(X,i)$ for each $u \in \mathrm{End}^0(X,i)$.

Remark 3.5. The definition of $T_\ell(X)$ as the projective limit of Galois modules X_{ℓ^m} implies that ${}^\sigma u(x) = \rho_{\ell,X}(\sigma) u \rho_{\ell,X}(\sigma)^{-1}(x)$ for all $x \in T_\ell(X)$. It follows easily that ${}^\sigma u(x) = \rho_{\ell,X}(\sigma) u \rho_{\ell,X}(\sigma)^{-1}(x)$ for all $x \in V_\ell(X), u \in \mathrm{End}^0(X), \sigma \in \mathrm{Gal}(K)$. This implies that for each $\sigma \in \mathrm{Gal}(K)$ we have $\rho_{\ell,X}(\sigma) \in \mathrm{Aut}_{E_\lambda}(V_\lambda(X))$ and therefore

$$\rho_{\ell,X}(\mathrm{Gal}(K)) \subset \mathrm{Aut}_{E_\lambda}(V_\lambda(X))$$

[7, pp. 767–768] (see also [11]). It is also clear that $\rho_{\ell,X}(\sigma) u \rho_{\ell,X}(\sigma)^{-1} \in \mathrm{End}^0(X) \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$ for all $u \in \mathrm{End}^0(X) \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$ and

$$\rho_{\ell,X}(\sigma) u \rho_{\ell,X}(\sigma)^{-1} \in \mathrm{End}^0(X,i) \otimes_{\mathbb{Q}} \mathbb{Q}_\ell \quad \forall u \in \mathrm{End}^0(X,i) \otimes_{\mathbb{Q}} \mathbb{Q}_\ell.$$

We refer to [18],[19], [21], [23] for a discussion of the following definition.

Definition 3.6. *Let V be a vector space over a field $\mathbb{F}$, let G be a group and $\rho : G \to \mathrm{Aut}_{\mathbb{F}}(V)$ a linear representation of G in V. We say that the G-module V is very simple if it enjoys the following property:*

If $R \subset \mathrm{End}_{\mathbb{F}}(V)$ is an $\mathbb{F}$-subalgebra containing the identity operator Id *such that $\rho(\sigma) R \rho(\sigma)^{-1} \subset R \quad \forall \sigma \in G$, then either $R = \mathbb{F} \cdot \mathrm{Id}$ or $R = \mathrm{End}_{\mathbb{F}}(V)$.*

Remark 3.7.

(i) If G' is a subgroup of G and the G'-module V is very simple, then obviously the G-module V is also very simple.

(ii) The G-module V is very simple if and only if the corresponding $\rho(G)$-module V is very simple. This implies that if $H \twoheadrightarrow G$ is a surjective group homomorphism, then the G-module V is very simple if and only if the corresponding H-module V is very simple.

(iii) Let G' be a normal subgroup of G. If V is a very simple G-module, then either $\rho(G') \subset \mathrm{Aut}_k(V)$ consists of scalars (i.e., lies in $k \cdot \mathrm{Id}$) or the G'-module V is absolutely simple. See [21, Remark 5.2(iv)].

(iv) Suppose F is a discrete valuation field with valuation ring O_F, maximal ideal m_F and residue field $k = O_F/m_F$. Suppose V_F a finite-dimensional F-vector space, $\rho_F : G \to \mathrm{Aut}_F(V_F)$ an F-linear representation of G. Suppose T is a G-stable O_F-lattice in V_F and the corresponding $k[G]$-module $T/m_F T$ is isomorphic to V. Assume that the G-module V is very simple. Then the G-module V_F is also very simple. See [21, Remark 5.2(v)].

Theorem 3.8. *Suppose that X is an abelian variety defined over K and $i(\mathcal{O}) \subset \mathrm{End}_K(X)$. Let ℓ be a prime different from $\mathrm{char}(K)$. Suppose that λ is the only maximal ideal dividing ℓ in $\mathcal{O}$. Suppose that the natural representation in the $k(\lambda)$-vector space X_λ is very simple. Then $\mathrm{End}^0(X,i)$ enjoys one of the following two properties:*

1. *$\mathrm{End}^0(X,i) = i(E)$, i.e., $i(E) \cong E$ is a maximal commutative subalgebra in $\mathrm{End}^0(X)$ and $i(\mathcal{O}) \cong \mathcal{O}$ is a maximal commutative subring in $\mathrm{End}(X)$. In particular, $i(E)$ contains the center of $\mathrm{End}^0(X)$.*
2. *The following two conditions are fulfilled:*
   - *(2a) $\mathrm{End}^0(X,i)$ is a central simple E-algebra of dimension r_X^2 and X is an abelian variety of CM-type over K_a.*
   - *(2b) If $\mathrm{char}(K) = 0$, then $[E : \mathbb{Q}]$ is even and there exist an $\frac{[E:\mathbb{Q}]}{2}$-dimensional abelian variety Z, an isogeny $\psi : Z^r \to X$ and an embedding $k : E \hookrightarrow \mathrm{End}^0(Z)$ that sends 1 to 1_Z and such that $\psi \in \mathrm{Hom}^0((Z^r, k^{(r)}), (X,i))$. In addition, X is defined over a number field.*
     If $\mathrm{char}(K) > 0$, then X is isogenous to an abelian variety defined over a finite field.

Proof. In light of Remark 3.7(ii), the $\mathrm{Gal}(K)$-module X_λ is very simple. In light of Remark 3.7(iv) and Remark 3.5, $\rho_{\ell,X} : \mathrm{Gal}(K) \to \mathrm{Aut}_{E_\lambda}(V_\ell(X))$ is also very simple. Let us put $R = \mathrm{End}^0(X,i) \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$. It follows from Remark 3.5 that either $R = E_\lambda \mathrm{Id}$ or $R = \mathrm{End}_{E_\lambda}(V_\ell(X))$. Now the result follows readily from Remarks 3.4 and 3.2. □

Let Y be an abelian variety of positive dimension over K_a and u a non-zero endomorphism of Y. Let us consider the abelian (sub)variety $Z = u(Y) \subset Y$.

Remark 3.9. Suppose that Y is defined over K and $u \in \mathrm{End}_K(Y)$. Clearly, Z and the inclusion map $Z \subset Y$ are defined over $K_a^{\mathrm{Gal}(K)}$, i.e., Z and $Z \subset Y$ are defined over a purely inseparable extension of K. By a theorem of Chow [2, Th. 5 on p. 26], Z is defined over K. Clearly, the graph of $Z \subset Y$ is an abelian subvariety of $Z \times Y$ defined over a purely inseparable extension of K. By the same theorem of Chow, this graph is also defined over K and therefore $Z \subset Y$ is defined over K.

Theorem 3.10. *Let Y be an abelian variety of positive dimension over K_a and δ an automorphism of Y. Suppose that the induced K_a-linear operator $\delta^* : \Omega^1(Y) \to \Omega^1(Y)$ is diagonalizable. Let S be the set of eigenvalues of δ^* and $\mathrm{mult}_Y : S \to \mathbb{Z}_+$ the integer-valued function which assigns to each eigenvalue its multiplicity.*

Suppose that $P(t)$ is a polynomial with integer coefficients such that $u = P(\delta)$ is a non-zero endomorphism of Y. Let us put $Z = u(Y)$. Clearly, Z is δ-invariant and we write $\delta_Z : Z \to Z$ for the corresponding automorphism of Z (i.e., for the restriction of δ to Z). Suppose that

$$\dim(Z) = \sum_{\lambda \in S, P(\lambda) \neq 0} \mathrm{mult}_Y(\lambda).$$

Then the spectrum of $\delta_Z^* : \Omega^1(Z) \to \Omega^1(Z)$ *coincides with*

$$S_P = \{\lambda \in S, P(\lambda) \neq 0\}$$

and the multiplicity of an eigenvalue λ *of* δ_Z^* *equals* $mult_Y(\lambda)$.

Proof. Clearly, u commutes with δ. We write v for the (surjective) homomorphism $Y \twoheadrightarrow Z$ induced by u and j for the inclusion map $Z \subset Y$. Notice that $u : Y \to Y$ splits into a composition $Y \overset{v}{\twoheadrightarrow} Z \overset{j}{\hookrightarrow} Y$, i.e., $u = jv$. We have

$$\delta_Z v = v\delta \in \mathrm{Hom}(Y, Z),\ j\delta_Z = \delta j \in \mathrm{Hom}(Z, Y),\ u = jv \in \mathrm{End}(Y),$$

$$u\delta = \delta u \in \mathrm{End}(Y).$$

It is also clear that the induced map $u^* : \Omega^1(Y) \to \Omega^1(Y)$ coincides with $P(\delta^*)$. It follows that $u^*(\Omega^1(Y)) = P(\delta^*)(\Omega^1(Y))$ has dimension

$$\sum_{\lambda \in S, P(\lambda) \neq 0} \mathrm{mult}_Y(\lambda) = \dim(Y)$$

and coincides with $\oplus_{\lambda \in S, P(\lambda) \neq 0} W_\lambda$, where W_λ is the eigenspace of δ^* attached to eigenvalue λ. Since $u^* = v^* j^*$, we have

$$u^*(\Omega^1(Y)) = v^* j^*(\Omega^1(Y)) \subset v^*(\Omega^1(Z)).$$

Since $\dim(u^*(\Omega^1(Y))) = \dim(Y) = \dim(\Omega^1(Z)) \geq \dim(v^*(\Omega^1(Z)))$, the subspace $u^*(\Omega^1(Y)) = v^*(\Omega^1(Z))$ and $v^* : \Omega^1(Z) \hookrightarrow \Omega^1(Y)$. It follows that if we denote by w the isomorphism $v^* : \Omega^1(Z) \cong v^*(\Omega^1(Z))$ and by γ the restriction of δ^* to $v^*(\Omega^1(Z))$, then $\gamma w = w\delta_Y^*$ and therefore $\gamma = w\delta_Y^* w^{-1}$. □

4 Cyclic covers and jacobians

We fix a prime number p and an integral power $q = p^r$ and assume that K is a field of characteristic different from p. We fix an algebraic closure K_a, a primitive qth root of unity $\zeta \in K_a$ and write $\mathrm{Gal}(K)$ for the absolute Galois group $\mathrm{Aut}(K_a/K)$.

Let $f(x) \in K[x]$ be a separable polynomial of degree $n \geq 4$. We write $\mathfrak{R}_f$ for the set of its roots and denote by $L = L_f = K(\mathfrak{R}_f) \subset K_a$ the corresponding splitting field. As usual, the Galois group $\mathrm{Gal}(L/K)$ is called the Galois group of f and denoted by $\mathrm{Gal}(f)$. Clearly, $\mathrm{Gal}(f)$ permutes elements of $\mathfrak{R}_f$ and the natural map of $\mathrm{Gal}(f)$ into the group $\mathrm{Perm}(\mathfrak{R}_f)$ of all permutations of $\mathfrak{R}_f$ is an embedding. We will identify $\mathrm{Gal}(f)$ with its image and consider it as a permutation group of $\mathfrak{R}_f$. Clearly, $\mathrm{Gal}(f)$ is transitive if and only if f

is irreducible in $K[x]$. Further, we assume that either p does not divide n or q does divide n.

If p does not divide n, then we write (as in [20, §3])

$$V_{f,p} := (\mathbb{F}_p^{\mathfrak{R}_f})^{00} = (\mathbb{F}_p^{\mathfrak{R}_f})^0$$

for the $(n-1)$-dimensional $\mathbb{F}_p$-vector space of functions

$$\{\phi : \mathfrak{R}_f \to \mathbb{F}_p, \ \sum_{\alpha \in \mathfrak{R}_f} \phi(\alpha) = \ 0\}$$

provided with a natural action of the permutation group $\mathrm{Gal}(f) \subset \mathrm{Perm}(\mathfrak{R}_f)$. It is the *heart* over the field $\mathbb{F}_p$ of the group $\mathrm{Gal}(f)$ acting on the set $\mathfrak{R}_f$ [3], [20].

Remark 4.1. If p does not divide n and $\mathrm{Gal}(f) = \mathbf{S}_n$ or $\mathbf{A}_n$, then the $\mathrm{Gal}(f)$-module $V_{f,p}$ is very simple (see [20, Lemma 3.5]).

Let $C = C_{f,q}$ be the smooth projective model of the smooth affine K-curve $y^q = f(x)$. So C is a smooth projective curve defined over K. The rational function $x \in K(C)$ defines a finite cover $\pi : C \to \mathbf{P}^1$ of degree p. Let $B' \subset C(K_a)$ be the set of ramification points. Clearly, the restriction of π to B' is an *injective* map $B' \hookrightarrow \mathbf{P}^1(K_a)$, whose image is the disjoint union of ∞ and $\mathfrak{R}_f$ if p does *not* divide $\deg(f)$ and just $\mathfrak{R}_f$ if it does. We write

$$B = \pi^{-1}(\mathfrak{R}_f) = \{(\alpha, 0) \mid \alpha \in \mathfrak{R}_f\} \subset B' \subset C(K_a).$$

Clearly, π is ramified at each point of B with ramification index q. We have $B' = B$ if n is divisible by q. If n is not divisible by p, then B' is the disjoint union of B and a single point $\infty' := \pi^{-1}(\infty)$. In addition, the ramification index of π at $\pi^{-1}(\infty)$ is also q. Using Hurwitz's formula, one may easily compute the genus $g = g(C) = g(C_{f,q})$ of C [1, pp. 401–402], [13, Proposition 1 on p. 3359], [6, p. 148]. Namely, $g = (q-1)(n-1)/2$ if p does *not* divide n and $(q-1)(n-2)/2$ if q does divide n.

Remark 4.2. Assume that p does not divide n and consider the plane triangle (Newton polygon)

$$\Delta_{n,q} := \{(j,i) \mid 0 \le j, \quad 0 \le i, \quad qj + ni \le nq\}$$

with the vertices $(0,0)$, $(0,q)$ and $(n,0)$. Let $L_{n,q}$ be the set of integer points in the interior of $\Delta_{n,q}$. One may easily check that $g = (q-1)(n-1)/2$ coincides with the number of elements of $L_{n,q}$. It is also clear that for each $(j,i) \in L_{n,q}$,

$$1 \le j \le n-1; \quad 1 \le i \le q-1; \quad q(j-1) + (j+1) \le n(q-i).$$

Elementary calculations [1, Theorem 3 on p. 403] show that

$$\omega_{j,i} := x^{j-1}dx/y^{q-i} = x^{j-1}y^i dx/y^q = x^{j-1}y^{i-1}dx/y^{q-1}$$

is a differential of the first kind on C for each $(j,i) \in L_{n,q}$. This implies easily that the collection $\{\omega_{j,i}\}_{(j,i)\in L_{n,q}}$ is a basis in the space of differentials of the first kind on C.

There is a non-trivial birational K_a-automorphism of C

$$\delta_q : (x,y) \mapsto (x, \zeta y).$$

Clearly, δ_q^q is the identity map and the set of fixed points of δ_q coincides with B'.

Remark 4.3. Let us assume that $n = \deg(f)$ is divisible by q say, $n = qm$ for some positive integer m. Let $\alpha \in K_a$ be a root of f and $K_1 = K(\alpha)$ be the corresponding subfield of K_a. We have $f(x) = (x-\alpha)f_1(x)$ with $f_1(x) \in K_1[x]$ a separable polynomial over K_1 of degree $qm-1 = n-1 \geq 4$. It is also clear that the polynomials $h(x) = f_1(x+\alpha), h_1(x) = x^{n-1}h(1/x) \in K_1[x]$ are separable of the same degree $qm-1 = n-1 \geq 4$. The standard substitution $x_1 = 1/(x-\alpha), y_1 = y/(x-\alpha)^m$ establishes a birational isomorphism between $C_{f,p}$ and a curve $C_{h_1} : y_1^q = h_1(x_1)$ (see [13, p. 3359]). In particular, the jacobians of C_f and C_{h_1} are isomorphic over K_a (and even over K_1). But $\deg(h_1) = qm-1$ is *not* divisible by p. Clearly, this isomorphism commutes with the actions of δ_q. Notice also that if the Galois group of f over K is $\mathbf{S}_n$ (resp. $\mathbf{A}_n$), then the Galois group of h_1 over K_1 is $\mathbf{S}_{n-1}$ (resp. $\mathbf{A}_{n-1}$).

Remark 4.4.

(i) It is well known that $\dim_{K_a}(\Omega^1(C_{(f,q)})) = g(C_{f,q})$. By functoriality, δ_q induces on $\Omega^1(C_{(f,q)})$ a certain K_a-linear automorphism

$$\delta_q^* : \Omega^1(C_{(f,q)}) \to \Omega^1(C_{(f,q)}).$$

Clearly, if for some positive integer j the differential $\omega_{j,i} = x^{j-1}dx/y^{q-i}$ lies in $\Omega^1(C_{(f,q)})$, then it is an eigenvector of δ_q^* with eigenvalue ζ^i.

(ii) Now assume that p does *not* divide n. It follows from Remark 4.2 that the collection $\{\omega_{j,i} = x^{j-1}dx/y^{q-i} \mid (i,j) \in L_{n,q}\}$ is an eigenbasis of $\Omega^1(C_{(f,q)})$. This implies that the multiplicity of the eigenvalue ζ^{-i} of δ_q^* coincides with the number of interior integer points in $\Delta_{n,q}$ along the corresponding (to $q-i$) horizontal line. Elementary calculations show that this number is $\left[\frac{ni}{q}\right]$; in particular, ζ^{-i} *is an eigenvalue if and only if* $\left[\frac{ni}{q}\right] > 0$. Taking into account that $n \geq 4$ and $q = p^r$, we conclude that ζ^i is an eigenvalue of δ_q^* for each integer i with $p^r - p^{r-1} \leq i \leq p^r - 1 = q-1$. It also follows easily that 1 is *not* an eigenvalue δ_q^*. This implies that

$$\mathcal{P}_q(\delta_q^*) = \delta_q^{*q-1} + \cdots + \delta_q^* + 1 = 0$$

in $\mathrm{End}_K(\Omega^1(C_{(f,q)}))$. In addition, one may check that if $\mathcal{H}(t)$ is a polynomial with rational coefficients such that $\mathcal{H}(\delta_q^*) = 0$ in $\mathrm{End}_K(\Omega^1(C_{(f,q)}))$, then $\mathcal{H}(t)$ is *divisible* by $\mathcal{P}_q(t)$ in $\mathbb{Q}[t]$.

Let $J(C) = J(C_{f,q})$ be the jacobian of C. It is a g-dimensional abelian variety defined over K and one may view (via Albanese functoriality) δ_q as an element of $\mathrm{Aut}(C) \subset \mathrm{Aut}(J(C)) \subset \mathrm{End}(J(C))$ such that $\delta_q \neq \mathrm{Id}$ but $\delta_q^q = \mathrm{Id}$, where Id is the identity endomorphism of $J(C)$. We write $\mathbb{Z}[\delta_q]$ for the subring of $\mathrm{End}(J(C))$ generated by δ_q.

Remark 4.5. Assume that p does not divide n. Let P_0 be one of the δ_q-invariant points (i.e., a ramification point for π) of $C_{f,p}(K_a)$. Then

$$\tau : C_{f,q} \to J(C_{f,q}), \quad P \mapsto \mathrm{cl}((P) - (P_0))$$

is an embedding of complex algebraic varieties and it is well known that the induced map $\tau^* : \Omega^1(J(C_{f,q})) \to \Omega^1(C_{f,q})$ is an isomorphism obviously commuting with the actions of δ_q. (Here cl stands for the linear equivalence class.) This implies that n_{σ_i} coincides with the dimension of the eigenspace of $\Omega^1(C_{(f,q)})$ attached to the eigenvalue ζ^{-i} of δ_q^*. Applying Remark 4.4, we conclude that if $\mathcal{H}(t)$ is a monic polynomial with integer coefficients such that $\mathcal{H}(\delta_q) = 0$ in $\mathrm{End}(J^{(f,q)})$, then $\mathcal{H}(t)$ is divisible by $\mathcal{P}_q(t)$ in $\mathbb{Q}[t]$ and therefore in $\mathbb{Z}[t]$.

Remark 4.6. Assume that p does not divide n. Clearly, the set S of eigenvalues λ of $\delta_q^* : \Omega^1(J(C_{f,q})) \to \Omega^1(J(C_{f,q}))$ with $\mathcal{P}_{q/p}(\lambda) \neq 0$ consists of *primitive* qth roots of unity ζ^{-i} $(1 \leq i < q, (i,p) = 1)$ with $\left[\frac{ni}{q}\right] > 0$ and the multiplicity of ζ^{-i} equals $\left[\frac{ni}{q}\right]$, thanks to Remarks 4.5 and 4.4. Let us compute the sum

$$M = \sum_{1 \leq i < q, (i,p)=1} \left[\frac{ni}{q}\right]$$

of multiplicities of eigenvalues from S.

First, assume that $q > 2$. Then $\varphi(q) = (p-1)p^{r-1}$ is even and for each (index) i the difference $q - i$ is also prime to p, lies between 1 and q and

$$\left[\frac{ni}{q}\right] + \left[\frac{n(q-i)}{q}\right] = n - 1.$$

It follows that

$$M = (n-1)\frac{\varphi(q)}{2} = \frac{(n-1)(p-1)p^{r-1}}{2}.$$

Assume that $q = p = 2$ and therefore $r = 1$. Then n is odd, $C_{f,q} = C_{f,2} : y^2 = f(x)$ is a hyperelliptic curve of genus $g = \frac{n-1}{2}$ and δ_2 is the hyperelliptic involution $(x, y) \mapsto (x, -y)$. It is well known that the differentials

$x^i \frac{dx}{y}$ $(0 \le i \le g-1)$ constitute a basis of the g-dimensional $\Omega^1(J(C_{f,2}))$. It follows that δ_2^* is just multiplication by -1. Therefore

$$M = g = \frac{n-1}{2} = \frac{(n-1)(p-1)p^{r-1}}{2}.$$

Clearly, if the abelian (sub)variety $Z := \mathcal{P}_{q/p}(\delta_q)(J(C_{f,q}))$ has dimension M, then the data $Y = J(C_{f,q}), \delta = \delta_q, P = \mathcal{P}_{q/p}(t)$ satisfy the conditions of Theorem 3.10.

Lemma 4.7. *Assume that p does not divide n. Let $D = \sum_{P \in B} a_P(P)$ be a divisor on $C = C_{f,p}$ with degree 0 and support in B. Then D is principal if and only if all the coefficients a_P are divisible by q.*

Proof. Suppose $D = \mathrm{div}(h)$, where $h \in K_a(C)$ is a non-zero rational function of C. Since D is δ_q-invariant, the rational function $\delta_q^* h := h\delta_q$ coincides with $c \cdot h$ for some non-zero $c \in K_a$. It follows easily from the δ_q-invariance of the splitting $K_a(C) = \oplus_{i=0}^{q-1} y^i \cdot K_a(x)$ that $h = y^i \cdot u(x)$ for some non-zero rational function $u(x) \in K_a(x)$ and a non-negative integer $i \le q-1$. It follows that all finite zeros and poles of $u(x)$ lie in B, i.e., there exists an integer-valued function b on $\mathfrak{R}_f$ such that u coincides, up to multiplication by a non-zero constant, to $\prod_{\alpha \in \mathfrak{R}_f} (x-\alpha)^{b(\alpha)}$. Notice that $\mathrm{div}(y) = \sum_{P \in B}(P) - n(\infty)$. On the other hand, for each $\alpha \in \mathfrak{R}_f$, we have $P_\alpha = (\alpha, 0) \in B$ and the corresponding divisor $\mathrm{div}(x-\alpha) = q((\alpha,0)) - q(\infty) = q(P_\alpha) - q(\infty)$ is divisible by q. This implies that $a_{P_\alpha} = q \cdot b(\alpha) + i$. Also, since ∞ is neither a zero nor a pole of h, we get the equality $0 = ni + \sum_{\alpha \in \mathfrak{R}_f} b(\alpha) q$. Since n and q are relatively prime, i must divide q. This implies that $i = 0$ and therefore the divisor $D = \mathrm{div}(u(x)) = \mathrm{div}(\prod_{\alpha \in \mathfrak{R}_f} (x-\alpha)^{b(\alpha)})$ is divisible by q.

Conversely, suppose a divisor $D = \sum_{P \in B} a_P(P)$ with $\sum_{P \in B} a_P = 0$ and all a_P are divisible by q. Let us put $h = \prod_{P \in B} (x - x(P))^{a_P/q}$. One may easily check that $D = \mathrm{div}(h)$. □

Lemma 4.8. *One has $1 + \delta_q + \cdots + \delta_q^{q-1} = 0$ in $\mathrm{End}(J(C_{f,q}))$. The subring $\mathbb{Z}[\delta_q] \subset \mathrm{End}(J(C_{f,q}))$ is isomorphic to the ring $\mathbb{Z}[t]/\mathcal{P}_q(t)\mathbb{Z}[t]$. The $\mathbb{Q}$-subalgebra $\mathbb{Q}[\delta_q] \subset \mathrm{End}^0(J(C_{f,q})) = \mathrm{End}^0(J(C_{f,q}))$ is isomorphic to $\mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t] = \prod_{i=1}^r \mathbb{Q}(\zeta_{p^i})$.*

Proof. If $q = p$ is a prime this assertion is proven in [6, p. 149], [8, p. 458]. So, further we may assume that $q > p$. It follows from Remark 4.3 that we may assume that p does *not* divide n.

Now we follow arguments of [8, p. 458] (where the case of $q = p$ was treated). The group $J(C_{f,q})(K_a)$ is generated by divisor classes of the form $(P) - (\infty)$ where P is a finite point on $C_{f,p}$. The divisor of the rational function $x - x(P)$ is $(\delta_q^{q-1} P) + \cdots + (\delta_q P) + (P) - q(\infty)$. This implies that $\mathcal{P}_q(\delta_q) = 0 \in \mathrm{End}(J(C_{f,q}))$. Applying Remark 4.5(ii), we conclude that $\mathcal{P}_q(t)$ is the minimal polynomial of δ_q in $\mathrm{End}(J(C_{f,q}))$. □

Let us define the abelian (sub)variety

$$J^{(f,q)} := \mathcal{P}_{q/p}(\delta_q)(J(C_{f,q})) \subset J(C_{f,q}).$$

Clearly, $J^{(f,q)}$ is a δ_q-invariant abelian subvariety defined over $K(\zeta_q)$. In addition, $\Phi_q(\delta_q)(J^{(f,q)}) = 0$.

Remark 4.9. If $q = p$, then $\mathcal{P}_{q/p}(t) = \mathcal{P}_1(t) = 1$ and therefore $J^{(f,p)} = J(C_{f,p})$.

Remark 4.10. Since the polynomials Φ_q and $\mathcal{P}_{q/p}$ are relatively prime, the homomorphism $\mathcal{P}_{q/p}(\delta_q) : J^{(f,q)} \to J^{(f,q)}$ has finite kernel and therefore is an isogeny. In particular, it is surjective.

Lemma 4.11. *Suppose that p does not divide n. Then*

$$\dim(J^{(f,q)}) = \frac{(p^r - p^{r-1})(n-1)}{2}$$

and there is a $K(\zeta)$-isogeny $J(C_{f,q}) \to J(C_{f,q/p}) \times J^{(f,q)}$. In addition, if $\zeta \in K$, then the Galois modules $V_{f,p}$ and $(J^{(f,q)})^{\delta_q} := \{z \in J^{(f,q)}(K_a) \mid \delta_q(z) = z\}$ are isomorphic.

Proof. We may assume that $\zeta \in K$. Consider the curve $C_{f,q/p} : y_1^{q/p} = f(x_1)$ and a regular surjective map $\pi_1 : C_{f,q} \to C_{f,q/p}$, $x_1 = x, y_1 = y^p$. Clearly, $\pi_1\delta_q = \delta_{q/p}\pi_1$. By Albanese functoriality, π_1 induces a certain surjective homomorphism of jacobians $J(C_{f,q}) \twoheadrightarrow J(C_{f,q/p})$ which we continue to denote by π_1. Clearly, the equality $\pi_1\delta_q = \delta_{q/p}\pi_1$ remains true in $\mathrm{Hom}(J(C_{f,q}), J(C_{f,q/p}))$. By Lemma 4.8, $\mathcal{P}_{q/p}(\delta_{q/p}) = 0 \in \mathrm{End}(J(C_{f,q/p}))$. It follows from Remark 4.10 that $\pi_1(J^{(f,q)}) = 0$ and therefore $\dim(J^{(f,q)})$ does not exceed

$$\begin{aligned}\dim(J(C_{f,q})) - \dim(J(C_{f,q/p})) &= \frac{(p^r-1)(n-1)}{2} - \frac{(p^{r-1}-1)(n-1)}{2} \\ &= \frac{(p^r - p^{r-1})(n-1)}{2}.\end{aligned}$$

By definition of $J^{(f,q)}$, for each divisor $D = \sum_{P\in B} a_P(P)$ the linear equivalence class of $p^{r-1}D = \sum_{P\in B} p^{r-1}a_P(P)$ lies in $(J^{(f,q)})^{\delta_q} \subset J^{(f,q)}(K_a) \subset J(C_{f,q})(K_a)$. It follows from Lemma 4.7 that the class of $p^{r-1}D$ is zero if and only if all $p^{r-1}a_P$ are divisible by $q = p^r$, i.e., all a_P are divisible by p. This implies that the set of linear equivalence classes of $p^{r-1}D$ is a Galois submodule isomorphic to $V_{f,p}$. We want to prove that $(J^{(f,q)})^{\delta_q} = V_{f,p}$.

Recall that $J^{(f,q)}$ is δ_q-invariant and the restriction of δ_q to $J^{(f,q)}$ satisfies the qth cyclotomic polynomial. This allows us to define the homomorphism $\mathbb{Z}[\zeta_q] \to \mathrm{End}(J^{(f,q)})$ that sends 1 to the identity map and ζ_q to δ_q. Let us put $E = \mathbb{Q}(\zeta_q), \mathcal{O} = \mathbb{Z}[\zeta_q] \subset \mathbb{Q}(\zeta_q) = E$. It is well known that $\mathcal{O}$ is the ring of

integers in E, the ideal $\lambda = (1-\zeta_q)\mathbb{Z}[\zeta_q] = (1-\zeta_q)\mathcal{O}$ is maximal in $\mathcal{O}$ with $\mathcal{O}/\lambda = \mathbb{F}_p$ and $\mathcal{O} \otimes \mathbb{Z}_p = \mathbb{Z}_p[\zeta_q]$ is the ring of integers in the field $\mathbb{Q}_p(\zeta_q)$. Notice also that $\mathcal{O} \otimes \mathbb{Z}_p$ coincides with the completion $\mathcal{O}_\lambda$ of $\mathcal{O}$ with respect to the λ-adic topology and $\mathcal{O}_\lambda/\lambda\mathcal{O}_\lambda = \mathcal{O}/\lambda = \mathbb{F}_p$.

It follows from Lemma 3.3 that

$$d = \frac{2\dim(J^{(f,q)})}{[E:\mathbb{Q}]} = \frac{2\dim(J^{(f,q)})}{p^r - p^{r-1}}$$

is a positive integer, the $\mathbb{Z}_p$-Tate module $T_p(J^{(f,q)})$ is a free $\mathcal{O}_\lambda$-module of rank d. Using the displayed formula (5) from Section 3, we conclude that

$$(J^{(f,q)})^{\delta_q} = \{u \in J^{(f,q)}(K_a) \mid (1-\delta_q)(u) = 0\} = J^{f,q}_\lambda = T_p(J^{f,q}) \otimes_{\mathcal{O}_\lambda} \mathbb{F}_p$$

is a d-dimensional $\mathbb{F}_p$-vector space. Since $(J^{(f,q)})^{\delta_q}$ contains $(n-1)$-dimensional $\mathbb{F}_p$-vector space $V_{f,p}$, we have $d \ge n-1$. This implies that

$$2\dim(J^{(f,q)}) = d(p^r - p^{r-1}) \ge (n-1)(p^r - p^{r-1})$$

and therefore

$$\dim(J^{(f,q)}) \ge \frac{(n-1)(p^r - p^{r-1})}{2}.$$

But we have already seen that

$$\dim(J^{(f,q)}) \le \frac{(n-1)(p^r - p^{r-1})}{2}.$$

This implies that

$$\dim(J^{(f,q)}) = \frac{(n-1)(p^r - p^{r-1})}{2}.$$

It follows that $d = n-1$ and therefore $(J^{(f,q)})^{\delta_q} = V_{f,p}$. Dimension arguments imply that $J^{(f,q)}$ coincides with the identity component of $\ker(\pi_1)$ and therefore there is an isogeny between $J(C_{f,q})$ and $J(C_{f,q/p}) \times J^{(f,q)}$. □

Corollary 4.12. *If p does not divide n, then there is a $K(\zeta_q)$-isogeny*

$$J(C_{f,q}) \to J(C_{f,p}) \times \prod_{i=2}^{r} J^{(f,p^i)} = \prod_{i=1}^{r} J^{(f,p^i)}.$$

Proof. Combine Corollary 4.11(ii) and Remark 4.9 with easy induction on r. □

Remark 4.13. Suppose that p does not divide n and consider the induced linear operator $\delta_q^* : \Omega^1(J^{(f,q)}) \to \Omega^1(J^{(f,q)})$. It follows from Theorem 3.10 combined with Remark 4.6 that its spectrum consists of primitive qth roots of unity ζ^{-i} $(1 \le i < q)$ with $[ni/q] > 0$ and the multiplicity of ζ^{-i} equals $[ni/q]$.

Theorem 4.14. *Suppose that $n \geq 5$ is an integer. Let p be a prime, $r \geq 1$ an integer and $q = p^r$. Suppose that p does not divide n. Suppose that K is a field of characteristic different from p containing a primitive qth root of unity ζ. Let $f(x) \in K[x]$ be a separable polynomial of degree n and $\mathrm{Gal}(f)$ its Galois group. Suppose that the $\mathrm{Gal}(f)$-module $V_{f,p}$ is very simple. Then the image $\mathcal{O}$ of $\mathbb{Z}[\delta_q] \to \mathrm{End}(J^{(f,q)})$ is isomorphic to $\mathbb{Z}[\zeta_q]$ and enjoys one of the following two properties:*

(i) $\mathcal{O}$ is a maximal commutative subring in $\mathrm{End}(J^{(f,q)})$;
(ii) $\mathrm{char}(K) > 0$ and the centralizer of $\mathcal{O} \otimes \mathbb{Q} \cong \mathbb{Q}(\zeta_q)$ in $\mathrm{End}^0(J^{(f,q)})$ is a central simple $(n-1)^2$-dimensional $\mathbb{Q}(\zeta_q)$-algebra. In addition, $J^{(f,q)}$ is an abelian variety of CM-type isogenous to a self-product of an absolutely simple abelian variety. Also $J^{(f,q)}$ is isogenous to an abelian variety defined over a finite field.

Proof. Clearly, $\mathcal{O}$ is isomorphic to $\mathbb{Z}[\zeta_q]$. Let us put $\lambda = (1-\zeta_q)\mathbb{Z}[\zeta_q]$. By Lemma 4.11(iii), the Galois module $(J^{(f,q)})^{\delta_q} = J^{(f,q)}_{\lambda}$ is isomorphic to $V_{f,p}$. Applying Theorem 3.8, we conclude that either (ii) holds true or one of the following conditions hold:

(a) $\mathcal{O}$ is a maximal commutative subring in $\mathrm{End}(J^{(f,q)})$;
(b) $\mathrm{char}(K) = 0$ and there exist a $\varphi(q)/2$-dimensional abelian variety Z over K_a, an embedding $\mathbb{Q}(\zeta_q) \hookrightarrow \mathrm{End}^0(Z)$ that sends 1 to 1_Z and a $\mathbb{Q}(\zeta_q)$-equivariant isogeny $\psi : Z^{n-1} \to J^{(f,q)}$.

Clearly, if (a) is fulfilled, then we are done. Also if $q = 2$, then $\varphi(q)/2 = 1/2$ is not an integer and therefore (b) is not fulfilled, i.e., (a) is fulfilled.

So further we assume that $q > 2$ and (b) holds true. In particular, $\mathrm{char}(K) = 0$. We need to arrive at a contradiction.

Since $\mathrm{char}(K) = 0$, the isogeny ψ induces an isomorphism $\psi^* : \Omega^1((J^{(f,q)})) \cong \Omega^1(Z^{n-1})$ that commutes with the actions of $\mathbb{Q}(\zeta_q)$. Since

$$\dim(\Omega^1(Z)) = \dim(Z) = \frac{\varphi(q)}{2},$$

the linear operator in $\Omega^1(Z)$ induced by $\zeta_q \in \mathbb{Q}(\zeta_q)$ has, at most, $\varphi(q)/2$ distinct eigenvalues. It follows that the linear operator in $\Omega^1(Z^{n-1}) = \Omega^1(Z)^{n-1}$ induced by ζ_q also has, at most, $\varphi(q)/2$ distinct eigenvalues. This implies that the linear operator δ_q^* in $\Omega^1((J^{(f,q)}))$ also has, at most, $\varphi(q)/2$ distinct eigenvalues. Recall that the eigenvalues of δ_q^* are primitive qth roots of unity ζ^{-i} with

$$1 \leq i < q, (i,p) = 1, \left[\frac{ni}{q}\right] > 0.$$

Clearly, the inequality $[ni/q] > 0$ means that $i > q/n$, since $(n,q) = (n,p^r) = 1$. So, in order to get a desired contradiction, it suffices to check that the cardinality of the set

$$B := \left\{ i \in \mathbb{Z} \mid \frac{q}{n} < i < q = p^r, (i,p) = 1 \right\}$$

is strictly greater than $(p-1)p^{r-1}/2$. Since $p \ge 2, n \ge 5$ and q/n is not an integer, we have

$$\frac{p}{n} \le \frac{p}{5} < \frac{p-1}{2}$$

and

$$\#(B) > \varphi(q) - \frac{q}{n} = (p-1)p^{r-1} - \frac{p^{r-1}p}{n} \ge \left(p - 1 - \frac{p}{5}\right) p^{r-1} > \frac{p-1}{2} p^{r-1}.$$

□

Corollary 4.15. *Suppose that $n \ge 5$ is an integer. Let p be a prime, $r \ge 1$ an integer and $q = p^r$. Assume in addition that either p does not divide n or $q \mid n$ and $(n,q) \ne (5,5)$. Let K be a field of characteristic different from p. Let $f(x) \in K[x]$ be an irreducible separable polynomial of degree n such that $\mathrm{Gal}(f) = \mathbf{S}_n$ or $\mathbf{A}_n$. Then the image $\mathcal{O}$ of $\mathbb{Z}[\delta_q] \to \mathrm{End}(J^{(f,q)})$ is isomorphic to $\mathbb{Z}[\zeta_q]$ and enjoys one of the following two properties:*

(i) $\mathcal{O}$ is a maximal commutative subring in $\mathrm{End}(J^{(f,q)})$;

(ii) $\mathrm{char}(K) > 0$ and the centralizer of $\mathcal{O} \otimes \mathbb{Q} \cong \mathbb{Q}(\zeta_q)$ in $\mathrm{End}^0(J^{(f,q)})$ is a central simple $(n-1)^2$-dimensional $\mathbb{Q}(\zeta_q)$-algebra. In addition, $J^{(f,q)}$ is an abelian variety of CM-type isogenous to a self-product of an absolutely simple abelian variety.

Proof. If p divides n, then $n > 5$ and therefore $n - 1 \ge 5$. By Remark 4.3, we may assume that p does not divide n. If we replace K by $K(\zeta)$, then still $\mathrm{Gal}(f) = \mathbf{S}_n$ or $\mathbf{A}_n$. By Remark 4.1, if $\mathrm{Gal}(f) = \mathbf{S}_n$ or $\mathbf{A}_n$, then the $\mathrm{Gal}(f)$-module $V_{f,p}$ is very simple. One has only to apply Theorem 4.14. □

Theorem 4.16. *Suppose $n \ge 4$ and p does not divide n. Assume also that $\mathrm{char}(K) = 0$ and $\mathbb{Q}[\delta_q]$ is a maximal commutative subalgebra in $\mathrm{End}^0(J^{(f,q)})$. Then $\mathrm{End}^0(J^{(f,q)}) = \mathbb{Q}[\delta_q] \cong \mathbb{Q}(\zeta_q)$ and therefore $\mathrm{End}(J^{(f,q)}) = \mathbb{Z}[\delta_q] \cong \mathbb{Z}[\zeta_q]$. In particular, $J^{(f,q)}$ is an absolutely simple abelian variety.*

Proof. Let $\mathfrak{C} = \mathfrak{C}_{J^{(f,p)}}$ be the center of $\mathrm{End}^0(J^{(f,p)})$. Since $\mathbb{Q}[\delta_q]$ is a maximal commutative subalgebra, $\mathfrak{C} \subset \mathbb{Q}[\delta_q]$.

Replacing, if necessary, K by its subfield (finitely) generated over $\mathbb{Q}$ by all the coefficients of f, we may assume that K (and therefore K_a) is isomorphic to a subfield of $\mathbb{C}$. So, $K \subset K_a \subset \mathbb{C}$. We may also assume that $\zeta = \zeta_q$ and consider $J^{(f,q)}$ as a complex abelian variety. Let $\Sigma = \Sigma_E$ be the set of all field embeddings $\sigma : E = \mathbb{Q}[\delta_q] \hookrightarrow \mathbb{C}$. We are going to apply Corollary 2.2 to $Z = J^{(f,q)}$ and $E = \mathbb{Q}[\delta_q]$. In order to do that we need to get some information about the multiplicities $n_\sigma = n_\sigma(Z,E) = n_\sigma(J^{(f,q)}, \mathbb{Q}[\delta_q])$. The displayed formula (1) in Section 2 allows us to do it, using the action of

$\mathbb{Q}[\delta_q]$ on $\Omega^1(J^{(f,q)})$. Namely, since δ_q generates the field E (over $\mathbb{Q}$), each $\Omega^1(J^{(f,q)})_\sigma$ is the eigenspace corresponding to the eigenvalue $\sigma(\delta_q)$ of δ_q and n_σ is the multiplicity of the eigenvalue $\sigma(\delta_q)$.

Let $i < q$ be a positive integer that is not divisible by p and $\sigma_i : \mathbb{Q}[\delta_q] \hookrightarrow \mathbb{C}$ be the embedding which sends δ_q to ζ^{-i}. Clearly, for each σ there exists precisely one i such that $\sigma = \sigma_i$. Clearly, $\Omega^1(J^{(f,q)})_{\sigma_i}$ is the eigenspace of $\Omega^1(J^{(f,q)})$ attached to the eigenvalue ζ^{-i} of δ_q. Therefore n_{σ_i} coincides with the multiplicity of the eigenvalue ζ^{-i}. It follows from Remark 4.13 that

$$n_{\sigma_i} = \left[\frac{ni}{q}\right].$$

The theorem follows from Corollary 2.2 applied to $E = \mathbb{Q}[\delta_q] \cong \mathbb{Q}(\zeta_q)$. □

Theorem 4.17. *Let p be a prime, r a positive integer, $q = p^r$ and K a field of characteristic zero. Suppose that $f(x) \in K[x]$ is an irreducible polynomial of degree $n \geq 5$ and $\mathrm{Gal}(f) = \mathbf{S}_n$ or $\mathbf{A}_n$. Assume also that either p does not divide n or q divides n. Then $\mathrm{End}^0(J^{(f,q)}) = \mathbb{Q}[\delta_q] \cong \mathbb{Q}(\zeta_q)$ and therefore $\mathrm{End}(J^{(f,q)}) = \mathbb{Z}[\delta_q] \cong \mathbb{Z}[\zeta_q]$. In particular, $J^{(f,q)}$ is an absolutely simple abelian variety.*

Proof. If $(n, q) \neq (5, 5)$, then the assertion follows from Corollary 4.15 combined with Theorem 4.16. The case $(n, q) = (5, 5)$ is contained in [22, Theorem 4.2]. □

Corollary 4.18. *Let p be a prime and K a field of characteristic zero. Suppose that $f(x) \in K[x]$ is an irreducible polynomial of degree $n \geq 5$ and $\mathrm{Gal}(f) = \mathbf{S}_n$ or $\mathbf{A}_n$. Let r and s be distinct positive integers. Assume also that either p does not divide n or both p^r and p^s divide n. Then $\mathrm{Hom}(J^{(f,p^r)}, J^{(f,p^s)}) = 0$.*

Proof. It follows from Theorem 4.17 that $J^{(f,p^r)}$ and $J^{(f,p^s)}$ are absolutely simple abelian varieties, whose endomorphism algebras $\mathbb{Q}(\zeta_{p^r})$ and $\mathbb{Q}(\zeta_{p^s})$ are not isomorphic. Therefore these abelian varieties are not isogenous. Since they are absolutely simple, every homomorphism between them is zero. □

Combining Theorem 4.16 and Theorem 4.14, we obtain the following statement.

Theorem 4.19. *Let p be a prime, r a positive integer, $q = p^r$. Suppose that K is a field of characteristic zero containing a primitive qth root of unity. Let $f(x) \in K[x]$ be a polynomial of degree $n \geq 5$. Assume also that p does not divide n and the $\mathrm{Gal}(f)$-module $V_{f,p}$ is very simple. Then $\mathrm{End}^0(J^{(f,q)}) = \mathbb{Q}[\delta_q] \cong \mathbb{Q}(\zeta_q)$ and therefore $\mathrm{End}(J^{(f,q)}) = \mathbb{Z}[\delta_q] \cong \mathbb{Z}[\zeta_q]$. In particular, $J^{(f,q)}$ is an absolutely simple abelian variety.*

Corollary 4.20. *Let p be a prime, and K a field of characteristic zero. Let $f(x) \in K[x]$ be a polynomial of degree $n \ge 5$. Assume also that p does not divide n and the $\mathrm{Gal}(f)$-module $V_{f,p}$ is very simple. If r and s are distinct positive integers such that K contains primitive p^rth and p^sth roots of unity, then $\mathrm{Hom}(J^{(f,p^r)}, J^{(f,p^s)}) = 0$.*

Proof. It follows from Theorem 4.19 that $J^{(f,p^r)}$ and $J^{(f,p^s)}$ are absolutely simple abelian varieties, whose endomorphism algebras $\mathbb{Q}(\zeta_{p^r})$ and $\mathbb{Q}(\zeta_{p^s})$ are not isomorphic. Therefore these abelian varieties are not isogenous. Since they are absolutely simple, every homomorphism between them is zero. □

5 Jacobians and their endomorphism rings

Throughout this section we assume that K is a field of characteristic zero. Recall that K_a is an algebraic closure of K and $\zeta \in K_a$ is a primitive qth root of unity. Suppose $f(x) \in K[x]$ is a polynomial of degree $n \ge 5$ without multiple roots, $\mathfrak{R}_f \subset K_a$ is the set of its roots, $K(\mathfrak{R}_f)$ is its splitting field. Let us put $\mathrm{Gal}(f) = \mathrm{Gal}(K(\mathfrak{R}_f)/K) \subset \mathrm{Perm}(\mathfrak{R}_f)$. Let r be a positive integer. Recall (Corollary 4.12) that if p does not divide n, then there is a $K(\zeta_{p^r})$-isogeny $J(C_{f,p^r}) \to \prod_{i=1}^{r} J^{(f,p^i)}$. Applying Theorem 4.19 and Corollary 4.20 to all $q = p^i$, we obtain the following assertion.

Theorem 5.1. *Let p be a prime, r a positive integer, $q = p^r$. Suppose that K is a field of characteristic zero containing a primitive p^rth root of unity. Let $f(x) \in K[x]$ be a polynomial of degree $n \ge 5$. Assume also that p does not divide n and the $\mathrm{Gal}(f)$-module $V_{f,p}$ is very simple. Then*

$$\mathrm{End}^0(J(C_{f,q})) = \mathbb{Q}[\delta_q] \cong \mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t] = \prod_{i=1}^{r} \mathbb{Q}(\zeta_{p^i}).$$

The next statement obviously generalizes Theorem 1.1.

Theorem 5.2. *Let p be a prime, r a positive integer and K a field of characteristic zero. Suppose that $f(x) \in K[x]$ is an irreducible polynomial of degree $n \ge 5$ and $\mathrm{Gal}(f) = \mathbf{S}_n$ or $\mathbf{A}_n$. Assume also that either p does not divide n or $q \mid n$. Then $\mathrm{End}^0(J(C_{f,q})) = \mathbb{Q}[\delta_q] \cong \mathbb{Q}[t]/\mathcal{P}_q(t)\mathbb{Q}[t] = \prod_{i=1}^{r} \mathbb{Q}(\zeta_{p^i})$.*

Proof. The existence of the isogeny $J(C_{f,q}) \to \prod_{i=1}^{r} J^{(f,p^i)}$ combined with Theorem 4.17 and Corollary 4.18 implies that the assertion holds if p does not divide n. If q divides n, then Remark 4.3 allows us to reduce this case to the already proven case when p does not divide $n-1$. □

Example 5.3. Suppose $L = \mathbb{C}(z_1, \cdots, z_n)$ is the field of rational functions in n independent variables $z_1, \cdots, z_n$ with constant field $\mathbb{C}$ and $K = L^{\mathbf{S}_n}$ is the subfield of symmetric functions. Then $K_a = L_a$ and $f(x) = \prod_{i=1}^{n}(x - z_i) \in K[x]$ is an irreducible polynomial over K with Galois group $\mathbf{S}_n$. Let Let $q = p^r$ be a power of a prime p. Let C be a smooth projective model of the K-curve $y^q = f(x)$ and $J(C)$ its jacobian. It follows from Theorem 5.2 that if $n \geq 5$ and either p does not divide n or q divides n, then the algebra of L_a-endomorphisms of $J(C)$ is $\prod_{i=1}^{r} \mathbb{Q}(\zeta_{p^i})$.

Example 5.4. Let $h(x) \in \mathbb{C}[x]$ be a *Morse polynomial* of degree $n \geq 5$. This means that the derivative $h'(x)$ of $h(x)$ has $n-1$ distinct roots $\beta_1, \cdots, \beta_{n-1}$ and $h(\beta_i) \neq h(\beta_j)$ while $i \neq j$. (For example, $x^n - x$ is a Morse polynomial.) If $K = \mathbb{C}(z)$, then a theorem of Hilbert ([10, Theorem 4.4.5, p. 41]) asserts that the Galois group of $h(x) - z$ over K is $\mathbf{S}_n$. Let $q = p^r$ be a power of a prime p. Let C be a smooth projective model of the K-curve $y^q = h(x) - z$ and $J(C)$ its jacobian. It follows from Theorem 5.2 that if either p does not divide n or q divides n, then the algebra of K_a-endomorphisms of $J(C)$ is $\prod_{i=1}^{r} \mathbb{Q}(\zeta_{p^i})$.

References

1. J. K. Koo – "On holomorphic differentials of some algebraic function field of one variable over **C**", *Bull. Austral. Math. Soc.* **43** (1991), no. 3, p. 399–405.
2. S. Lang – *Abelian varieties*, Springer-Verlag, New York, 1983, Reprint of the 1959 original.
3. B. Mortimer – "The modular permutation representations of the known doubly transitive groups", *Proc. London Math. Soc. (3)* **41** (1980), no. 1, p. 1–20.
4. D. Mumford – *Abelian varieties*, 2nd. edition, Oxford University Press, 1974.
5. F. Oort – "The isogeny class of a CM-type abelian variety is defined over a finite extension of the prime field", *J. Pure Appl. Algebra* **3** (1973), p. 399–408.
6. B. Poonen and E. F. Schaefer – "Explicit descent for Jacobians of cyclic covers of the projective line", *J. Reine Angew. Math.* **488** (1997), p. 141–188.
7. K. A. Ribet – "Galois action on division points of Abelian varieties with real multiplications", *Amer. J. Math.* **98** (1976), no. 3, p. 751–804.
8. E. F. Schaefer – "Computing a Selmer group of a Jacobian using functions on the curve", *Math. Ann.* **310** (1998), no. 3, p. 447–471.
9. I. Schur – "Gleichungen ohne Affect", *Sitz. Preuss. Akad. Wiss., Physik-Math. Klasse* (1930), p. 443–449.
10. J.-P. Serre – *Topics in Galois theory*, Research Notes in Mathematics, vol. 1, 3rd. edition, Jones and Bartlett Publishers, Boston, MA, 1992.
11. — , *Abelian l-adic representations and elliptic curves*, Research Notes in Mathematics, vol. 7, A K Peters Ltd., Wellesley, MA, 1998.
12. G. Shimura – *Abelian varieties with complex multiplication and modular functions*, Princeton Mathematical Series, vol. 46, Princeton University Press, Princeton, NJ, 1998.
13. C. Towse – "Weierstrass points on cyclic covers of the projective line", *Trans. Amer. Math. Soc.* **348** (1996), no. 8, p. 3355–3378.

14. Yu. G. Zarhin – "Endomorphism rings of certain Jacobians in finite characteristic", *Mat. Sb.* **193** (2002), no. 8, p. 39–48 and *Sbornik Math.* **193** (2002), no. 8, p. 1139–1149.
15. — , "Homomorphisms of abelian varieties", `arXiv.org/abs/math/0406273`, to appear in: *Proceedings of the Arithmetic, Geometry and Coding Theory - 9 Conference*.
16. — , "Non-supersingular hyperelliptic jacobians", *Bull. Soc. Math. France*, **132** (2004), p. 617–634.
17. — , "Hyperelliptic Jacobians without complex multiplication", *Math. Res. Lett.* **7** (2000), no. 1, p. 123–132.
18. — , "Hyperelliptic Jacobians and modular representations", Moduli of abelian varieties (Texel Island, 1999), (G. van der Geer, C. Faber, F. Oort eds.), Progr. Math., vol. 195, Birkhäuser, Basel, 2001, p. 473–490.
19. — , "Hyperelliptic Jacobians without complex multiplication in positive characteristic", *Math. Res. Lett.* **8** (2001), no. 4, p. 429–435.
20. — , "Cyclic covers of the projective line, their Jacobians and endomorphisms", *J. Reine Angew. Math.* **544** (2002), p. 91–110.
21. — , "Very simple 2-adic representations and hyperelliptic Jacobians", *Mosc. Math. J.* **2** (2002), no. 2, p. 403–431.
22. — , "The endomorphism rings of Jacobians of cyclic covers of the projective line", *Math. Proc. Cambridge Philos. Soc.* **136** (2004), no. 2, p. 257–267.
23. — , "Very simple representations: variations on a theme of Clifford", Progress in Galois Theory (H. Völklein and T. Shaska, eds.), Kluwer Academic Publishers, 2004, p. 151–168.